Fri, 10/11/2019 - 14:11

Application 3: Supporting FAIR data discoverability in clinical research: providing a global metadata repository (MDR) of clinical study object

Principal Investigator and optional collaborators Sergey Goryanin – ECRIN Christian Ohmann, Steve Canham, Serena Battaglia – ECRIN Stefano Nicotri, Giacinto Donvito – INFN

Email sergei.gorianin@ecrin.org

Describe the proposed Project
In order to fully assess and review the evidence generated in clinical research, it is necessary to have access not only to the published results but also to the source individual participant data and related study documents (e.g. study protocol, statistical analysis plan, case report form). As data and document sharing becomes more and more common, however, the researcher is faced with a bewildering mosaic of possible source locations and access modalities. There is an urgent need to develop a central resource that can catalogue all the diverse data and documents associated with a clinical study, and then make that information searchable by using a central web portal. In the EU H2020-funded project eXtreme DataCloud (XDC; grant agreement 777367) such a service is developed under the coordination of ECRIN-ERIC (European Clinical Research Infrastructure Network) and will be available to open public till the end of 2019. Currently, the MDR instance in production contains more than 700,000 data objects from one trial registry (ClinicalTrial.gov) and 6 repositories (PubMed, ZENODO, Data Dryad, WWARN, Edinburgh DataShare, BioLINCC).
The first objective of the proposed project is to extend the MDR demonstrator to run in production in the EOSC environment and be part of the EOSC catalogue and to complete the data base by integrating all major data sources dedicated to clinical research.
The second objective is to include others EOSC services not already included into MDR: we plan to include also EGI Fed Cloud Resources to host the distributed repositories.

Description of the services and the technical environment that you have already in place
The web application is in place with an easy-to-use user interface, which provides flexible and scalable solutions to i) map metadata from different data sources, ii) save it into ECRIN MetaData Repository (MDR) CORE database and iii) convert it to JSON format and structure, based on the common ECRIN Metadata schema (https://zenodo.org/record/1312539), including the algorithms, developed by ECRIN team, to link studies and data objects.
The user interface (platform) is developed by OneData, which also provides the metadata management system leveraging an instance of ElasticSearch on-top of the standard Onedata solution. The service at the moment is in production at INFN-Bari and integrated into the XDC platform.
Tools used during the development:
Anaconda 3 Environment;
Python 3.6
Django 2.2

Description of the services and resources that you need and expected benefits
ECRIN MetaData Repository (MDR) CORE database and metadata conversion tool (developed and maintained by ECRIN).
OneData service will manage the metadata (leveraging an ElasticSearch service) and provide the user interface. At the moment this service is hosted at INFN, with a professional approach in

software deployment, continuous delivery and integration and server capacity.
The services and resources requested will allow to extend the scalability of the service running in production with a larger hardware resources using EOSC service provider. This will allow to cover the main data sources dedicated to clinical research and providing a user-friendly and efficient platform.
The service is intrinsically distributed (thanks to the Onedata architecture) so being part of EOSC could allow other providers to join the service, and to increase the scaling possibility.

Science Area 3.3 Health Sciences: Clinical Research, Epidemiology, Public health

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners
The service will have major scientific impact for the clinical research community: it will become a major source for identifying study material for secondary analyses (re-analysis, meta-analysis, further analysis) and for supporting the planning of clinical studies. It will also represent a major resource for translational research, supporting transferability and generalisability from experimental studies via pre-clinical studies to medical care.

Contribution to Open Access and FAIR
The service will provide a major contribution to the findability (F in FAIR) of studies and related data objects in clinical research, covering a wide range of study types, such as interventional trials, observational studies, epidemiological studies based on registries and cohorts. Indirectly, the service will also support accessibility (A in FAIR) to data objects in clinical research by providing evidence which object can be fully accessed and how.

Expected duration (from 6 to 12 months) 12 months

Minimal Compute and Storage capacity needed for sustaining the Project
Hardware requirements:
16-32 Gb of RAM (Total amount of RAM);
8-16 (v)CPU (v - if we'll use CloudServer, CPU - traditional server);
500 Gb of disk space (including web application, all necessary software and packages, structured database and storage for converted JSON files);
GPU - not needed
Software requirements:
Ubuntu server 18.04 LTS

Compute and Storage capacity to fully scale-up the Project after the completion of the pilot
Hardware requirements:
32 Gb of RAM (Total amount of RAM);
16 (v)CPU (v - if we'll use CloudServer, CPU - traditional server);
4.5 Tb of disk space (structured, semi-structured and non-structured databases + space for converted JSON files);
GPU - not needed
Software requirements:
Ubuntu server 18.04 LTS

Minimal storage capacity for long-term archiving for sustaining the Project
2 Tb (Structured database - PostgreSQL with all imported data + storage for JSON files)

Long-term data management policies and long-term archiving capacity required by the Project
The MDR service will be run by ECRIN-ERIC.
ECRIN is a sustainable public, non-profit organisation that links scientific partners and networks across Europe (http://www.ecrin.org/), to facilitate multinational clinical research.
ECRIN is currently increasing its data services offer and the MDR will be part of the 'data sharing toolbox' which provides data management policies and procedures for clinical trials data.

A long-term archiving capacity is needed for the correct implementation of the project (see above).

Mention any classified and/or privacy-sensitive data The service will be based upon open and public metadata.

Any other requirements
Deep learning algorithms (AI) for linking clinical studies to related data objects (under development).

Mon, 10/14/2019 - 08:43

Application 4: Open AiiDA lab platform for cloud computing in Materials Science

Principal Investigator and optional collaborators Giovanni Pizzi, EPFL; Nicola Marzari, EPFL; Leopold Talirz, EPFL

Email giovanni.pizzi@epfl.ch

Describe the proposed Project
The AiiDA lab brings the AiiDA workflow manager for computational science (www.aiida.net) to the cloud.
While domain experts can install AiiDA on their own hardware, the AiiDA lab web platform gives novice users access to their personal pre-configured AiiDA environment in the cloud. Interactive apps provide a graphical user interface for running and managing workflows in the browser.
By lowering the technical barriers of entry, the AiiDA lab gives more researchers access to advanced workflow management and provenance tracking capabilities.

Description of the services and the technical environment that you have already in place
AiiDA (aiida.net) is a workflow manager for computational science with a strong focus on provenance, performance and extensibility. When executing a workflow, AiiDA records the provenance – calculations performed, codes used and data generated – in a directed acyclic graph tailored to provide full reproducibility of any given result. The AiiDA engine relies on a message queue in order to support high-throughput use cases of up to 50k calculations per hour, and the relational database backend enables performant queries on graphs of tens of millions of nodes.
AiiDA (TRL 7-8) is used in production for high-throughput calculations (see e.g. the >150 citations of the AiiDA paper [1,2]) and, following more than 5 years of development (and over 3 years of stable versions used in production for publications), we will be releasing a new version 1.0 by October 31st, 2019. Besides development of the core, the AiiDA ecosystem comprises a large number of plugins, e.g. for connecting to various materials science simulation codes (full list on the AiiDA plugin registry [3]).

For the AiiDA lab (TRL 6-7, http://materialscloud.org/aiidalab), we are currently operating 4 instances, two on Openstack VMs and two on kubernetes. One of the kubernetes instances is deployed on limited EOSC-hub test resources in the CESNET Czech computing centre, and already uses EGI check-in for authentication.
The AiiDA lab uses docker for user containers, and kubernetes for orchestration. Users get persistent home volumes for active use (no long-term storage component).

[1] AiiDA paper https://www.sciencedirect.com/science/article/pii/S0927025615005820, [2] open-access preprint: https://arxiv.org/abs/1909.00433
[3] AiiDa plugin registry: https://aiidateam.github.io/aiida-registry/

Description of the services and resources that you need and expected benefits
After collecting experience and feedback on internal, smaller-scale instances of the AiiDA lab over the last 2 years, we would like to use a kubernetes cluster in order to be able to offer the AiiDA lab as an open service to the European computational (materials) science community with "no questions asked".

We are convinced that this will significantly lower the barrier of entry to using efficiently HPC codes and resources, and give more researchers access to advanced workflow management and provenance tracking capabilities.
In terms of resources, we will need compute, memory and persistent volumes. We may be interested in access to an S3-compatible object store (AiiDA may switch from storing file-based research data on the file system to an object store: S3-compatible, OpenStack Swift or other) but not in the first 6 months.

Science Area
1.3 Physical sciences
1.4 Chemical sciences
2.5 Materials engineering

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners
Reproducible calculations, automated workflows for materials properties calculations (turnkey solutions), FAIR sharing of research data.

Contribution to Open Access and FAIR
Reproducibility and seamless sharing of resources (not only of research data, but also the workflows to produce them) are two cornerstones on which the AiiDA ecosystem is built.

In order to make reproducibility possible in the age of high-throughput computations and complex workflows, the AiiDA workflow manager automatically tracks the full provenance of calculations in the form of a provenance graph, including not only inputs and outputs of calculations but the links from one calculation to the next, including metadata on computers, software, etc.

When researchers publish the AiiDA graphs on repositories like the Materials Cloud Archive [1,2] (Findable), they provide access not only to the results of calculations, but to every step along the way. Peers can explore the database interactively in the browser (Accessible - for example, use the provenance browser in [1] to explore the graph), download individual files (Interoperable) or the whole database (e.g. [4]), and start their research right from where the original author left off (Reusable).
The Materials Cloud Archive is indexed by re3data.org, FAIRsharing.org and recommended repository for materials science by Nature Scientific Data. Datasets receive DOIs and are indexed by Google Dataset search.

[1] archive.materialscloud.org
[2] https://www.materialscloud.org/policies#archive

[3] https://www.materialscloud.org/explore/2dstructures/details/e7db98c1-9d25-4872-8236-68559c5b0702
[4] https://doi.org/10.24435/materialscloud:2017.0008/v3

Expected duration (from 6 to 12 months) 12 months

Minimal Compute and Storage capacity needed for sustaining the Project

Jupyter notebooks require ~3GB of RAM and ~2 CPU cores per active user. At minimum, the AiiDA lab needs to be able to support introductory AiiDA tutorials, which have been ~40-70 participants in size in 2019 (already a few tutorials planned for 2020).
At the same time, the cluster should be available for people to go and give AiiDA a try. In order not to block these users, some additional margin is necessary. Therefore, for a minimum viable solution we ask resources to support ~100 concurrent users: 300GB RAM, 200 VCPUs.
We note that this capacity will not be needed continuously, especially in the first few months, even if we expect usage to increase, and we will have peak usage during AiiDA tutorials.

In terms of storage, having ~10GB of persistent storage per user will be sufficient for most simple use cases. We would like to support of the order of ~1000 (non-concurrent) users in total over the next 12 months, so we expect a need of 10TB. We emphasize that new storage will be needed as new users log in to the platform for the first time and the volumes are created for them. Therefore, the 10TB are not required from the very beginning. If the required size is too large, we could start with a lower initial quota (e.g. 2TB).

Compute and Storage capacity to fully scale-up the Project after the completion of the pilot
Compute, memory and storage needs scale linearly with the number of users. Therefore, the number of future users will determine the future needs. It is possible that the resources asked above can also be sufficient after the first 12 months.

Minimal storage capacity for long-term archiving for sustaining the Project
None for this service. If users use the data and want to publish their results, they can export their AiiDA databases to existing long-term storage repositories, either generic ones like Zenodo, or field-specific ones like the Materials Cloud Archive [http://archive.materiaslcloud.org/] (maintained by us) that provides intuitive interactive visualisations of the AiiDA graph.

Long-term data management policies and long-term archiving capacity required by the Project
None for this service (see above).

Mention any classified and/or privacy-sensitive data None.

Any other requirements
I'm not sure whether this is the right place to mention this, but I feel that the login/registration procedure via the EGI checkin takes too many steps and may deter prospective users from entering the platform. Could this be improved?
Happy to provide more input on this (e.g. the user should not be logged out after registration; the fact that an email has been sent should be displayed more clearly, ... Given that EGI checkin is relying on authentication by external identity providers, perhaps one could even skip the verification of the email address if it is already provided by the identity provider).
In my view, one should aim to reduce the time needed to register at least by a factor of 2.

Finally, when I clicked on the "B2ACCESS" option for signing in, I got this:

ERROR
SAML service got an invalid request.

If you are a user then you can be sure that the web application you was using previously is either misconfigured or buggy.

If you are an administrator or developer, here the details of the error follows:

eu.unicore.samly2.exceptions.SAMLRequesterException: Issuer is not among trusted: https://aai-demo.egi.eu/proxy/module.php/saml/sp/metadata.php/sso

Caused by: eu.unicore.samly2.exceptions.SAMLRequesterException: Issuer is not among trusted: https://aai-demo.egi.eu/proxy/module.php/saml/sp/metadata.php/sso

Tue, 10/15/2019 - 11:54

Application 6: VESPA-Cloud

Principal Investigator and optional collaborators

Baptiste Cecconi, LESIA, Observatoire de Paris, CNRS, PSL, France (PI, see below for collaborators)

Email

baptiste.cecconi@obspm.fr

Describe the proposed Project

VESPA (Virtual European Solar and Planetary Access) is a network of interoperable data services covering all fields of Solar System Sciences. It is a mature project, developed within EUROPLANET-FP7 and EUROPLANET-2020-RI. The latter ended in Aug. 2019. It will be further supported under the EUROPLANET-2024-RI project (starting in Feb. 2020).

Proposed collaborators (all funded through Europlanet-2024-RI):

- Pierre Le Sidaner, DIO, Observatoire de Paris, CNRS, PSL, France
- Stéphane Erard, LESIA, Observatoire de Paris, CNRS, PSL, France
- Angelo Pio Rossi, Jacobs Uni, Bremen, Germany
- Markus Demleitner, Heidelberg Uni, Germany
- Marco Molinaro, OATF-INAF, Trieste, Italy
- Albert Shih, DIO, Observatoire de Paris, CNRS, PSL, France
- Cyril Chauvin, DIO, Observatoire de Paris, CNRS, PSL, France
- Nicolas André, IRAP, Université de Toulouse, CNRS, France

The VESPA data providers are using a standard API (based on the Table Access Protocol of IVOA (International Virtual Observatory Alliance) and EPNcore, a common dictionary of metadata developed by the VESPA team). The VESPA services consist in searchable metadata tables, with links (URLs) to science data products (files, web-services...). The VESPA metadata includes relevant keywords for scientific data discovery, such as data coverage (temporal, spectral, spatial…), data content (physical parameters, processing level…), data origin (observatory, instrument, publisher…) or data access (format, URL, size…). VESPA hence provides a unified data discovery service for Solar System Sciences.

The architecture of the VESPA network is distributed (the metadata tables are hosted and maintained by the VESPA providers), but it is not redundant. The hosting and maintenance of VESPA provider's servers has proved to be a single point failure for small teams with little IT support. The VESPA-Cloud project with EOSC-Hub would greatly facilitate the sustainability of data sharing from small teams as well as teams, whose institutions have restrictive firewall policies (like labs hosted by space agencies, e.g., DLR in Germany). Each VESPA data provider is using the same server software, namely DaCHS (Data Centre Helper Suite), developed by the Heidelberg team included in the project.

VESPA-Cloud will provide a cloud-hosted facility to host VESPA compliant metadata tables in a controlled and maintained software environment. The VESPA providers will focus on the science application configuration, whereas the VESPA core team will support them with the maintenance of the deployed instances. The development of the VESPA provider's data service will be done using a git versioning system (github or institute gitlab).

An instance of the VESPA query interface portal will also be implemented on an EOSC-hub provided virtual machine.

In the course of the VESPA-Cloud project, we will implement in the DaCHS framework cloud-storage API connectors (such as Amazon S3, iRODS, etc.) to reading data in the cloud and ingesting metadata. Since DacHS is used worldwide by many datacenters to share astronomical and solar system data collections, many teams will benefit from this development.

We also propose to setup a Europlanet Research Community, which will include VESPA-Cloud, and other Europlanet-2024-RI and Europlanet-Society related projects (such as SPIDER - Sun Planet Interaction Digital Environment on Request, or the previous services developed within PSWS - Planetary Space Weather Services).

Further development plans for VESPA-Cloud are listed below:
• New VESPA portal architecture.
a new VESPA portal architecture based on Lucene-like technologies, will be developed in the frame of the upcoming EUROPLANET-2024-RI project. This would greatly enhance the portal search interface, especially for complex queries dealing with several services, where SQL-like queries are difficult set up and to generalize. This would also allow VESPA to be interoperable with NASA/PDS4 (Planetary Data Archive) Search Engine.
As the development didn't started yet, we don't have quantitative elements for sizing our needs.
The architecture of the search portal will be split into 3 elements:
• a data ingestion server, which will harvest VESPA provider's servers (on VESPA-Cloud and on the classical VESPA network) regularly and update the Elastic database;
• an Elastic nodes cluster (possibly using "Elastic Cloud Compute Cluster") with the VESPA network data
• a front-end web query portal with the user interface querying the Elastic cluster.
This new architecture is required with the growing number of services and data products served by the VESPA providers.
• JupyterHub.
Access to VESPA data services through community based python scripts (astropy, pyvo...) with a JupyterHub facility (with "EGI Notebooks" applications). At the moment, we distribute jupyter notebook tutorials, which should be run locally by users on their own machine.
• Run-on-demand.
On-demand computing services (models, cutouts, resampling...), using UWS (Universal Worker Service, an IVOA standard) as a job submission manager.
The VESPA providers can serve data products as well as data services (like cutout or resampling services on both data or simulation runs), or even direct calls to numerical codes through REST interfaces. The implementation of a UWS application, based on OPUS (Observatoire de Paris UWS System: https://uws-server.readthedocs.io/en/latest/) will enhance the overall interoperability between the VESPA network (and other IVOA based frameworks) with the EOSC resources.
• Federated Authentication
User and Group management using federated authentication is not yet implemented in the VESPA

network. Such capabilities will allow team to work with the VESPA infrastructure before their data is publicly released. This leads to a wider adoption of the VESPA network in the community as well as provision of better services, since the provider's will also be users of the data services.

Description of the services and the technical environment that you have already in place
VESPA is a mature project, with 50 VESPA providers distributing open access datasets throughout the world (EU, Japan, USA). In October 2019, the current number of data products available within the VESPA network reaches 18.3 millions (among which 5 millions products from the ESA Planetary Science Archive).

Each VESPA provider is hosting and maintaining a server (physical or virtualized) with the same software distribution (DaCHS, Data Centre Helper Suite), which implements the interoperability layers (from IVOA and VESPA) and following FAIR principles. Each server hosts a table of standardized metadata with URLs to data files or data services. Data files can be hosted by the VESPA provider team, or in an external archive (e.g., ESA/PSA - Planetary Science Archive).

The VESPA query interface portal is developed and maintained at the Observatoire de Paris (Paris, France).

VESPA:
- http://www.europlanet-vespa.eu (project)
- http://vespa.obspm.fr (query portal)
- https://voparis-wiki.obspm.fr (wiki and documentation)
o VESPA/EPNcore metadata dictionary:
https://voparis-wiki.obspm.fr/display/VES/EPNcore+v2
o Tutorials for implementing VESPA services:
https://voparis-wiki.obspm.fr/display/VES/Implementing+a+VESPA+service
DaCHS:
- https://dachs-doc.readthedocs.io/

A preliminary prototype of a DaCHS instance on the EOSC-hub infrastructure has been tested earlier in 2019. This instance has been ordered through the EOSC-hub marketplace (https://marketplace.eosc-portal.eu), using EGI Cloud container compute BETA. This has been running for several weeks, with success. We could install and run the DaCHS framework, as well as serve a VESPA metadata table. This instance is now in undeployed status.

Description of the services and resources that you need and expected benefits
The VESPA architecture relies on the assumption that data provider's servers are up and running continuously. The VESPA network is distributed but not redundant. For small teams with little or no IT support is available locally, the services are down regularly. We thus need a more stable and manageable platform for hosting those services. The EOSC-hub "cloud container compute" service would solve this problem.

We propose to use the EOSC infrastructure to host VESPA provider's servers (through a controlled deployment environment with git-managed containers).

The VESPA providers would be able to:

- order a VM with all the server framework installed,
- configure the server for their science application,
- co-administrate the server packages with the VESPA team,
- update the content and the tables.

In a second phase, we will implement an EOSC-hub hosted VESPA portal, using the web interface developed at Observatoire de Paris. The portal will be deployed from a git-based repository.

Science Area 1.3 Physical sciences (Astronomy)

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners
VESPA is a distributed (although not redundant) data discovery and access framework. Hosting VESPA services in the cloud ensures their availability on the long term, and in turn the reliability of the full VESPA network. Starting in 2020, we propose to use the VESPA-Cloud infrastructure for the new VESPA providers selected after the yearly VESPA implementation workshop AO. This would add at 3 to 5 new services in 2020 from external teams. Science teams with the VESPA project will also use the VESPA-Cloud facility. We thus estimate to open about 10 services during year 2020.

The current VESPA network connects 50 data services, serving about 1.8M data products, with an average monthly visitor count of 100. With an enhanced visibility of the VESPA network through VESPA-Cloud, we expected to see a wider adoption.

VESPA-Cloud is a proof of concept, which will demonstrate the use and the efficiency of the EOSC-Hub infrastructure to the Solar System science community. Furthermore, in the course of the Europlanet-2024-RI project, the VESPA team will build strong interfaces with the planetary surfaces team building on GIS technologies (GMAP work package), as well as Space Weather (SPIDER work package) and Machine Learning (ML work package). This opens doors for reaching out similar communities focusing on Earth sciences. The VESPA-Cloud team has also existing contacts with the several ESCAPE work packages (e.g., on the interoperability, radio-astronomy, or solar physics topics).

Contribution to Open Access and FAIR
The goal of VESPA is to make data Solar System Findable and Accessible through interoperable interfaces, and is recommending standard data and metadata formats, ensuring reusability.

All software developed for VESPA are open-source (mostly GPLv3, or Apache).

VESPA-Cloud enhances the accessibility, by providing a sustainable access for VESPA dataset

Expected duration (from 6 to 12 months)12 months for setting up services and finding a sustainable approach for further operations.

Minimal Compute and Storage capacity needed for sustaining the Project
- 10 VM instances
- 2 CPU per VM,
- 4GB RAM per VM,
- 20 GB disk per VM,
- 1 fixed IP per VM

- 5 remote ssh-key access per VM
- deployment of containers from git-managed repository
- 10 TB of storage accessible from every VM

The minimal individual VESPA-Cloud provider’s instance is: 2 CPU, 4GB RAM, 20 GB disk.
We estimate to have about 10 instances to deploy in the first stage. Each instance must have a fixed and public IP address (customizable DNS names preferred). The instances are expected to be up and running all the time. Short unavailability of the services is acceptable, if the instance can be relaunched automatically.

Another need is a Storage Buffer for data ingestion. This is a global and temporary storage volume, which can be mounted on any VESPA-Cloud provider's instance for metadata extraction and ingestion. Data are pushed onto this volume for initial metadata extraction and ingestion and removed after this task is finished. This storage size should be 10TB. For providers needing a permanent storage capability, we will investigate with EUDAT, Zenodo or other EOSC partners.

For the current version of the VESPA portal, the compute and storage specifications are the same as for the individual VESPA-Cloud provider’s instances.

For future prospects, as listed in the project description section, it is too early to propose a sizing of needs. We will work in parallel of the VESPA-Cloud project, with the Europlanet-2024-RI work program on this infrastructure sizing.

Compute and Storage capacity to fully scale-up the Project after the completion of the pilot
With the new Europlanet-2024-RI program, we can expect at least 5 new providers per year, each with the same compute and storage needs.

Among the further developments foreseen after the completion of the pilot program:
- Next level data portal (lucent-like).
- JupyterHub access.
- On-demand computing services (models, cutouts, resampling...)

However, they are not yet not completely defined and sized, since this is part of the work to be accomplished in Europlanet-2024-RI.

Minimal storage capacity for long-term archiving for sustaining the Project

Not fully applicable, see below.

Long-term data management policies and long-term archiving capacity required by the Project
The Data Management Plan of the Europlanet-2020-RI/VESPA is available on the JRA2 page (https://voparis-wiki.obspm.fr/pages/viewpage.action?pageId=559857), under the "issued documents" section.

The VESPA-Cloud services are hosting metadata tables of data products hosted on independent facilities. Each metadata table can be rebuilt from a resource descriptor script maintained with a git versioning system. The archiving of the computed metadata tables is not planned yet, but any discussion on this specific point is welcome. The archiving of the resource descriptor script should also be discussed, but it may contain private data (such as logins and tokens), which shall not be openly available. The provider's teams are currently in charge of their own data preservation, outside of the VESPA or VESPA-Cloud projects.

However, the VESPA-Cloud provider’s instances are expected to run as long as the data providers are sharing their data. A long-term sustainability plan has to be prepared during the Pilot program, together with the new-born Europlanet Society (https://www.europlanet-society.org).

Mention any classified and/or privacy-sensitive dataNo sensitive data

Any other requirements

October 15, 2019 - 14:42

Application 8: OpenBioMaps data management service for biological sciences and biodiversity conservation

Principal Investigator and optional collaborators: Dr. Miklós Bán, University of Debrecen, Dept. of Evolutionary Zoology

Email: banm@vocs.unideb.hu

Describe the proposed Project:
The OpenBioMaps (OBM) is a free and open-source database framework for scientific and conservation purposes. It is a grass-root initiative to build bridges between science, conservation and education. In the current state, there are ~30 active scientific and conservation projects in three countries which use OBM. These projects are independent of each other, the only connection (yet) among them that they use the same tool for data management (OBM). These projects collected (observations or descriptive data: occurrence data of specimens of different animal and plant species in nature; DNA sequences or literary data) data from about 50 countries around the world. There are spatial distribution data for about 12,000 species in these databases. Currently, there are about 6 million data records on five servers in 2 countries. The servers are also independent of each other with different rules and user communities (governmental, civil, scientific).

Through the EaP project, we have two aims, first, would like to expand the capabilities of the OBM platform (create a new OBM node) to provide more computing capacity for larger projects, and after introducing the new OBM node, we would like to develop and introduce a new computing-intensive service layer. By creating a new, high-capacity node in a stable IT environment, we would be able to expand our current Central and Eastern European user community with new Western European links which is especially important in conservation science area, because the countries of Central and Eastern Europe are currently the most endangered in terms of biodiversity loss in the EU. Furthermore, high computational capacity will enable us to develop a new service layer which will provide - to be at the level of individual projects and above the project level - new data connection and interpretation services.

Description of the services and the technical environment that you have already in place:
We already have a network of OBM servers. These servers are operated independently of each other but there are few services which connect them and these servers as nodes forming a Network of OBM servers. The OBM network is open and new servers can join freely and can share resources with the other servers. Currently, shared resources are always storage capacity.
Each server provides web services which include database services and map services. The map services using standard protocols, therefore these are integratable into other not-OBM services. The database service can be accessed directly or integrated with web services. There is also available database management (user and API) interfaces.
The Technology Readiness Level of OBM is about 8 and 9 because it is used in “production” environment by several research institutes and National Parks, but we and the community continuously developing new components/modules for the system.
After the first six months of the running of the new node, we would like to introduce a new service layer module which is still in the development stage. Its TRL is 3 currently, although we have previous experience in the analysis required to produce the scientific part of the module: https://www.dunaipoly.hu/uploads/2017-04/20170419104310-rosalia-9-tomori... Chapter 2. pp. 443-612.

The list of the OBM based projects of all known OBM servers are available on the following link:
- https://openbiomaps.org/projects/

The map of OBM servers is available here:
- https://openbiomaps.org/projects/openbiomaps_network/

The technical documentation of OBM is available here:
- https://openbiomaps.org/documents/en/

The source code of the OBM service components are accessible here:
- Web applications: https://gitlab.com/openbiomaps/web-app
- Mobile application: https://gitlab.com/openbiomaps/openbiomaps-mobile
- Containerized web server edition: https://gitlab.com/openbiomaps/docker

Scientific papers about the OBM system:
- https://www.biorxiv.org/content/10.1101/010405v1 doi: https://doi.org/10.1101/010405
- https://content.sciendo.com/view/journals/hacq/18/2/article-p179.xml doi: https://doi.org/10.2478/hacq-2019-0007

Some published database and long long term stored queries in different OBM based projects with DOI:
- https://search.datacite.org/works?query=OpenBioMaps

Description of the services and resources that you need and expected benefits:
Basically, we only need computing capacity and storage space to build the new high-performance server node and develop the planned service layer. The high CPU capacity allows us to run giant projects within the OBM system and test the development needed for it.

Science Area: Biology, conservation biology, ecology, biodiversity. Related codes from the revised FOS: 1.5, 1.6, 1.7, 4.1

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners:
OpenBioMaps' main scientific role is to provide tools, services and interfaces to improve data management and shaping collaboration between science, education and conservation. We already have important results to achieve this goal: There are several conservational institutes and research projects who use OBM services. The next objective is to establish active links between these projects and areas. To highlight the covered subject and geographical areas of OBM based projects we present a list here of the largest OBM partners/users with geographical coverage of their projects: Duna-Ipoly National Park (Hungary), Duna-Dráva National Park (Hungary), Bükk National Park (Hungary), Fertő-Hanság National Park (Hungary), Őrség National Park (Hungary), Hortobágy National Park (Hungary), Milvus Group Association (Romania), BalkanHerps (Albania, Greece, Macedonia, Bosnia-Herzegovina, Serbia, Turkey, Bulgaria, Romania, Slovenia), Debrecen-DNA Bank (Hungary, Romania, Greece, Ukrajna, Russia, Kazakhstan, Mongolia, Azerbaijan, China,..), Kurgan Database (Eastern European and Asian countries).

The new OBM node’s main scientific impact will be to provide a common place for large scientific and conservational projects. "Large" means, 2-30 million biological/spatial data records in one projects (One record can contain 200 attributes in several tables). Solving the problems of biotic data management in large projects and sharing these experiences is an extremely important output of this EaP project. This node also will be an important interconnection place for Eastern Central and Western European projects. Currently there only a few scientific links between Western and Eastern Europe in this scientific area. The large Western European data projects do not contain data from Eastern Europe and there are only a few examples of usage international database-related tools in Eastern Europe.

The introduction of the new (recently developed) service layer is scheduled for the second half of the planned EaP project. This service layer after validating the spatial distribution models with large datasets will be available in each OBM project, and the newly received data in the project will improve individually the service layers. About 10000 species distribution modelling maps will be publicly available for the OBM community and to anyone outside of OpenBioMaps. This service layer also includes a new database-based online validating method that has never been used before. It will also be freely usable by anyone and can be integrated as an optional feature into any OpenBioMaps database project. We expect that this module of the new service layer will improve data quality in several projects, especially civil science projects.

Introducing this new service layer, we will provide immediate access to scientific tools and results for nature conservation.

Although the above list shows that OpenBioMaps-based projects have wide geographical coverage, most of them are based in Hungary. We expect greater international visibility for OpenBioMaps with using EOSC-hub services.

Contribution to Open Access and FAIR: OpenBioMaps provides a database framework for a variety of projects, as well as a free service for hosting many scientific and citizen science and conservational projects. These projects tend to share more data after the OBM system was introduced, due to its flexible interface and more data sharing options. The OpenBioMaps system supports data and database findability, accessibility, interoperability and reusability in an original and complex way. However, these expectations are met independently through the individual functions of the freely configurable modules. That is, they are present in different projects to varying degrees depending on user needs and goals.

Expected duration (from 6 to 12 months): 12

Minimal Compute and Storage capacity needed for sustaining the Project :
16 CPU cores
128 Gb RAM (4Gb RAM per cores)
2 Tb HD

Compute and Storage capacity to fully scale-up the Project after the completion of the pilot:
32 CPU cores
256 Gb RAM (8Gb RAM per cores)
4 Tb HD

Minimal storage capacity for long-term archiving for sustaining the Project:
The text information is stored in SQL format and does not take up much storage space, but the raster-based map files take much more. However there is no point in archiving the whole thing, so our minimal long-term archiving capacity for the proposed project is quite small, only 2Tb.

Long-term data management policies and long-term archiving capacity required by the Project : Long-term data management and archiving are more of an issue for OpenBioMaps as a whole than for this EaP project. Our calculated long-term archive storage requirement is 12Tb based on current utilized capacity.

Mention any classified and/or privacy-sensitive data:
Spatial and temporal occurrence data of strictly protected species - that is conservation sensitive data. Scientific data in a before publishing state and any licenced data.

Any other requirements:

Tuesday, October 15, 2019 - 14:56

Application 9: AGINFRA+: Virtual Research Environments to Support Agriculture and Food Research Communities

Principal Investigator and optional collaborators: Leonardo Candela (CNR), Collaborators: AGINFRA+ project partners

Email: leonardo.candela@isti.cnr.it

Describe the proposed Project:
AGINFRA+ addresses the challenge of supporting user-driven design and prototyping of innovative e-infrastructure services and applications. It particularly tries to meet the needs of the scientific and technological communities that work on the multi-disciplinary and multi-domain problems related to agriculture and food. It uses, adapts and evolves existing open e-infrastructure resources and services, in order to demonstrate how fast prototyping and development of innovative data and computing- intensive applications can take place.

The AGINFRA+ project is exploiting the Virtual Research Environments (VREs) paradigm for three research communities. VREs are a prominent existing cloud-based solution provided by the D4Science Initiative. VREs are web-based, community-oriented, user-friendly, open-science-compliant working environments for scientists and Evaluation practitioners working together on a research task. These research communities are (a) the Agro-climatic and economic modelling research community (b) The Food safety risk assessment research community and (c) the Food security research community.

Description of the services and the technical environment that you have already in place:
The AGINFRA+ VREs are offering a large array of facilities as-a-Service by an integrated environment including: (i) a data & semantics facilities enacting the creation and or- ganisation of semantically rich metadata (by VocBench); (ii) a shared workspace to store, organise and share any version of a research artefact; (iii) a social networking area to have discussions on any topic and be informed on happenings, e.g. the availability of a research outcome; (iv) a data analytics platform to execute analytics tasks either provided by the user or provided by others. It is endowed with importing and sharing facilities for analytics methods implemented in forms including R, Java, Python, and KNIME (largely used by the food safety community). The platform enacts tasks execution by a distributed and hybrid computing infrastructure; (v) a catalogue-based publishing platform to make the existence of a certain artefact “public” according to FAIR principles; (vi) a Jupyter notebook based environment for documenting and recording analytics processes; (vii) a scholarly publishing platform integrated with Pensoft infrastructure to enact the creation of innovative papers including datasets and methods hosted by AGINFRA; (viii) a platform for data visualisation enabling to create smart graphs and share them.

System in TRL 9 and in production, information about the existing VREs is available at https://aginfra.d4science.org/explore. The services are described in conference contributions like: https://doi.org/10.1016/j.future.2018.10.035, http://ceur-ws.org/Vol-2363/paper3.pdf, and has supported the scientific research behind the publications with DOIs listed below:
https://doi.org/10.1016/j.mran.2018.09.001
https://doi.org/10.1109/eScience.2018.00124
https://doi.org/10.18174/FAIRdata2018.16273
https://doi.org/10.1016/j.future.2019.05.063

https://doi.org/10.3897/fmj.1.46561
https://doi.org/10.2903/sp.efsa.2019.EN-1701
https://doi.org/10.1109/eScience.2018.00124

Description of the services and resources that you need and expected benefits:
AGINFRA+ VREs already integrate with several of the EOSC services: EOSC-portal AAI for authentication of users, EGI Notebooks for providing an interactive coding platform, and an OpenAIRE dashboard for the project.

AGINFRA+ VREs require a IaaS platform to deploy its DataMiner analytics facility which provides big-data analytics features for our users. A typical setup of the DataMiner facility requires two clusters of 15 VMs with 16 vCPUs/32 GB RAM each, that can be managed via APIs (preferably OpenStack). Both EGI Cloud Compute and OCRE services can be considered in this case. The use of these services will, on the one hand, facilitate the support of new communities with enough computing resources to perform relevant data analytics for their research and, on the other hand, will help on the sustainability of the project results by establishing stable service level agreements to support the operation of the AGINFRA+ services.

AGINFRA+ can also benefit from B2FIND to improve the discoverability of the research objects of the VREs.

Science Area:
4. Agricultural sciences
4.1 Agriculture, Forestry, and Fisheries
4.4 Agricultural biotechnology

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners:
The use of EOSC services will contribute to the sustainability of the AGINFRA+ VREs beyond the life of the project thus allowing researchers to produce relevant scientific results in a stable platform. The integration with B2FIND will attract new users to the EOSC interested in the agri-food areas.

Contribution to Open Access and FAIR:
AGINFRA+ has a strong focus on Open Access and FAIR:
All VREs include a catalogue-based publishing platform [1] that allows users to share their research objects with other researchers within the VRE or the general public. following FAIR principles.
The data analytics platform of the VREs provide provenance records of every computation, contributing to the Reusability and Reproducibility of the results.
The project supports the creation of two data journals based on the ARPHA publishing platform targeting Food Modelling and Viticulture Data that allow researchers to publish research assets like models, datasets as linked-open-data papers. These journals can be easily feed with the publishing platform mentioned above.

[1] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi, F. Sinibaldi (2019) Enacting Open Science by D4Science. Future Generation Computer Systems doi: 10.1016/j.future.2019.05.063

Expected duration (from 6 to 12 months): 12 months

Minimal Compute and Storage capacity needed for sustaining the Project : The pilot will require the following resources at the selected IaaS providers: > 4 VMs with 16vCPUs/32GB RAM and at least 1 proxy VM with public IP that redirects requests to the cluster. All these VMs need to be located on a single data center.
Compute and Storage capacity to fully scale-up the Project after the completion of the pilot: The typical configuration of the big-data analytics of AGINFRA+ require 2 clusters (which can be deployed at different providers) with 15 VMs with 16 vCPUs/32 GB RAM each.

The adaptation of the VRE tools to the EOSC-hub services and deployment on new resources is feasible within 6 months, however AGINFRA+ looks towards the establishment of long-term SLAs with EOSC-hub providers to ensure the sustainability of its VREs (12 months and beyond).

Minimal storage capacity for long-term archiving for sustaining the Project:
No long-term archiving is required at the moment.

Long-term data management policies and long-term archiving capacity required by the Project :
No long-term archiving is required at the moment.

Mention any classified and/or privacy-sensitive data: Not applicable

Any other requirements:

Tue, 10/15/2019 - 17:02

Application 11: EOSC DevOps framework and virtual infrastructure for ENVRI-FAIR common FAIR data services

Principal Investigator and optional collaborators Zhiming Zhao (University of Amsterdam), Alex Vermeulen (University of Lund)

Email z.zhao@uva.nl

Describe the proposed Project
The ENVRI-FAIR IT development WP (WP7) is responsible for co-designing, developing, testing, and validating the common services developed in the ENVRI-FAIR RIs, together with WP8-11 for 4 specific environmental sub-domains. The proposed project aims to use EOSC services (Cloud, DevOps, Jupyter, and Storage) to enhance the development of ENVRI common operations and their future deployment in the EOSC environment.

Scientific and technical challenges:
1. The development of ENVRI common data services are in different private test beds and development environments, which make the future integration difficult.
2. The ENVRI common data services lack of common integration, deployment pipeline to standardize their lifecycle management. It makes the development, testing, integration and deployment time consuming.
3. The developers of ENVRI common services often have to interact with users to co-develop science cases. The prototypes of those cases are often less portable and shared.
4. The ENVRI community generates lots of data; the data sharing is quite hard across RIs.
With the project, the community will get the following benefits
1. A virtual infrastructure based on IaaS (Infrastructure-as-a-Service) allows RI development communities to access elastic resources, and perform the integration.
2. A DevOps management service can enhance the collaboration among developers, testing teams and service operators.
3. A notebook based environment to access and integrate data services for the community.
4. A cloud based storage service for the community.

Description of the services and the technical environment that you have already in place
Currently, our development relies on private infrastructure from RIs (e.g., test bed in ICOS), institutes (e.g., ExoGENI from UvA), EGI FedCloud (e.g., for specific use cases).
Most of the applications are: 1) data management services/databases, e.g., catalogues, web portals; 2) data analytics, using Hadoop/Spark/Storm; and 3) scientific workflows (via specific engine or Jupyter notebooks). To better support the development and deployment of common operations/services ENVRI community require, we plan to study EOSC services (see above) for several purposes:
1. Earlier services developed in ENVRIplus (part of them are at TRL 7), or using Open Source software packages (can be TRL higher than 8, e.g., SPARK, CKAN/GeoNetwork, ElasticSearch, etc.) need to be tested in Cloud infrastructures, in order to be integrated with the resources offered by EOSC.
2. A continuous DevOps framework will enable the automation of integration and deployment those services during their lifecycle.

3. We envision demand for data processing services (e.g., Jupyter) for developing the specific science cases within ENVRI-FAIR.
4. The ENVRI community need cloud storage for storing data and secondary results of the data processing workflows.

Description of the services and resources that you need and expected benefits
1. A virtual infrastructure based on IaaS provided by EOSC EGI FedCloud (or 100 Percent IT Trusted Cloud) allows development teams to: 1) deploy automated testing environment, 2) perform specific software evaluation, 3) system integration and performance benchmark. The results will support the ENVRI-FAIR EOSC integration WP (WP5), specifically to deploy services and test them in the EOSC ecosystem.
2. A DevOps management service, possibly provided by Jelastic Platform-as-a-Service (https://jelastic.com/), will enable the development teams in ENVRI-FAIR to: 1) speed up the testing, integration and deployment processes in software lifecycle and 2) enhance the collaboration among different teams for programming, testing and operation.
3. Computing services, specifically those provided by EGI Notebooks, will 1) enable rapid code development for data processing and visualization using Jupyter and 2) enhance the sharing and collaboration among team members.
4. A storage solution for the ENVRI community is urgently needed that will 1) enable data sharing and 2) deliver to the community the benefits of elastic storage.

Science Area
Serving the software development from different subdomains of ENVRI (environmental and earth sciences) (1.5 in OECD), namely atmosphere, marine, solid earth, ecosystem and biodiversity.

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners
The proposed project will accelerate the development of the common services in ENVRI RIs and will promote the interoperability and reuse of the software results. More specifically:
1. The virtual infrastructure based on IaaS provided by EOSC EGI FedCloud (or 100 Percent IT Trusted Cloud) will have impacts on: 1) helping the ENVRI community to familiarize with the EOSC infrastructure and 2) accelerating the integration of the ENVRI solution with the EOSC ecosystem.
2. The DevOps management service will have impacts on: 1) shortening the lifecycle of ENVRI service delivery and improve the interaction between ENVRI developers and the user community and 2) improving the agility in responding to demands from the ENVRI scientific user communities.
3. Computing services will promote the visibility of ENVRI services in scientific communities.
4. The storage service will improve the data sharing among scientific communities, and improve their collaboration. It will improve the capacity of the experiments.

Contribution to Open Access and FAIR
Accelerate FAIRness implementation by developing FAIRness for data, FAIRness software, services and knowledge, curated and managed as FAIR resources.

Expected duration (from 6 to 12 months) 12

Minimal Compute and Storage capacity needed for sustaining the Project
We would minimally need 4 VMs, each preferably with 4 cores and 8G memory. They will be used to configure different testing environment needed by development teams. If possible, resources should be extensible based on user demands. More specifically: 1) 2 VMs for automated testing, 2)

2 VM for workflow tasks (will be configured as container cluster).
B2SAFE Storage space of approximately 10 TB is considered sustainable.

Compute and Storage capacity to fully scale-up the Project after the completion of the pilot
The DevOps framework and Jupyter notebooks will be scaled out to enable use within the entire ENVRI community.

Minimal storage capacity for long-term archiving for sustaining the Project
The data produced from the project will be mainly from the software, runtime logs, simulation results, and the other content generated by the software results. After the project,
- The developed software (including source and containers) will be further curated in the archive like Zenodo.
- The other data will need stored in the cloud storage. The storage space needed is about the several TBytes.

Long-term data management policies and long-term archiving capacity required by the Project
Long term preservation will be done jointly with the RIs in ENVRI. It will be based on the ENVRI-FAIR data management plan.

Mention any classified and/or privacy-sensitive data Not at the moment.

Any other requirements No.

Fri, 11/01/2019 - 14:51

Application 14: Integration of toxicology and risk assessment services into the EOSC marketplace

Principal Investigator and optional collaborators Thomas Exner and Barry Hardy, Edelweiss Connect GmbH, Technology Park Basel, Hochbergerstrasse 60C, CH-4057 Basel, Switzerland

Email thomas.exner@edelweissconnect.com

Describe the proposed Project
This proposal is additionally supported by:
Iseult Lynch and Anastasios Papadiamantis, School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK (on behalf of the NanoCommons project, grant agreement no. 731032)

Stefan Kramer, Johannes Gutenberg Universität Mainz
Chris Evelo, Egon Willighagen, Department of Bioinformatics - BiGCaT, Maastricht University
Danyel Jennen, Department of Toxicogenomics, Maastricht University
Haralambos Sarimveis, National Technical University of Athens
Costas Charitidis, National Technical University of Athens
Marc Jacobs, Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V.
Ola Spjuth, Uppsala Universitet
Tim Dudgeon, Informatics Matters ltd.
Paul Jennings, Vrije Universiteit Amsterdam
Lee Walker and Marianne Matzke, Natural Environment Research Council
Vladimir Lobaskin, University College Dublin
Socorro Vázquez-campos, Acondicionamiento Tarrasense Associacion
Andrea Haase, Bundesinstitut für Risikobewertung
Andreas Falk, BioNanoNet Forschungsgesellschaft mbH
Martin Himly, Paris Lodron University of Salzburg
Antreas Afantitis, Novamechanics ltd.
Dieter Maier, BioMax Informatics AG
Manuel Pastor, Universitat Pompeu Fabra
Igor Tetko, BIGCHEM GmbH
Urban Fagerholm, Prosilico
Frank von Delft, Diamond Light Source Ltd.

Toxicology is undergoing a paradigm shift, from a phenomenological to a mechanistic discipline. Traditional approaches to risk assessment rely heavily on animal testing. However, ethical issues concerning animal welfare have resulted in new legislations aiming at drastically reducing or completely banning animal testing. Additionally, new emerging materials like nanomaterials and nano-enabled products are adding more challenges for proper safety assessment since toxic effects are associated with their complex structure and may vary upon nanoparticle evolution (coating, dissolution, agglomeration, etc.) throughout the entire testing procedure. In vitro testing complemented by in silico modelling (machine learning to predict adverse effect, bioinformatics, modelling of exposure scenarios, atomistic simulations and many more) represents a viable alternative to classical approaches for chronic and systemic toxicity as demonstrated in different research projects. Notably, these projects have produced a significant wealth of data and have provided a panel of tools for processing and analysing the toxicity mechanisms involved including procedures for simulation and modelling. However, since the data was in a large extent collected by independent researchers, it appears fragmented over a large set of databases or even kept in local repositories and not publicly and openly available in a FAIR way. The same is true for the software tools lacking harmonization and interoperability.

To satisfy the demand of the European research communities, OpenRiskNet (funded under the call EINFRA-22-2016 “User-driven e-infrastructure innovation” of the Horizon 2020 work programme 2016-2017) created an openly accessible e-infrastructure provided by a combination of research-intensive academic groups and SMEs. Research communities involved in (nano-)safety assessment include the EU’s chemical manufacturing industries, e.g. pharmaceutical companies, chemical and agrochemical industries and cosmetic industries, and the corresponding regulatory agencies, e.g. the European Medicines Agency (EMA), the European Chemicals Agency (ECHA), the Scientific Committee on Consumer Safety (SCCS), the European Food Safety Authority (EFSA) and the Organization for Economic Co-operation and Development (OECD) for a standardized, reproducible and interoperable way to access all available data, knowledge and analysis and modelling tools. The safety assessment encompasses toxicology and especially predictive toxicology, systems and structural biology, bioinformatics and its subtopics toxicogenomics, cheminformatics, biophysics and computer science.

OpenRiskNet combines the achievements from earlier projects, which generated modelling and validation workflows, knowledge integration and data management. It furthermore integrates ongoing projects and important stakeholders through an associated partner programme. The main components of the infrastructure are an interoperability layer added to every service to describe the functionality guaranteeing technical and semantic interoperability, a discovery service, deployment options based on container technology, and packaging of the infrastructure into virtual instances. This is complemented by training and support on integration of specific services based on prototype implementation, usage of standard file formats for data sharing including the generation of templates for data and metadata, as well as the harmonized usage of ontologies.

Developments of OpenRiskNet have already been implemented by several research projects (EU-ToxRisk, ACEnano, in3) and in commercial settings. Furthermore, they are a central part of the infrastructure work packages of new proposals. A direct follow-up is the NanoCommons infrastructure project, which is particularly relevant in nanosafety assessment. Its major intention is to create a community framework and infrastructure for reproducible science, and in particular for in silico workflows by:
1) integration and federation of existing NMs characterisation and interaction mechanisms knowledge, protocols and data (beyond simple toxicity), along with quality assurance criteria and underpinning ontologies;
2) compilation and development of a user-friendly interface for a suite of computational tools for mechanistic and statistical modelling, read-across, grouping, safe-by-design and life cycle assessment, and bench-marking of their predictive power;
3) and provision of (typically remote) access to its KnowledgeBase, modelling toolbox (predictive, grouping, risk assessment) and workflow optimisation, and the supporting expertise, to the broader user community via the Transnational Access model.

In this way, it addresses the complete data lifecycle, starting with (i) planning the experimental workflow followed by (ii) data processing/analysis and (iii) toxicity evaluation/prediction until (iv) data storage compliant with the FAIR data principles.

As just described, the mentioned projects have, on the one hand, developed, and still are developing and will continue to develop data management and sharing tools and services as well as processing and analysis software specifically for the risk assessment and heighbour communities and good links to other data and software providers have been established, many joining the common infrastructure development. However, on the other hand, demographic data from the survey of workshop participants showing that still a low share of external researchers, risk assessors and regulators has been reached so far. Therefore, there is still the need to enlarge the outreach to the complete target audience mentioned in the paragraph above and to even more simplify and harmonize the access to have the full benefit of integrating data across disciplines. This exactly aligns with the community building, user guidance, open data and open science ambitions of EOSC and we hope that we can profit from the high visibility of EOSC and tools on the EOSC market place. To be able to do so, the goal of the project proposed for the Early Adopter Programme is:
* to establish guidelines and workflows how to link risk assessment services to existing EOSC services;
* to harmonize access; and in this way,
* to prepare them to become part of the EOSC marketplace.

Description of the services and the technical environment that you have already in place
The main concept of the e-infrastructure are virtual research environments (VRE) - standardized computational environments based on stock solutions for containerized software and microservices (Kubernetes/OpenShift). Such VREs can be operated either on individual end user machines (Minikube/Minishift) or deployed to in-house or public cloud computing infrastructure and interface with HPC systems. This infrastructure allows integrating data, analysis, modelling and simulation services for all areas of risk assessment. For advanced usage of the OpenRiskNet and NanoCommons toxicology and risk assessment tools, workflow managers like jupyter as well as nextflow and Squonk developed and maintained by OpenRiskNet partners are fully integrated allowing access to all features via REST API calls and SPARQL queries. Additionally, single-sign-on and continuous integration / continuous delivery (CI/CD) options are provided to support system administrators and software developers. Listings of all available services and their API specifications can be found at
1) OpenRiskNet: https://openrisknet.org/e-infrastructure/services/
2) NanoCommons: https://infrastructure.nanocommons.eu/

A reference instance of such a VRE is available at https://home.prod.openrisknet.org/ currently running on the SNIC Science Cloud (https://www.snic.se/allocations/ssc/) but soon to be moved to the cloud infrastructure at the Johannes Gutenberg Universität Mainz. The reference infrastructure was used by project partners and external partners (associated partners, implementation challenge winners (https://openrisknet.org/associated-partner-programme/implementation-challenge/) and other third-parties) and its operation is already secured for the next two years. Other instances are running e.g. at the Diamond light source and will be set up at the University of Birmingham for the NanoCommons project. This demonstrates that a TRL of 8 has been reached. All technical documentation providing evidence of the TRL level can be found on github (https://github.com/OpenRiskNet/home) as well as in the Resources and Training library (https://openrisknet.org/library/) and support offerings of OpenRiskNet (accessible via https://openrisknet.org/e-infrastructure/) and NanoCommons (accessible via https://www.nanocommons.eu).

Description of the services and resources that you need and expected benefits
The Goal of the project proposed here is to integrate the data management and sharing services of OpenRiskNet and NanoCommons as well as to constantly increase the suite of processing, analysis, modelling and simulation tools provided by the consortium partners of the two projects but also other associated partners from EU-funded research projects into EOSC and especially the EOSC marketplace. In this way, we plan to provide a central entry point to all available data sources and software tools to the toxicology and risk assessment community as well as to neighbouring scientific areas such as drug development, personalized medicine, environmental protection and innovative material production.

For this integration it is important to guarantee harmonization and interoperability with existing EOSC services to allow integrated usage of all these services. Toxicology as a very interdisciplinary area will then be able to profit from data coming from a variety of disciplines including system and structural biology, medicine and even more distant areas like social science to understand population differences based on lifestyle and environmental influences. In the same way, neighboring disciplines can profit from the data provided by our community to include safety-related aspects during the complete lifecycle of products starting from their development, production up to disposal (safe-by-design).

To reach the presented aims the project will explore how standard EOSC functionality like data storage (short-term and long-term), data sharing following the FAIR principles including central registrations of persistent identifiers, resource listings, compute infrastructure and single-sign-on and licensing options could be exploited to optimize access and applicability of the provided services and guarantee long-term sustainability of the risk assessment e-infrastructure as part of the global, pan-European infrastructure offered by EOSC.

Science Area
1.2 Computer and information sciences: bioinformatics, machine learning, big data
1.4 Chemical sciences: polymer science, nanomaterials, physical chemistry, computational chemistry, drug design
1.5 Earth and related Environmental sciences: environmental sciences, water resources
1.6 Biological sciences: cell biology, microbiology; biochemistry and molecular biology; biochemical research methods; biophysics; reproductive biology; developmental biology; theoretical and computational biology
2.10 Nano-technology
3.3 Health sciences
3.4 Medical biotechnology
4.1 Agriculture, Forestry, and Fisheries

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners

The increasing amount of data and software tools available in a harmonized and interoperable form will address existing gaps in chemical and nanomaterial risk assessment. In many cases, new approach to address the goal of the 3Rs (Replacement, Reduction and Refinement) of animal testing like integrating approaches to testing and assessment, read across and group are hindered by sufficient data for training and validation of these methods. The central access point on EOSC for all the data, tools to analyse it and to sustain them as part of the EOSC marketplace will give individual researchers and research projects the chance to re-use the data and tools and, in this way, avoid duplication of work and expenses. Additionally, it will provide the infrastructure for these projects to store and share their data according to FAIR principles and consequently fulfill their obligations with respect to open data and open science.

Contribution to Open Access and FAIR
The community represented by OpenRiskNet and NanoCommons and their associated partners are fully committed to Open Access and the FAIR principles. Even if OpenRiskNet was not producing new data, it created new ways to find and access data easier and in an interoperable way and demonstrated workflows how to re-use the data. NanoCommons is taking this one step further by actively working together with finished and ongoing research projects to make data openly accessible, in the case where it is not available publicly, or improve its FAIRness by providing guidelines and checklists on FAIR data management specifically adapted to the nanosafety area.

Expected duration (from 6 to 12 months) 12

Minimal Compute and Storage capacity needed for sustaining the Project
3 small servers (e.g. 8 cores, 16GB RAM)
3 medium servers (e.g. 16 cores, 64GB RAM)
2 TB storage available as volumes that can be attached to those servers.

Compute and Storage capacity to fully scale-up the Project after the completion of the pilot
3 small servers (e.g. 8 cores, 16GB RAM)
5 medium servers (e.g. 16 cores, 128GB RAM)
5 TB storage available as volumes that can be attached to those servers

Minimal storage capacity for long-term archiving for sustaining the Project
Resources for long-term archiving of raw big data (omics) used in toxicology and risk assessment are available already partly as part of EOSC provided by Elixir and/or EBI. The additional data sources added by this project would be a collection of smaller sources of raw data, literature curated data as well as processed data with the goal to make them interoperable with the larger sources and sustain them in the future after the end of the corresponding research projects. These data sources will provide customized user interfaces and access options via APIs but will be based on harmonized approaches of OpenRiskNet, NanoCommons and via these of EOSC hub.

During the time of the project, databases of a combined size of approximate 1.5TB are planned to be deployed onto the EOSC infrastructure. This includes e.g.
1) ToxCast dataset (100 GB)
2) TG-GATEs and DrugMatrix processed data (700 GB)
3) Pre-reasoned datasets for specific case studies (e.g. TGX, 100 GB)
4) ACEnano data warehouse (100 GB)
5) NanoFase data warehouse (200 GB)

Long-term data management policies and long-term archiving capacity required by the Project

The above-described data source will only be the beginning of toxicology data provided via EOSC. More data sources will be added over time and covered by the harmonization efforts of NanoCommons and more generally EOSC. Since these sources are managed by their own data management teams, policies will be enforced by these. However, everyone deciding to become part of the infrastructure has to commit completely to the FAIR principles and open data and open science and implement the relevant guidelines developed by EOSC in collaboration with the relevant research communities. NanoCommons will be in charge of supervision and monitoring that legal, ethical and quality control requirements are fulfilled. Through this mechanism, we hope to increase the amount of data available via EOSC marketplace to 10 TB over the next two years and to a much higher amount thereafter.

Mention any classified and/or privacy-sensitive data

The storage of classified data is not anticipated for the project. However, the underlying databases will provide mechanisms to keep data confidential for an embargo period until the results have been published and the data can be publicly released. With respect to privacy-sensitive data, the only information collected is person data on data providers and users (name, affiliation) since they are part of the metadata needed for quality control of the uploaded data and avoiding misuse of the infrastructure. Corresponding terms of use and privacy policies are already in place.

Any other requirements

Fri, 11/01/2019 - 16:10

Application 15: Towards a Global Federated Framework For Open Science Cloud: Three Use Cases

Principal Investigator and optional collaborators Hussein Sherief , Almaahad Almutagadem (AASCTC),CODATA-AOSP; Jianhui Li, China Network Information Center (CNIC-CAS)

Email hussein.sherief@aasctc.com

Describe the proposed Project
The ultimate goal of the project is to empower the scientific researchers from Africa and China to interact with EOSC services and data -- publish data sets and data analysis services to EOSC Marketplace and access them from outside of Europe. We also aim to enable our scientific communities users to develop their own applications to analyse data on remote cloud servers offered by both EOSC/EGI and CNIC CAS in China. The technical challenge is working towards a so-called Global Open Science Cloud (GOSC) so that service providers can validate the identity credentials of the users globally and the global federation manager will manage the global federated instances of membership, policy, resources monitoring, portability and interoperability.

We identified three user cases within the scope of the Early Adoption Programme:.
1) Disaster Risk: New dataset are made available by Data Cloud of CAS, www.csdb.cn/pageDataResource, that providing high resolution (8 m) satellite data for the simulation of tsunami, hurricane, earth quakes , typhoon, floods and extreme weather. We want to integrate this dataset from CASEarth with the analysis tools developed by the EOSC-Hub Disaster Mitigation Competence Centre, and the service portal developed by OPENCoasts Thematic Center.
2) Smart City: A Smart City tools, Snap4City, https://www.snap4city.org, is developed. CAS provides high resolution data and sensor data for the city of Shenzhen in Guangzhou province, China. We want to integrate this service with EOSC so scientists from Europe and Africa can test the service and provide the evaluation feedback. The EGI Check-in service will enable the access by the global scientific community.
3) Precision Medicine: Beijing Genomics Institute , Data bank and Beijing Institute of Genomics (BIG) provides datasets for analysing genetic make up of mentally disordered patients. We want to test this data with the service developed by EOSC-hub Elixir Competence Center. We will involve researchers at H3Bionet in South Africa, https://h3abionet.org/, to test the service.

Description of the services and the technical environment that you have already in place
The existing Chinese research communities are affiliated to CAS, universities and the industrial researches. This community is set up in the form of alliance which is led by CNIC-CAS. We also have African research communities.
The Disaster Risk community is led by Institute of Remote sensing CAS
The Smart City community is lead by Wuhan University
The Precision Medicine is lead by Beijing Institute of Genomics CAS

The services and technical environment already in place are as follows:
1) Disaster Risk:
For Disaster Risk the existing deployed CSTCloud services is called geodata (CASEarth). The input to CASEarth include social Economic Data, Airborne Remote sensing Data, Remote Sensing Satellite

Data, Navigation Satellite Data and Field investigation Data which is processed to give cloud service for Biological Ecosystem, Beautiful China (environment), Digital Silk Rood(environment of the Silk Rood countries) three Pole environment and Coastal and Deep Sea. These deployments include atmospheric and ocean simulations.
From EOSC , OpenCoasts and AGROS can be interoperable with CASEarth using a workflow software Kubernetes. This will allow OpenCoasts and ARGOS communities to access the data from CASEarth as well as deploy CASEarth services

We have docker containers, CentOS 7 Linux, S3 sorage/POSIX, Hadoop HDFS, and CASEarth databases and analytic tools from EOSC-Hub Disaster Mitigation Competence Centre, and the service portal developed by OPENCoasts Thematic Center.

2) Precision Medicine:
For precision medicine the existing Deployed CSTCloud service is called Big Data. This service gives analysis of genomics data. It includes Assembly tools, Evolution tools, , Methylation tools, Pangenome tools, Sequence Analysis tools and Transcription Profiling tools as well as databases from Beijing Institute of Genomics. The other Analysis services include DNA and RNA sequencing , Molecular and cellular structures, and Proteins from Beijing Genomics Institute along with the genomics Data Bank

Elixir gives services in genes and Genomics, Molecular and cellular structures, Evolution and phylogeny, Proteins and proteomes, and Bioinformatics.

Elixir can be made interoperable with CSTCloud Big Data service including the BIG and BGI databases.
H3Bionet from South Africa can use this interoperable services. H3Bionet is a project for whole Human genomic sequencing specialized for Africa. A sub section of his project is precision medicine as it applies for Africans.

We have docker containers, CentOS 7 Linux, S3 sorage/POSIX, Hadoop HDFS, and CSTCloud Big Data which includes csdb data bases (BIG and BGI data bases), Elixir and CSTCloud Big Data tools.

3) Smart City:
For smart city , CSTCloud has existing deployed services AI and IoT. AI has machine learning, tensorflow, deep learning, etc services. IoT services gives integration to senors and their data. CSTClod has Shenzhen city sensor and satellite data.

Snap4city is EOSC service for smart city.

CSTCloud AI , CSTCloud IoT and Snap4city can be made interoperable by developing a workflow based on Kubernetes. This will allow Snap4City to access the services of CSTCloud AI , CSTCloud IoT and the Shenzhen smart city data.

We have docker containers, CentOS 7 Linux, S3 sorage/POSIX, Hadoop HDFS, and databases of csdb , and tools from CSTCloud IoT , CSTCloud AI and geodata, Snap4City

Description of the services and resources that you need and expected benefits
The global federated framework needs to be implemented using EOSC and CSTCloud forming GOSC. The benefit for the user community is access to new data sets and very advanced data processing tools that would have a seamless interoperability for the further analyses of data and development of even more advanced tools in Disaster Risk, Smart City and Precision Medicine.

EOSC services to be integrated include:

EOSC-hub Disaster Mitigation CC services: to integrate data analysis tool for disaster risk research
EOSC-hub OPENCoasts portal
EOSC-hub ELIXER CC services
EGI Check-in: to enable users from China and Africa to access EOSC datasets and services
EGI Data transfer: to transfer dataset from CAS Cloud to EGI
EGI Cloud: for data processing and analysing
EGI Notebook: to enable researchers from China and Africa to analysis data
EUDAT B2FIND : to publish Chinese (genomics) dataset to EOSC

Science Area
1.2 Computer and Information Sciences
1.5 Earth and Related Environmental Science
1.6 Biological Sciences
2.7 Environmental Engineering
3.1 Basic Medicine
3.4 Health Biotechnology
5.7 Social and Economic Geography

Expected scientific impact of using services, data and other research outputs from EOSC-hub and its partners
Disaster Risk:

The simulation of Typhoon using OpenCoasts, AGROS and CASEarth services and data will allow the high resolution verification of the simulation in the forecasting, characteristics and trajectory of the typhoon because of the usage of
high resolution Satellite data (8 m) .

Smart City:

A full and comprehensive study of smart city affects the society, the economy and the health of the smart city dwellers. It also gives them control of the city by interacting with the governance of the city as active members impacting the level of smartness and efficient management of their time. This use case contributes in a fundamental way to Social and economic geography.

Using the CSTCloud services and Snap4City services and the sensor and satellite data , it is possible to compute the traffic, the transportation efficiency, the power efficiency of the city. In an interactive way, the citizens of the city can use this information to manage their activities.

Precision Medicine:

Elixir services and CSTCloud services can be used to map the mental disorder diseases genetic pattern. These will deviate from the genetic pattern of normal people. Identifying the section of the genetics pattern that is responsible for the mental disorder disease will help in finding ways of correcting these abnormal patterns of the disease to that of the normal pattern. This results in treatment that is precisely tailored for the specific patient.

There is a lot of stress and trauma globally which is affecting the youth in Europe , Asia and Africa leading to chronic mental disorders which the current science of psychiatry fails to remedy. Genetic based precision psychiatry gives a genetic image to it so that treatment that corrects the troubled section of the genetic make up can be adjusted. This is computationally very intensive. This use case contributes to science of psychiatry in a fundamental way.

Contribution to Open Access and FAIR

F for federated. This pilot project introduces globally federated concept

A for accessible . This pilot project introduced globally accessible concept

I for interoperability . This pilot project introduces global interoperability concept

The idea of FAIR is extended to embrace globality which allows us to have global open access.

Expected duration (from 6 to 12 months) 12 months

Minimal Compute and Storage capacity needed for sustaining the Project
Disaster Risk
For each use case provide information about
Number of VMs needed
3000
Number of CPU cores per VM
8
Amount of RAM per VM
128
Amount of storage per VM
200 GB HDD
Any additional storage requirements (e.g. Block Storage)
4 PB Long term storage for archiving results
Any additional technical requirements

Smart city
For each use case provide information about

Number of VMs needed
2000
Number of CPU cores per VM
8
Amount of RAM per VM
128
Amount of storage per VM
100 GB HDD
Any additional storage requirements (e.g. Block Storage)
20 TB Long term storage for archiving results
Any additional technical requirements

Precision Medicine
For each use case provide information about
Number of VMs needed
300
Number of CPU cores per VM
8
Amount of RAM per VM
128
Amount of storage per VM
100 GB HDD

Compute and Storage capacity to fully scale-up the Project after the completion of the pilot
For the fully scale-up of the project all cases should be using PFLops speed and Peta Bytes storage.

Minimal storage capacity for long-term archiving for sustaining the Project 1 PB Long term storage for archiving results

Long-term data management policies and long-term archiving capacity required by the Project
For the long term Open Data / Open access , FAIR policies must be implemented

Mention any classified and/or privacy-sensitive data None

Any other requirements