Call: INFRAEOSC-02-2019

Topic: Prototyping innovative services

INFRAEOSC-02-2019

Building Open Science Services on European E-infrastructure

Acronym: BOSSEE

Date of Preparation: January 29, 2019

Coordinator: Benjamin Ragan-Kelley

e-mail: benjaminrk@simula.no

tel/fax:

Keywords: Open Science, reproducibility, reusability, education, accessibility, Jupyter, Binder, notebooks, cloud, EOSC, FAIR data, astronomy, geosciences, health sciences, photon science, XFEL

#	Participant organisation name	Short name	Country
1	Simula Research Laboratory (coordinator)	Simula	NO
2	CNRS-Observatoire astronomique de Strasbourg	CNRS-ObAS	FR
3	École polytechnique	EP	FR
4	EGI	EGI	NL
5	European XFEL GmbH	EuXFEL	DE
6	INSERM	INSERM	FR
7	QuantStack	QuantStack	FR
8	University of Oslo	UiO	NO
9	Université Paris-Sud	UPSud	FR
10	University of Silesia	Silesia	PL
11	Wild Tree Tech	WildTree	CH

Abstract

To truly achieve the societal goals of Open Science, we must make progress beyond the 'mere availability' of scientific results, to the practical usability and exploitation of such data once it is made available, an area where there is much room for improvement. The Jupyter ecosystem shows great promise as a collection of tools for bridging this gap; for making Open Science useful and accessible to all, from researchers to educators to public citizens. Jupyter is of increasing importance in computational science, data science, academia, industry, governments, and service providers, and used by millions worldwide. Jupyter notebooks have great potential to push Open Science forward because they can encapsulate a computational study that can be turned into a publication or produce part of a publication, such as a figure, a major part of making complex tasks reproducible. The Jupyter-based Binder project adds a means to execute notebooks in specified computational environments, an aspect of reproducibility not yet widely supported.

We will (i) extend the capabilities of the Jupyter tools and ecosystem to add functionality that we view as essential to and providing great value for EOSC and Open Science, focused on accessibility, interactivity, and reproducibility. Based on this framework of improved Jupyter tools, it will be possible to Build Open Science Services on European E-infrastructure (BOSSEE), and (ii) build a range of diverse innovative open services on EOSC as part of this project, both to demonstrate and ensure that our developments serve real-world Open Science use cases.

Many BOSSEE partners have longstanding experience and leadership roles in the Jupyter ecosystem, and in deploying services built on Jupyter to many users across the globe. Complementary to this core expertise, we integrate partners focusing on the application of these tools to a wide range of scientific disciplines and communities, for which EOSC hosted demonstrator services are developed.

Contents

1 Excellence

In many scientific disciplines, it is common for researchers to rely on heterogeneous computational tools and technologies to collect data, explore the input data sets, run simulations, visualise the outcome, and share their result with peers or with a larger audience. Often, such data analysis cycles are iteratively refined.

For simple datasets, processes may remain manageable. However, when dealing with larger and more complex use cases, including big data from research facilities or High Performance Computing resources, the complexity makes iteration cycles slower for the researchers. A complex iteration cycle also makes research results more difficult to reproduce. Results that cannot be reproduced make research ineffective: they create barriers towards re-using the results in future research work, a critical aspect of Open Science. This situation is exacerbated by the current and accelerating increase of the amount of scientific data being available, including the data becoming accessible through the EOSC-Hub. But this growing availability of data also provides a massive opportunity for Open Science.

Project Jupyter has developed as one piece of various solutions to the data deluge, by enabling the construction of computational services accessible from anywhere, any web-browser-enabled device, with access to any data. Jupyter-based tools such as Binder and repo2docker show great promise for enabling researchers to better perform **Reproducible and Open Science**. Jupyter was recognised for its contribution to data analysis in research with the prestigious 2017 *ACM Software System Award*, of which previous winners include TCP/IP, UNIX, and the World Wide Web. It is widely used today in research, education, and industry. We will build on these tools, both improving their capabilities and expanding their accessibility to new communities, both academic and demographic, in order to **further the mission of Open Science**.

In this proposal, core team members of Jupyter projects – including a number of recipients of the *ACM Software System Award* – and key contributors to the open source scientific computing ecosystem, detail improvements to the capabilities of Project Jupyter to **provide a framework on which innovative EOSC services can be created**. By collaborating with a wide variety of stakeholders from diverse scientific and educational domains, we aim to demonstrate and ensure that such innovative EOSC services – built on Project Jupyter – are feasible, valuable, and effective in furthering Open Science. The goal is to **improve the accessibility of EOSC resources to researchers and the general public, and improve the accessibility, interactivity, reproducibility, and re-usability of computational research and Open Science.**

1.1 Objectives

The aims of BOSSEE are to:

Aim 1: Enable a **sustainable**, **community-developed**, general purpose, **interoperable** toolbox for interactive computing, data processing, and visualization **that facilitates the entire life-cycle of Open Science**, from initial exploration to **reproducible publication**, research and development in industry, teaching, and outreach. This is by supporting and steering the Jupyter software ecosystem, which exists to develop open source software, open standards, and services for interactive computing across dozens of programming languages.

Aim 2: Leverage this technology for all scientists, across borders, domains, disciplines, and demographics, through **free public distributed collaborative services** tightly integrated into the European Open Science Cloud (EOSC), in collaboration with a federation of related services operated by the wider community.

Aim 3: Demonstrate the value and versatility of such services through **innovative co-designed tailored applications** in a variety of disciplines and contexts.

Aim 4: **Support Open Science** and maximize impact through development and dissemination of best practices, **training**, and **community building** around the usage and development of the above toolbox, with a focus on **interoperability**, **reproducibility** and **reusability**.

We will achieve our aims through the following objectives:

Objective 1: **Infrastructure and services for Jupyter on EOSC** — Contribute a distributed infrastructure to the European Open Science Cloud (EOSC) and the wider Open Science community that can be tailored to provide a multitude of generic or specialized services that facilitate open science in a **wide range of scientific domains** and projects. This infrastructure will build on the Jupyter project and ecosystem, taking the form of a federation of JupyterHub/Binder instances, tightly integrated into the EOSC-Hub. To maximize **impact, outreach, and sustainability**, the federation will include and encourage instances operated by external partners, whether free or non-free, public or private, general purpose or custom built for specific needs – e.g. providing access to specialized or large data sets, or specific hardware. This objective will be supported by improvements to the Jupyter deployment toolbox which fosters **reuse and interoperability beyond the Jupyter ecosystem**.

Objective 2: **Improving interactive computing** — Improve the interactive computing capabilities of students, researchers, educators, and the public through contributions to the Jupyter environment, in the form of developments of interactive widgets, visualization tools, collaboration features, dashboards, teaching tools, and expanded support for more language communities, such as interactive C++. While Jupyter is already widely used, there are many areas of interactive exploration that can be developed further.

Objective 3: **Reproducibility and FAIR data** — Extend facilities for **reproducibility of computational environments** and facilitating **FAIR data practices**. We will contribute to the recording and reproducibility of environments with repo2docker and Binder, and extend capabilities to better support FAIR data requirements. In particular, the archival of execution environments to support **reusability** of notebooks in the future needs attention. Such notebooks may, for example, be published alongside traditional publications to detail the computation of published data and figures, and address the Reusablity requirement of FAIR data.

Objective 4: **Demonstrators in science and education** — We will demonstrate and ensure the versatility and value of the components and the services built from them, through applications to a number of domains in academic research, education, research infrastructures, SMEs, and for the public sector, driven through our project partners. In particular, we will contribute demonstrators in the following areas: astronomy (**T4.2**), education (**T4.3**), fluid dynamics (**T4.4**), geosciences (**T4.5**), health (**T4.6**), mathematics (**T4.7**), and photon science (**T4.8**), involving universities, research infrastructure facilities, and SMEs.

Objective 5: **Outreach, engagement, and sustainability** — Reach out to scientists and the wider Open Science and Open Data communities to encourage engagement and exploitation of the EOSC-Hub and the Jupyter-based Open Science Services for their research domains and interests. Engaging a larger community will help **ensure the sustainability** of the services and underlying infrastructure by distributing its development, hosting, and maintenance over stakeholders from a variety of institutions and backgrounds, from the private sector to public research, education and open government.

Table 1.1.1: Each objective and the tasks which further that goal.

Objective	Tasks
1	**T2.1**: “Maintenance of Jupyter and JupyterHub” **T2.2**: “JupyterHub / BinderHub convergence”, **T5.1**: “Prototype European Binder instance and global federation”, **T5.2**: “Integration with EOSC”, **T5.3**: “Easy deployment of JupyterHub and BinderHub on a variety of infrastructure”
2	**T2.3**: “Accessibility in Jupyter”, **T2.4**: “Multi-device Real-time Collaboration”, **T3.2**: “Interactive C++ in Jupyter with XEUS”, **T3.3**: “Jupyter Interactive Widgets”, **T3.5**: “Teaching tools, infrastructure, and best practices”
3	**T3.1**: “Further development of repo2docker and Binder”, **T3.4**: “Archiving software environments for reproducible computation”
4	**T4.2**: “Demonstrator: Astronomy”, **T4.3**: “Demonstrator: enriched teaching with Jupyter”, **T4.4**: “Demonstrator: Visualisation and control of fluid dynamics in Jupyter notebook”, **T4.5**: “Demonstrator: Geosciences”, **T4.6**: “Demonstrator: Nuclear Medicine dosimetry”, **T4.7**: “Demonstrator: Interactive Mathematics with Jupyter Widgets”, **T4.8**: “Demonstrator: Reproducible photon science workflows at European XFEL”
5	**T6.2**: “Training Workshops and community building”, **T6.3**: “Online resources for open science”, **T6.4**: “Local Help Desk”

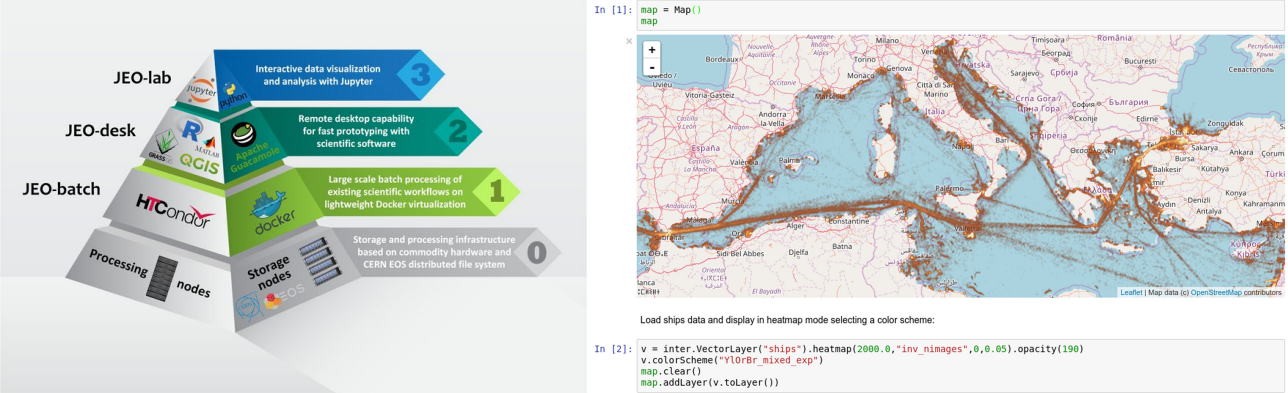


Figure 1.2.1: *Left*: The Joint Research Centre (JRC) Earth Observation Data and Processing Platform (JEODPP) is a heavy user of the Jupyter Notebook (source: https://cidportal.jrc.ec.europa.eu/home/), where it features at the top of the pyramid to help users with interactive data visualisation and analysis. *Right*: An example service in which an interactive visualisation is provided through the Jupyter Notebook rendering of the density map of the ships detected from Sentinel-1 images over the Mediterranean sea during the period October 2014 to September 2016. [Soi+18, Figure 6].

1.2 Relation to the Work Programme

The BOSSEE project addresses the challenges of the "Prototyping new innovative services" call (ID: INFRAEOSC-02-2019).

Our strategy is based on taking the increasingly popular Jupyter Notebook and Jupyter Ecosystem: we want to evolve and improve them so that new innovative services based of the Jupyter tools can be developed for the EOSC.

There is evidence that the Jupyter Notebook is an e-infrastructure that is useful across many domains: it is already widely adopted in numerous communities and used by millions of researchers and educators worldwide [PG15].

- *Journalists* and practitioners of *data-driven journalism* at the LA Times, BuzzFeed News, Columbia Journalism School [Lat] [Han18] [Hir16],
- *Research institutions* such as CERN, JRC, and many more, operating institution-wide Jupyter deployment,
- *Universities* using Jupyter as a teaching platform,
- *Large cloud providers* building commercial products on the top of Jupyter (Google DataLab and Colaboratory, Amazon Sagemaker, Microsoft Azure Notebooks),
- *Other EOSC projects*. Jupyter is already planned to become an important service on the European Open Science Cloud (for example the EOSC-04-funded PaNOSC project [Pan]).
- *Data scientists*: some argue that the Jupyter Notebook is *the* tool of choice for data scientists across domains [Per18].
- Over 3 million notebooks are deposited on GitHub [Par19].

A particular example is the Joint Research Centre Earth Observation Data and Processing Platform (JEODPP) shown in Fig. 1.2.1, illustrating the interactive data exploration within an environment that allows to save and communicate the data exploration conveniently. These projects are building upon Jupyter as it is available at the moment.

Within this context of a common software platform, we address the call for prototyping innovative services:

- Our BOSSEE proposal is focussed on developing the next generation of the Jupyter tools to enable all domains to develop such services on the EOSC.
- Through the BOSSEE-improved Jupyter, we will provide an agile and fit-for-purpose *EOSC-enabled framework* within which innovative services can be built by the scientific communities.
- We believe this is realistic as a multitude of (non-EOSC) services and use cases based on Jupyter have been developed, as sketched above.
- We co-design the framework and the service demonstrators with core Jupyter developers, EGI and domain specialists from the scientific communities.
- We will provide some particular services as part of the project (in the domains of astronomy, geosciences, health, mathematics, education, and photon science) to stimulate the design of other novel innovative services to address the evolving needs of the scientific community.

1.3 Concept and Methodology

1.3.1 Concept

Open Science is the principle that science, in order to be most **impactful** and **socially responsible**, should be done **publicly**, with as much of the scientific process and products **accessible, reviewable, and reusable** by as many members of the global community as possible. In the modern age of computational science, almost all academic fields, from humanities to social sciences to biology and astronomy are presented with exciting opportunities for Open Science. As more and more research takes the form of code and/or data, the opportunity to share, reproduce, and reuse scientific work is greater than ever, even enabling new forms of **interdisciplinary collaboration**.

At the same time as we share in these exciting opportunities, there are corresponding challenges, technical and social, to making Open Science a practical reality. We face big questions: If a researcher has code and/or data to publicise, how is that best done? How do researchers learn **Open Science best practices** in their field? How do previously disconnected fields benefit from each other's work as the same computational challenges are faced again and again by different communities?

These are the questions that guide BOSSEE. With so much research being done that wants to be Open, how can we make Open Science

1. as **easy** as possible to share?
2. as **useful** as possible to other researchers and the public?

Our plan for **improving access and effectiveness of Open Science** can be summarised as:

1. improve and maintain **common software infrastructure** used for Open Science,
2. develop the Jupyter ecosystem to improve capabilities to **better serve Open Science**,
3. **guide, validate, and demonstrate** our developments through collaboration with a wide variety of application domains,
4. enable students and researchers to perform Open Science through **training and education**, and improving inclusiveness by focusing these on under-served and under-represented communities, and
5. operate services to facilitate Open Science collaborations with Jupyter software.

1.3.2 Project Jupyter and the surrounding ecosystem

Jupyter ecosystem as the root of BOSSEE

BOSSEE has chosen to centre its efforts on the Jupyter software ecosystem. Figure 1.3.1 summarises a typical use case of Jupyter Notebook and Binder; both are described in more detail below.

The Jupyter notebook and Jupyter ecosystem are of increasing importance in computational science and data science, in academia, industry, and services. In addition to supporting high productivity of researchers, they have great potential to push Open Science forward: the notebook provides a complete description of a computational and data science study (Step 1 in figure 1.3.1), and the notebook can – in principle – be turned into a publication, or can be used to provide the required computation for a part of a publication, such as a figure (Step 2 in figure 1.3.1). Once the researcher has specified what software is required to execute the notebook (Step 3 in figure 1.3.1), the study is completely reproducible by anyone (Step 4 in figure 1.3.1).

In this way, the notebook enables reproducibility of complex tasks with hardly any additional effort on the user side. The Binder project allows to execute such notebooks in tailored computational environments; an aspect of reproducibility that is not widely supported yet, and a great opportunity for improving best practices in Open Science.

Furthermore, for users wanting to connect to a local Jupyter notebook server on their machine, or to connect to a server somewhere else on the Internet, the users only need a web-browser to display and use the notebook regardless of the location of the notebook server, allowing computation to run anywhere from a local laptop to a remote supercomputer or in the cloud. Because of these characteristics, the Notebook is already planned to become an important service on the European Open Science Cloud (EOSC) (for example in [Pan]), and is an ideal component to use when building Open Science Services.

Project Jupyter

Project Jupyter [Jup], which has grown increasingly popular in the scientific computing community, has become the *lingua franca* of interactive computing in both academia and industry. The main goal of Project Jupyter is to provide a consistent set of tools to improve researchers' workflows from the exploratory phase of the analysis to the communication of the results [Klu+16].

Split in 2014 from the *IPython Project* [Ipya], Jupyter has grown rapidly in popularity and adoption both in the industry and academia. We estimate the user base of the Jupyter notebook to be in the millions [PG15]. Users range from data scientists to researchers, educators, and students from many fields, including journalists and librarians. In 2017, the Jupyter team was awarded the *ACM Software System Award*, an annual award that honors people or an organization *"for developing a software system that had a lasting influence"*. Prior recipients include *Unix*, *TCP/IP*, and the *World Wide Web* [Acm].

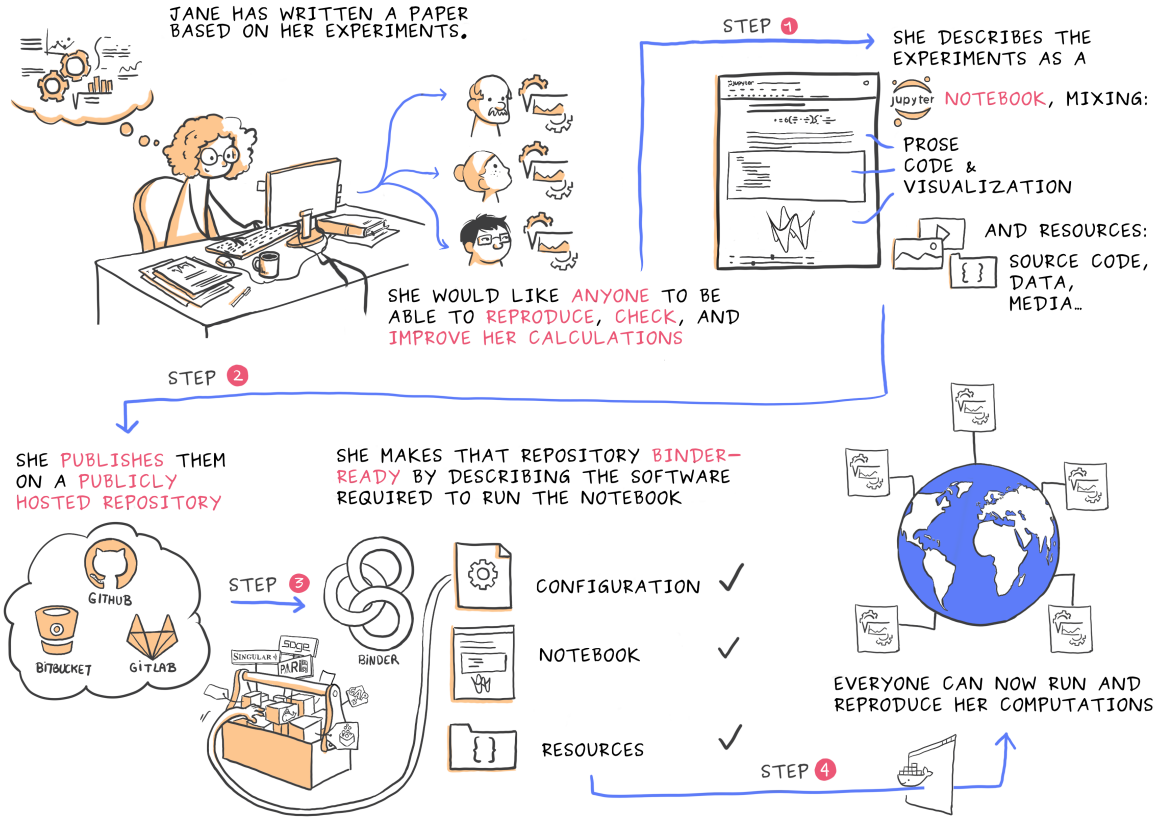


Figure 1.3.1: A typical use case for Jupyter notebooks in research. Image by Juliette Belin for the OpenDreamKit project, used under CC-BY-SA.

A large number of discrete software components make up Project Jupyter. While these interact with one another, many can be installed separately to serve various use cases. For this proposal, we loosely divide the software involved into *Jupyter core* developed under the guidance of the developers who started the project, and the broader *Jupyter ecosystem* including software developed by third parties, which may interact or build upon core Jupyter components. Some of the components and concepts important to BOSSEE are detailed below.

Jupyter core

- The **Jupyter Notebook** is the flagship application of Project Jupyter. It allows the creation of notebook documents, containing a mixture of text and interactively executable code, along with rich output from running that code. Figure 1.3.2 shows an open notebook including graphs from an audio processing example. Notebook documents are readily shareable, providing a popular way to describe and illustrate computational methods and tools. **JupyterLab** is the new, modular, extensible client application for Jupyter notebooks, but the document format, server, and user model are the same.

- **Jupyter kernels** are the backend software which allow Jupyter to execute code in many different programming languages. The **IPython** kernel is the reference kernel, supporting the Python programming language, and is developed by the Jupyter core team. Kernels for other languages are maintained by third parties

- **nbconvert** converts notebook files to a variety of other file formats, including HTML and PDF, so that the content of a notebook can easily be shared with people who don't have Jupyter software. nbconvert also powers **nbviewer**, a web service which provides static HTML views of publicly accessible notebooks.

- **JupyterHub** is a multi-user extension of the Jupyter Notebook. It runs on one or more notebook servers, for example at a research institution. Users can log in to author and run notebooks securely through their web browser, without needing to install any special software on their own computer.

Jupyter ecosystem

While Jupyter is a large, distributed, coordinated project, the wider community of Jupyter users develops a great deal of software with Jupyter integration, providing increased or domain-specific functionality, building on top of Jupyter, or integrating core Jupyter components in some aspect. We call this the **Jupyter ecosystem**. The broader Jupyter ecosystem includes many more projects than we will describe here, but a selection of projects which are relevant to BOSSEE includes:

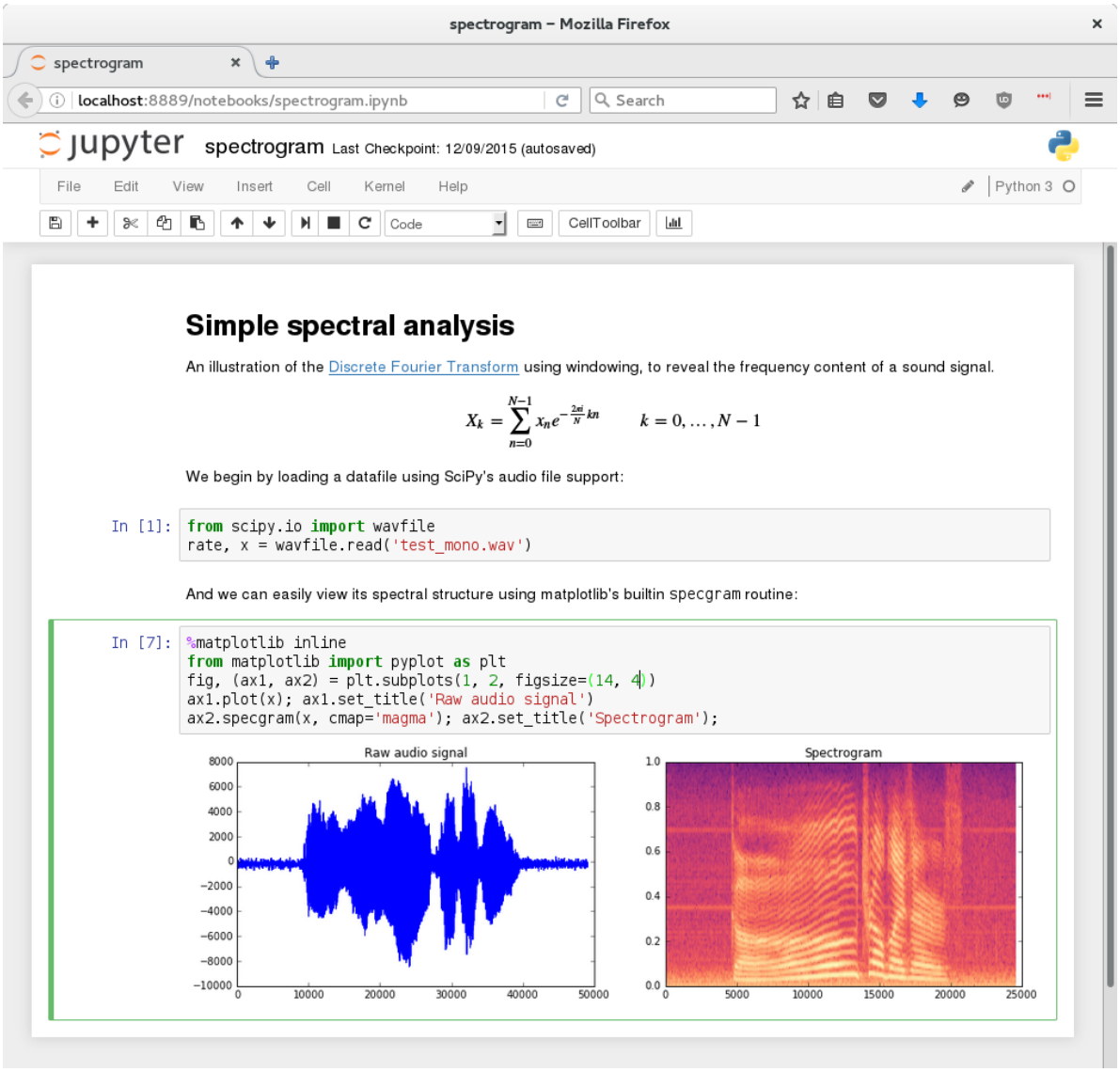


Figure 1.3.2: A notebook document in the Jupyter Notebook interface.

- **Binder** builds on JupyterHub to allow sharing executable environments along with data files and a description of the software components required to run the notebooks. When someone accesses a Binder repository, the service builds the computational environment on demand, allowing them to execute and modify a copy of the notebooks. **repo2docker** [For+18] and **BinderHub** [Jup+18] are components of the Binder software.

- **nbsphinx** [Nbs] integrates notebooks with the *Sphinx* documentation system, which is widely used for software documentation, especially but not only for software written in Python. This allows developers to write notebooks showing how to use their software, then seamlessly make those notebooks part of their main documentation.

- **nbval** [Nbv] is a plugin for the popular *pytest* testing framework to automatically execute notebooks and optionally check that the output matches that saved in the file. While this is not a subsitute for a test suite, it's valuable for documentation with code examples in notebooks. If changes to the underlying tools mean the example no longer works, testing with nbval will quickly show this, so that either the software or the example can be corrected. This ensures that example code and documentation don't get outdated.

- **nbdime** [Nbd] provides tools for comparing and merging notebooks. These integrate with version control systems such as *git*, which are designed for plain text files and typically don't handle notebook files well.

- **Widgets** allow interactive output in the notebook which can communicate with the kernel, updating values in the kernel and updating the displayed output as code runs. **ipywidgets** [Ipyc] provides the main implementation for the IPython kernel, while other packages such as **bqplot** [Bqp], **ipyvolume** [Ipyb] and **K3D** [K3d] extend the framework to provide 2D and 3D visualisations. Figure 1.3.3 shows a simple example of interactive widgets in use.

- The **Voila** package [Voi] enables the sharing of notebook-based interactive dashboards for non-technical users.

- The **Xeus** instrastructure [CM17] supports writing kernels in C++. **xeus-cling** is one such kernel, running user code in C++, and built upon CERN's C++ interpreter, "cling" [Vas+12], which has significant adoption in the High Energy Physics community. xeus-cling is already in use for teaching the C++ programming language.

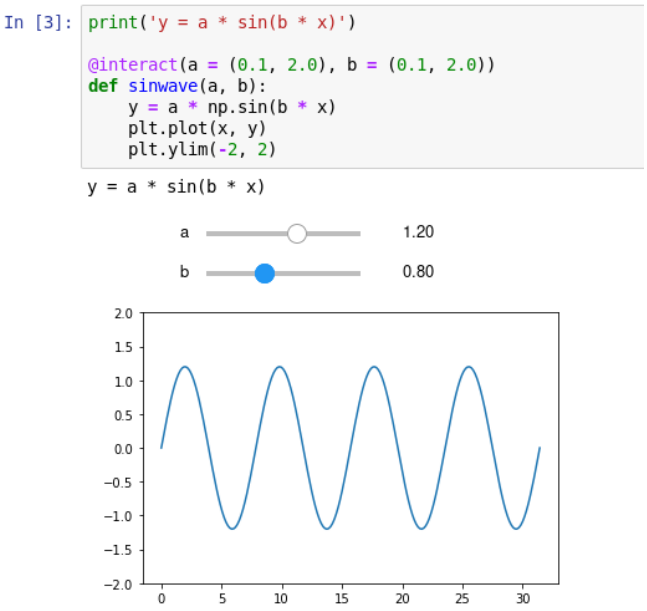


Figure 1.3.3: An example of using two simple slider widgets to explore the parameter space of a function. The `@interact` decorator creates the widgets and connects them to the function.

Jupyter as a basis for web services

Because the Jupyter notebook is a web-based application, it can be deployed at computational facilities or in the cloud, and can function as the basis for services exposing computational resources of all kinds to researchers and the public. Because Jupyter is **interactive**, it enables making scientific results and communications more interactive than static publications. The audience can follow their own initiative and ask their own questions of published data without needing support from the publishing author, greatly facilitating the **practicality of Open Science**.

Jupyter is generic

BOSSEE chose Jupyter because it is Generic. Jupyter makes no domain-specific or even language-specific assumptions. Any application where mixing description, code, and results is valuable can make use of Jupyter. This broad applicability makes investment in the Jupyter ecosystem extremely effective, because improvements to Jupyter can serve many communities simultaneously.

Jupyter is built from a collection of standard protocols and file formats. Jupyter is not just a single, monolithic piece of software, but a description of how such software can be built. The result is the ability for a variety of communities and applications to use components of Jupyter for their purposes, and/or reimplement pieces to meet their needs. For example:

1. The notebook file format is a well-specified JSON document, which can be interpreted by many systems. This has facilitated the development of different services providing rendering of notebooks, e.g. the code hosting website GitHub, which renders notebooks for easy viewing by anyone, without Jupyter software.

2. The Jupyter protocol describes how execution is performed, which has enabled the development of over one hundred kernel implementations in dozens of languages[1].

3. Output in the Jupyter protocol uses web-standard MIME types, enabling any possible format to be an output in a Jupyter notebook.

4. The JupyterLab extension system provides a system for building applications from Jupyter components and others.

5. The Jupyter Widgets provide a system for customizing and extending interactivity in Jupyter-based environments.

The popularity of Jupyter, with millions of users and hundreds of open source contributors, is an indicator of the value and impact of this approach.

Improvement to the Jupyter ecosystem

The benefits of focusing our work on a mature system like Jupyter include:

[1] https://github.com/jupyter/jupyter/wiki/Jupyter-kernels

- vibrant community ensures health and sustainability,
- large existing user base maximises impact of contributions,
- mature software ecosystem maintains quality software through industry standards such as version control, tests, continuous integration, stable release cycles, roadmaps, and user support.

The Jupyter community aims to be inclusive, and BOSSEE fully embraces and supports that approach. Jupyter is inclusive across a number of axes. By being applicable across numerous domains, Jupyter and BOSSEE encourage participation from individuals of various interests and backgrounds, and has taken action to improve diversity in the project by participating in "Outreachy," a program of paid internships for individuals from groups that face under-representation, systemic bias, or discrimination. Jupyter has also operated workshops focused on training contributors from under-represented groups. In being free, public, open source software, Jupyter and BOSSEE are accessible to as many individuals as possible, and invites users and contributors beyond origin, nationality, beliefs, orientation. One area where Jupyter has lacked in this regard is in the User Interface accessibility, and we will help improve this in **T2.3**. Additionally, the project will focus some of its workshops in **T6.2** on under-represented communities.

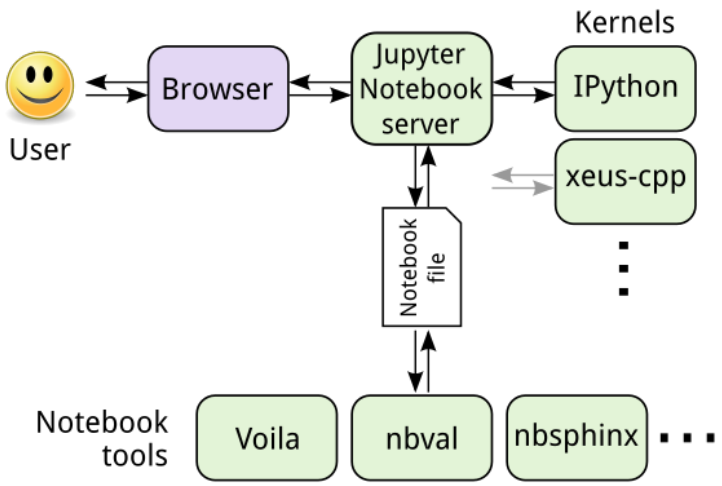


Figure 1.3.4: The architecture of the Jupyter Notebook, kernels, and tools which operate on notebook files

Related projects

EOSC-hub is a 33 million Euro H2020 project that started in January 2018 with the involvement of over 100 institutes. In three years the project is establishing the first elements of the European Open Science Cloud. EOSC-hub defines, creates and operates the integration and management system of the EOSC. This integration and management system (the Hub) builds on mature processes, policies and tools from the leading European e-infrastructures to cover the whole life-cycle of services from planning to delivery. Through this management system online and 'human' services, software and data are delivered towards researchers via a single EOSC Portal. The Marketplace already includes nearly 50 services from EOSC-hub provided by 3 e-infrastructure communities (EGI, EUDAT, INDIGO-DataCloud), and from 18 Research Infrastructures and scientific service providers. The catalogue of services is expected to radically grow in the next years through national, regional and EU initiatives.

Integrating Jupyter-based services into EOSC provides an excellent opportunity for facilitating interoperability of EOSC services, bringing data and computation together in a flexible environment.

1.3.3 Methodology

Proposed improvements to core components of Jupyter (WP2)

We plan to make technical changes to Jupyter software to better support real-time collaboration (**T2.4**), so that two or more people in different places or working on different devices can work together on the same notebook. This would significantly enhance the value of notebooks for collaborative research. We will also work on making Jupyter software accessible to as broad a range of users as possible (**T2.3**).

Further work to bring the code behind JupyterHub and Binder closer together (**T2.2**) will bring a range of benefits, allowing more flexible sharing of notebooks along with access to remote computing resources such as those available through EOSC.

Finally, we are explicitly allocating time in WP2 for maintaining Jupyter software, as well as new development (**T2.1**). Maintenance is crucial to creating reliable, sustainable software, but its cost is often swept under the rug in funding applications

because of the perceived pressure to focus on novelty. Being up front and explicit about this cost is critical to the sustainability of open source open science.

Proposed improvements to the Jupyter ecosystem (WP3)
We further propose improvements to the wider Jupyter ecosystem for better scientific workflows. In particular, we have identified possible improvements to:

- Binder and its crucial software component *repo2docker* (**T3.1**).
- Xeus, to better support the C++ programming language in notebooks (**T3.2**).
- Interactive widgets, including tools for 3D visualisation to help people make sense of large amounts of data (**T3.3**).
- Archiving of computational environments to allow reproducible research with a focus on the long term (**T3.4**).
- Tooling and guidelines for using notebooks in education (**T3.5**).

We may create new open source software projects in these tasks, but we will carefully review existing software, both in the Jupyter ecosystem and beyond, to avoid unnecessary duplication of effort.

Beyond the improvement to the Jupyter Project (WP4, WP5, WP6)
Beyond the improvement to the Jupyter core and ecosystem software for EOSC, we plan on

- Design, implementation, application, demonstration and evaluation of new innovative EOSC services in multiple demonstrators, that cover research fields such as health, astrophysics, photon and neutron science, geosciences and mathematics, and also interests of participating SMEs (WP4).
- Operating a *European Binder Service* on the EOSC-Hub and enabling provision of Jupyter Services through the EOSC-Hub (WP5).
- Producing *training and education material* to disseminate the ability to do reproducible computational science using the tools we develop, among others (WP6).

The science demonstrators

We describe the context and challenges for each demonstrater in this section. The particular planned activities are shown in the corresponding tasks in WP4.

Demonstrator: Astronomy (T4.2)
The Strasbourg Astronomical Data Center (CDS) is a scientific data center hosted by the Observatory of Strasbourg. The CDS plays a unique and essential role in astronomy by adding value to published and reference data. CDS runs astronomical services that provide data for the world-wide astronomy research community. Its three main services (SIMBAD, VizieR and Aladin) are heavily used with up to one million queries per day. These services can be accessed through web interfaces, mainly for human interaction, as well as through programmatic interfaces, including the standardized protocols defined by the International Virtual Observatory Alliance [Ivo].

Python and notebooks are rapidly increasing in importance for astronomy research. Indeed, Python for Astronomy software ecosystem has known a constant steady growth in the latest years, as shown in figure 1.3.5. As Python and notebooks integrate well together, the Jupyter notebook as an analysis tool is becoming a hot topic in the astronomical world: large surveys like the LSST (Large Synoptic Survey Telescope) have endorsed the usage of the Jupyter platform for their data access portal [JCDF17].

We will develop a Jupyter-based framework to efficiently access, explore, visualize and analyze reference data that are available through CDS services as a real example of using open astronomy data. We will provide scientific users with a set of customizable Jupyter notebooks for visualization and analysis tasks, providing a new level of interoperability with python libraries and notebooks as is highly demanded by the astronomy research community.

The focus is on the two following user stories:

- analysis of catalogue data results, up to billions of rows. Tabular data is the typical output of SIMBAD and VizieR data.
- modular dashboard-like interface providing a top level interactive view of the available data for a given astronomical object and enabling loading and analysis of those data.

Access to the notebooks will be provided as a one-click action option from SIMBAD and VizieR results pages. Thus, providing with a one-click way of visualizing, filtering and analyzing these potentially large tables will bridge the gap between access and analysis of the data, with zero installation for the user. For specific science cases, we will explore rendering of notebooks with interactive widgets through Voila [Voi], as to allow users not familiar with Python to benefit from the Jupyter notebook framework. Figure 1.3.6 depicts typical data objects we want to analyse and interact with in the notebooks: images, catalogue data, datasets coverages.

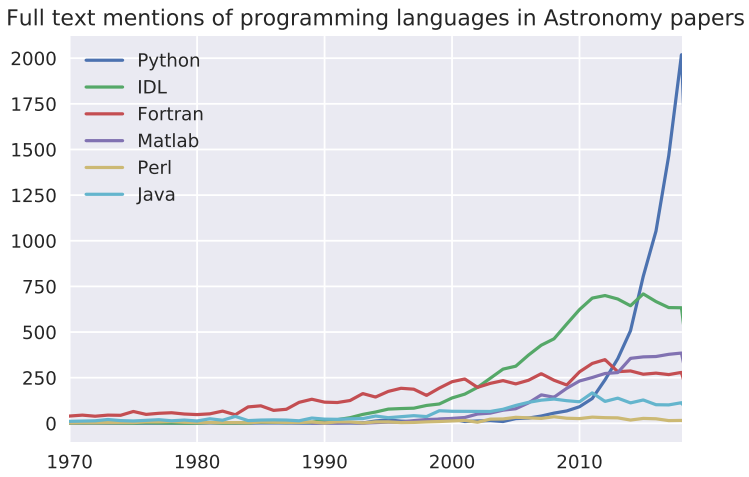


Figure 1.3.5: Mentions of programming languages in refereed Astronomy papers, extracted from ADS. Python usage has increased dramatically in the recent years.

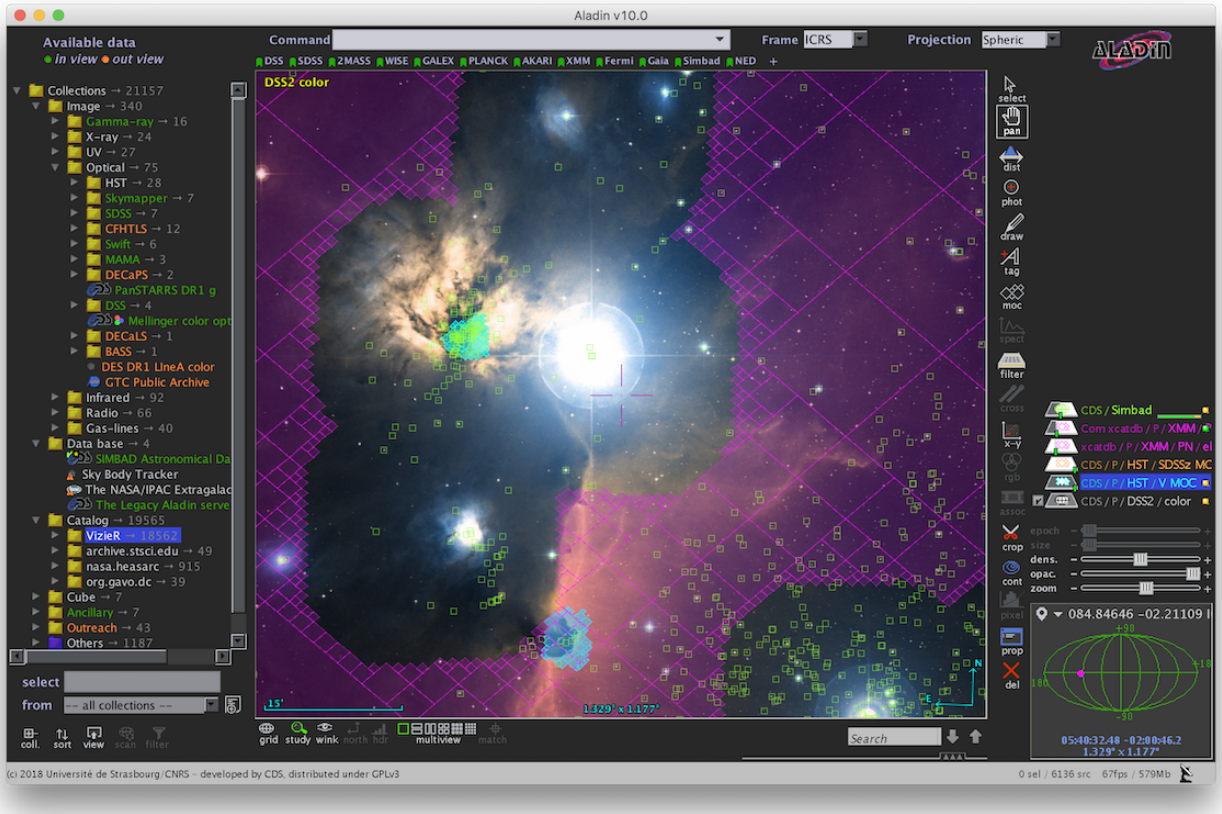


Figure 1.3.6: Example of astronomical data objects: Simbad sources, XMM and Hubble coverages overlaid on Digital Sky Survey imagery in the vicinity of the Horsehead nebula, and visualized in Aladin Desktop software.

These new developments will be highly visible to the large number of astronomers who use the CDS services (50,000 unique visitors per month) and such tools are in high demand by these users.

The CDS expertise in astronomy data and interfaces will be profitably combined with expertise of BOSSEE partners to ensure the deployment of high quality widgets (Simula, WildTree Tech, QuantStack).

The particular activities for this demonstrator are shown in WP4 in **T4.2**.

Demonstrator: Enriched education with Jupyter (T4.3)

In the recent years, Jupyter technologies have been widely adopted worldwide in higher education – and even in high

schools – for teaching in all areas of sciences. The Jupyter notebook indeed provides a very versatile environment – with a smooth learning curve – for authoring interactive material such as class notes, exercise sheets, dedicated applets; all the way to complete books such as those produced by OpenDreamKit for biology, physics, and mathematics. The interactivity engages the students to take an active role, for example playing with code, exploring the effect of tweaking the parameters in a simulation, changing visualizing tools, adding personal notes. This lets them take progressively ownership of the material and better understands the issues, and encourages them to create their own documents and share their experience with colleagues and teachers.

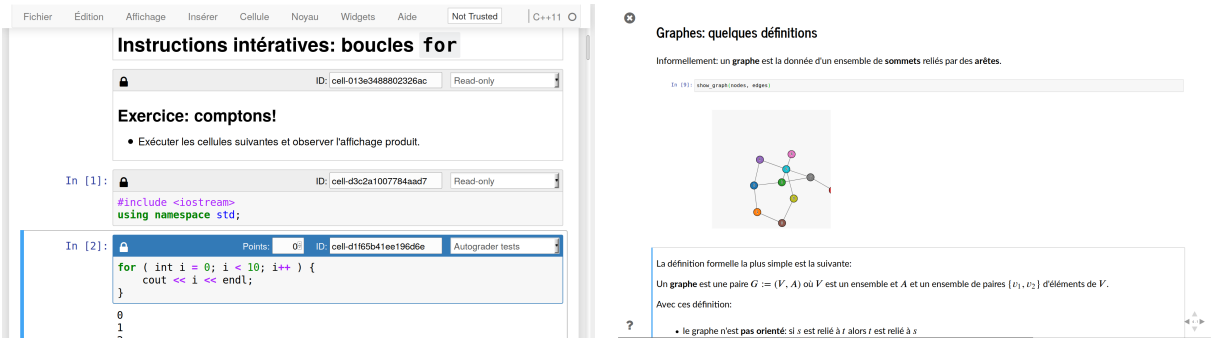


Figure 1.3.7: Jupyter based teaching material from Paris Sud. On the left: an exercise sheet for the course *Introduction to programming*; this instructor version showcases interactive C++ execution and automatic grading configuration menus. On the right: interactive slides for a graph theory course.

Success stories include:

- Berkeley's "Data 8: The foundations of Data Science" open course (data8.org) which is delivered yearly to thousands of lower undergraduate students in all majors, scientific or not,
- QuantEcon' open interactive book "Lectures in Quantitative Economics" (https://lectures.quantecon.org/), entirely authored with Jupyter.
- 100+ IPython/Jupyter based MOOC's (Massively Open Online Courses) on Coursera (https://www.coursera.org/courses?query=ipython)
- Paris Sud's first-year course "Info 111: Introduction to Computer Science" where each year 400 students write their first lines of code in C++ in a Jupyter notebook (see Figure 1.3.7).

École Polytechnique, Université Paris-Sud, and other participants from this project have been early adopters of these tools (see the description of EP and UPSud, and also task **T3.5**). We learned the hard way that deploying the Jupyter environment at a large scale (e.g. for a university) requires specialized expertise (DevOps, software development, ...) which impedes its adoption by the greatest number of people. High quality hosted solutions (e.g. CoCalc[Coc], Gryd[Gry]) do exist but are not the final solution when it is desired to exert greater control on private data, integration with the local infrastructure (authentication, shared drive, e-learning environment, dedicated hardware, ...), or to use available local computing resources rather than paid services.

Further improving the Jupyter environment for education, while leveraging it to the greatest degree, are therefore key motivations for the following tasks of this proposal:

- Tasks **T2.2** and **T5.3** will greatly ease the deployment of Jupyter environments, with tight integration in the existing local infrastructure and full customizability by the teachers.
- Task **T3.5** will improve the interoperability with existing e-learning systems, and further develop teaching aids for, e.g., material sharing, (self)-evaluation, and grade management.
- Task **T4.7** will support teaching in mathematics through better support for real-time interactivity.
- Task **T3.2** will support teaching in computer-science and scientific programming through better C++ integration in the notebook and will allow to first class students to focus on the syntax of the language without distractions such as compiling and linking a program.
- Task **T5.2** will ease publication and FAIR access to course material, which in turn will promote sharing and collaboration in the education community.

The particular activities for this demonstrator are shown in WP4 in **T4.3**.

Demonstrator: Visualisation and control of fluid dynamics in Jupyter notebook (T4.4)
In recent years, the lattice Boltzmann method (LBM) emerged as an interesting alternative to more established methods

for fluid flow simulations. Sailfish-cfd [JK14] is an open source implementation of the LBM on General Purpose Graphical Processing Unit (GPGPU) devices. It is written in Python with real-time generation of CUDA-C code. In order to harvest capabilities of GPGPUs one needs to access the specialized hardware, which usually is available to researchers as remote HPC resources. The typical fluid dynamics research workflow consists of three stages: preparing boundary conditions, running a simulation, and data analysis. The first and last stage require capable and responsive user interface for maniputation and inspection of 3d data. The Jupyter 3d visualisation widgets developed in **T3.3** can fulfil such needs.

Based on previous experience with K3D-jupyter [K3d] widgets we know that web browser based software can display moderate dataset during the simulation. As the dataset is becoming larger the visualisation in the browser turns out to be nontrivial due to limitations of the browser itself and required large data transfers. It is an open question how much of data processing should be performed on server-side and what can be done on the client hardware (i.e. in the widget in the browser side of the user). Our experience suggests that there is no clear answer and it depends on the size of the data and its nature. For example, volume rendering technique can be very effective on the browser side but infers large data transfers. One can perform it the server-side, in a distributed way if the simulation uses many nodes, but the interactivity is limited by network latency. We will attempt to provide practical solutions to this issue.

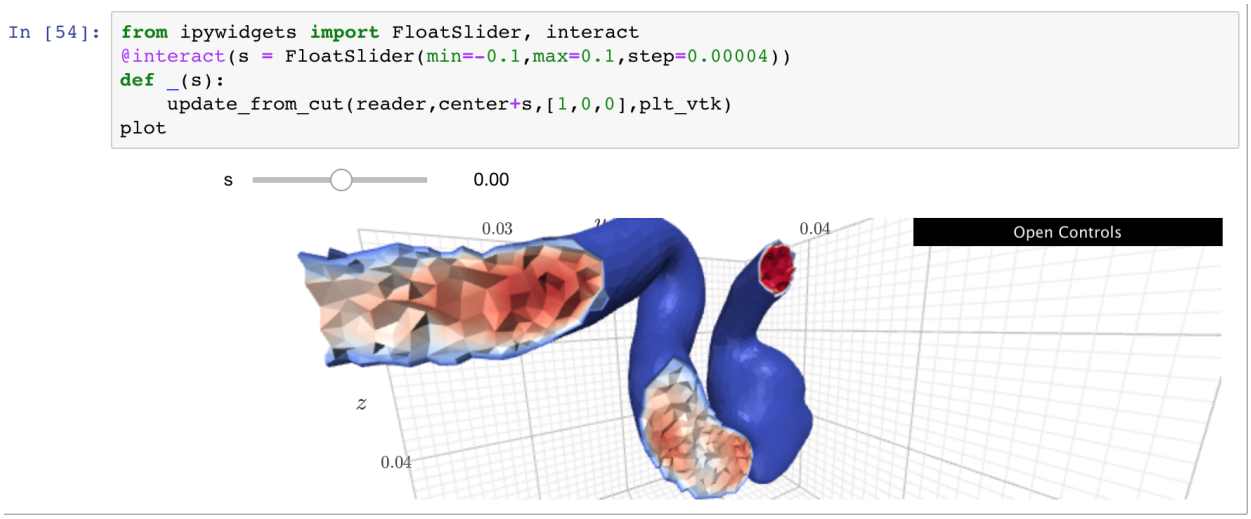


Figure 1.3.8: An example showing Jupyter based visualization of velocity magnitude in the blood flow through an aneurysm. It demonstrates the use of small interactive widget for selecting the cutting plane. Visualization is done by a K3D-jupyter widget [K3d]. In this case most of computations are done using VTK library on the server side, and the K3D-jupyter widget is used to display the colored surface mesh.

The particular activities for this demonstrator are shown in WP4 in **T4.4**.

Demonstrator: Geosciences (T4.5)
The amount of geospatial data from a variety of sources, including satellite observations, 4D simulations and in-situ observations, contributed by volunteers or state agencies keeps increasing. In many disciplines, managing this large volume has become a challenge, and the old approach of downloading datasets for local analysis has become intractable.

The heterogeneity of the tools used in different institutions to deal with large geographical datasets makes it difficult for researchers to share the outcome of their work in a reproducible or interoperable fashion.

In this context, Jupyter is now emerging as a standard exploration tool for geospatial analysis, climate science, geology and by data providers in these areas.

To mention a few,

- the *PanGeo* platform [HRA18] (Funded by the NSF, NASA, and the Alfred P. Sloan Foundation) is built upon Jupyter, JupyterHub, Binder, and Dask.
- the *Joint Research Centre Earth Observation Data and Processing Platform* (JEODPP) [Soi+18] relies on Jupyter, JupyterHub and ipyleaflet as its main user interface (see also figure 1.2.1 on page 5).
- the *Google Earth Engine* platform also offers a Jupyter-based user interface allowing the visual exploration of the data with ipyleaflet [Eri+17].

In these three cases, deferred processing is used to restrict computation to the extent of the area displayed in the map viewer, which allowed these platforms to scale up to petabytes of data. In all examples, interactive visualization is a key feature of the platform. Beyond tile-based 2-D visualization, the ability to efficiently process and visualize vector or 3-D data is also becoming critical.

The BOSSEE team, which comprises the main authors of the technologies upon which these platforms are built (Jupyter, JupyterHub, Binder, ipyleaflet), together with the Department of Geosciences of the University of Oslo, are in a unique position to bring these technologies together in the context of EOSC.

This demonstrator will focus on tools for two transversal research projects

- LATICE (Land-Atmosphere Interactions in Cold Environments)
- EarthFlows (Interface Dynamics in Geophysical Flows)

The work items for this demonstrator fall in two main categories: visualization and geographical data processing tools. Data will not be produced as part of BOSSEE. The University of Oslo follows the ”open as standard” policy and all data used for BOSSEE will be publicly available through:

- the Norwegian Research Data archive
- Zenodo for smaller datasets and for datasets used for teaching
- SQL requests for in-situ observations stored in local databases (for instance collected during field campaigns)
- the Earth System Grid Federation (ESGF) for all climate data,
- Copernicus data portal for satellite observations.

Beyond their use in scientific research, these development will be used in the class room for teaching master’s students with best practices in open science.

The particular activities for this demonstrator are shown in WP4 in **T4.5**.

Demonstrator: Nuclear Medicine dosimetry (T4.6)

Nuclear Medicine is a field of medicine where radioactive material (radiopharmaceutical) is used for diagnostic and therapy. The OpenDose project [Cha+17] is a collaborative effort to generate a reference database, freely available, proposing dosimetric data applicable in a context of nuclear medicine dosimetry. A major aspect of the project is the development of tools ensuring traceability and reproducibility of generated results.

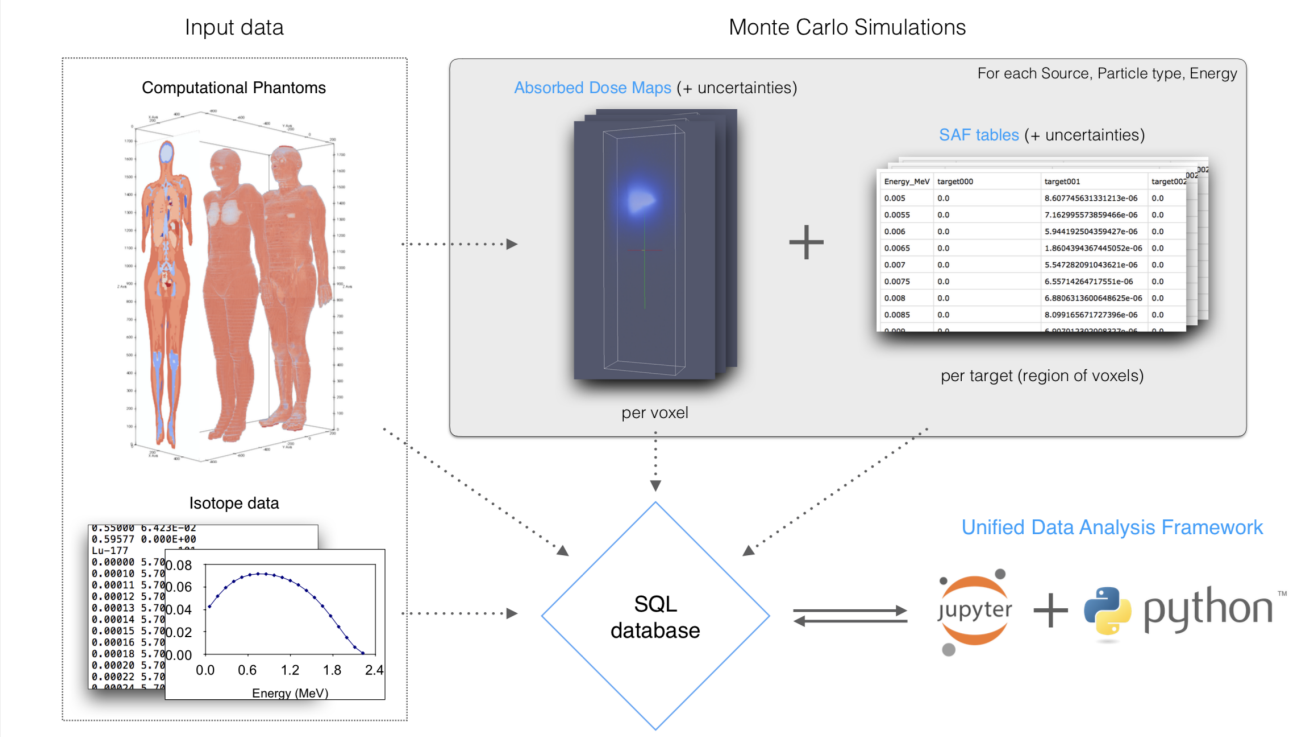


Figure 1.3.9: OpenDose project overall framework including the unified data analysis to be developed in this demonstrator.

OpenDose data is produced using the five most represented Monte Carlo simulation software tools in medical applications: Geant4/GATE, MCNP, EGS, PENELOPE and Fluka. Each simulation consists of calculating radiation transport in anthropomorphic models for specific parameters (source organ, particle type, energy, model and number of primaries to simulate). Every simulation produces binary (3D matrices) and ASCII files for a total of ∼150MB / simulation. The 3D matrices contain energy deposited per voxels, and ASCII files contain pre-processed data corresponding to energy deposited per regions such

as organs and tissues. These raw outputs are later processed into dosimetric data such as Specific Absorbed Fractions (SAFs) and S-values.

Producing data for one model (ex. Adult female) requires ∼30,000 simulations, with the workload shared between the different teams and software.

The data produced by all the teams is currently centralised at the Cancer Research Center of Toulouse (CRCT), processed and fed into a local SQL database at CRCT.

This collaborative effort raises some challenges:

- Data production: a total of 750,000 hours of CPU time is needed per model.
- Volume of data: one model represents TBs of raw data that can be heterogeneous from the different teams.
- Data analysis: raw data has to be processed into dosimetric data in a robust and reproducible way.
- Database: has to be efficient and handle all the data (raw and processed).
- Visualization: display and compare results from all teams.

Figure 1.3.9 shows the overall framework of the project and how data will be managed.

By building a set of tools to access and process data within the Jupyter ecosystem, we will ensure the production of traceable and reproducible dosimetric data for the OpenDose project members.

Another major aspect of the OpenDose collaboration is to provide open access to the generated dosimetric data. For that purpose a website is under development to allow data download and simple dosimetry calculations. For users who need more advanced calculations, a dedicated Jupyter workspace will provide a set of tools to easily access, process and display the OpenDose data.

The particular activities for this demonstrator are shown in WP4 in **T4.6**.

Demonstrator: Interactive Mathematics with Jupyter Widgets (T4.7)

Computations have played a long time and ever increasing role for research and teaching in (pure) mathematics, to explore, search and check for conjectures, or better understand algorithmic ideas. This led to the development of a whole ecosystem of mathematical software, many of which are open source. Given the huge variety of mathematical objects and workflows, the Read-Eval-Print-Loop (REPL) paradigm – on which Jupyter is based – is particularly suitable: the user interacts with the system by typing commands that use its library of mathematical features, often combined with personal code. In fact, the REPL and notebook paradigms of Jupyter as well as some of its interactive features were largely inspired by that of computer algebra systems such as Maple, Mathematica, or SageMath.

One major action of the OpenDreamKit project was to foster the convergence between the Jupyter and math software ecosystems: nowadays Jupyter can be used as a uniform user interface for most major systems: e.g. GAP, OSCAR, Pari/GP, SageMath, Singular, and even for C++ libraries. This interface is being widely adopted: for example, Jupyter has become the standard user interface for SageMath, enabling to phase out its former bespoke notebook; by now, thousands of Jupyter notebooks for SageMath are publicly shared (6000+ on GitHub alone).

Thanks to this prior work, the mathematical community will immediately enjoy all the benefits brought by EOSC-based generic Jupyter services, including eased collaboration, sharing, archival, and reproducibility.

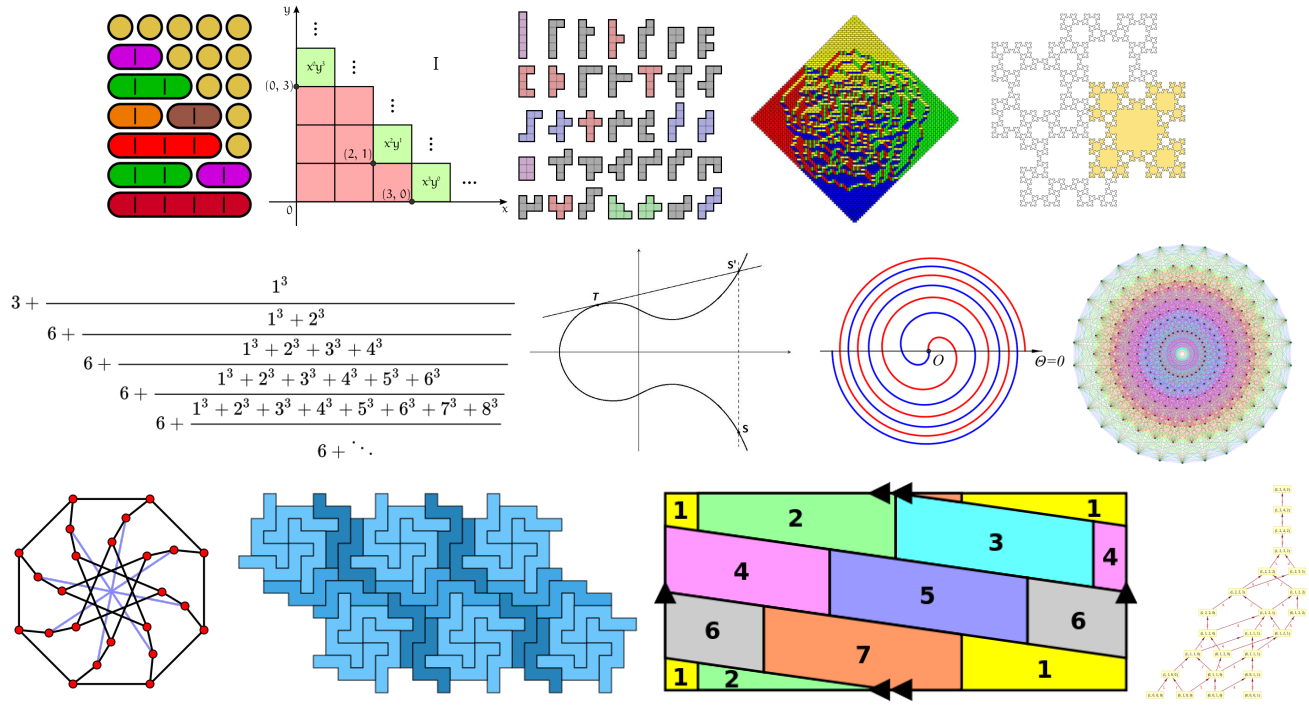


Figure 1.3.10: Graphical visualizations of a variety of mathematical objects

The next step to maximise attractivity and impact in the mathematical community, and this is the aim of this task, is to go beyond the REPL paradigm, and **leverage the real time interactivity and flexibility brought by Jupyter widgets for Mathematical purposes**. This will, for example, make it straightforward for a teacher or researcher to build and disseminate via the EOSC mini applications or dashboards enabling the graphical exploration of a whole range of mathematical inputs, with real-time visualization of the associated outputs.

The unique challenge comes from the huge variety of mathematical objects that the user may want to visualize and interact with, and the variety of graphical representations (see Figure 1.3.10). Co-design is central here, as building a bespoke interactive visualization entails a combination of technology skills (e.g. javascript development) and business knowledge (designing the interaction and visualization). The role of Research Software Engineers is to leverage the technology by encapsulating the technical difficulties into flexible and easy to use tool boxes from which mathematicians can build mini-applications as innovative services that are tailored to their needs.

Within OpenDreamKit, we conducted experiments to explore this venue [BT18]. One specific focus was to enable not only *interactive visualization*, but also *interactive editing*: being able to graphically modify the mathematical object being visualized; this enables the interactive exploration of how the modifications affect its properties, or to use the editor as an input widget for a larger application or dashboard. The outcome of this task is the development of two prototypes in SageMath (sage-combinat-widget, a library of widgets for combinatorics, and sage-explorer a generic dashboard for interactive browsing and introspection of mathematical objects), and contributions to Francy, an Interactive Discrete Math Framework for GAP and SageMath.

The particular activities for this demonstrator are shown in WP4 in **T4.7**.

Demonstrator: Reproducible photon science workflows at European XFEL (T4.8)

European XFEL is a research facility that provides X-ray Free Electron Laser (XFEL) light to image structures at the nanoscale. It is currently the world's most brilliant laser, created in a 3.4km long tunnel, and supporting user experiments since September 2017. These imaging capabilities of European XFEL and similar services available from synchrotron and neutron sources, underpin lots of fundamental and applied research, in domains ranging from physics and material science to biochemistry and drug design. Some example data is shown in figure 1.3.11.

All of the data recorded at European XFEL will be made freely available after an embargo period of three years [Eux]. This provides scientific transparency and is expected to enable better exploitation of the data, as more researchers than those conducting the experiments have access to the results. If the analysis steps are not carefully recorded, there is a risk that the necessary understanding of the data is lost by the time it is made public or subsequently, greatly reducing its scientific value.

We are keen to complement this open data access to the actual data with open access to reproducible data analysis, to confirm conclusions drawn and to significantly lower the barriers for re-analysis with new tools or for new research purposes.

A task in the EC funded project Photon and Neutron Open Science Cloud (PaNOSC) is using the Jupyter Ecosystem tools as they are in 2019 to provide interactive data analysis services to complement the data: through use of Jupyter Notebook and exploitation of the mybinder.org service, this activity will reduce the barrier for interactively exploring the data, understanding and making use of the data, and to do this through a central portal such as EOSC.

Here, we combine and use the new developments (WP2, WP3) of this proposal to enable new qualities of open science services, and to demonstrate the potential impact of these improvements for a wide set of EOSC services through a demonstrator in Photon Science.

Context: The very first experiments at European XFEL produced as little as 45 terabytes of data on average, but as the facility develops, the amount of data produced per time is expected to grow substantially: Given the rate of light pulses, there is the potential to produce up to a petabyte of data within the beam time of one experiment (typically one week). These significant amounts of data need to be complemented by complicated workflows to convert the data into insight through data analysis. Derived results of such data analysis are typically much smaller in size and useful to archive together with the raw data. To explain how they have been obtained, the particular workflow of data analysis also needs to be archived.

Vision: At European XFEL, it is proposed to use Jupyter notebooks to facilitate this workflow: the simplest model would be to use one notebook per workflow. Once the data capture from the experiment is completed, this notebook can be executed (without being displayed in a web browser) to start processing the data. When the notebook has completed execution, it is saved, and contains the analysis results (it may of course also created files on disk as part of the process).

A particularly useful aspect of the notebooks is that they mix data analysis commands with outputs, and that the notebook provides a complete (and thus reproducible) summary of the data analysis when it succeeds with the execution. Should the execution fail, for example half-way through the notebook, then derived results obtained prior to the error occurring are preserved and can be inspected. The error is embedded in the notebook and appears after the command that has triggered the error; which helps with debugging the process.

This is of particular interest as the data analysis processes at European XFEL may fail not because of software errors but due to variation in the data that require (manual) expert adjustments of parameters. The "failure" of such an analysis workflow (represented through the Notebook) is thus not exceptional, but a common occurrence. The scientist conducting the experiment is sufficiently skilled to modify the parameters and wants to either re-execute the notebook from the beginning or to continue from the point of failure. The notebook caters for both use cases. The modified notebook would need to be preserved of course to provide reproducibility of the derived results that the notebook has computed.

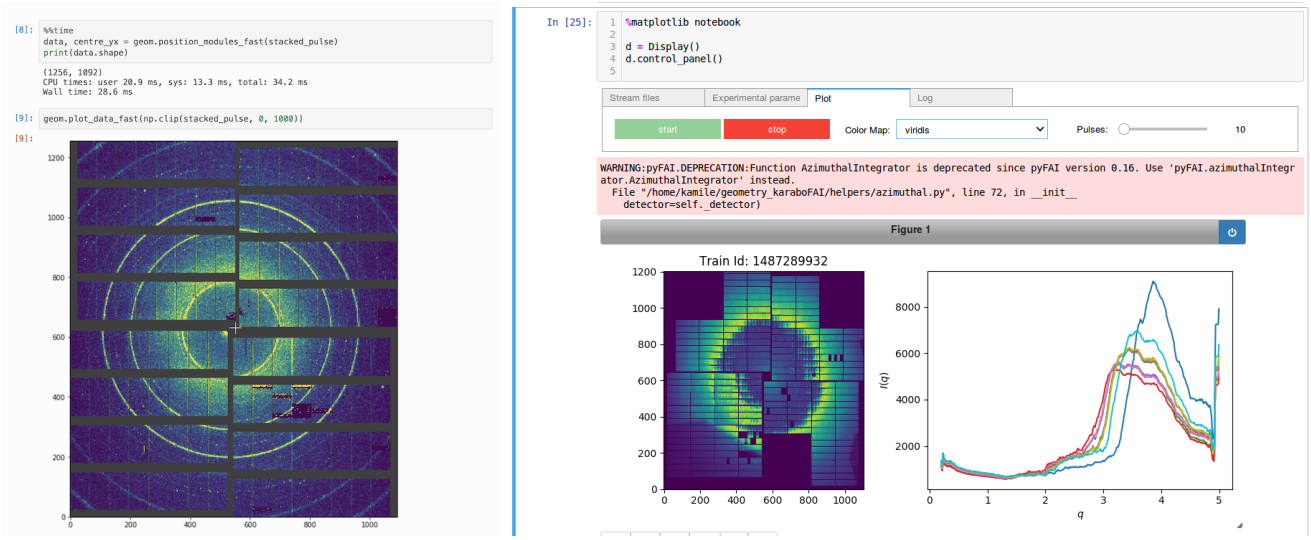


Figure 1.3.11: Prototypes for data analysis of 2d x-ray detector images in the Jupyter notebook, relating to the photon science use case. *(Left)* Data from crystallography scattering experiment. *(Right)* Azimuthal integration of detector data as one step in the data analysis workflow.

We are aiming for re-executability of the notebook for the lifetime of the data. The lifetime of the archived data at European XFEL is currently guaranteed for 5 years and aimed to be 10 years [Eux]. It is possible though, that data used for publications will be preserved for longer, and it would be highly desirable to keep the data analysis re-executable for the same period of time, potentially well exceeding 10 years.

The particular activities for this demonstrator are shown in WP4 in **T4.8**.

1.4 Ambition

BOSSEE's ambition is to **improve the global accessibility** of scientific tools, results, ideas, data and data analysis; **enabling collaboration** among researchers and between researchers and the public.

The world's computational resources are constantly growing and science is producing ever-more useful and interesting data. But **how do we enable the European and global communities to make use of that data?** And **not just researchers**, but the public as well? Open science is a principle of making research results as broadly accessible and useful as possible. The first, minimal step for this is making publications free to access. The second step for computational research is to make code and data publicly accessible, enabling transparency and facilitating reproducibility and verifiability of results. But **merely making these resources technically available is not the best we can do**. There can be many challenges with software, such as environment specifications and resource requirements, that can be an impediment to the transition from 'technically available' to 'practically useful.' With the tools of the open source open science community and the resources of the European Open Science Cloud, we can do better.

The Jupyter ecosystem consists of a **large, global community** of developers and researchers producing software focused on interactive computation and communication, and is **widely used by millions** of individuals in numerous scientific fields, ranging from molecular biology [WM16] to materials science [Hug+14], astronomy [BP17] and climate science [Lak15b; Lak15a]. Jupyter software is permissively licensed under the Berkeley Software Distribution (BSD) license, allowing anyone to use Jupyter software for free, and even build derivative commercial products, as has been done in the cases of Google Colaboratory, Microsoft Azure Notebooks, IBM Watson Studio, and others. By contributing to the Jupyter ecosystem, BOSSEE maximises its impact, immediately benefiting the existing large Jupyter community, and increasing the likelihood that BOSSEE's results will be maintained by the community after the end of formal funding.

When it comes to Jupyter and Open Science for EOSC, we aim to improve the *status quo* by bringing the two together:

- Improve software in the Jupyter ecosystem to **better serve Open Science**.
- Enable researchers and the public to **better perform and benefit from Open Science**, through software, services, and education.

Open Science that truly benefits society must be more than merely technically accessible. Individuals must be able to find the resources they want and interact with them. Ideally, they should be able to ask new questions of models and data published by those that came before. This is where BOSSEE fits in. Excellent research is being performed in all scientific fields, but making those results practically accessible and engaging to others is a challenge. Jupyter notebooks enable publishing code and data in a form that is interactive, where readers can see code, run it, and see results. They can then modify the code and produce new data and charts that the first authors may not have considered. Jupyter does not solve the software installation problem, however, which can be significant for scientific software. For a publication to be truly interactive or reproducible, it must include a computational environment (or a sufficiently precise description of one such that it can be recreated) in order to reliably be able to run for another individual. Services and tools such as Binder and repo2docker have started to demonstrate that this can be facilitated: by publishing a description of the requirements to run the software, repo2docker is able to recreate a computational environment with everything needed to run the software, including a Jupyter environment for interactively exploring the resource. Binder wraps this in a web service, enabling immediate, free sharing of computational results on the web, with no requirements of readers other than a web browser. By developing these tools further, and deploying Binder or similar services on EOSC infrastructure, EOSC enables researchers to (i) **make their results available** through EOSC, (ii) allows other researchers to **extend that research within minutes** or hours, and (iii) **enables the public to interact with the science they are funding**.

By cooperating with specific applications from diverse domains in science and education, we ensure and demonstrate that the work is valuable to a broad community of researchers, students, educators, and the public. All together, Jupyter and Binder enable the migration of the open science community from static publication to **truly interactive, reproducible publications**.

The Jupyter notebook application is already TRL 8, while the Binder software and service prototype at mybinder.org is TRL 6. We will bring Binder to at least TRL 8 during the course of the project.

2 Impact

The central impact of the BOSSEE project will be a significant improvement and extension of the Jupyter tools and ecosystem described in section 1.3.2 to facilitate open science and reproducible research, and their wide availability on the European Open Science Cloud. While this will initially be developed alongside applications for our own institutions, we expect the tools we develop to be useful to a much larger group of researchers, as data analysis and simulation are now crucial parts of most academic disciplines.

Jupyter-based technology will be especially valuable given the decentralised nature of EOSC. Large-scale experiments often produce so much data that it is impractical to transfer the data to another site for analysis. At European XFEL, for instance, hundreds of terabytes of data may be recorded for a single user experiment. The analysis steps thus need to run where the data is stored, even if the scientist has returned to their home institution far away. As the Jupyter Notebook interface runs in a web browser, it is readily usable for remote access.

In addition, BOSSEE will produce highly visible demonstrations of notebooks for open science in targeted scientific disciplines. The result will be innovative new prototype services that will provide direct benefits to the early adopters in various research fields. Moreover it will serve as a demonstration of a strategy for open science using notebooks that will be applicable across many domains.

EGI supports the e-Infrastructure uptake of several ESFRIs in the EOSC-hub project. Several of these already experiment with Jupyter (either using the EGI's JupyterHub service, or community specific installations). These communities will be reached out to by EGI directly in the EOSC-hub project, and through the annual events organized by EGI around EOSC: EOSC-hub weeks, the DI4R conferences and EGI conferences. During these events EGI will contribute to promote the adoption of the project results.

2.1 Expected Impacts

The expected impact of BOSSEE with respect to the work program is detailed in the table below.

Expected impact	
Integrating co-design into research and development of new services to better support scientific, industrial and societal applications benefiting from a strong user orientation	The Jupyter tools have always been driven by a close connection to users; since the project began as IPython in 2001, many of the developers have been scientific researchers using the tools as they developed them. More recently, when Jupyter has benefited from dedicated developer time, developers have remained in academic institutions, in the kind of role now referred to as 'research software engineers', allowing day-to-day interactions with researchers using Jupyter in a wide range of fields. By supporting developers in various research institutions where the improvements will be used as they are developed, BOSSEE will continue this invaluable collaboration. The improvements and extensions of the core parts of the Jupyter system are being co-designed by technical, industrial and scientific experts in the BOSSEE project, so that they will be widely applicable in new innovative services across most academic disciplines. The impact of this approach for enabling scientific use of notebooks is expected to be very high because it is a direct response to the strong demand from scientists for improving the productivity and reproducibility of their work. The notebook approach is being embraced in many scientific disciplines, so the proposed services to be developed in BOSSEE are strongly oriented to the user needs.
Supporting the objectives of Open Science by improving access to content and resources, and facilitating interdisciplinary collaborations	Jupyter notebooks have seen rapid uptake in many kinds of research, because they bring together the essential elements of the modern scientific computational workflow (from data collection to publication and open sharing) in the familiar format of a scientific notebook, with powerful functionality for access to scientific content for analysis and visualisation. Notebooks also embody the core concepts of open science by providing a mechanism to reproduce results in publications, and collaborative sharing of not just scientific results, but of the code that produced them. We expect the use of notebooks in EOSC to improve access to scientific code: digital documents and notebooks encourage publishing workflows, whereas code in scripts or manual interactive workflows are often kept by the researchers who performed them. The focus on clarity and reproducibility also helps to ensure that data is meaningfully accessible, by preserving essential understanding to make sense of the raw data. We have already seen a good example of the Jupyter ecosystem facilitating an interdisciplinary collaboration: the LIGO scientific collaboration shared notebooks detailing the data processing steps which led to the discovery of gravitational waves, using the Binder service to allow anyone to re-compute the published plots. Scientists with no background in gravitational waves studied these notebooks and improved the signal processing. In this proposal, we want to provide this ability to a wider audience through EOSC, including for disciplines which rely on processing much larger volumes of data [Lig]. The astronomy application in BOSSEE (**T4.2**) is designed to provide a new level of interoperability of reference astronomy data within Jupyter notebooks. By connecting new notebook capabilities to existing and highly used services, we expect to have impacts for the users and also for the service provider. The scientific users will have access to new capabilities, and we anticipate adoption of new innovative ways of using the data. We also expect an impact on the services themselves, in terms of usage, but also in terms of capturing precious information and feedback on how to evolve these services to best support open and collaborative use of the data and services.
Fostering the innovation potential by opening up the EOSC ecosystem of e-infrastructure service providers to new innovative actors	Jupyter is a collection of open source software built around openly documented protocols and formats, along with familiar technologies such as HTML and the Python programming language. It's easy for third parties to create new tools and services using and integrating Jupyter, as evidenced by the thriving ecosystem of tools already in development, both by commercial and non-commercial actors. To highlight just one example, the first version of the popular Binder service was developed by a group at the Howard Hughes Medical Institute, working independently of the core Jupyter maintainers, but building on the powerful capabilities provided by Jupyter. By bringing the diverse expertise of the BOSSEE partners together in this common project, we expect a high impact in terms of enabling a new level of integration of scientific, technical and industrial interests for the common goal of open science, and building a toolkit from which others may build innovative services, for commercial or public use.

2.1.1 Measuring impact

As we are building tools and services for Open Science, the best measure of our impact is in the adoption and use of these tools and services, which can be observed qualitatively (positive anecdotal feedback) and quantitatively (counting visitors to a service, for example). Much of our work will be in the form of contributions to existing public projects, such as Jupyter and Binder, which can be measured in our participation in those projects, such as code and documentation contributions, bug reports, and roadmap contributions.

We can measure our progress toward aims and objectives in 1.1 via the following Key Performance Indicators (KPIs):

KPI 1: Attendees at Open Science workshops organised by BOSSEE participants.
KPI 2: Open publications for which the authors have made a reproducible version available through BOSSEE services.
KPI 3: Visitors to BOSSEE services, engaging with open, interactive communications.
KPI 4: Publications and presentations by BOSSEE documenting the use of BOSSEE services for Open Science.
KPI 5: Contributions by BOSSEE and the wider community to Jupyter software and others, including issues reported, bugs fixed, features added, and roadmaps developed.

2.1.2 Barriers, Obstacles and Framework conditions

The BOSSEE project will certainly face a number of challenges as it undertakes the ambitious program of work described by this proposal. We can identify a number of potential barriers and obstacles but overall these are assessed to be minor and planning is in place to mitigate the identified risks.

While a number of the partners have worked closely together in previous projects, the integration of new partners from different disciplines will require dedicated efforts for communication within the project.

A detailed assessment of risks and mitigations can be found in 3.2.7.

2.2 Measures to maximise impact

BOSSEE is contributing to tools for Open Science and for building and operating Open Science services. Tools only have impact if and when they are used, so it is important that we disseminate our work in order to reach and support user communities for our software and services. This section outlines how the project will establish and organise the dissemination and communication actions to promote the project and the adoption of its outcomes beyond the project's lifetime.

The dissemination and communication plan is outlined in the following sub-sections. Therein we distinguish:

- Dissemination as the public disclosure of the results of the project through a process of promotion and awareness-raising right from the beginning of a project. It makes research results known to various stakeholder groups (like research peers, industry and other commercial actors, professional organisations, policymakers) in a targeted way, to enable them to use the results in their own work.

- Communication as the strategic and targeted measures for promoting the project and its results to a multitude of audiences, including the media and the public, and possibly engaging in a two-way exchange. The aim is to reach out to society as a whole and in particular to some specific audiences while demonstrating how EU funding contributes to tackling societal challenges.

2.2.1 Dissemination and exploitation of results

WP6 is focused on dissemination of BOSSEE work. Our goal is to facilitate Open Science through the development and use of open and freely available tools. All BOSSEE software will be made publicly and freely available under open source licenses, and hosted on public code hosting sites such as GitHub. Most BOSSEE work will be in the form of contributions to existing projects, which will be governed by the licenses of those projects. All Jupyter and Binder software is released under the permissive BSD license, which specifically allows commercial exploitation, as has proven successful in enabling collaborations with industrial partners such as Google, Microsoft, IBM, and more. This means that all BOSSEE software will be available and accessible to all who find it, at no cost to BOSSEE, enabling long-term access beyond the funding of BOSSEE. Similarly, non-code products such as dissemination works (workshop materials, etc.) will be made freely available under open Creative Commons licenses.

As a result, the primary dissemination effort is to:

1. make sure that prospective users are **aware of the work**, and

2. enable them to use the tools through **learning resources, training, and services**.

Our focus for dissemination will be on **T6.2**, operating workshops, training various communities in the availability, purpose, development, and use of BOSSEE software and services. We will make a particular effort to use these workshops as an opportunity to **support diversity and inclusion in the Open Science community**, by running workshops for under-served and under-represented groups in the academic and open source communities. Additionally, for long-term resources available to the wider community who will not be able to attend workshops, we will produce **free, online materials for training** in the

use of BOSSEE software and services. These resources will be hosted on free, public hosting services, such as GitHub Pages, enabling long-term access to the work of BOSSEE, even after the end of funding.

The operation of services in WP5 is also a dissemination activity, as services like Binder not only enable Open Science by facilitating interactive publications, they also enable **interactive demonstration of tools and functionality** developed in BOSSEE, e.g. Xeus (**T3.2**). We have budgeted € 5000 per month for the operation of services in WP5, to be spent by EGI on cloud computing resources for service hosting. This cost estimate is based on the operation costs of mybinder.org. The involvement of hosting provided via EGI helps the project reach a sustainable setup, because we can negotiate hosting conditions with institutes that are dedicated for the support of open science and can co-fund the operational cost from their budget. BOSSEE, in collaboration with the operators of mybinder.org, will explore sustainability plans for covering long-term costs of operating such services, including institutional subscription models, donations, and others.

Data management plan

Except for the usage data described below, BOSSEE activities will not generate or collect data. While we have many demonstrators that interact with data, they do not generate or collect that data themselves, but rather provide analytical mechanisms or access to data governed by existing data management plans and data policies of project partners at each site, as well as publicly accessible open data.

Service usage data

Any data collected through the operation of public services such as Binder (**T5.1**) (e.g. popularity data for public open science repositories) will be fully anonymised to the satisfaction of relevant best privacy practices and regulations, such as GDPR, and made publicly available in the standard JSON Lines format, as is done already for mybinder.org [Myb]. This is very small data and easily archived on free hosting services such as GitHub, and will be made available under the Creative-Commons Universal Public Domain Dedication (CC0). There is no cost to the project associated with archiving this data long-term.

2.2.2 Communication activities

The main remaining goal for dissemination is making sure that potential users are aware of the tools and services developed by BOSSEE.

In order to maximise this impact, it is vital to address the audience as one project and ensure the immediate recognition of information stemming from it. Together with all partners involved, BOSSEE will therefore build **a strong project identity**. The following design and communication elements will be used to strengthen the project uniformity and identity and to deliver clear messages to our audience: BOSSEE naming, logo, presentations template, templates for reports and letters, project posters/leaflets etc.

In addition to BOSSEE-organised workshops in **T6.2**, the primary mechanism by which we will communicate our results is through publications and conferences. All publications funded by BOSSEE will be **Open Access**, and sites expecting publications have budgeted funds for paying Open Access fees. We will identify and attend appropriate conferences for sharing our work, including running tutorials at conferences in historically interested communities such as PyData and SciPy. Also, we will identify and attend conferences from complementary communities such as ROpenSci, Mozilla Science, and Julia as well as domain specific conferences to maximise the impact of BOSSEE and to broaden its audience outside the traditionally included communities.

We will operate a **website** (**T6.1**) for collecting and sharing information about BOSSEE and its progress. It will provide a centralised way to access the various publicly available deliverables, publications and articles related to the project. The site will be regularly updated over the lifetime of the project with the project publications and public materials, such as flyers, posters and public deliverables, organized workshops, available services, news, etc. Site analytics will be associated with the project website, in order to provide useful insight on how to improve its impact. In addition, the project intends to develop its presence on **the social and content networks**, such as Twitter and Facebook. The channels will be used for interaction with the professional community as well as the general public (differentiation on the content per channel based on the target group wishing to address). As part of the project?s communication plan, BOSSEE will develop a social media strategy in order to increase outreach and social impact, which can be summarised as follows: (a) identifying target audience and key stakeholders, (b) updating social media content and sparking discussion in social media/tweeting, (c) measuring social impact and reassessing social media strategy as required.

3 Implementation

3.1 Work Plan — Work packages, deliverables and milestones

3.1.1 Overall Structure of the Work Plan

As shown in Table 3.1.1, the work plan is broken down into six work packages: WP2 about maintaining the core Jupyter software infrastructure, WP3 for developing the ecosystem of software surrounding Jupyter, WP4 for building applications to demonstrate the efficacy and guide the development of core infrastructure, WP5 for operating services built on these components and collaborating with existing EOSC stakeholders, WP6 for educating the public on Open Science best practices with the project's tools and fostering diversity in the research and software communities. This is complemented by the usual management work package (WP1). The Gantt chart on Page 25 illustrates the timeline for the various tasks for these work packages.

WP	Title	Simula	CNRS-ObAS	EP	EGI	EuXFEL	INSERM	QuantStack	UiO	UPSud	Silesia	WildTree	total
WP1	Management	*24*	3	3	3	3	3	3	3	3	3	3	**54**
WP2	Core	*30*		13		16		15		14		8	**96**
WP3	Ecosystem	30		20		54		*20*		2	4	6	**136**
WP4	Demonstrators	9	12	3	7	*36*	24	6	12	20	12	3	**144**
WP5	EOSC	18		10	12	6				0		*13*	**59**
WP6	Education	13	3	3	2	8	*12*	4	12	3	5	3	**68**
totals		**124**	**18**	**52**	**24**	**123**	**39**	**48**	**27**	**42**	**24**	**36**	**557**

Efforts in PM; WP lead efforts light gray italicised

Table 3.1.1: Work Packages

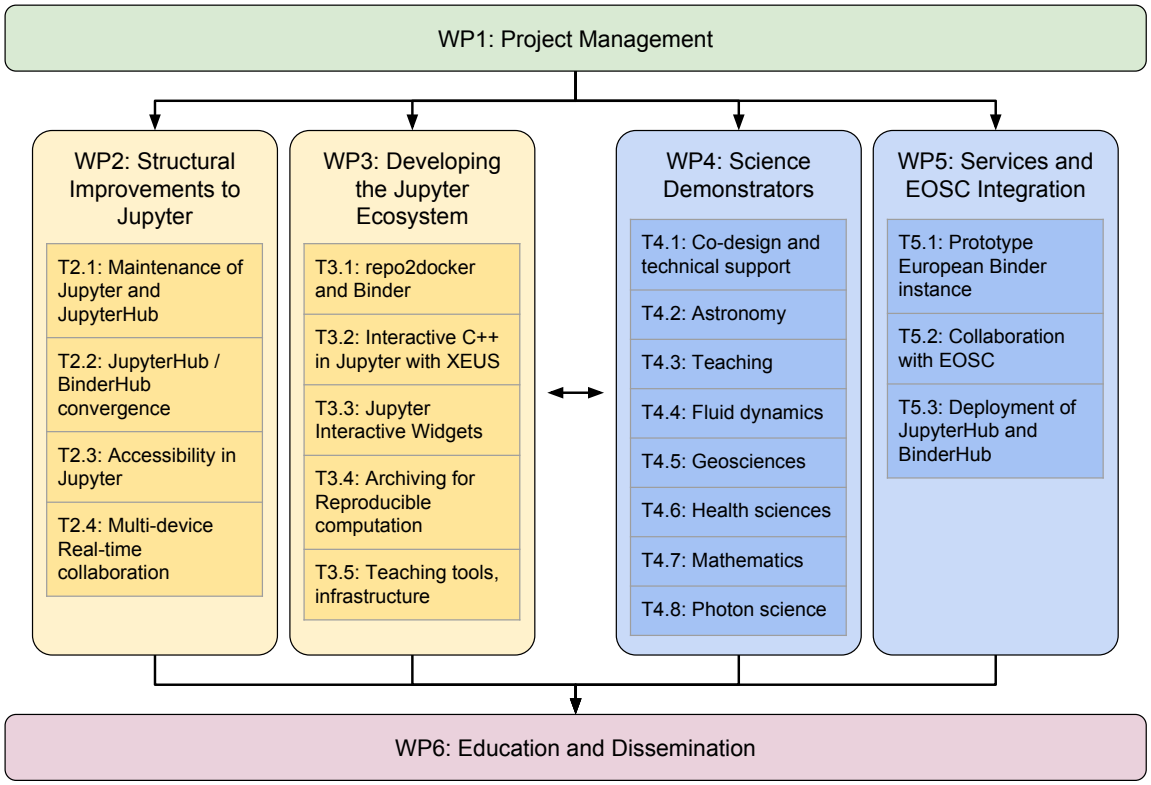


Figure 3.1.1: The relationships and interactions of the work packages, broken up into four main categories: Management (WP1), Development of new functionality surrounding Jupyter (WP2, WP3), Demonstrators and Services (WP4, WP5), and Education and Dissemination (WP6). Ultimately, all work packages benefit from and feed back to all other work packages.

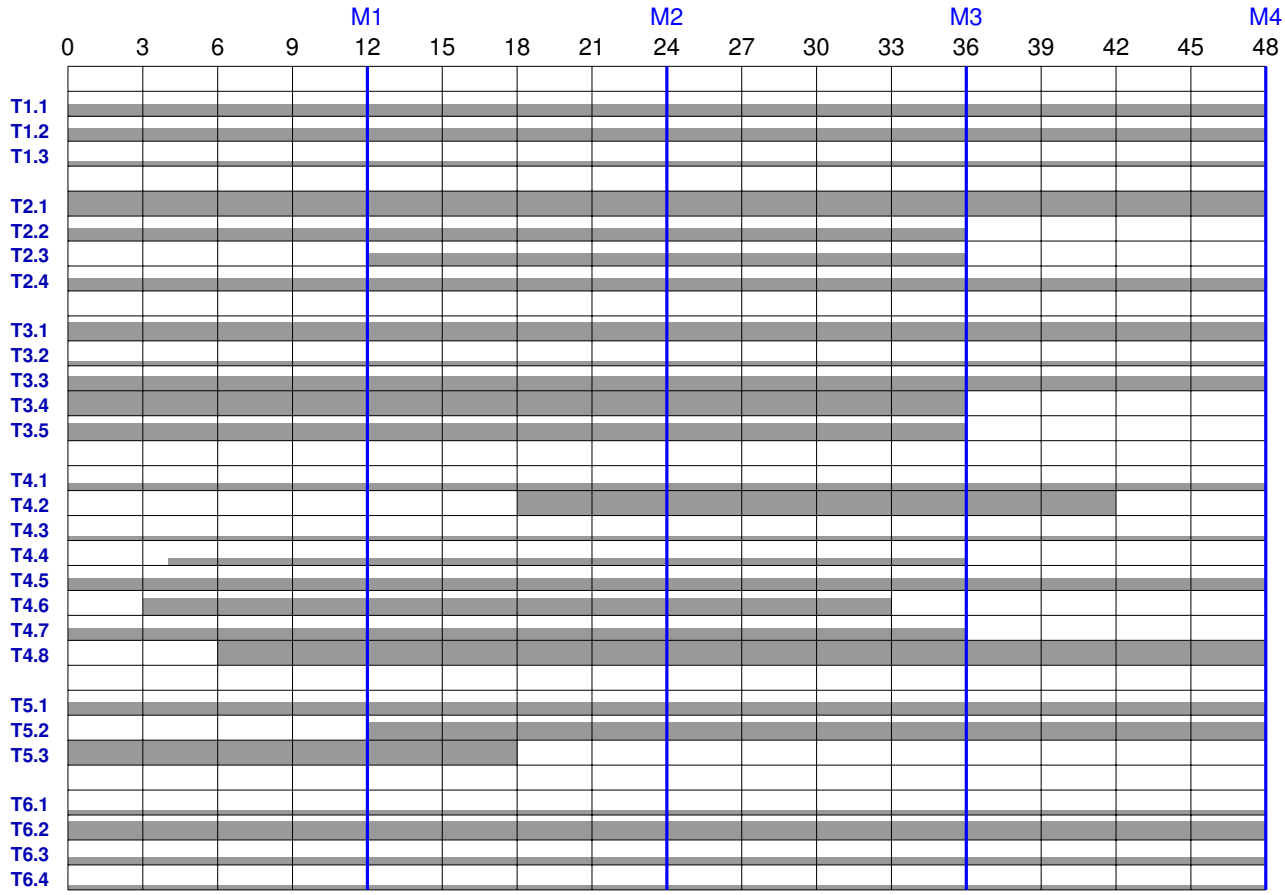


Figure 3.1.2: Gantt Chart: Overview Work Package Activities – *Bars shown at reduced height (e.g. 50%) indicate reduced intensity during that work phase (e.g. to 50%).*

3.1.2 Deliverables

#	Deliverable name	WP	Lead	Type	Level	Due
D1.1	Basic project infrastructure (websites, wikis, issue trackers, mailing lists, repositories)	WP1	Simula	DEC	PU	1
D5.1	A report describing the technical and service provisioning integration activities required to ensure the provisioning of BOSSEE services through EOSC	WP5	EGI	R	PU	6
D1.2	Data Management Plan draft	WP1	Simula	R	PU	9
D1.3	Innovation Management Plan	WP1	Simula	R	CO	9
D3.1	Guidelines for Binder use to improve reproducibility of environments, based on existing technology such as repo2docker	WP3	EuXFEL	DEC	PU	12
D3.2	Study of the practices of using Jupyter for teaching and the needs to effectively manage classes and associated courses in the education community	WP3	EP	R	PU	12
D4.1	Initial requirements for the demonstrators	WP4	Simula	R	PU	12
D2.4	Report and plan for Jupyter accessibility	WP2	Simula	R	PU	18
D5.2	Report on the technical and service provisioning integration activities and their status, reporting period 1	WP5	EGI	R	PU	18
D5.3	Documentation of how to deploy easily JupyterHub and BinderHub on OpenStack and OpenShift	WP5	EP	R	PU	18
D6.1	Community building: Report on impact of development workshops, dissemination and training activities, reporting period 1	WP6	INSERM	R	PU	18
D2.1	Contributions to core Jupyter and JupyterHub software	WP2	Simula	OTHER	PU	24
D2.3	Guidelines for a JupyterHub/Binder convergence	WP2	EP	R	PU	24
D4.2	Report on the developments of the demonstrator services deployed locally	WP4	EP	R	PU	24
D5.4	Documentation and tools of how to easily deploy a public BinderHub on EU cloud infrastructure	WP5	Simula	R	PU	24
D2.5	Improved accessibility of Jupyter software	WP2	Simula	OTHER	PU	36
D3.3	Unified framework to effectively manage classes and associated courses using Jupyter technology	WP3	EP	OTHER	PU	36
D3.4	Long-term reproducibility: Computational environment software archive system that extends lifetime of computational environments used in Binder service.	WP3	EuXFEL	OTHER	PU	36
D3.5	Implement interoperable 3d visualization widget based on K3D-jupyter code	WP3	Silesia	OTHER	PU	36
D4.3	Demonstrators based on locally developed services made accessible through EOSC. Demonstrators may be developed further subsequently	WP4	EGI	DEM	PU	36
D5.5	Report on the technical and service provisioning integration activities and their status, reporting period 2	WP5	EGI	R	PU	36
D6.2	Community building: Report on impact of development workshops, dissemination and training activities, reporting period 2	WP6	INSERM	R	PU	36
D1.4	Data Management Plan	WP1	Simula	R	PU	48
D2.2	Public releases of core Jupyter and JupyterHub software supporting BOSSEE services	WP2	Simula	OTHER	PU	48
D2.6	Real-time collaborative notebook supporting multiple devices and selective aggregation and distribution	WP2	UPSud	OTHER	PU	48
D4.4	Evaluation of demonstrators and case studies. Report on feasibility and user feedback to guide EOSC service design	WP4	EuXFEL	R	PU	48
D5.6	Documentation on best-practices of operating a federation of BinderHubs	WP5	Simula	R	PU	48
D5.7	Report on the technical and service provisioning integration activities and their status, reporting period 3	WP5	EGI	R	PU	48
D6.3	Community building: Report on impact of development workshops, dissemination and training activities, reporting period 3	WP6	INSERM	R	PU	48
D6.4	Interactive textbook on stochastic processes in physics	WP6	Silesia	DEM	PU	48

3.1.3 Milestones

General Milestones

1. **Milestone M1 (Month12) Startup, requirements, and prototype generic Jupyter service** By milestone 1, we will have established the infrastructure for the project and begun prototyping development and deployment of services, engaging with the existing communities, coordinating plans for BOSSEE with those of the wider Jupyter and open science communities, and prototyping operation of a generic Jupyter service for Open Science. EGI Infrastructure-as-a-Service (IaaS) cloud resources will be available with a Service Level Agreement (SLA) for BOSSEE.

2. **Milestone M2 (Month24) Generic Jupyter service and early local demonstrators** By milestone 2, we will have deployed a generic Jupyter service for Open Science, and begun to experiment with early-adopter users and local demonstrators to guide further development of BOSSEE, ensuring that development serves the needs of the community. An initial BOSSEE Service Management System shall be available, and BOSSEE services will be integrated with the EOSC Marketplace, and AAI Cloud services.

3. **Milestone M3 (Month36) Demonstrator services and community engagement** By milestone 3, demonstrator services should be useful and accessible to a broad range of users. By this point, we will have training materials and run workshops to train users in Open Science, making use of operational BOSSEE services, gathering feedback to guide the further development of BOSSEE. The BOSSEE Service Management System will be fully compliant with the EOSC Service Management practices.

4. **Milestone M4 (Month48) Full EOSC integration, adoption, sustainability, and evaluation** By this point, all BOSSEE services should be operational, available via EOSC-hub, and the generic Jupyter service TRL 8. Through community engagement via workshops, conferences, and other media, the services should have established groups of users, benefiting from these services and improving the Open Science landscape on EOSC and beyond. At the end of the project, we will have engaged with the community to evaluate the prototype EOSC services, identified which services and tools shall be sustained beyond the life of the project, and developed a sustainability plan for how this may be achieved under community support and leadership.

3.1.4 Work Package Descriptions

Work Package 1: Project Management												**Start:** 0
Site	*Simula*	CNRS-ObAS	EP	EGI	EuXFEL	INSERM	QuantStack	UiO	UPSud	Silesia	WildTree	**all**
Effort	*24*	3	3	3	3	3	3	3	3	3	3	**54**

Objectives

The main objective of WP1 is to establish and maintain an effective contract, project, and operational management approach ensuring:

- Timely and successful implementation of the project; including administrative and legal coordination
- Technical management and quality assurance
- Risk and innovation management of the project as a whole; including data and IPR management
- Smooth communication and interaction with the EC and other interested parties

Description

The project will be managed by Simula, which has extensive experience in administering and leading EU funded and national projects. The coordinator together with the WP leaders, will be responsible for monitoring WP status, coordination of work plan updates and annual internal progress reports. The project management structure and roles of partners in the consortium are presented in 3.2.

T1.1 Administrative Management M0-M48@.5; Sites: Simula (lead), CNRS-ObAS, EGI, EP, INSERM, QuantStack, UiO, UPSud, Silesia, WildTree, EuXFEL

The task includes the following activities:

(a) Preparation, distribution and maintenance of all contractual documents (Consortium Agreement, Grant Agreement and all other legal frameworks)

(b) Establishment of appropriate communication and collaborative environment for the consortium, as well as the EC and other relevant academic and industry stakeholders (the project website, intranet and communication procedures) to organise transfer of knowledge, present and promote project results (D1.1);

(c) Organisation of project review and progress meetings;

(d) Performing qualitative and quantitative risk analysis, planning risk mitigation and control

(e) Progress and Financial Reporting to the EC;

(f) Data and IPR Management will be managed in accordance with agreed rules stated in the Consortium Agreement and in accordance with the Data Management Plans (D1.4, D1.3).

T1.2 Technical Project Management M0-M48@.5; Sites: Simula (lead), CNRS-ObAS, EGI, EP, INSERM, QuantStack, UiO, UPSud, Silesia, WildTree, EuXFEL

The project scientific and technical management ensures coherent quality and soundness of the work and results. A quality assurance plan will be developed by Simula, involving all partners, and will be followed up regularly. It will address the reviews and approval of technical reports and deliverables. In addition, the Project Coordinator with the help of the coordination team will regularly review technological risks and recommend mitigation plans to minimise or remove them. This will be reported on at each Reporting Period in the project's Technical Report.

T1.3 Innovation Management M0-M48@.2; Sites: Simula (lead), CNRS-ObAS, EGI, EP, INSERM, QuantStack, UiO, UPSud, Silesia, WildTree, EuXFEL

One of the most important criteria for success for the BOSSEE project is to bring the project results into use. Therefore, exploitation routes will be sought whenever possible. In order to create a common understanding within the Consortium of how we can best shepherd an idea all the way from conception to realisation and exploitation, the Coordinator will be responsible for the preparation and realisation of an Innovation Plan. This plan will assure that research activities meet the required milestones and produce outputs fully aligned with the project objectives. All research activities will go through an initial process where the exploitation opportunity is identified along with the main stakeholders for the exploitation opportunity and an IP owner (D1.3).

Deliverables:

D1.1 (Due: 1, Type: DEC, Dissem.: PU, Lead: Simula) *Basic project infrastructure (websites, wikis, issue trackers, mailing lists, repositories)* ⇝M1

D1.2 (Due: 9, Type: R, Dissem.: PU, Lead: Simula) *Data Management Plan draft* ⇝M1

D1.3 (Due: 9, Type: R, Dissem.: CO, Lead: Simula) *Innovation Management Plan* ⇝M1

D1.4 (Due: 48, Type: R, Dissem.: PU, Lead: Simula) *Data Management Plan* ⇝M4

Work Package 2: Structural improvements to Jupyter							**Start:** 0
Site	*Simula*	EP	EuXFEL	QuantStack	UPSud	WildTree	**all**
Effort	*30*	13	16	15	14	8	**96**

Objectives

- to support and maintain core Jupyter infrastructure in order to keep it healthy and useful for open science
- to develop new features in the core of Jupyter to bring it to a wider community
- to develop new features in the core of Jupyter to make it more effective in facilitating open science

Description

Community-led open source software is critical to a sustainable future for open science. Commonly used tools make up a shared infrastructure, where investment in core components benefits the widest user community. BOSSEE is centred around the Jupyter project, which is a collection of projects for interactive computing and communicating computational ideas.

This work package is focused on developing and maintaining the core of Jupyter. In particular, we will help maintain these projects to meet the needs of the Jupyter community, with a focus on needs for open science. To serve the needs of BOSSEE, Jupyter core infrastructure will need improvements to security, performance, and scalability, which will be provided in **T2.1**. In addition, we will develop new features in the core of Jupyter to bring it to a wider audience, and to improve its usefulness to those working toward open science practices, including via collaboration features (**T2.4**) and accessibility (**T2.3**).

T2.1 Maintenance of Jupyter and JupyterHub M0-M48; Sites: Simula (lead), QuantStack, UPSud, EuXFEL

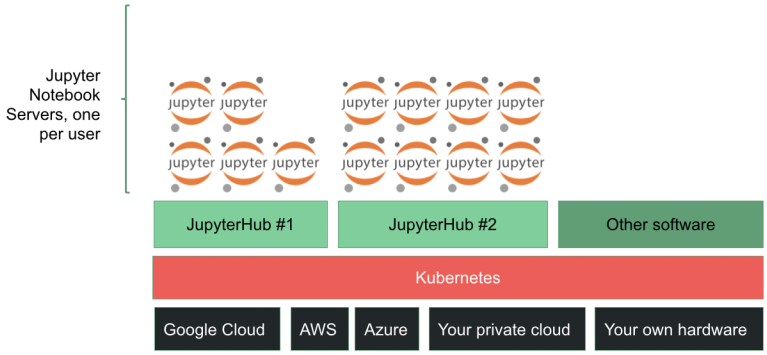


Figure 3.1.3: Jupyter notebooks deployed for many users on shared infrastructure using JupyterHub and Kubernetes.

Developing software that people will use requires maintenance of that software, not just new development. Millions of people rely on Jupyter software, including all participants in BOSSEE, and with this proposal, we will fund general support of the Jupyter infrastructure.

Maintenance of core software is often an implicit and un-paid cost, or one hidden in over-describing the resources required to deliver proposed developments. In BOSSEE, we make it clear and explicit that we will spend a significant amount of time developing and maintaining the core Jupyter and JupyterHub e-Infrastructure to respond to the needs of BOSSEE and others, and contribute towards the sustainability and health of the community.

We will provide support to the Jupyter e-Infrastructure software, ensuring that it meets the needs (D2.1) of BOSSEE, and aid in the release process to ensure that stable releases of Jupyter software can be used in mature BOSSEE services (D2.2).

BOSSEE will need improvements to core Jupyter functionality, including areas of:

(a) ease of deployment
(b) security
(c) scalability of JupyterHub
(d) performance

We will contribute improvements in these areas, meeting the needs of BOSSEE and benefiting the wider Jupyter community.

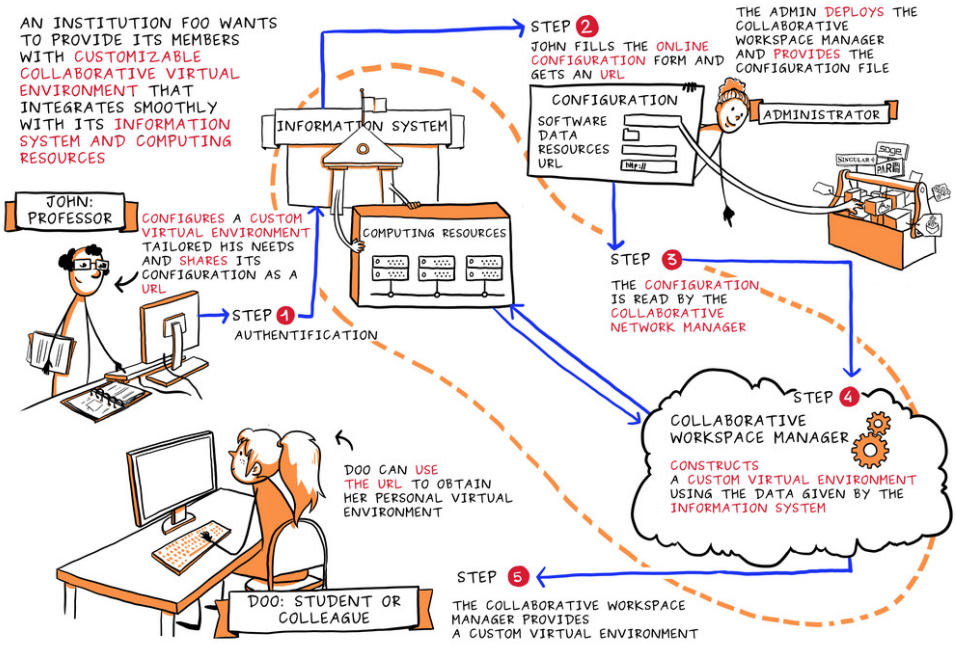


Figure 3.1.4: A use case in education for the JupyterHub - Binder convergence

T2.2 JupyterHub / BinderHub convergence M0-M36@.5; Sites: EP (lead), WildTree

Scenario

An institution – typically a university, a national lab, a transnational research infrastructure such as the European XFEL, or transnational infrastructure provider like EGI, or even an NGO or SME – wishes to provide its members and users with a Jupyter service.

The service lets user spawn and access personal or collaborative virtual environments: namely a web interface to a light weight virtual machine, in which they can use Jupyter notebooks, run calculations, etc.

To cater for a large variety of use cases in teaching and research, the main aim of the upcoming specifications is to make the service as versatile as possible. In particular, it should empower the users to customize the service (available software stack, storage setup, ...), without a need for administrator intervention.

State of the art JupyterHub already provides authentication, persistent storage and some default environments for its users. On the other hand, BinderHub offers the possibility to define more precisely the requirements for your teaching or research environment which makes it very flexible. However, BinderHub is still missing the authentication and persistent storage capabilities provided by JupyterHub.

This task The purpose of this task is to have the same services offered by JupyterHub (authentication, persistent storage, ...) with the flexibility of BinderHub (construction of your own environment for teaching or research).

The task includes the following activities:

- Extend where needed JupyterHub's authentication features
- Credential management
- Customizable persistence at the admin, and user level
- Choice of container registry at the admin and user level
- Runtime resource configuration

We will provide a report explaining how to make JupyterHub/Binder convergence a reality (D2.3).

T2.3 Accessibility in Jupyter M12-M36@.5; Sites: Simula (lead)

Improving the accessibility of Jupyter software is important to ensuring the value of BOSSEE reaches and appropriately supports the broadest community. In particular, we should take advantage of established guidelines (such as a11y) and technologies to ensure that people with visual impairments can make full use of the software. This includes considerations such as font size and contrast ratios, as well as ensuring that the interface can be used with a screen reader, for people who cannot use a conventional display screen.
Working with the community and UI accesibility experts, we will audit the accessibility of Jupyter software and assemble a roadmap for improved accessibility (D2.4), and ultimately work to improve the accessibility of Jupyter software, starting with JupyterHub (D2.5).

T2.4 Multi-device Real-time Collaboration M0-M48@.5; Sites: UPSud (lead), EP, QuantStack, EuXFEL

Collaboration is tremendously important in both research and industry, so the ability to collectively edit and run notebooks would be invaluable. It's also increasingly common for each person to use multiple devices to manipulate or present content, and seamlessly transferring updates between them presents similar technical challenges. Jupyter already recognizes the distributed nature of the digital environment by supporting remote kernels for computation, but the interface assumes that one user at a time works on a notebook from one device. BOSSEE will embrace this multi-device world and facilitate the distribution and real-time collaborative editing of content across multiple devices for presentation, interaction and collaboration purposes.
There are already some services providing real-time collaboration atop notebooks. Google offers *Colaboratory* [Col], a proprietary web application using its *Google Drive* backend. *CoCalc* [Coc] provides a suite of tools for collaborative mathematical computing, including an open-source mechanism for real-time collaboration on notebooks, but this is tightly integrated with that service's infrastructure. We want to make realtime collaboration available to as many Jupyter users as possible, however they choose to install and use the notebook software.
We will create a real-time collaborative notebook or JupyterLab-like application by leveraging well-known synchronization techniques such as Operational Transformation (OT) or Conflict-free Replicated Data Types (CRDTs), with server-side hosting of the document state. We will build on our experience with Webstrates (`http://webstrates.net`), a web-based environment that supports real-time sharing of web content. The CodeStrates extension to Webstrates uses the layout of Jupyter notebooks, and we have created a proof-of-concept showing that it can work with Jupyter kernels. However CodeStrates, and there are interesting unresolved questions about code execution in such a shared environment: Should it be synchronized or not among participants? Should there be a single kernel or one per participant? etc.
We will also enable selective distribution, aggregation and control of content across devices. We have used Webstrates to distribute and synchronize content across multiple devices such as a tablet, a laptop and a large wall-sized display. Yet this does not cover all use cases. For example, in a meeting, the participants should each be able to run their own notebook and pick which content to share with the group on a large display. We will create an environment where a notebook can collect specific cells from another notebook, or where a JupyterLab widget aggregates data from a collection of widgets running on each user's device. We also want to support remote interaction using one device to control another, e.g. a widget on a smartphone to control a parameter in a computation taking place in a particular notebook, whose result is shown on a large shared display.

Deliverables:
D2.1 (Due: 24, Type: OTHER, Dissem.: PU, Lead: Simula) *Contributions to core Jupyter and JupyterHub software* ⇝M2
D2.2 (Due: 48, Type: OTHER, Dissem.: PU, Lead: Simula) *Public releases of core Jupyter and JupyterHub software supporting BOSSEE services* ⇝M4
D2.3 (Due: 24, Type: R, Dissem.: PU, Lead: EP) *Guidelines for a JupyterHub/Binder convergence* ⇝M2
D2.4 (Due: 18, Type: R, Dissem.: PU, Lead: Simula) *Report and plan for Jupyter accessibility* ⇝M2
D2.5 (Due: 36, Type: OTHER, Dissem.: PU, Lead: Simula) *Improved accessibility of Jupyter software* ⇝M3
D2.6 (Due: 48, Type: OTHER, Dissem.: PU, Lead: UPSud) *Real-time collaborative notebook supporting multiple devices and selective aggregation and distribution* ⇝M4

Work Package 3: Developing the Jupyter Ecosystem								**Start:** 0
Site	Simula	EP	EuXFEL	*QuantStack*	UPSud	Silesia	WildTree	**all**
Effort	30	20	54	*20*	2	4	6	**136**

Objectives

- develop projects for creating open science services built out of Jupyter components and exploring new models for such services
- develop tools for interactive visualization in Jupyter
- develop workflows for data science using Jupyter software

Description

Open source software in general, and Jupyter in particular, is developed not as a monolithic application, but rather as a collection of related components, which can be assembled in numerous combinations to meet diverse needs. The Jupyter community is no different. Jupyter itself is composed of several projects, but there are even more projects that build on top of Jupyter to create things like cloud services or data pipelines. The goal of BOSSEE is to facilitate open science through Jupyter, and this includes working with projects all around the Jupyter ecosystem. We will focus this work package on developing Jupyter ecosystem projects with an emphasis on open science.

repo2docker is a project for creating reproducible environments in which Jupyter notebooks (and other user interfaces) can be run. It reads a number of common formats to list required software packages, and prepares a Docker container with those packages installed. BinderHub is software for operating a web service using repo2docker, which enables sharing of interactive and reproducible Jupyter (and Rstudio) environments on the web with a single link. We will develop repo2docker and BinderHub further to meet the needs of the open science community.

Widgets are an extension to Jupyter, which can define new kinds of interactivity. 3D visualisation of data is important to many kinds of science, and there is lots of room for development of the 3D visualisation landscape in Jupyter.

In addition to the interactive aspects of Jupyter, notebooks can be used in a ”workflows” style, where job systems run analyses and produce reports, either on a scheduled basis or triggered by events. There is a great deal of interest in using notebooks in this way, and much room for development of tools supporting workflows in data-driven open science.

T3.1 Further development of repo2docker and Binder M0-M48@.75; Sites: Simula (lead), WildTree, EuXFEL

Running someone else’s analyses is a particularly difficult problem. There are differences between operating systems, versions of installed software and access to the required data sets. These challenges mean that is currently considered to be beyond the scope of an expert peer reviewer to verify data science analysis codes before publication. BinderHub, part of Project Jupyter, enables one-click running of git repositories. BinderHub provides a web interface to the repo2docker tool.

The task includes the following activities

- extend repo2docker with support for execution on cloud resources
- extend repo2docker with support for execution on HPC resources with Docker support
- improved ”first use” experience of running repo2docker locally
- add support for using archives such as Zenodo as source for repo2docker and BinderHub
- define procedures and recommendations for long term reproducibility and sustainability of repo2docker compatible repositories
- create educational material describing repo2docker and its benefits to researchers
- Enable Openshift based deployments of BinderHub
- User surveys about pain points using BinderHub
- User authentication in BinderHub

T3.2 Interactive C++ in Jupyter with XEUS M0-M48@.2; Sites: QuantStack (lead), EP

(a) cling & xeus-cling

i. further development of xeus & xeus-cling for a fully-fledged **Jupyter** experience with the C++ programming language.

ii. improved **packaging** of cling and related packages for conda and other package managers, including full **windows** support.

iii. continued development of C++ **interactive widgets** and backends for ipyvolume, bqplot, ipyleaflet (xvolume, xplot, xleaflet), and better testing for the feature parity with ipywidgets.

iv. production of **teaching material** for C++ with xeus-cling and interactive widgets.

v. integration with read-only notebook viewers such as **voila** to produce a fully compiled application from a notebook interacting with Jupyter frontend components such as interactive widgets.

vi. enable different levels of **compiler optimization** for code interpreted by cling.
vii. improve the **magics** plugin system to simplify authoring of custom magics for xeus-cling.

(b) core xeus library
i. improve language bindings for the C++ interactive widgets to enable interactive widgets in all xeus-based kernels.
ii. enable pluggable history managers, with specialized implementations for both SQLite-based history and in-memory history.

(c) cling support for common C++ scientific computing packages
i. cling offers the possibility to automatically load the runtime of libraries upon the inclusion of its headers using special pragmas. Several libraries such as `xtensor`, and `symengine` now make use of this possibility. We propose to generalize this approach to the main C++ libraries for scientific computing that may accept such pragmas to be included upstream.

Scientists, educators, and engineers not only use programming languages to build software systems, but also in **interactive workflows**, using the tools at hand to explore and reason about problems.
Running some code, looking at a visualization, loading data, and running more code. Quick iteration is especially important during the exploratory phase of a project.
While C++ is ubiquitous in scientific computing for close-to-the-metal performance number crunching, *we lack a good story for interactive computing in C++*. This hurts the productivity of C++ developers:

- Progress in software projects often comes from **incrementalism**. Obstacles to fast iteration hinder progress.
- his also makes C++ more **difficult to teach**. The first hours of a C++ class are rarely rewarding as the students must learn how to set up a small project before writing any code. And then, a lot more time is required before their work can result in any visual outcome.

The cling interpreter fills the gap of interactivity for the C++ programming language and its use at scale in at LHC proves that C++ can be a language for interactive scientific computing.
The xeus-cling kernel
The goal of the xeus-cling project is to improve the integration with Project Jupyter and make C++ a **first-class citizen** of the ecosystem.
By leveraging the rich ecosystem of tools built upon Project Jupyter, which is language agnostic, we can lift C++ from a language reserved to high-performance performance computing to a high-level language of data science like Python, R, or Julia.
Teaching the C++ programming language
Since September 2017, the 400 first-year students at Paris-Sud University who take the “Info 111: Introduction to Computer Science” class write their first lines of C++ in a Jupyter notebook, with the xeus-cling kernel.
The use of project Jupyter for teaching C++ is especially useful for the first classes where students can focus on the syntax of the language without distractions such as compiling and linking a program.
The availability of **Jupyter interactive widgets** in that environment offers a simple means to obtain a **visual outcome** in a few lines of code, for a more **rewarding learning experience**. It is also not typical for the C++ programming language.
Finally, the **cloud hosting** of the environment removes the hurdle of installing a development environment on a large variety of student’s machines.
C++ as a common denominator
The fragmentation of the ecosystem between the main languages of data science causes a lot of duplication of work and harms sustainability in the long run. Often, implementation of standard protocols, file formats, and reference implementations of numerical methods are duplicated in each language.
A common denominator of the three main languages of data Science (Julia, Python, R, forming the Jupyter name) is their ability to call into natively built libraries. All three interpreters have a clean C API that can also be used from the C++ programming language.
However, a solid C++ implementation can always be exposed to Julia, Python and R, making the common implementation more sustainable in the long run. The static typing of the language may make the initial implementation less immediate, but for greater stability.
For this reason, we strive to provide solid C++ implementation of

(a) standard protocols, such as the Jupyter kernel protocol (with xeus), now used to make other kernels than the C++ kernel.
(b) standard in-memory and file formats such as apache arrow, FITS, NetCDF, HDF5.
(c) data structures, such as N-D arrays and dataframes
(d) reference implementation of numerical methods.

Reference scientific Python projects such as Sympy have moved their core to a solid C++ engine (such as symengine), which has then been exposed to Julia and R.

T3.3 Jupyter Interactive Widgets M0-M48@.6; Sites: QuantStack (lead), Silesia, EuXFEL

The task includes the following activities:

(a) Improvements to the core Jupyter Widgets package
   i. Improve the testing tools for Widget libraries, including means to test messaging, schemas, and actual rendering of widgets with a headless browser.
   ii. Modernize of the *@jupyter-widgetsbase* JavaScript package: drop *backbone.js* and adopt a more modern MVC framework, support streaming of widgets messages to multiple frontends for future support of live collaboration.
   iii. Iterate on the core *@jupyter-widgetscontrols* package. This may involve the adoption of a modern framework such as *React.js* for the widget view implementation, rather than the current custom implementation.
   iv. Create new controls for the core Jupyter widgets package such as token inputs, typeahead, and tree views. These common controls tend to be implemented in several downstream packages, which causes unnecessary duplication of work and harms sustainability.

(b) Dealing with large widget state
   i. Create a generic mechanism for widgets to refer to external data services or companion files rather than storing their state entirely in the notebook format for offline view.

(c) Simplify the authoring of complex GUIs with Jupyter widgets.
   i. Provide pre-defined Jupyter widget layouts based on the recently introduced CSS Grid Layout, with named areas to which widgets can be assigned, such as a centra area, left and right tab bars, footer and headers, or layouts including areas for logging results on long-running tasks.
   ii. Experiment with the generation of a Jupyter-widgets based GUI from a declarative configuration file, or by introspecting a configuration object.

(d) Develop interoperable Jupyter widgets for 3d data visualisation.
   i. Adopt existing 3d jupyter widget new mechanisms and API of Jupyter widgets
   ii. Experiment with large dataset visualization and inspection in Jupyter widgets. This will include work on the flexible split of preprocessing between server and widget and the development of data formats for special use cases.
   iii. Improve and standardize the interoperablity of different components and dataformats.
   iv. Create comprehensive documentation of 3d widgets with scientifically relevant examples.
   v. Create developer documentation and engage the community for contributions implementing specialisations of the 3d widget.(D3.5)

T3.4 Archiving software environments for reproducible computation M0-M36; Sites: EuXFEL (lead), EGI, QuantStack, Simula, UPSud, WildTree

Reproducible research can inspire greater confidence in scientific results, and make it easier for future research to build on those results.
Reproducibility is seen as an essential pillar of scientific truth, nevertheless there is a real shortcoming of truly reproducible research in the areas of computational and data science. In part, this is a cultural matter. However, there is also a lack of computational e-infrastructure supporting reproducibility.

Jupyter Notebooks combine explanation with code and output and are thus valuable tools for making scientific computing more reproducible. However, the code in a notebook invariably relies on external code and hidden dependencies: libraries and programs which are not saved as part of the notebook.
This task concerns ways to record the versions of these tools in use, and to make them available for practical reproduction of the computation.
Binder and its tool repo2docker are a first step in the right direction: given the description of a computational environment, they allow to create that computational environment as a docker container automatically on demand, which in turn allows to execute a given notebook within this container environment. By archiving the notebooks together with the environment specification, the container computation environment can be created on demand. We see a number of publications being complemented by such Git repositories that allow reproducing figures and results from papers by re-executing notebooks; often archiving these repositories via the Zenodo service (for example publication [CO+18b]) with github repository [CO+18a]).
Container technologies, such as Docker, offer exciting possibilities for capturing a computational environment, but much of the development of these tools is focused on short-term operational uses, not long-term preservation.
There are currently at least two limitations in the existing repo2docker approach:

(a) the environment specifications need to be written carefully and need to explicitly define particular version numbers of operating systems, libraries, and software to be used in the environment. While there are no guidelines (yet) for best practice in writing such specifications, in principle users can do this correctly.

(b) when repo2docker builds a container environment, it relies on the required software being available on the Internet: Commands that clone software from GitHub assume that the software is actually available on GitHub. If a relevant repository disappears (or GitHub disappears), it will be impossible to clone that software from there, and this will break the binder execution and thus reproducibility. Some environments are specified through Dockerfiles, and often start from an Ubuntu Linux distribution container, followed by an `apt-get update` command. This

command will fail once the age of the specified distribution exceeds the support period, and similarly subsequent `apt-get install` commands will fail. As the Binder service matures, we start to see such failures occur.

The task addresses these problems and includes the following activities:

- Literature review and technology exploration: research Binder model and horizon scan for related technology to support long term reproducibility.
- Establish and document best practice for Binder use: Create public guidelines for building containers for scientific computing purposes so that they remain useful in the longer term, building on existing technology such as repo2docker. (→ D3.1)
- Facilitating reproducible creation and long-term archiving of container images for reproducibility: Develop new software to provide long-term executable computational environments that support the Binder model.
  There is a trade-off between preserving binary container images, and preserving the source code and instructions to build a container. Preserving sources is more transparent, and makes it easier to modify the code to explore a result, but without special care, the instructions may not continue to work in the future, or may not build an equivalent container. We will explore both approaches, with a particular interest in how to make build instructions that can still work many years in the future. (→ D3.4)
  It may be necessary for the build process to make use of alternative sources for source code and packages to install, such as snapshots of github repositories that are available on Zenodo, or dedicated software archive servers that provide selected pieces of software that are required to build particular containers.
  The developed software will be integrated into the EOSC Binder (**T5.1**).

This technology will be co-developed with a real scientific use case at European XFEL (see **T4.8** in WP4).
We note that there is a wide variety of use cases for this kind of technology and advances towards long term reproducibility, which we expect to grow in importance over time and with the increasing acceptance of open science: journals do increasingly (and rightly so) request from authors that they can reproduce their studies, or even submit corresponding code with the submission of the papers. As Binder and notebooks are a popular way of achieving this, the long term executability becomes a challenge. The same is true for universities and other academic institutions that take FAIR data seriously, and want to support the reproducibility and re-usability aspect of data comprehensively through providing data analysis software that allows access to the meaning of the data.

T3.5 Teaching tools, infrastructure, and best practices M0-M36@.7; Sites: EP (lead), UPSud

This task is devoted to improving the Jupyter ecosystem for education. See page 12 for context and a list of other tasks that will contribute to better teaching.
Setting up a comfortable working environment for both teachers and students requires tools for easy sharing, collecting, self assessment, and semi-automatic grading of course material, class management, and integration with the local e-learning infrastructure such as, e.g., Moodle or OpenEDX.
We will **review the state of the art**: existing tools, within the Jupyter ecosystem (e.g. nbgrader [Ham16] or OK[Okp]) and outside; course services (e.g. Gryd[Gry] or CoCalc[Coc]); course infrastructure that have been designed and deployed at many institutions (Berkeley, École Polytechnique, Université Paris Sud), etc.
To build and share a better vision of the needs, we will **conduct a survey in the education community about the usage of Jupyter**. In particular we will seek feedback from Jupyter-based MOOCs (Massively Open Online Courses), e.g. on Coursera[2], and Fun[3].
The collected requirements will be exposed in a first report D3.2 and largely disseminated.
The core of the task will then be to **further develop teaching tools, infrastructure, and course templates to contribute to the emergence of versatile solutions and best practices around them**.
The outcome will be put into production by the participants of the BOSSEE project (and beyond!) who will deliver a large number of courses using Jupyter technology (see **T4.3**). The variety of use cases and infrastructure will provide a rich test bed and immediate feedback at each iteration, ensuring that the developments are informed and steered by demand (co-design), and battle field tested.
At this stage, we already envision specific development in the following directions:

- Collaborative grade management
- Insulation through container of the automatic grading
- Integration with e-learning platforms (e.g. Moodle, OpenEDX (Coursera/Fun)), through an LTI connector.
- Develop course templates for various use cases.
- Disseminate tutorials on all of the above.

A second report D3.3 will review the developments, our in-class experience with them, and best practices.

Deliverables:
D3.1 (Due: 12, Type: DEC, Dissem.: PU, Lead: EuXFEL) *Guidelines for Binder use to improve reproducibility of environments, based on existing technology such as repo2docker* ⇝M1

[2] https://www.coursera.org/courses?query=jupyter
[3] https://www.fun-mooc.fr/courses/course-v1:inria+41016+session01bis/about

D3.2 (Due: 12, Type: R, Dissem.: PU, Lead: EP) *Study of the practices of using Jupyter for teaching and the needs to effectively manage classes and associated courses in the education community* ⇝M1

D3.3 (Due: 36, Type: OTHER, Dissem.: PU, Lead: EP) *Unified framework to effectively manage classes and associated courses using Jupyter technology* ⇝M3

D3.4 (Due: 36, Type: OTHER, Dissem.: PU, Lead: EuXFEL) *Long-term reproducibility: Computational environment software archive system that extends lifetime of computational environments used in Binder service.* ⇝M3

D3.5 (Due: 36, Type: OTHER, Dissem.: PU, Lead: Silesia) *Implement interoperable 3d visualization widget based on K3D-jupyter code* ⇝M3

Work Package 4: Science demonstrators												**Start:** 0
Site	Simula	CNRS-ObAS	EP	EGI	*EuXFEL*	INSERM	QuantStack	UiO	UPSud	Silesia	WildTree	**all**
Effort	9	12	3	7	*36*	24	6	12	20	12	3	**144**

Objectives

The objectives of this work package are

- to guide the development of core tools by simultaneously developing and using applications in diverse fields with active scientists from these fields, and
- to demonstrate that the tools we develop are valuable to diverse fields of science, thus ensuring we develop e-infrastructure and services which can cater for a broad customer base of EOSC.

Description

Whilst the components issued from work packages WP2 and WP3 will be made available as generic building blocks for EOSC services, this work package aims at building and deploying bespoke EOSC services targeting real-world cases.

We have selected a number of applications in a variety of domains to demonstrate the broad impact of BOSSEE, in particular in the areas of astronomy (**T4.2**), education (**T4.3**), fluid dynamics (**T4.4**), geosciences (**T4.5**), health (**T4.6**), mathematics (**T4.7**) and photon science and imaging (**T4.8**). The context and vision for each of the demonstrators is described in section 1.3.3 on page 11.

Working closely with the core developers of the Jupyter ecosystem will make it possible to go way beyond what is normally available ”out-of-the-box” and to offer better solutions, thereby guiding further development of the core features.

Our demonstrators will typically undergo two-stages: (i) development and testing of the services locally at the developing partner site. (ii) Making the service available through the European Open Science Cloud (EOSC).

All demonstrators will deliver base-line services by making the relevant notebooks executable in the *European Binder Service* instance that this project will deploy on EOSC (**T5.1**). This will demonstrate the Jupyter service capabilities such as reproducibility, interactive widget use and visualisation, and show how these can enable new open science on EOSC.

The particular workflows, data infrastructures and data policies for FAIR[i] sharing of data vary from one community and use-case to the other, or may not be fully defined yet. Therefore, this proposal does not enforce a specific way of handling data. Instead we will explore in the demonstrator tasks how existing data policies, infrastructure and workflows can be respected and integrated with authentication and authorisation, data management, and JupyterHub/Binder services on EOSC. EGI is a partner for all the tasks in this work package and will work with us to find the best integration solutions in the evolving EOSC infrastructure.

In the EOSC-hub project EGI operates a Jupyter Hub service which is deployed in a scalable mode on EGI IaaS Cloud. This Jupyter Hub is already integrated with the EUDAT B2DROP and OneData data management services of EOSC, and will be integrated, in the next 12 months, with the EUDAT B2SHARE service. The integrations enable users to move data between Jupyter notebooks and storage sites of the EGI Federation (with Onedata), between Jupyter and storage sites of the EUDAT federation (B2SHARE), and between Jupyter and their personal cloud storage hosted in EUDAT (B2DROP). The WP4 use cases will evaluate these data management integrations and EGI will bring the respective technology from EOSC-hub into the services operated by BOSSEE.

For some of the demonstrators, authentication and authorization and/or data management are being advanced outside BOSSEE. This is for instance the case for the photon science and astronomy demonstrators via PaNOSC and ESCAPE projects, respectively.

[i] Findable, Accessible, Interoperable and Reusable

T4.1 Co-design and technical support M0-M48@.3; Sites: Simula (lead), CNRS-ObAS, EGI, EP, INSERM, QuantStack, Silesia, UiO, UPSud, WildTree, EuXFEL

This task coordinates and supports the work of the other WP4 tasks. It will help us exploit synergies and coordinate the gathering and formulation of requirements and the preparation of the deliverables. It will also support demonstrator efforts to speed-up the development process and ease the deployment of innovative services. Finally, it will drive the co-design cycle between WP4 and all the technical work packages (WP2, WP3 and WP5).

- Assess co-design efforts and distribute workload across the core developers
- Regularly feedback information to WP6 to adapt trainings and dissemination
- Offer technical support to the demonstrators throughout all the steps until final deployment of the services
- Liase with **T5.2** for authentication, authorisation, data management and further EOSC integration.

T4.2 Demonstrator: Astronomy M18-M42; Sites: CNRS-ObAS (lead), EGI, INSERM, QuantStack, Simula, WildTree, EuXFEL

This task (see page 11 for context) will build on existing Python libraries (astroquery cds, simbad, vizier and xmatch packages) to access CDS – Centre de Données Astronomiques de Strasbourg – data. A representative sample of astronomical data is pictured in figure 1.3.6 describes the variety of data distributed by CDS.
For visualization, we will use proven tools like *GLUE* and *ipyvolume*, which are now built upon the Jupyter stack. We will also make significant improvement to existing Jupyter widgets (*ipyaladin*, interactive sky atlas running in the notebook) and develop a new widget to offer a tree-like view of available astronomical datasets, making use of the existing IVOA [Ivo] protocols.
We will also develop Python libraries to allow integration and usage in notebook of existing CDS infrastructure services, namely CDSLogin (which provides authentication) and CDS MyData (remote storage space for tabular data). This will allow the user to interact with one's personal storage space from the notebook. It will also allow for advanced customisation of the interface to fit user needs.
The work is organised with a two stage approach. Firstly, the generated notebooks will run locally on user machines (representing a milestone for this task). Following the Binder development in WP5, we will aim to run these notebooks on the European Binder Service. The aspiration is that this contributes to the development of innovative services for the EOSC. The deliverable of this task will be a demonstrator available to the scientific user community (D4.3).
By milestone 2, astronomical data services based on reference astronomy data from CDS will be made available in Jupyter notebooks. Based on developments in WP5, we will assess how to run these notebooks on BOSSEE EOSC services.
The work carried out in this task will be reported on in D4.4.

T4.3 Demonstrator: enriched teaching with Jupyter M0-M48@.2; Sites: EP (lead), EGI, INSERM, UiO, UPSud, EuXFEL

In this task (see page 12 for context), we will **deliver and help deliver Jupyter-based courses at a large scale** in our own institutions, as a mean to **inform, evaluate, provide feedback on, and demonstrate the value of the work performed** in this project in the context of higher education, as well as to **develop and share best practices** and **demonstrate and disseminate** Jupyter's full potential for teaching.
École polytechnique and Université Paris-Sud are particularly well suited for this task because they

(a) host a variety of local infrastructure (dedicated servers, local cloud, computer labs, ...);

(b) host a reactive community with highly qualified research software engineers (DevOps, software developers), researchers, professors, and students that have been working together on this topic for several years, with close collaboration between the two sites;

(c) offer very diverse courses, in many disciplines, and ranging from large lower undergraduate courses to specialized classes for graduate students and top notch engineers;

(d) have strong support from their respective teaching departments.

The task will include the following activities

- Reinforce the use of Jupyter technology in courses at all levels, notably in Mathematics and Data Science, in close collaboration between Ecole polytechnique and Université Paris Sud;
- Test the new developments and feed back to tasks **T2.2**, **T3.2**, **T3.5**, **T4.7** and **T5.3**;
- Follow up on a successful Jupyter day in 2018 [4] by organizing a yearly Jupyter event showcasing the latest advances for teaching and research;
- Foster sharing of experience, best practices and course material, at the local level, and then worldwide, through meetups, blogs, etc.
- Publish selected teaching material for interactive use on BOSSEE's EOSC services (D4.3).

The work carried out will be reported on in D4.4.

T4.4 Demonstrator: Visualisation and control of fluid dynamics in Jupyter notebook M4-M36@.3; Sites: Silesia (lead), EGI

In this task (see page 13), we will construct tools for editing and inspecting boundary conditions in fluid dynamics simulation as well as capable and optimized visualization utilities. Having such tools available as Jupyter widgets will allow to complete the typical CFD workflow without leaving Jupyter notebook. We plan the following activities

- Implement advanced widgets for data visualisation of large fluid dynamics simulations.
- Implement widgets for inspection and editing boundary conditions in LBM.
- Develop Jupyter notebooks demonstrating the full workflow of fluid dynamics simulation based on the high-performance Sailfish-cfd solver.
- Publish selected demonstration notebooks for interactive use on GPU-enabled BOSSEE's EOSC services (D4.3).

This work will closely interact in with the task **T3.3**: it will both provide guidelines and inspirations for the development to **T3.3** and serve as test case for implemented features in **T3.3**. It will be reported on in D4.4.

[4] http://www.cmap.polytechnique.fr/~massot/Personal_web_page_of_Marc_Massot/JupyterX.html

T4.5 Demonstrator: Geosciences M0-M48@.5; Sites: UiO (lead), EGI, QuantStack, Simula, UPSud

The aim of this task (see page 14) is to build on the Jupyter ecosystem to create a standardized and shareable computing, data analysis and visualization framework for Geosciences. This task will focus on filling gaps that hinder open science and will include the following activities:

Visualization

- Improvement upon existing mapping tools for specialized visualization of in-situ and model-generated data arizing in specific use cases (Land, river-runoff, ocean, ice, wave and atmosphere models, particle dispersion models, oil spill models, etc.).
- Improvements of the tooling for 3-D visualization of geographical datasets in the Jupyter notebook, for use cases such as displaying clouds, volcanic plumes, atmospheric rivers.

Collaboration with Jupyter with specialized tools for earth sciences

- adding the ability to interactively integrate information or corrections observed during field trips, correspdonding to specific geographical locations.
- adding the ability to deploy Jupyter-based applications together with the correspdonding execution environment, both in the form of a runnable notebook with *Binder* or as a read-only yet interactive *Voila* dashboard.

This work will be carried out in two stages with first local development and deployment of BOSSEE EOSC services for Geosciences (such as a BinderHub for Big data geosciences and *voila* innovative interactive *App*) and then deployment of these services on EOSC (D4.3).

T4.6 Demonstrator: Nuclear Medicine dosimetry M3-M33@.7; Sites: INSERM (lead), EGI, EuXFEL

The objective of this task (see page 15 for context) is to build on the Jupyter ecosystem to create a unified data analysis framework for the OpenDose project. The task includes the following activities:

- Developing tools to work seamlessly on the SQL database holding the dosimetric data.
- Developing data analysis tools using the Python data science ecosystem where possible.
- Developing visualization tools, exploring Widgets inside the Notebook for interactivity.
- Evaluate the relevance of Jupyter Notebooks from user feedback
- Disseminating results.
- Providing support to users.

First, these developments will be available to users as Jupyter Notebooks to be run locally on their machines. In a second time, these Notebooks will be ported on the BOSSEE's EOSC services following developments of Task **T5.1** (D4.3), contributing to the development of innovatives services for the scientific community. The reproducibility of generated results will be ensured by archiving the software environment with developments of Task **T3.4**. The evaluation of the developed tools from user feedback will contribute to the deliverable D4.4.

T4.7 Demonstrator: Interactive Mathematics with Jupyter Widgets M0-M36@.5; Sites: UPSud (lead), EGI, EP, QuantStack

The aim of this task (see page 16 for context) is to build on this experience to further develop and promote the use of Jupyter widgets for interactive Mathematics. This will include the following actions:

- Engage with the community through tutorials, workshops, online discussions, for codesign and for dissemination of the outcomes.
- Tackle hurdles to real-time interactivity, typically by modernizing the existing 2D and 3D visualization tools in SageMath.
- Bring sage-combinat-widgets and sage-explorer from usable prototypes to standard tools, and further contribute to the development of the Francy framework.
- Develop other generic mathematical widgets according to the users popular requests.
- Demonstrate the value all of the above through applications in research and teaching.
- Publish selected demonstration notebooks for interactive use on BOSSEE's EOSC services (D4.3).

The work carried out will be reported on in D4.4.

T4.8 Demonstrator: Reproducible photon science workflows at European XFEL M6-M48; Sites: EuXFEL (lead), EGI, INSERM, Simula, UPSud

This task (see page 17 for context) includes the following activities:

- Use the software archive for reproducible computation (as co-developed in **T3.4**), with the aim to provide reproducible computation environments for data analysis at European XFEL that remains executable for the same duration as the data is kept (currently aiming at 10+ years, at least 5 years).
  As is common in computational science, software used at XFEL often relies on specific combinations of libraries, in many cases with particular version requirements. Thus we will need a dedicated software archive that holds all relevant packages and source codes that are required to build the required computational environments (see **T3.4**) to ensure they are available even if an open source software provider decides to remove their repositories, or changes the API of a package, or GitHub decides to terminate their business.
  Applying the work from **T3.4** in the context of a production system will demonstrate its true utility, and provide

important feedback for the design. There will be iterative feedback and refinement of the service.

- Extend the use of notebooks from *interactive* data exploration and analysis at European XFEL to also provide computational work flows via (semi-)automatic execution of notebooks as described above. The work done in **T2.4** will allow us to execute notebooks in the background, and to connect to the running notebook process to display or inspect progress, or to modify such a notebook if the science requires it.
  By doing so, we can make the standard analysis that is carried out by the facility available on EOSC as a service. By using one tool (the notebook) we simplify processes for users and for the research facility.
- Use the work from **T3.3** on state-preserving widgets to provide GUI-like elements in notebook where interactive user input, data exploration or parameter modification is required.
- Explore use of the Voila capability to provide data exploration dash-boards to lower barriers of working with the data (will only be possible for somewhat standard experiments).
- Work with the PaNOSC project [Pan] to evaluate and use these new and EOSC-enabled services for other Photon and Neutron Science research facilities.
- Develop a demonstrator (Deliverable D4.3).
- Evaluate the chosen workflow design and experience from using it in a real-world context; make this available as a report and through presentations/workshops to interested organisations and facilities. (D4.4).

Deliverables:
D4.1 (Due: 12, Type: R, Dissem.: PU, Lead: Simula) *Initial requirements for the demonstrators* ⇝M1
D4.2 (Due: 24, Type: R, Dissem.: PU, Lead: EP) *Report on the developments of the demonstrator services deployed locally* ⇝M2
D4.3 (Due: 36, Type: DEM, Dissem.: PU, Lead: EGI) *Demonstrators based on locally developed services made accessible through EOSC. Demonstrators may be developed further subsequently* ⇝M3
D4.4 (Due: 48, Type: R, Dissem.: PU, Lead: EuXFEL) *Evaluation of demonstrators and case studies. Report on feasibility and user feedback to guide EOSC service design* ⇝M4

Work Package 5: Services and EOSC Integration						**Start:** 0
Site	Simula	EP	EGI	EuXFEL	*WildTree*	**all**
Effort	18	10	12	6	*13*	**59**

Objectives

- to support and develop BinderHub infrastructure based in the EU for science and education applications
- to scale this service by migrating to EOSC

Description

This work package is focussed on developing and operating a publicly accessible BinderHub service based in the EU.

The team of Project Binder provide a service at https://mybinder.org which is public infrastructure powered by the software they develop. mybinder.org is used to launch around 80000 ”binders” every week and is being used by researchers and teachers to make their work reproducible.

Although mybinder.org is very popular, there are currently only two additional public BinderHubs worldwide. This means that there is a limited expertise in operating and maintaining such public infrastructure.

In particular, we will help maintain the tools and documentation required to operate public infrastructure like Binder to meet the needs of the open-science community in the EU. In addition we will develop new features for the BinderHub project that will bring it to a wider audience and improve its usefulness for scientists based in the EU.

T5.1 Prototype European Binder instance and global federation M0-M48@.5; Sites: Simula (lead), EGI, UPSud, WildTree

The purpose of this task is to create a community of practice of BinderHub operators and operate a European Binder instance.

Re-executing the code associated with all European Open Access publications requires a very large amount of very diverse compute resources. It is unrealistic that one BinderHub instance could provide all of these. This is why in this task we will work on building a community of practice around operating public Binder instances at a research institution. Constraints from data protection and privacy mean not all code supporting a publication can be made public and/or that data can not be moved outside of the EU. This task will work on operating a EU based Binder instance accessible to the public. This will mirror the current mybinder.org but be hosted in the EU. This European Binder instance will serve as the first partner in the global federation of Binder instances.

We shall begin operating this service as a prototype from the beginning of the project, so that it is ready for testing by month 12 and stable and available for local demonstrators development in WP4 for month 24.

- Build a community of practice of BinderHub operators. Create a forum for Binder operators to exchange experience and tools.
- Operate a Binder instance hosted in Europe.
- Develop a federation of Binder instances and produce a report on best practices for coordinating several federated BinderHub instances across institutes. (D5.4)

T5.2 Integration with EOSC M12-M48@.7; Sites: EGI (lead), Simula, WildTree, EuXFEL

The task includes the following activities:

- Identify Infrastructure as a Service (IaaS) cloud compute, providers from EOSC to support the development of a BindeHub Infrastructure in EOSC,
- Support EOSC integration of science demonstrators from WP4,
- Operate the services in the EOSC IaaS cloud compute infrastructure,
- Harmonise access to the BOSSEE services with the EOSC AAI solution,
- Validate services with EOSC requirements,
- Register the resulting services in the EOSC services catalogue and marketplace,
- Set-up and align a Service Management System with EOSC,

All these activities will be reported in the periodic reports on EOSC operation during each reporting period (D5.2, D5.5, D5.7).

T5.3 Easy deployment of JupyterHub and BinderHub on a variety of infrastructure M0-M18; Sites: EP (lead), UPSud, Simula, WildTree

The JupyterHub or BinderHub installation tutorials mostly focuses on deployment on corporate cloud services (mainly GoogleCloud). For many academic institutions and SMEs, a deployment on their own infrastructure is preferable for a combination of reasons (better control on private data, easier integration with the local information system, the difficulty of funding the costs of external services, ...).

There are many pre-existing academic clouds managed by people with a high level of expertise and a thorough knowledge of their infrastructure and associated tools. These are often more cost-effective for research and education. This is also true for SMEs. The purpose of this task is to make the deployment of JupyterHub or BinderHub easier upon these academic cloud computing and to provide scalable and high-quality infrastructure for education and research. Most of them are based on two open-source softwares which can manage this kind of infrastructure: OpenStack and OpenShift. We will then focus on these two tools.
While platforms such as GoogleCloud already offer effective tools to deploy a Kubernetes cluster, this is not the case when we deploy our own infrastructure. The tools available for the two solutions mentioned above are numerous and it is often difficult to choose which ones are the most suitable and best maintained. We will therefore first take an interest in making a state of the art and explaining our choices. We already began this study at Ecole polytechnique (see the post in medium[5]) and we want to deepen it. At each step, we will use infrastructures available locally (OpenShift at Ecole polytechnique and OpenStack at University of Paris-Sud), available in France (OpenShift at Mathrice which is a "Groupements de Service" at CNRS of the research laboratories IT in French mathematics), available in Europe (EGI). We will document all the process to allow the community to easily deploy a JupyterHub or BinderHub on their own infrastructure based on OpenStack and/or OpenShift.
Once the infrastructure is in place, even if JupyterHub or BinderHub provide efficient teaching or research environments, one issue remains: the lack of choice in terms of storage devices and data persistence. Thus, we will provide a unified environment and focus on the possibility to mount various storage devices on the containers deployed on the Kubernetes cluster and provide persistence storage for Binder. This will be a key issue for the task **T2.2**.
The task includes the following activities:

- Prototypes / POC deployment on OpenStack and/or OpenShift
- Partial / Full automation of the deployment
- Documentation (→ D5.3)

Deliverables:
D5.1 (Due: 6, Type: R, Dissem.: PU, Lead: EGI) *A report describing the technical and service provisioning integration activities required to ensure the provisioning of BOSSEE services through EOSC* ⇝M3
D5.2 (Due: 18, Type: R, Dissem.: PU, Lead: EGI) *Report on the technical and service provisioning integration activities and their status, reporting period 1* ⇝M2
D5.3 (Due: 18, Type: R, Dissem.: PU, Lead: EP) *Documentation of how to deploy easily JupyterHub and BinderHub on OpenStack and OpenShift* ⇝M1
D5.4 (Due: 24, Type: R, Dissem.: PU, Lead: Simula) *Documentation and tools of how to easily deploy a public BinderHub on EU cloud infrastructure* ⇝M2
D5.5 (Due: 36, Type: R, Dissem.: PU, Lead: EGI) *Report on the technical and service provisioning integration activities and their status, reporting period 2* ⇝M3
D5.6 (Due: 48, Type: R, Dissem.: PU, Lead: Simula) *Documentation on best-practices of operating a federation of BinderHubs* ⇝M4
D5.7 (Due: 48, Type: R, Dissem.: PU, Lead: EGI) *Report on the technical and service provisioning integration activities and their status, reporting period 3* ⇝M4

[5] https://blog.jupyter.org/how-to-deploy-jupyterhub-with-kubernetes-on-openstack-f8f6120d4b1

Work Package 6: Education and Dissemination												**Start:** 0
Site	Simula	CNRS-ObAS	EP	EGI	EuXFEL	*INSERM*	QuantStack	UiO	UPSud	Silesia	WildTree	**all**
Effort	13	3	3	2	8	*12*	4	12	3	5	3	**68**

Objectives

The objective of this work package is to develop the community at the European scale, foster cross team collaboration, spread the expertise, and engage the greater community to participate in the definition and refinement of the requirements, and the implementation and use of the produced solutions. This includes:

- Ensure awareness of the results in the user community;
- Train researchers in best practices for open and reproducible science
- Educate the community on the value of open science
- Produce training and education material to disseminate the ability to do reproducible computational science using the tools we develop.
- Define individual exploitation plans;

Description

Open science is entirely dependent on researchers adopting open practices. In BOSSEE, we are developing tools to facilitate these practices, but they only work if researchers actually adopt them. Going further, it is also clear that open science is not just of value to researchers: one of the largest benefits of open science is that it makes science accessible to the broader public who may not be members of the research community.

To this end, in addition to training researchers, we will also train the public in how to make use of the open science research and services facilitated by BOSSEE. This will be done through regular open dissemination and training workshops, as well as by producing and maintaining material for online courses and documentation.

BOSSEE will develop, through WP4, a number of applications and demonstrators that will be disseminated in different ways. We will also participate in the concertation activities, consultations and other meetings and events of the European E-Infrastructure projects.

All the code, documents, test and build infrastructure produced as part of the project will be made available as open source. Open access to all publications resulting from the project will be ensured.

T6.1 Dissemination and communication activities M0-M48@.2; Sites: Simula (lead), INSERM, QuantStack, Silesia, UPSud, WildTree, EuXFEL

This task comprises all forms of direct dissemination and public communication activities such as press releases, creation of the project web-site including visitor analysis and monitoring tools, scientific and technical publications, outreach activities (seminars, keynote talks, media interviews, press releases), promotion through social media (e.g. Twitter, Facebook, LinkedIn), creation of advertisement materials such as flyers, posters, and electronic feeds as well as their distribution. We will use standard community building technology such as mailing lists, Wikis and Forums, to ensure dissemination and engagement of the community to support this. We will also generate press releases at appropriate moments.

T6.2 Training Workshops and community building M0-M48@.75; Sites: UiO (lead), CNRS-ObAS, EGI, EP, INSERM, QuantStack, Silesia, Simula, UPSud, WildTree, EuXFEL

This task will be in charge of:

- Defining and implementing a strategy to enable a shared vision of the Jupyter ecosystem across all the actors from developers, users to every stakeholder: the current misalignment hinders the full exploitation of Open Software practices where co-design is a de facto approach.
  For instance, the official Jupyter documentation (https://jupyter.org/documentation) solely reflects the view of developers where the Jupyter ecosystem is defined as a set of software packages (jupyter-core, jupyter-client, kernels, widgets (ipywidgets, ipyleaflet, etc.). The user vision is relegated to examplars (blogs, newsletters, etc.) which inevitably tend to be restrictive but often become de facto standards. This can lead to misconceptions and makes it more difficult for on-boarding novices and new communities.
- Triggering a cultural change to help under-represented groups to actively participate to the development of open source project to ensure the sustainability of the BOSSEE services deployed on EOSC-HUB.
- Foster Open innovation by collaborating with others from different background and activities (school, universities, industries, journalists, artists, etc.)

To achieve these goals, the following actions/activities will take place:

- co-design hackathons: the co-design efforts between domain scientists, BOSSEE developers and service providers will be carried out at any point in time of the project and will be registered in a co-design register to help for future engagement with new communities of users. To be fully effective, co-design hackathons will be organized to set the stage, define rules for co-design interactions and more importantly align all actors into a common user-centred vision of BOSSEE services and associated development towards a successful EOSC deployment.
- Workshops on Findable, Accessible, Interoperable and Reusable (FAIR) software and data to facilitate the adoption of Open Science and Open Scholarship best practices (transparent, sharable and collaborative Science): this would not be restricted to the Jupyter ecosystem and will teach users how to make data, lab notes and other research processes freely available, under terms that enable reuse (licensing), redistribution and reproducibility of methods and/or results.
- Trainings on how to use BOSSEE software and services to fully exploit BOSSEE developments for EOSC: develop training materials and organize training events for researchers and the public to enable Open Science and maximise the usefulness of BOSSEE developments.
- BOSSEE Admin trainings: training event for learning on how to deploy BOSSEE services such as BinderHub.
- Open call for open innovation mini-projects: mentored by BOSSEE staff and targeting SMEs, municipalities, journalists, artists, etc.
- Dissemination during conferences, such as EWASS (European Week of Astronomy and Space Science), ADASS (Astronomical Data Analysis and Software and Systems), and IVOA (International Virtual Observatory Alliance) meetings (for the astronomy demonstrator **T4.2**).

The work will be done in collaboration with CodeRefinery project which strongly support BOSSEE proposal and will be committing staff time for organizing and running workshops on Open Science best practices.

T6.3 Online resources for open science M0-M48@.3; Sites: INSERM (lead), CNRS-ObAS, EP, QuantStack, Silesia, Simula, UPSud, WildTree, EuXFEL

The aim of this task is to provide communities with online resources for Open Science and support **T6.2**. This task includes the following activities:

- Collaboration with the CodeRefinery project for the development and maintainance of the online lesson materials on open science best practices (JupyterLab, version control, collaboration and peer review, documentation, testing, software licensing, and reproducible research). Following CodeRefinery's tradition, the aim will be to contribute the lessons to Software Carpentry and Data Carpentry.
- Development of an interactive book on applied stochastic processes in Physics (D6.4). Unlike classical books in this subject, it will be supplemented by interactive numerical examples solving real life problems. This will be the occasion to emphasize the role of high performance computing in solving problems modeled by stochastic differential equations (SDE). The book will be available via the EOSC hub, and showcase the use of HPC hardware (GPU).
- Production of a video tutorial for the astronomy application on the CDS YouTube channel (`https://www.youtube.com/channel/UCUESQl7rNupLlV_VcceE0Ng`).

All material will be licensed under an open license such as CC BY-SA 4.0.

T6.4 Local Help Desk M0-M48@.2; Sites: UPSud (lead), CNRS-ObAS, EP, INSERM, QuantStack, Silesia, Simula, WildTree, EuXFEL

Dissemination events and tutorials are very effective tools for engaging scientists and giving them the desire to acquire new technologies and best practices. The next barrier to adoption comes when, back home, the scientists start using them on their daily problem. Having access to a local expert – even for a small amount of time – makes a huge difference, saving on the wasted time and frustration on the inevitable rough corners, and getting first hand advice and guidance in the rich landscape of available tools that could otherwise soon feel overwhelming.

At several of our sites, our Research Software Engineers will dedicate some fraction of their time to deliver such help to the local community, experimenting with various formats: help desk hours where scientists can drop by to get help; regular meet-ups where scientists can reconvene to work on their problems or on-demand tutorials with expert supervision and mutual help; in-lab visits to the scientists for more in-depth discussions; etc.

An explicit aim of this task is to foster the creation of sustainable Research Software Engineer groups within institutions to support their scientists.

This will be the occasion for our Research Software Engineers to witness first hand how users adopt or struggle with the projects technologies and services, and escalate the hurdles and barriers to adoptions as well as success stories. The sites will keep in close contact to exchange on the effectiveness of the various formats, and the outcome will be reported on in deliverables D6.1, D6.2, D6.3.

Deliverables:
D6.1 (Due: 18, Type: R, Dissem.: PU, Lead: INSERM) *Community building: Report on impact of development workshops, dissemination and training activities, reporting period 1* ⇝M2
D6.2 (Due: 36, Type: R, Dissem.: PU, Lead: INSERM) *Community building: Report on impact of development work-*

shops, dissemination and training activities, reporting period 2 ⇝M3

D6.3 (Due: 48, Type: R, Dissem.: PU, Lead: INSERM) *Community building: Report on impact of development workshops, dissemination and training activities, reporting period 3* ⇝M4

D6.4 (Due: 48, Type: DEM, Dissem.: PU, Lead: Silesia) *Interactive textbook on stochastic processes in physics* ⇝M4

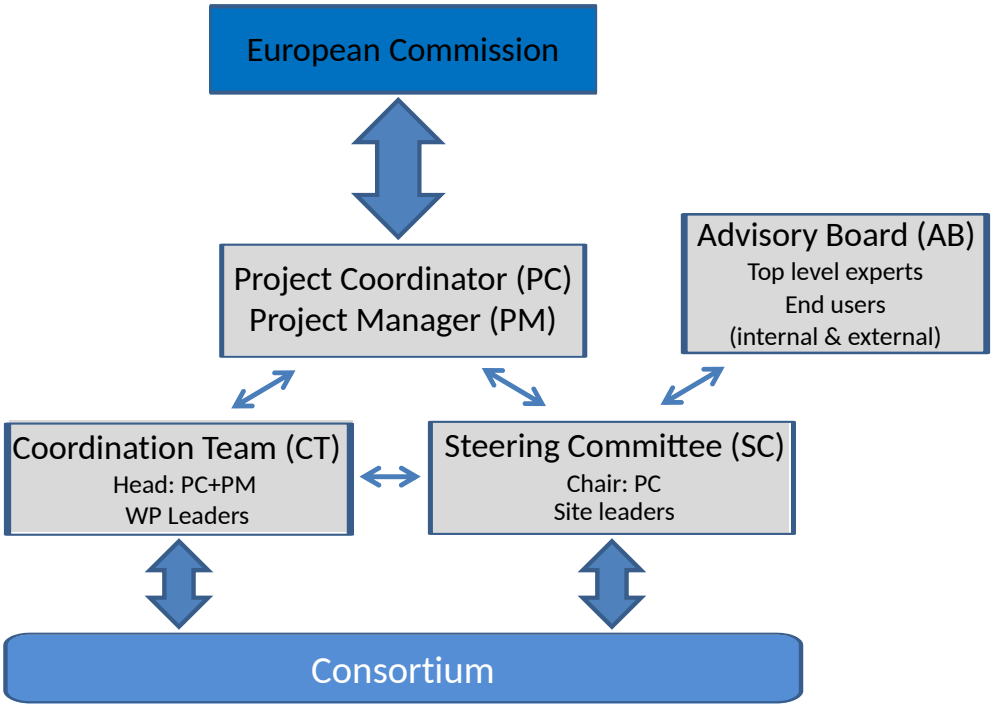


Figure 3.2.1: Management structure

3.2 Management Structure and Procedures

The management and decision-making approach of BOSSEE has been tailored to the real needs of the project and *overcomplexity has been avoided in favour of an efficient management organisation able to effectively guarantee project success*. To this end, the number of committees has been kept to the minimum and the rules will be flexible whilst still providing a good basis for efficient implementation of the project.

Project management activities in BOSSEE comprise a wide array of activities, including scientific and administrative management, guidance on decision making, contractual management, financial management, supervision of and compliance with ethical standards, management of knowledge and IPR issues, and coordination of communication activities. Most partners have extensive experience with EU funded projects, while two SMEs are recent start-ups. BOSSEE management will thus be tailored to varying needs and requirements of individual partners.

3.2.1 Management

The project will be coordinated by Simula, represented by Benjamin Ragan-Kelley (Project Coordinator), who has been a developer and leader in the Jupyter project since its inception in 2014, and IPython before that, and a site and work package leader in prior H2020 project, OpenDreamKit, and currently leads the JupyterHub and Binder open source projects in the Jupyter community.

The Project Coordinator will be assisted by a part-time (50%) Project Manager, Katarina Subakova, who has significant experience with Horizon 2020, both as a project manager and as a head of Horizon 2020 Helpdesk. Additional feedback and expertise will be brought by Financial, Legal and European affairs officers from Simula.

In addition, Hans Fangohr will act as Project Deputy, being constantly updated by the coordinator and manager on the project evolution in order to be able to temporarily take over the coordination in case the coordinator would be incapacitated.

3.2.2 Organisational structure and decision-making

The organisational structure, shown in the Figure 3.2.1, has been designed to enable efficient coordination of the project — the development and evaluation of an Open Science toolkit integrating several previously separated tools and software and involving both academic actors and industrial stakeholders. It is jointly agreed on by BOSSEE consortium and adapted to its size and composition, the tasks and duties of all partners involved.

We have designed the management structure and procedures to deal in a flexible manner with the following challenges:

- to integrate all consortium members and to mobilise their expertise, knowledge and networks at every stage of the project;
- to give the maximum attention to the end-users needs and requirements;
- to continuously involve expertise and knowledge of relevant stakeholders and their networks, and
- to efficiently coordinate the project implementation in a collaborative environment and ensure its sustainability.

The design has been largely adapted from that of OpenDreamKit which has proved very effective for this type of project, with some simplification as suggested by past experience:

- Suppression of the Quality Review Board: its role can effectively be subsumed by the steering committee; also the head of OpenDreamKit's Quality Review Board, Hans Fangohr, will carry over the accumulated experience and best practices, and the lessons learned from the OpenDreamKit quality review board will be shared within the BOSSEE consortium.
- Suppression of the End User Group: as in OpenDreamKit, all participants are either end users themselves, or in close contact with such end users. To guarantee the effectiveness of this simplified structure, we include end users in the advisory board.

The coordinator acts as an intermediary between the Partners and the European Commission. The coordinator will oversee the project planning, monitor that execution is carried out in time and that the objectives are achieved and closely interact with the project officer for project monitoring and delivery of the performance indicators. The Project Manager will ensure efficient day-to-day management of the project, reporting, feedback to partners on administrative, financial and legal issues, tracking of resource allocation and consumption, and communication inside and outside the consortium.

The resources of all partners will be mobilised by decentralisation of responsibilities through the assignment of leadership for work packages. Clear distribution of tasks, efficient decision making mechanisms and a sound financial management will safeguard the achievement of the project's objectives.

3.2.3 Milestones

For a description of the milestones and their motivations see Section 3.1.3; a tabulation of the milestones, which work packages are involved, and a means of verification can be seen in Table 3.2.1.

#	Name	WPs[6]/Deliverables involved	Mo	Means of Verif.
M1	*Startup, requirements, and prototype generic Jupyter service*	D1.1 D1.2 D1.3 D3.1 D3.2 D4.1 D5.3	12	Completed all corresponding deliverables and preparation for deployment of prototype services is underway
M2	*Generic Jupyter service and early local demonstrators*	D2.1 D2.3 D2.4 D4.2 D5.2 D5.4 D6.1	24	Completed all corresponding deliverables and early users are able to access and test prototype services
M3	*Demonstrator services and community engagement*	D2.5 D3.3 D3.4 D3.5 D4.3 D5.1 D5.5 D6.2	36	Completed all corresponding deliverables and services are operational and ready for public testing
M4	*Full EOSC integration, adoption, sustainability, and evaluation*	D1.4 D2.2 D2.6 D4.4 D5.6 D5.7 D6.3 D6.4	48	Completed all corresponding deliverables and reported progress at the final project review

Table 3.2.1: Milestones, Deliverables, and Verification

3.2.4 Project roles

The following bodies will form the organisational structure of the BOSSEE project : Coordination Team (CT), Steering Committee (SC), Advisory Board (AB).

Coordination Team (CT)

Members: The CT is composed of the Work Package leaders and headed by the Project Coordinator, assisted by the Project Manager.

Responsibilities: The CT is an executive body in charge of the project implementation and monitoring. It takes operational decisions necessary for the smooth execution of the project.

Tasks:

1. Monitoring the timely execution of the tasks and achievement of the objectives;
2. Preparation of scientific and financial progress reports;
3. Controlling Work Package progress by assessing it through technical reports developed by the partners;

[6]The work package number is the first number in the deliverable number.

4. Making proposals to the Steering Committee of re-allocation of tasks, resources and financial needs for the fulfilment of the work plan;
5. Preparing the drafts and validating the project deliverables to be submitted to the Commission.

Meetings: The coordination team will meet every 6 months. Meetings can be through video-conference on occasions. If necessary, extra meetings will be arranged.

Steering Committee (SC)

Members: The SC is chaired by the Project Coordinator and includes one representative from each partner organisation.

Responsibilities: The SC is the decision-making body in charge of the strategic orientation of the project. It takes decisions on scientific directions, re-allocation of resources, consortium changes and intellectual property rights.

Meetings: Annual meetings. Meetings can be through video-conference on occasions. If necessary, extra-meetings will be arranged. Written minutes of each meeting will be produced, which shall be the formal record of all decisions taken. A procedure for comment and acceptance is proposed.

Voting procedure: The SC shall not deliberate until a quorum of three-fourth (3/4) of all Members are present (possibly through video-conference) or represented. Each Member shall have one vote. The SC will work on consensual decisions as much as possible and resort to voting only if unavoidable. Voting decisions shall be taken by a majority of two-thirds (2/3) of votes with quorum two-thids (2/3) of the whole set of members. Exceptional decisions (large changes to the budget ($\geq$ 100k euros), evolution to the consortium, firing the coordinator, resolving ambiguity about whether something is a hard question) shall be taken by a majority of three-fourth (3/4) of votes with quorum three-fourth (3/4) of the whole set of members. Votes can be electronic.

Advisory board (AB)

Members: top level experts and end-users from partner and external organisations, from a variety of disciplines and both from academic and industrial sector. Together, they have a deep understanding of both market and technical problems, and an awareness of opportunities

Responsibilities: to give an independent opinion on steering, scientific and innovation matters, in order to guaranty quality implementation of the project, adequateness to the end-users needs, efficient innovation management, and project sustainability.

Meetings: at the request of the Steering Committee, including attending annual project meetings.

3.2.5 Project management tools and procedures

Project partners and management bodies will communicate through a dedicated project web platform, maintained by the Project Manager. WP leaders will monitor progress of participants of their WP at least monthly, and participants will inform their WP leaders when problems are encountered. Major problems will be discussed in (teleconference) meetings with the Project Coordinator and Project Manager. Each WP leader will be free to organise extra meetings with WP partners, if necessary. Scientific and financial progress reports will be collected, assembled and transmitted to the Project Coordinator by the WP leaders through the web platform. On basis of the Progress Reports, the Coordination Team will monitor progress of the project, identify bottlenecks and find solutions for these problems. Where needed, adaptations to the project plan will be made, with the aim of ensuring the delivery of the project results as agreed with the EC. Major adaptations need to be approved by the Steering Committee.

The Coordination Team, will ensure efficient innovation management. They will carefully monitor new opportunities in order to suggest, if necessary, to new directions to the Steering Committee. For legal aspects, the latter will have a feedback from legal officers from the Coordinator's European affairs office, specialised in Intellectual Property.

Our management structure and procedures will ensure that our network of partners from both academic and industrial sectors is focused at achieving the promised tasks and deliverables, efficiently managing the innovation process and largely opening the developed tools, insights and services to the EOSC and its final users. The partners will sign a Consortium Agreement, in which operational rules and decision making procedures will be laid down.

CONFLICT RESOLUTION - For the cases where contentious decisions need to be made, or conflicts are arising which cannot be solved bilaterally, steps to prevent further escalation and procedures to resolve disputes and decision-making procedures are laid down in the Consortium Agreement. In case of conflicts or a needed decision in a task or work package, the WP Leader shall resolve the conflict and agree on a solution or decision. If a solution of the conflict cannot be found or if a Party does not accept the proposed solution, the issue must be forwarded to the Steering committee (SC), which in its turn will try to resolve the issue together with the conflicting parties within 10 working days. If no agreement can be reached, the Project Coordinator (PC) can base its decision on the 2/3 majority vote of the SC or take a different decision. In any case, the Coordinator will take the final decision. All parties will make efforts that decisions are resolved at a level as low as possible.

3.2.6 Quality and Innovation management

BOSSEE Project Quality Management will ensure that the project meets a high quality level as well as satisfying the EU commission requirements. The basic approach to quality management is compatible with the International Project Management standards and guidelines (PMI).

- Quality planning: internal standards from participants and national and international standards and regulatory requirements will be considered to create the quality metric.
- Quality control: the overall project performance will be evaluated on a regular basis to provide confidence that the project satisfies the expectation and EU provisions.
- Quality assurance: specific project results will be monitored to determine if they comply with expectations and eliminate causes of unsatisfactory performance. To this aim, deliverables have been planned to give the necessary information and evidence that quality standards have been achieved.

In the context of the Responsible Research Innovation paradigm promoted by the European Commission in the H2020 framework, BOSSEE is willing to respond to societal desirability and societal acceptability question identifying new opportunities in the Open Science hub through innovation that is in the public interest. BOSSEE consortium recognises innovation as an essential element for driving business growth and maintaining competitive advantage. The innovation management work will prevent that BOSSEE results suffer from missing expected functionalities, which would lead to user dissatisfaction, and will ensure that the ethical and societal aspects are incorporated into the design process from the beginning.

3.2.7 Risks and risk management strategy

The risk in the project execution as planned is carefully assessed and managed. We base our plans on long standing experience, and we bring together the world's experts in the relevant tools and techniques.

A key feature of this project is the involvement of a wide set of partners from multiple domains. While this ensures complementary coverage of a wide set of skills and provides robustness in different ways, we will have to ensure that all the partners work together as a closely knit team.

Our open source approach means that all our code and outputs will be open and visible to anybody at sites like GitHub and Bitbucket throughout the project. It is common for some users to run the latest development versions of computational and infrastructure software, thus beta-testing code between major releases. This reduces the risk of developing software which people won't use: where our design decision or technical approaches are controversial, this will be detected early by those users, giving the consortium useful feedback to consider.

As part of the Management Work Package, and with support from the Coordination Team, the project coordinator will maintain and regularly update a Risk Management Plan; at the end of each Reporting Period, this updated plan will be included in the project's Technical Report. It will identify and categorise all potential strategic risks (legal, financial, human resources risks, etc.) to the successful delivery of the project, their probability and impact. For each risk area, mechanisms for risk mitigation will be identified and contingency actions will be proposed.

Risks will be evaluated in terms of project goals and objectives, according to the following four steps:

1. Identification of risks using a structured and rational approach to ensure that all areas are addressed.
2. Quantitative assessment and ranking of the risks.
3. Definition of procedure to reduce (or minimize) risk.
4. Monitoring and management of risks throughout the project life with milestone review and reassessment.

Finally, as reported above, a conflict resolution mechanism will be put in place, whereby decision making divergence and conflicts that cannot be solved at the Steering Committee (SC) level will be submitted to the Coordinator. The mediation and resolution process used is the following:

- Case presented by the involved parties.
- Development of a fact-based and neutral report by the coordinator to be provided to the conflicting parties and SC.
- Final decision to resolve the conflict made by SC.

An initial risk assessment appears as Table 3.2.2.

Risk	Level without / with mitigation	Mitigation measures
General technical / scientific risks		
Implementing infrastructure that does not match the needs of end users	High/Low	Many of the members of the consortium are themselves end-users with a diverse range of needs and points of views; hence the design of the proposal and the governance of the project is naturally steered by demand; besides, because we provide a toolkit, users have the flexibility to adapt the infrastructure to their needs. In addition, the open source nature of the project facilitates and promotes the involvement of the wider community in terms of providing feedback and requesting additional features via platforms such as GitHub and Bitbucket on a regular basis.
Lack of predictability for tasks that are pursued jointly with the community	Medium/Low	The PIs have a strong experience managing community-developed projects where the execution of tasks depends on the availability of partners. Some tasks may end up requiring more manpower from BOSSEE to be completed on time, while others may be entirely taken care of by the community. Reallocating tasks and redefining work plans is common practice needed to cater for a fast evolving context. Such random factors will be averaged out over the large number of independent tasks.
Reliance on external software components	Medium/Low	The non trivial software components BOSSEE relies on are open source. Most are very mature and supported by an active community, which offers strong long run guarantees. The other components could be replaced by alternatives, or even taken over by the participants if necessary.
EOSC validation and integration	High/Low	In order to deploy services on EOSC and make them available, we must make sure our services are validated and operational to to the satisfaction of EOSC. By including EGI as a partner, we have the support and expertise of a core EOSC participant, to help us navigate EOSC requirements for BOSSEE services.
Management risks		
Recruitment of highly qualified staff	High/Medium	Great care was taken coordinating with currently running projects to rehire personnel with strong track record, and identifying pool of candidates to hire from, notably in the developers community of software related to the project. This was favoured by the partners' long history of training and outreach activities. In addition, we have a critical mass of qualified staff in the project enabling us to train and mentor new recruits.
Different groups not forming effective team	Medium/Low	The participants have a long track record of working collaboratively on code across multiple sites. Aggressive planning of project meetings, workshops and one-to-one partner visits will facilitate effective teamwork, combining face-to-face time at one site with remote collaboration.
Partner leaves the consortium	High/Low	If the GA requires a replacement in order to achieve the project's objectives, the consortium will invite a new relevant partner in. If a replacement is not necessary, the resources and tasks of the departing partner will be reallocated to the alternative ones within the consortium.
Dissemination risks		
Impact of dissemination activities is lower than planned.	Medium/Low	Partners in the consortium have a proven track record at community building, training, dissemination, social media communication, and outreach, which reduces the risk. The Project Coordinator will monitor impact of all dissemination activities. If a deficiency is identified, the consortium will propose relevant corrective actions.

Table 3.2.2: Initial Risk Assessment

3.3 Consortium as a Whole

The BOSSEE consortium spans the broad spectrum of actors required for successfully developing an apt and easy-to-navigate sustainable service accessible through the EOSC hub catering to the needs of the European scientific community. It is composed of four academic institutions, four research centers, one e-Infrastructure federation, and two SMEs based in six different countries (Norway, France, Netherlands, Germany, Poland, Switzerland). The Consortium ensures a critical mass of scientific expertise and excellence in key areas (astronomy, geosciences, health sciences, mathematics, photon science, education) with research organisations and SMEs of recognised international reputation. Namely, BOSSEE consortium brings in:

- A set of use cases that cover several application domains and users, and that impose very diverse requirements on EOSC infrastructure (European XFEL, CNRS-ObAS, UiO, INSERM, Paris Sud);
- Lead developers in the Jupyter Ecosystem, including IPython, the Jupyter Notebook, JupyterLab, JupyterHub, Binder, MyBinder.org, Jupyter Widgets located at Simula, European XFEL, QuantStack, and WildTreeTech, as referenced in section 1.3.2.
- Experts and major promoters of the Jupyter collaborative user interfaces for interactive and exploratory computing in a variety of scientific domains (Centre de Donnees astronomiques de Strasbourg, European XFEL, INSERM, QuantStack, Paris-Sud, École Polytechnique, Silesia, University of Oslo).
- A long experience and proven track record of success with large and complex collaborative projects, including European E-Infrastructure projects (XFEL, Simula, UPSud, Silesia), projects focused on large-scale infrastructures and large experimental services (EGI, XFEL), as well as experience in running large scale open source projects (Jupyter project).
- A comprehensive range of skill sets and competencies in several relevant domains, from applied research to standardisation to business analysis.

The consortium has developed through collaborations and common interests over recent years. Some partners have been working together on different aspects of Jupyter and software for education for many years (European XFEL, QuantStack, Simula, Wild Tree Tech). Meanwhile, others joined together during a previous successful H2020 European Research Infrastructure project OpenDreamKit #676541 (European XFEL, Paris-Sud, Silesia, Simula). OpenDreamKit's review meetings and participation to H2020 E-infrastructure events, as well as community Jupyter workshops (organized by e.g. by École Polytechnique, Paris-Sud, Simula, UiO) led to connections with UiO, INSERM, and EGI, and to collaborations such as a prototype deployment of Binder on EGI's infrastructure. The connection with CNRS-ObAS grew out of a workshop coorganized back in 2013 by Paris-Sud on mathematical databases, where CDS's head Françoise Genova was invited speaker; this later led to her joining OpenDreamKit's Advisory Board.

Finally, we note that the project partners are long time passionate advocates of Open Science; building on highly successful past experience with OpenDreamKit, they *have chosen to write this proposal fully in the open* on GitHub (http://github.com/bossee-project/proposal) for maximum transparency and engagement of the community. We have used the same open source collaboration tools and practices as the Open Source Open Science community.

In addition to the joint planning and writing period for this proposal, the project partners have been interacting through a number of other activities, including:

1. Joint software development
    - Jupyter Notebook (EuXFEL, Simula, WildTree, QuantStack)
    - Jupyter Widgets (QuantStack, Silesia, UPSud, Simula)
    - Binder and repo2docker (Simula, WildTree)
    - thebelab (QuantStack, Simula, UPSud)
    - nbval (EuXFEL, Simula)
2. Joint projects
    - OpenDreamKit (EuXFEL, Simula, UPSud, Silesia)
    - PaNOSC (EGI, EuXFEL)
    - Computing in high school science education - iCSE4school, Erasmus+ Strategic Partnerships, (Simula, Silesia), 2014-2017
    - Computers in Science Education:iCSE (Silesia, UiO), funded by EFS, 2011-2014.
3. Joint publications
    - *Jupyter Notebooks – a publishing format for reproducible computational workflows* [Klu+16] (EuXFEL, Simula, QuantStack)
4. Collaboration
    - MyBinder for teaching and reproducibility (EuXFEL, Simula, WildTree)

- Widgets for computational science (EuXFEL, QuantStack, Silesia)
- OpenGATE, Monte Carlo platform for medical applications (INSERM, UPSud)
- Life Science Grid Community (EGI, INSERM)
- JupyterDays events organised at École Polytechnique, Paris-Sud, and other sites, with participation from other sites (CNRS-ObAS, EP, UPSud, Simula).
- Research Bazaar workshops on Jupyter, Binder, and reproducibility (Simula, UiO)

Table 3.3.1 shows a summary of the links between partners.

	Simula	CNRS-ObAS	EP	EGI	EuXFEL	INSERM	QuantStack	UiO	UPSud	Silesia	WildTree
Simula			●		@@@●★		★	●	●●	@●	@@@
CNRS-ObAS			●						●		
EP	●	●					@●		●●		●
EGI					●	●					
EuXFEL	@@@●★			●			★		●	@●	@
INSERM				●					●		
QuantStack	★		@●		★				@●		
UiO	●									●	
UPSud	●●	●	●●		●	●	@●			●	●
Silesia	@●				@●			●	●		
WildTree	@@@		●		@				●		
joint	★≙ publication, ●≙ project, ○≙ organization, @≙ software/resource dev, ☺≙ supervision										

Table 3.3.1: Previous Collaboration between BOSSEE members

3.4 Resources to be Committed

BOSSEE applies for a total budget of **€ 5,956,255.00** as the amount required achieving the objectives. The total budget is described in the subsequent sections together with the staff effort necessary to implement the action (Table 3.4.1). The necessary physical resources, the quantities of each and when they would be needed has been carefully determined on the base of the following criteria:

1. **Historical information** the partners involved are well experienced in the Jupyter development, software for education, processing and scale-up; past experiences from each partner has been taken into consideration to evaluate what, and in what quantities, different resources will be needed in the project.

2. **Work plan structure** the deliverables and milestones identified in the project work plan.

3. **The inputs necessary to resource planning**; further, the timetable of activities has helped to identify when each resource will be needed in the project.

4. **Resource pool description** the resources available in the consortium have been carefully analysed in order to avoid any duplicating of existing resources and allocate efficiently and effectively the resources necessary.

5. **Cost estimating** The approximated costs of the resources needed to complete successfully the project activities has been estimated on the base of: a) Resource planning results, b) Activity duration estimation (Gantt and effort form), c) Commercially available data on durables and consumables needed, d) Preliminary Risk Assessments. The financial allocation among the 11 partners reflects the tasks committed by each partner and the collaborative nature of the project itself. On the whole, the financial allocation is well-balanced and homogeneous.

3.4.1 Management Level Description of Resources and Budget

Staff efforts The BOSSEE project is gathering sites with core developers from the Jupyter project and history in open source software development, and brings them together with domain specialists from a range of domains. The major investment of the project is in software development, which is realised through person time and displayed in table 3.4.1.

WP	Title	Simula	CNRS-ObAS	EP	EGI	EuXFEL	INSERM	QuantStack	UiO	UPSud	Silesia	WildTree	total
WP1	Management	*24*	3	3	3	3	3	3	3	3	3	3	**54**
WP2	Core	*30*		13		16		15		14		8	**96**
WP3	Ecosystem	30		20		54		*20*		2	4	6	**136**
WP4	Demonstrators	9	12	3	7	*36*	24	6	12	20	12	3	**144**
WP5	EOSC	18		10	12	6				0		*13*	**59**
WP6	Education	13	3	3	2	8	*12*	4	12	3	5	3	**68**
totals		**124**	**18**	**52**	**24**	**123**	**39**	**48**	**27**	**42**	**24**	**36**	**557**

Efforts in PM; WP lead efforts light gray italicised

Table 3.4.1: Summary of Staff Efforts

Travel, dissemination, and outreach The nature of this proposal – of providing a framework that allows design and deployment of innovative services – means that the project has the potential to have high impact for EOSC. At the same time, it requires input from and engagement with a significant number of stakeholders, including potential users of the services such as scientists, developers of other services and other EOSC-funded projects, other open science software projects, and the developing EOSC itself. Consequently, requirements capture, networking, feedback, training and education workshops and outreach activities are all important, and the second highest cost for this project.

Guidelines for travel and dissemination We use the following guidelines for expected travel expenses: €2500 for attendance of a typical one week international conference outside Europe (including travel, subsistence, accommodation and registration), €1250 for a corresponding conference in Europe, €750 for a one-week visit of a project partner, for instance for coding sprints and one-to-one research visits. We expect a similar cost per week while hosting visitors. For the half-yearly project meetings, we expect on average a cost of €500 for travel, accommodation and subsistence.

Anticipated activities:

1. *Project meetings*: For the 9 project meetings that take place every 6 months, we expect the PI from each site to attend all of them (cost of 9 * 500 = €4500). For a researcher, we also expect that they attend all such project meetings (€4500). [7] We include in this item local expenses for the organization and catering of meetings, travel, accommodation and subsistence of attendants.

2. *Hosting visitors*: We expect that the site spends €2000 per year to host external visitors contributing to the project (total €8000).

3. *Site visits*: We expect the researcher to carry out 3 one-week visits to other sites (each at €750) every year, totalling 3 * 4 * 750 = €9000 over 4 years).

4. *Conference dissemination*: We expect the researcher to attend on average 1 international conference and 1 European meeting per year (cost of 4 * 2500 + 4 * 1250 = €15000) and the investigator to attend the equivalent of one international or two European gatherings (totals €10000).

5. *Advisory board*: For organisation, catering and attendance of 5 advisory board members to 5 meetings (kickoff and then annually), we budget €750 per person and meeting, and allocate the total of €18750 to Simula's travel budget only.

Where there are multiple investigators per site, they will share the travel and associated costs outlined above. Where there are multiple researchers, or researchers not employed for the full 48 months, the travel budget is adapted accordingly.

Guidelines for outreach costs *Publication charges*: We also request €3000 per partner to pay for open access publication charges. (Some partners have other means do pay these costs, and for them these are not needed.)

Workshops: We request funds for dissemination and outreach activities such as workshops that facilitate community building, provide training and disseminate best practice and encourages sustained contributions of the community to the project and beyond the lifetime of the funding. For a one-week workshop that we organise, we will typically use meeting rooms of the project partners to minimise cost, and assume a cost of 400 EUR per participant to provide location and subsidise accommodation and catering for attendees. A workshop for 10 people will thus cost about €4000. Participants donate their time and need to fund their travel from other sources. By partially contributing to the attendance cost, we hope to enable PhD students to engage with the project and expect positive effects on the sustainability of the activities, by embedding the tools and knowledge with the next generation of scientists. To support dissemination of the project, we also expect modest expenses for websites and logo design, flyers, posters, goodies, explainer comics or hiring professional support when needed.

[7] PI and researcher roles are defined for each site in section 4.1

Details are given in the tables in section 3.4.2 below and in the work packages.

3.4.2 Resource summaries for consortium member sites

In this section we briefly describe the requested resources. See the participant descriptions in the description of the consortium for the specific role of each member.

Some partners do not require all the costs outlined above in section 3.4.1 and have accordingly reduced their requirements below.

Resources Simula Research Laboratory Simula requests 124 person months to provide the effort required.

Simula	**Cost (€)**	**Justification**
Travel	98250	Travel for PI and 2 researchers and the advisory board (see 3.4.1)
Workshops	16000	Workshops (40 attendees in total) (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Other goods and services	3500	Financial audit
	7500	Consumables (3 High Performance laptops for workshops, sprints, dissemination)
Total	128250	

Table 3.4.2: Overview: Non-staff resources to be committed at Simula Research Laboratory (all in €)

Resources CNRS Observatoire Astronomique de Strasbourg CNRS-ObAS requests 18 person months to provide the effort required.

CNRS-ObAS	**Cost (€)**	**Justification**
Travel	13750	Travel for PI and ≈0.4 researchers (see 3.4.1)
Workshops	4000	Workshop (10 attendees in total) (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Total	20750	

Table 3.4.3: Overview: Non-staff resources to be committed at CNRS-ObAS (all in €)

Resources École Polytechnique EP requests 52 person months to provide the effort required.

EP	**Cost (€)**	**Justification**
Travel	45000	Travel for PI and 1 researchers (see 3.4.1)
Workshops	8000	Workshop (20 attendees in total) (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Other goods and services	5000	Financial audit
Total	61000	

Table 3.4.4: Overview: Non-staff resources to be committed at École Polytechnique (all in €)

Resources EGI EGI requests 24 person months to provide the effort required.

EGI	**Cost (€)**	**Justification**
Travel	20750	Travel for PI and 0.5 researchers (see 3.4.1)
Other goods and services	4500	Financial audit
	180000	Cloud computing (180k; 5k/month) for operation of services including **T5.1**.
Total	205250	

Table 3.4.5: Overview: Non-staff resources to be committed at EGI (all in €)

Resources European XFEL EuXFEL requests 123 person months to provide the effort required.

EuXFEL	Cost (€)	Justification
Travel	79500	Travel for PI and 2 researchers (see 3.4.1)
Workshops	16000	Workshops (40 attendees in total) (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Other goods and services	6500	Financial audit
	7500	Consumables (3 High Performance laptops for workshops, sprints, dissemination)
Total	112500	

Table 3.4.6: Overview: Non-staff resources to be committed at European XFEL (all in €)

Resources INSERM INSERM requests 39 person months to provide the effort required.

INSERM	Cost (€)	Justification
Travel	37300	Travel for PI and ≈0.8 researchers (see 3.4.1)
Workshops	4000	Workshop (10 attendees in total) (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Other goods and services	2500	Consumables (1 High Performance laptop for workshops, sprints, dissemination)
Total	46800	

Table 3.4.7: Overview: Non-staff resources to be committed at INSERM (all in €)

Resources QuantStack QuantStack requests 48 person months to provide the effort required.

QuantStack	Cost (€)	Justification
Travel	51000	Travel for PI and 1 researcher (see 3.4.1)
Workshops	8000	Workshops (40 attendees in total) (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Other goods and services	4000	Financial audit
	2500	Consumables (1 High Performance laptop for workshops, sprints, dissemination)
Total	68500	

Table 3.4.8: Overview: Non-staff resources to be committed at QuantStack (all in €)

Resources University of Oslo UiO requests 27 person months to provide the effort required.

UiO	Cost (€)	Justification
Travel	28060	Travel for PI and ≈0.5 researchers (see 3.4.1)
Workshops	8000	Workshop (20 attendees in total) (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Other goods and services	2500	Consumables (1 High Performance laptop for workshops, sprints, dissemination)
Total	41560	

Table 3.4.9: Overview: Non-staff resources to be committed at University of Oslo (all in €)

Resources University Paris-Sud UPSud requests 42 person months to provide the effort required.

UPSud	Cost (€)	Justification
Travel	35375	Travel for PI and ≈0.75 researchers (see 3.4.1)
Workshops	13500	2 workshops (34 attendees in total) (see the guidelines 3.4.1)
Total	48875	

Table 3.4.10: Overview: Non-staff resources to be committed at University Paris Sud (all in €)

Resources University of Silesia Silesia requests 24 person months to provide the effort required.

Silesia	Cost (€)	Justification
Travel	25750	Travel for PI and 0.5 researchers (see 3.4.1)
Publication charges	3000	Open access publication charges (see 3.4.1)
Other goods and services	52000	Contracted development
	2500	Consumables (1 High Performance laptop for workshops, sprints, dissemination)
Total	83250	

Table 3.4.11: Overview: Non-staff resources to be committed at University of Silesia (all in €)

Resources Wild Tree Tech WildTree requests 36 person months to provide the effort required.

WildTree	Cost (€)	Justification
Travel	36145	Travel for PI and 0.75 researchers (see 3.4.1)
Other goods and services	4000	Financial audit
	2500	Consumables (1 High Performance laptop for workshops, sprints, dissemination)
Total	42645	

Table 3.4.12: Overview: Non-staff resources to be committed at Wild Tree Tech (all in €)

4 Members of the Consortium

4.1 Participants

4.1.1 Simula: Simula Research Laboratory (NO)

Simula is an internationally-leading Norwegian research institute in the key ICT areas: communication systems, scientific computing and software engineering. Simula's research areas have been evaluated with the highest score by international expert panels in several national evaluations.

Dedicated to tackling scientific challenges with long-term impact and of genuine importance to real life, Simula offers an environment that emphasises and promotes basic research. This translates into numerous projects funded by the EU, Norwegian government or regional institutions, that Simula was involved in. In 2017, it successfully concluded Norwegian Centre of Excellence for Biomedical Computing and is currently hosting the Centre for Research-based Innovation, Certus. In addition, Simula is deeply involved in research education with 35 PhD students, 40 master's students, and 20 postdoctoral fellows supervised annually; and application-driven innovation and commercialisation, where it owns parts of 16 start-up companies with 110 employees.

The Department for Numerical Analysis and Scientific Computing (SCAN) aims to develop mathematical methods and scientific tools to reach new understanding of complex physical processes. It targets fundamental medical and industrial problems where new insights from mathematical modelling can advance today's knowledge. The department has hosted a ten-year Norwegian Centre of Excellence in Biomedical Computing (2007-2017), one of the most prestigious research environments in Norway, targeting ambitious and groundbreaking research. The department received top scores in all six evaluations carried out by the Research Council of Norway and is running a multitude of national and international research projects, including one ERC Starter Grant project.

Curriculum vitae

Benjamin Ragan-Kelley (leadPI, male, 28 PM) Benjamin Ragan-Kelley is one of the core maintainers and developers of the Jupyter and IPython projects, and currently leads the JupyterHub and BinderHub development teams. He has been a contributor to these projects since 2006, prior to the establishment of Jupyter as a separate project from IPython. He is an expert in all levels of Jupyter development, especially the aspects of deploying Jupyter-based services, which is the focus of this proposal. Benjamin will lead BOSSEE.

Beyond Jupyter, Benjamin has contributed widely to open source software, especially in the scientific Python community. He is a maintainer of numerous scientific packages in the conda-forge package management system, building packages used widely in education and research, such as PETSc, MPICH, and FEniCS.

Benjamin is a Research Engineer in the department of Scientific Computing and Numerical Analysis at Simula Research Laboratory in Oslo, Norway, where his primary responsibility is developing and maintaining the Jupyter software ecosystem, as well as supporting research scientists in diverse fields, including biomedical computing.

Prior to his current position at Simula, Benjamin received his Bachelor's degree *Magne cum Laude* in Engineering Physics in 2007 from Santa Clara University and his PhD in Applied Science and Technology from the University of California, Berkeley in 2013. He worked as a postdoctoral fellow at Simula Research Laboratory prior to becoming a permanent Research Engineer. He was honored along with the rest of the Jupyter steering council with the 2017 ACM Software System Award for Jupyter.

Katarina Subakova (PM, female, 24 PM) Katarina Subakova has 9 years of experience within the EU research agenda. Currently, she holds the position of EU Funding Manager at Simula Research Laboratory in Oslo, Norway, where her responsibilities include financial and administrative management of all H2020 projects, supporting scientist in identifying funding and developing proposals, and identifying future research challenges. In addition, she serves as an external consultant to Telenor Norway for all projects funded under Horizon 2020 scheme. Prior to her current position, she headed the Horizon 2020 Helpdesk service offered by the European Commission.

She holds the IAPP certification on General Data Protection Regulation (GDPR).

Katarina will lead the administrative management of the BOSSEE and she will act as the project's Data Protection Officer (DPO).

Research Engineer x2 (R, 72 PM) We will hire two postdoctoral-level research engineers to carry out the work at Simula, under the leadership of and together with Dr. Ragan-Kelley. The fellow will have a background in computational science, combined with IPython and Jupyter Notebook experience, and past experience of software engineering. An ideal candidate will also have good communication skills and team working abilities, and in particular interest and skill in the development and operation of software services to best support this part of the project.

Publications, products, achievements

1. 2017 ACM Software System Award for Jupyter
2. M. Bussonier, J. Forde, J. Freeman, B. Granger, T. Head, C. Holdgraf, K. Kelley, G. Nalvarte, A. Osheroff, M. Pacer et al. Binder 2.0 - Reproducible, interactive, sharable environments for science at scale In Python in Science ConferenceProceedings of the 17th Python in Science Conference. Austin, Texas: SciPy, 2018.

3. J. Forde, T. Head, C. Holdgraf, Y. Panda, G. Nalvarte, M. Pacer, F. Perez, B. Ragan-Kelley and E. Sundell. Reproducible Research Environments with Repo2Docker In ICML 2018 Reproducible Machine Learning. ICML, 2018.
4. T. Kluyver, B. Ragan-Kelley, F. Perez, B. Granger, M. Bussonier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay et al. Jupyter Notebooks: a publishing format for reproducible computational workflows In 20th International Conference on Electronic Publishing. IOS Press, 2016.

Relevant projects or activities

1. OpenDreamKit (GA No. 676541) Open Digital Research Environment Toolkit for the Advancement of Mathematics, participant, Work Package lead
2. Jupyter - collaboration with UC Berkeley, Cal Poly, funded by Gordon & Betty Moore Foundation, Alfred P. Sloan Foundation, and Helmsley Trust
3. Binder - collaboration with UC Berkeley, funded by Gordon & Betty Moore Foundation

Significant infrastructure

The fully owned Simula subsidiary Simula Innovation handles pre-commercial innovation projects, creation and follow-up of company spin-offs, and general support for entrepreneurs.

4.1.2 CNRS-ObAS: CNRS-Observatoire astronomique de Strasbourg (FR)

The Observatoire Astronomique de Strasbourg (ObAS) is a Joint Research Unit (UMR7550) of the CNRS and of the Université de Strasbourg. ObAS hosts the Centre de Données astronomiques de Strasbourg (Strasbourg astronomical Data Centre CDS, http://cds.unistra.fr). Since its creation in 1972, the CDS has been providing reference services which are widely used by the world-wide astronomical community with more than 1 million queries/day on average in 2017. The CDS is labelled as a Research Infrastructure in the French national Research Infrastructure Roadmap. Since 2006 CDS has been the coordinator, on behalf of the CNRS, of all the projects funded by the European Commission to support the implementation of European Virtual Observatory.

Curriculum vitae

Mark Allen (leadPI, male, 0 PM) Mark Allen is the Director of the Centre de Données astronomiques de Strasbourg (Strasbourg astronomical Data Centre CDS, http://cds.unistra.fr).

He is a CNRS director of research at the Observatoire Astronomique de Strasbourg, He has 17 years experience implementing e-Science projects in Astronomy within the CDS and EC funded European Virtual Observatory (Euro- VO) projects. He is the Chair of the IVOA Executive Board, and has served as the chair of the IVOA Committee for Science Priorities, the chair of the IVOA Applications Working Group and as IVOA Executive Secretary. As Euro-VO project scientist he has engaged and coordinated astronomical data centres, software developers and scientists in the development and use of the Virtual Observatory framework including support to the astronomy community via leading schools and workshops at the national and European levels. His astronomical interests include active galactic nuclei and comparison of theoretical plasma models to multi-wavelength observations.

Thomas Boch (R, male, 3 PM) Thomas Boch is a Research Engineer in charge of service integration at CDS.

He is the developer of Aladin Lite, an interactive sky atlas running in the browser, used and deployed by more than 50 professional astronomy sites. He collaborated with large agencies (ESA - European Space Agency, ESO - European Southern Observatory) to help them integrate Aladin Lite into their own web portal.

He developed the CDS portal which provides with a single entry point to CDS services. He is also co-author of several Virtual Observatory standards, including HiPS and MOC.

He is actively developing and supervising the development of several Python packages to allow for access and visualisation of astronomical data: mocpy, hipspy, astroquery.cds, ipyaladin (Jupyter widget enabling Aladin Lite in the Jupyter notebook).

Sébastien Derriere (R, male, 3 PM) Sébastien Derriere works as Astronome Adjoint for the CDS.

He has a long experience in the management of astronomical metadata, and the cross-identification and statistical classification of astronomical sources. He has also been involved in the Virtual Observatory (VO) project, for the definition of standardized vocabularies, and implementation of the Registry for CDS services.

He has contributed to defining and developing some astronomy portals based on widgets, at CDS, or for the ASTRODEEP FP7 project (Grant Agreement n.312725).

He has participated in many technical workshops, VO schools and training events, to disseminate the usage of VO tools to the community. For example, in 2018, he led a tutorial during the ADASS conference, including the usage of a Python notebook (http://cds.unistra.fr/adass2018/), and contributed to the 4th ASTERICS (Horizon 2020, Grant Agreement n.653477) School. He also created a number of YouTube video tutorials on the usage of CDS services.

Software Engineer (R, 12 PM) We intend to hire a software engineer for 12 months during the project to be be supervised at CDS to work on developments defined in Task **T4.2**. We aim to hire someone in the period of months 12-24.

Publications, products, achievements

1. F. Genova, M. G. Allen, C. Arviset, A. Lawrence, F. Pasian, E. Solano, J. Wambsganss, Euro-VO - Coordination of Virtual Observatory activities in Europe, Astronomy and Computing, 2015, Volume 11, p. 181-189
2. F. Genova et. al 2017, Building a Disciplinary, World-Wide Data Infrastructure. Data Science Journal, 16: 16, pp. 1-13, DOI: https://doi.org/10.5334/dsj-2017-016
3. P. Fernique, M. G Allen, T. Boch, A. Oberto, F-X. Pineau, D. Durand, C. Bot, L. Cambresy, S. Derriere, F. Bonnarel, F. Genova, Hierarchical Progressive Surveys - Multi-resolution HEALPix data structures for astronomical images, catalogues and 3-dimensional data cubes. 2015, A & A, 578, A114
4. M. Baumann, T.Boch, New Python developments to access CDS services, Proceedings of The Astronomical Data Analysis Software and Systems conference 2018.

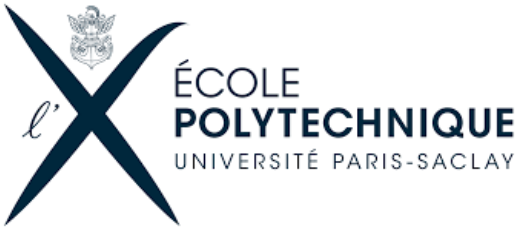
Relevant projects or activities

1. ESCAPE (EC funded project, 2019-2023, #824064, INFRAEOSC-0402018)
2. ASTERICS (EC funded project, 2015-2019, #653477, Research and Innovation Action)
3. AENEAS (EC funded project, 2017-2020, #731016, Research and Innovation Action)
4. CoSADIE - Collaborative and Sustainable Astronomical Data Infrastructure for Europe (EC funded project #312559, 2012-2015, CSA)
5. RDA Europe - the European plug-in to the Research Data Alliance (RDA) (EC funded project #632756, 2014-2016, CSA)

Significant infrastructure

CDS is data centre for reference astronomy data. CDS runs physical infrastructure for 1Petabyte. Connected to French national network RENATER. Computing power for X-Match and generation of all-sky survey data. Currently running prototype notebook servers.

4.1.3 EP: École polytechnique (FR)



École Polytechnique (l'X), is the leading member of Grandes Écoles in France for science and technology according to all French rankings last year, ranked 2nd best small university in the world by Times Higher Education in 2018, and 16th best European university by QS World University Rankings in 2018. With fewer than 3000 Bachelor, Masters, and PhD students enrolled at any time, the institution has produced 4 Fields Medalists, 4 Nobel prize winners, and 3 French presidents, ranking it 6th worldwide in terms of number of Nobel prize recipients. L'X is composed of 22 laboratories supporting research in physics, engineering, mathematics, biology, and chemistry (among others) and is connected to several French national research institutions such as CNRS, CEA, and INRIA. L'X prides itself on promoting multidisciplinary research through collaboration between its many labs and with external, national and international partners and a strong connection with industry and start-up.

The Mathematics department at Ecole polytechnique has started a reform of their various teaching offer based on Jupyter. Many courses from the Bachelor program, Engineering school, and the Master program have already begun relying on a strong use of Jupyter notebooks / JupyterHub[8] and this will continue with a strong support of the Dean of undergraduate studies and of graduate studies. In addition, several software and research engineers in applied mathematics have been recruited and participate, together with a system engineer, to this effort in order to help in terms of building an infrastructure dedicated to Jupyter. This infrastructure together with support from the head of the Ecole polytechnique, fosters the dissemination into various other departments (Physics, Mechanical Engineering, Biology...), where some Jupyter-based courses are already starting. A community is emerging, notably through the strong engagement of students in the computer science club of École polytechnique (Binet Réseau).

Curriculum vitae

Loic Gouarin (leadPI, male, 8 PM) Loic Gouarin is Research Engineer in scientific computing at CMAP (Centre de Mathématiques Appliquées) which is a Joint Research Unit (UMR7641) of the CNRS and École polytechnique. He works on several scientific computing open-source projects in different fields such as Lattice-Boltzmann methods, Stokes solvers for fluid particles interaction, adaptive mesh refinement, ...

He is also director of the "GdR Calcul" where his role is to animate the scientific and high performance computing community in France, in particular by organising conferences, meetings, and seminars. In this context, he organises himself 3 to 4 training and development workshops per year, and promotes the use of Python and c++ for teaching and research in France.

For several years, he has been very involved in promoting the Jupyter project and its use for teaching and research in the French community. He is one of the core developers of xeus-cling and he is working on the possibility of easily deploying a JupyterHub or BinderHub on academic clouds. He also believes that reproducible research is an essential part of promoting new computation codes, new numerical methods, ... introduced in related publications and therefore he uses Jupyter as a first approach to achieve this.

David Delavennat (R, male, 4 PM) David Delavennat is Research Engineer in scientific infrastructures at CMLS (Centre de Mathématiques Laurent Schwartz) which is a Joint Research Unit (UMR7640) of the CNRS and École polytechnique. He works on several french scientific infrastructures projects using microservices, virtualization, container and cloud technologies.

He is coordinator of ARGOS, the Paris-South business network targeted to the academic System and Network Administrators. He animates this community, organising conferences and meetings He represents ARGOS at RESINFO, the french business network of regional SNA business networks. In this context, he participate each year to the organization of National Teaching Actions.

For several years, he has been very involved in promoting the DevOps paradigm and Continuous Integration for deploying production infrastructures on academic Openstack and Kubernetes environment for the French Mathematical community. He is working on simplifying the deployment of a JupyterHub or BinderHub onto academic PaaS and Iaas. He also believes that reproducible infrastructures is an essential part of reproducible science.

Marc Massot (R, male, 2 PM) Marc Massot obtained his PhD in Applied Mathematics from Ecole Polytechnique, France, in 1996. After a year at Yale University, Department of Mechanical Engineering, he obtained a CNRS position in the Applied Mathematics Laboratory of the University of Lyon, France, where he stayed until 2005. He was offered an Associate Professor position at Ecole Centrale Paris when he installed a mathematics team in the EM2C mechanical engineering laboratory. From 2008 to 2010, he had the responsibility of structuring scientific computing at INSMI (French National Mathematics Institute of CNRS) at a national level in the French Mathematical Community. He was Visiting Professor for one year at the Center for Turbulence Research, Stanford University, in 2011-2012 and created and chaired the Fédération de Mathématiques de l'Ecole Centrale Paris between 2013 and 2016. He initiated the Computing Center (Mésocentre) of Ecole Centrale Paris in 2010, of which he was the deputy director until 2016 and he is scientific adviser at ONERA DMPE and scientific collaborator at Maison de la Simulation since 2013. Full Professor since 2011, he has been recruited on a full Professor position at Ecole Polytechnique, Centre de Mathématiques Appliquées in 2017 and co-chairs the Initiative HPC@Maths, which foster the interaction of mathematics, scientific computing and HPC with industry and in particular SMEs.

His main fields of research are mathematical modeling and numerical analysis, analysis of PDEs and dynamical systems for multi-scale systems, scientific computing and high performance computing with applications in combustion, two-phase flows, plasma physics and biomedical engineering.

[8] http://www.cmap.polytechnique.fr/~massot/Personal_web_page_of_Marc_Massot/MAP551

Since his arrival at Ecole polytechnique, he has been very involved in promoting the Jupyter project, creating courses entirely relying on Jupyter notebooks (such as "Dynamical systems for the modeling and simulation of multi-scale reacting media" MAP551) and making the link with the various departments and students of Ecole polytechnique, as well as Paris-Saclay community. One of the leaders of the new Computing Center in the process of being created, he is also involved with the direction of Ecole polytechnique for the project of building an infrastructure for Jupyter at the level of the school, for its use in both research and teaching.

Laurent Series (R, male, 2 PM) Laurent Series is Research Engineer in scientific computing at CMAP (Centre de Mathématiques Appliquées) which is a Joint Research Unit (UMR7641) of the CNRS and École polytechnique. He works on several scientific computing open-source projects in different fields like finite element method for solid mechanics, multiresolution for reaction-diffusion equations modeling multi-scale reaction wave ...

He was technical responsible of Computing Center (Mésocentre) of Ecole CentraleSupélec from 2009 to 2018. In this context, he organises and participates in the user support team (assistance in porting codes, parallelization, vectorisation, optimisation). Since his arrival at École Polytechnique, he is involved in the process of creation of a new Computing Center.

For several years, he has been very involved in promoting the Jupyter project by writing Jupyter notebooks for course (such as "Dynamical systems for the modeling and simulation of multi-scale reacting media" MAP551) and by promoting, among researchers, its use to share their work. He is also involved with the direction of École polytechnique for the project of building an infrastructure for Jupyter at the level of the school, for its use in both research and teaching.

DevOps Engineer (R, 18 PM) We intend to hire a DevOps engineer for 18 months during the project to be supervised at École polytechnique to work on developments defined in Tasks **T2.2** and **T5.3**.

Software Engineer (R, 18 PM) We intend to hire a DevOps engineer for 18 months during the project to be supervised at École polytechnique to work on developments defined in Task **T3.5**.

Publications, products, achievements

1. D. Delavennat, L. Gouarin, G. Philippon, *Deploying JupyterHub with Kubernetes on OpenStack*
   https://blog.jupyter.org/how-to-deploy-jupyterhub-with-kubernetes-on-openstack-f8f6120d4b1
2. JupyterDay at École polytechnique in 2018
   http://www.cmap.polytechnique.fr/~massot/Personal_web_page_of_Marc_Massot/JupyterX.html
3. L. Gouarin, *C++ course using xeus-cling*
   https://github.com/gouarin/cours_cpp_moderne
4. M.Massot, L. Series, *Systèmes dynamiques pour la modélisation et la simulation des "milieux réactifs" multi-échelles*
   http://www.cmap.polytechnique.fr/~massot/Personal_web_page_of_Marc_Massot/MAP551

Relevant projects or activities

1. Through a previous position at Paris Sud, Loïc Gouarin is a participant of OpenDreamKit (GA No. 676541) Open Digital Research Environment Toolkit for the Advancement of Mathematics.

Significant infrastructure

EP has strongly contributed to the deployment of JupyterHub on UPSud's local OpenStack based cloud infrastructure Cloud@VD. EP invested last year in the purchase of 100 cores for this infrastructure in order to study how to deploy JupyterHub and BinderHub and start to offer to researchers and students this kind of service. In 2019, EP will have its own OpenShift cloud infrastructure with 200-300 cores dedicated to teaching with Jupyter.

4.1.4 EGI: EGI (NL)

The EGI Foundation (also known as Stichting EGI and abbreviated as EGI.eu) is a not-for-profit foundation established under the Dutch law to coordinate the EGI Federation (abbreviated as EGI), an international collaboration that federates the digital capabilities, resources and expertise of national and international research communities in Europe and worldwide. The main goal is to empower researchers from all disciplines to collaborate and to carry out data- and compute-intensive science and innovation. The EGI Foundation coordinates areas such as overseeing infrastructure operations, user community support, contact with technology providers, strategy and policy development, flagship events and dissemination of news and achievements. As part of its mandate, the EGI Foundation actively represents the EGI federation at European level with policy makers and funding agencies, it provides expert advice to shape policies and funding programs and also support the implementation of the policy priorities. The EGI Foundation holds certifications in both ISO/IEC 9000 “Quality Management” and ISO/IEC 20000 “IT Service Management”.

Curriculum vitae

Gergely Sipos (leadPI, male) Gergely Sipos works as Technical Outreach Manager for EGI.eu. He coordinates user engagement and supports activities in the EGI community, supporting scientific communities in exploiting EGI services to push scientific boundaries.

He holds MSc and PhD in computer science with specialisation in management, from the University of Miskolc (Hungary).

Publications, products, achievements

1. David, M.; Borges, G.; Pina, J. et al., ”Validation of Grid Middleware for the European Grid Infrastructure”, (2014) DOI: 10.1007/s10723-014-9301-z
2. Ferrari, T.; Gaido, L., ”Resources and Services of the EGEE Production Infrastructure”, Journal of Grid Computing, June 2011, Volume 9, Issue 2, pp 119-133, DOI: 10.1007/s10723-011-9184-1, June 2011.
3. Matti Heikkurinen, Sandra Cohen, Fotis Karagiannis, Kashif Iqbal, Sergio Andreozzi, Michele Michelotto, ”Answering the Cost Assessment Scaling Challenge: Modelling the Annual Cost of European Computing Services for Research”, Journal of Grid Computing, May 2014, DOI:10.1007/s10723-014-9302-y
4. Sy Holsinger, Sergio Andreozzi, ”EGI: Implementing service management in a large-scale e-Infrastructure”, Proceedings of the IEEE Network Operations and Management Symposium (NOMS) Conference, 2014, Krakow, Poland, DOI: 10.1109/NOMS.2014.6838371
5. Sergio Andreozzi, Sy Holsinger, Damir Marinovic, Steven Newhouse, ”EGI: an Open e-Infrastructure Ecosystem for the Digital European Research Area”, Proceedings of eChallenges e-2012 Conference, Lisbon, Portugal, ISBN: 978-1-905824-35-91
6. Ohmann, C.; Canham, S.; Danielyan, E.; Robertshaw, S.; Legré, Y.; Clivio, L. & Demotes, J., ”Cloud computing and clinical trials: report from an ECRIN workshop”, Commentary Trials, Springer, December 2015, 16:318, DOI: /10.1186/s13063-015-0835-6
7. Wallom, D.C.H.; Turilli, M., Drescher, M.; Scardaci, D. & Newhouse, S., ”Federating Infrastructure as a Service Cloud Computing Systems to Creates a Uniform EInfrastructure for Research”, IEEE 11th International Conference on e-Science 2015, DOI:10.1109/eScience.2015.51
8. Cloud, Fernandez, E.; Sipos, G.; Scardaci, D.; Wallom, D.C.H. & Chen, Y., ”The user support programme and the training infrastructure of the EGI Federated Cloud”, International Conference on High Performance Computing & Simulation (HPCS) 2015, DOI: 10.1109/HPCSim.2015.7237016
9. Sergio Andreozzi, Owen Appleton, Sara Coelho, Tiziana Ferrari, Sy Holsinger, Yannick Legré, ”Open Science Commons”, Jan 2015, http://go.egi.eu/oscwp
10. Kimmo Koski, Kristiina Hormia-Poutanen, Prof. Mike Chatzopoulos, Yannick Legré, Bob Day, ”European Open Science Cloud for Research”, Oct 2015, https://zenodo.org/record/32915

Relevant projects or activities

1. EGI-Engage: Engaging the Research Community towards an Open Science Commons
2. EGI-InSPIRE: Integrated Sustainable Pan-European Infrastructure for Researchers in Europe, Project Coordinator, RI-261323
3. AARC (from May 2015)
4. BioMedBridges (Nr 284209)
5. BioVeL (Nr 283359)
6. Civic Epistemologies: Development of a Roadmap for Citizen Researchers in the Digital Culture, RI-632694
7. CloudWATCH: A European cloud observatory supporting cloud policies, standard profiles and services, RI-610994
8. DCH-RP (Nr 312274)
9. e-Fiscal: Financial Study for Sustainable Computing e-Infrastructures, RI-283449
10. EDISON: Creating the Data Science Profession (Nr 675419)
11. ENVRI (Nr 283465)
12. ENVRIplus (Nr 654182)

13. ER-Flow (Nr 312579)
14. eScienceTalk: Supporting grid and high performance computing reporting across Europe, Project Coordinator, RI-260733
15. FedSM: Implementing Service Management in Federated e-Infrastructures, RI-312851
16. Helix Nebula Science Cloud Project, the (from January 2016)
17. HelixNebula: Big science teams up with big business, RI-312301
18. INDIGO-DataCloud (from May 2015)

Significant infrastructure

EGI Infrastructure: The EGI Foundation coordinates the delivery of the EGI Infrastructure that brings together more than 300 data centres worldwide and also includes the largest community cloud federation in Europe with 21 cloud providers across 12 European countries offering IaaS cloud and storage services. The Infrastructure includes a Jupyter Hub that is open for researchers on the scalable EGI IaaS cloud.

4.1.5 EuXFEL: EUROPEAN XFEL GMBH (DE)

European X-Ray Free-Electron Laser Facility GmbH is a limited liability company under German law. At present, 12 countries are participating in the project: Denmark, France, Germany, Hungary, Italy, Poland, Russia, Slovakia, Spain, Sweden, Switzerland, and the United Kingdom. The company is in charge of the operation and construction of the European XFEL, a 3.4 km long X-ray free-electron laser facility extending from Hamburg to the neighbouring town of Schenefeld in the German federal state of Schleswig-Holstein. Civil construction started in early 2009, and the user operation in September 2017. With its repetition rate of 27,000 pulses per second and a peak brilliance a billion times higher than that of the best synchrotron X-ray radiation sources, the European XFEL will allow the investigation of still open scientific problems in a variety of disciplines (physics, structural biology, chemistry, planetary science, study of matter under extreme conditions and many others).

European XFEL has a data policy in place [Data] which opens up facility data for open access after an embargo period of 3 years.

Curriculum vitae

Hans Fangohr (leadPI, male, 5 PM) Hans Fangohr is an academic at the University of Southampton in the United Kingdom since 2002 (full professor since 2010), and leading the data analysis services at European XFEL in Germany since 2017.

He has been a long term proponent of Open Science, and in particular involved with the the use and further development of the Jupyter Notebook to enable this. He has hosted Thomas Kluyver at the University of Southampton since 2015 from where he contributed as a core developer of the Jupyter team. As a PI in the EC-funded e-INFRA OpenDreamKit project (2015-2019), he has pushed forward the use of Jupyter Notebooks for reproducible computational science, and started the notebook validation tool (NBVAL). He made use of the Jupyter Ecosystem for research and education at graduate and postgraduate level at the University of Southampton, and shared resources widely, including a text book provided through Jupyter Notebooks, which can be executed interactively online [1].

Since 2017, he is designing data analysis services and infrastructure at the European XFEL research facility. European XFEL is using IPython and the Jupyter Notebook as core utilities in their large scale experiment control, data capture and data analysis. Within the e-INFRA project PaNOSC (Photon and Neutron Science Open Cloud, 2018-2021, [Pan]), he is leader of the Work Package 4, which is focused on data analysis services for the EOSC Hub, and the use of the Jupyter notebook with its existing features on the EOSC hub.

In this project (BOSSEE), where new capabilities for the Jupyter notebook and ecosystem are being designed, Hans' wide experience and interaction with different science groups will be beneficial to ensure the outcome is of value to open science in many domains. This includes him chairing the interdisciplinary computational modelling group at the University of Southampton (200 academics, 2008-2017), chairing the national EPSRC scientific advisory committee on High Performance Computing in the UK (2014-2017) and interacting with a large variety of science users at European XFEL in his role to lead the data analysis service provision.

Sandor Brockhauser (PI, male, 1 PM) Sandor Brockhauser is the head of the Control and Analysis Software Group at the European XFEL. He received his M.Sc. in Informatics from Technical University of Budapest, Hungary, earned a Ph.D. from University of Leoben, Austria, and received his HDR degree in physics at the University of Joseph Fourier, Grenoble in France.

Since 2004, when he joined the European Molecular Biology Laboratory (EMBL) in Grenoble, France, he worked in Macromolecular Crystallography and became the scientist in charge of the Multi- Wavelength Anomalous Dispersion Beamline ID14-4 at European Synchrotron Radiation Facility (ESRF), France. Between 2013-15, he moved to Szeged, Hungary where he joined the Extreme Light Infrastructure, ELI-ALPS, and has established and built up its Scientific Engineering Division. In the same time, he also established the X-ray Crystallography Laboratory at a European Center of Excellence, the Biological Research Center, Szeged of the Hungarian Academy of Sciences. Moving to the European XFEL in 2016, he became responsible for the full control system of the beamlines and scientific instruments, Karabo, that enables the integration of Experiment Control, Data Acquisition and Analysis. During the last two years, Karabo has been deployed, photon beamlines were successfully commissioned and two of the initial scientific instruments have been put in operation producing 0,5PT of data in 5 weeks of experiments. Jupyter tools are embedded in the Karabo system and European XFEL analysis activities.

Krzysztof Wrona (PI, male, 1 PM) Krzysztof Wrona has a background in computer physics. As the group leader of IT and Data Management at European XFEL, he is in charge of the management of scientific data in the frame of the user program of the European XFEL facility. He has more than 15 years of experience in data storage, processing, and in general IT issues.

Thomas Kluyver (R, male, 48 PM) Thomas Kluyver is one of the core maintainers of the Jupyter and IPython projects, which form a central part of this proposal. He has been part of the core IPython development team since 2011, before Jupyter was split out as a separate project. He is closely familiar with key parts of the code, and was one of 15 core developers to receive the 2017 ACM Software System Award.

Besides Jupyter, Thomas has contributed to a wide range of open source software projects in the Python ecosystem, ranging from other scientific computing tools such as h5py to packaging utilities such as Flit, as well as improvements in the Python standard library.

Thomas is part of the Data Analysis team at European XFEL. He has a specific remit to facilitate the use of Jupyter by internal groups and visiting researchers, but he also contributes to the general software engineering effort working to allow convenient, reproducible analysis of data collected at European XFEL.

Before working as a software engineer, Thomas studied plant biology, gaining a PhD from the University of Sheffield in 2013. His work since then has continued in close connection with research, at the University of California, Berkeley, and the University of Southampton, before joining European XFEL.

Research Engineer x2 (R, 68 PM) We will hire two postdoctoral-level research software engineers (for 68 person months in total) to carry out the required work for this project at European XFEL. They will work under supervision of Hans Fangohr, with support from Sandor Brockhauser and Krzysztof Wrona for particular aspects. The employees will either have a scientific background and significant software engineering expertise, or an education in computer science and an aptitude to work with scientists on computational science and data science problems.

Publications, products, achievements

1. H.Fangohr, Python for Computational Science and Engineering (2018) DOI: 10.5281/zenodo.1411868 https://github.com/fangohr/introduction-to-python-for-computational-science-and-engineering
2. H.Fangohr et al., "Data Analysis support in Karabo at European XFEL", Proceedings of International Conference on Accelerator and Large Experimental Physics Control Systems 2017, ISBN 978-3-95450- 193-9, Data Analytics, Barcelona, Spain, TUCPA01 (2017) DOI: 10.18429/JACoW-ICALEPCS2017-TUCPA01
3. H.Fangohr. *A Comparison of C, Matlab and Python as Teaching Languages in Engineering* Lecture Notes on Computational Science **3039**, 1210-1217 (2004)
4. T. Kluyver, B. Ragan-Kelley, F. Perez, B. Granger, M. Bussonier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay et al. Jupyter *Notebooks: a publishing format for reproducible computational workflows* In 20th International Conference on Electronic Publishing. IOS Press, 2016.

Relevant projects or activities

1. OpenDreamKit (GA No. 676541) Open Digital Research Environment Toolkit for the Advancement of Mathematics, participant
2. EOSCpilot (GA No. 739563) The European Open Science Cloud for Research Pilot Project, participant
3. PaNOSC (GA No. 823852) Photon and Neutron Open Science Cloud, participant
4. CALIPSOplus (GA No. 730872) Convenient Access to Light Sources Open to Innovation, Science and to the World, participant
5. ATTRACT (GA No. 777222) breAkThrough innovaTion pRogrAmme for a pan-European Detection and Imaging eCosysTem, participant

Significant infrastructure

European XFEL, represented as a landmark on the ESFRI Roadmap, is a single site X-ray research infrastructure. In the first phase of operation, 3 beamlines and 6 experiments will be available for users. The SASE1 beamline comprises the instruments Single Particles, clusters, and Biomolecules and Serial Femtosecond Crystallography (SPB/SFX) and Femtosecond X-ray Experiments (FXE), SASE 2 includes Materials Imaging and Dynamics (MID) and High Energy Density Science (HED) and SASE3 Small Quantum Systems (SQS), and Spectroscopy and Coherent Scattering (SCS).

European XFEL's HPC system contains 86 compute nodes with in total 7632 CPUs (i.e. 3816 hyperthreaded physical CPUs) and in total 49 TB of RAM. 7 nodes feature GPGPUs (4 Nvidia Tesla 40k and 3 Nvidia P100). This system is part of the larger Maxwell cluster operated by DESY, totalling 13232 CPUs and 84 TB of RAM.

4.1.6 INSERM: INSERM (FR)

Inserm, the French National Institute of Health & Medical Research is the only public sector research institution in France exclusively dedicated to human health. Under the dual aegis of the Ministries of Health and Research, Inserm has a budget of 900 M euros and employs 15,000 scientists, engineers and technicians all with one shared objective, namely to promote health - by advancing knowledge about living organisms and their diseases, developing innovative treatment modalities and conducting research on public health.

Inserm is represented within the BOSSEE consortium through the Cancer Research Centre of Toulouse (CRCT) and Inserm's Computing Department (Département du Système D'Information, DSI).

CRCT gathers academic, scientific, medical, clinical, technological and pharmaceutical research on cancer on a 220-hectares site next to Toulouse, France. Its missions are to improve fundamental knowledge on all aspects of cancer biology and to provide patients with rapid access to innovative and individualized treatments. On these premises, CRCT comprises 21 teams affiliated to Inserm, the University of Toulouse and the CNRS (National Centre for Scientific Research). Team 15 of CRCT led by M. Bardiès aggregates Medical Physics resources available in Toulouse around a common research theme: the optimization of radiotherapy through the development of innovative dosimetric approaches at various scales (cell, tissue, patient).

Inserm's IT department (DSI) defines and coordinates IT and information systems aspects across the whole institution. It designs and operates Inserm's information system to support research activities of the institute, and provides counselling and support to research units on information technologies. It also plays a strong role in coordinating IT security and risk management policies for the French health science community.

Curriculum vitae

Manuel Bardiès (leadPI, male, 3 PM) Manuel Bardiès, PhD, obtained his doctorate on radiopharmaceutical dosimetry from Paul Sabatier University (Toulouse III) in 1991. He has been developing his research in radiopharmaceutical dosimetry within INSERM (National Institute of Health and Medical Research), since 1992, in Nantes then in Toulouse (2011) within the Cancer Research Centre of Toulouse (CRCT). He is the responsible of CRCT Team 15 entitled "Multi- resolution dosimetry for radiotherapy optimization".

Dr. Bardiès has been appointed to several international positions. He was one of the founders of the EANM Dosimetry Committee (member from 2001 to 2013, chair 2009-2011). He also chaired of EFOMP Science Committee (2014-2016).

Dr. Bardiès is also involved in education and is currently member of the Board of the European School for Medical Physics Expert (ESMPE) and member of the European School of Multimodality Imaging and Therapy (ESMIT).

The team led by Manuel Bardiès in Toulouse (CRCT Team 15) is primarily involved in radiopharmaceutical dosimetry, at various scales (cell, tissue, organs). This requires the ability to assess radiopharmaceutical pharmacokinetics in vivo, through quantitative SPECT or PET small-animal imaging. An important part of research activity is related to Monte Carlo modelling of radiation transport through biological structures of interest, in order to give account of energy deposition within tumour targets - or critical non-tumour tissues/organs. The objective is to improve molecular radiotherapy by allowing patient-specific treatments, as an important application of personalized medicine.

Maxime Chauvin (R, male, 24 PM) Maxime Chauvin, PhD, obtained his doctorate in astrophysics from Paul Sabatier University (Toulouse III) in 2011. He has been working in the field of astrophysics and particle physics from 2007 to 2016. He was involved in several instrument development and their data analysis. Dr. Chauvin made breakthrough discoveries in the observation of polarised X-rays from neutron stars and black holes.

Since 2017 he applies his expertise in numerical simulation and data analysis to optimise radiotherapy in the CRCT team of Manuel Bardiès at Inserm. He initiated the OpenDose project and he is the principal coordinator of the collaboration.

Isabelle Perseil (PI, female, 6 PM) Isabelle Perseil, PhD, is the Head of the Computational Science Coordination and e-infrastructures of Inserm. Dr. Perseil manages a group of 3 experts which provides the best practices in software engineering, Data Management, Big data, deep learning, HPC, Grids, Cloud Computing, parallel computing to 300 research units (1200 research teams).

The Computational Science Coordination is working with 13 regional administrations and 23 regional Mesocenters to pool the computational resources (grids and HPC) and train more than 1000 engineers and researchers to HPC (OpenMP, MPI and now ORWL) and Big data (MapReduce, Hadoop, Spark, Flink, Storm).

Gilles Mathieu (R, male, 6 PM) Gilles Mathieu is a research engineer, specialist in distributed architectures, grid and cloud computing. Within Inserm IT department, Gilles gained a strong experience in providing trainings and promoting technical solutions to the Health Science communities.

Before joining Inserm-DSI in June 2014, Gilles was Technical Director of the French National Grid Initiative (France Grilles) at CNRS.

He has been involved in large European projects since 2004 (EGEE I, II and III, EGI-InSPIRE) and has a strong experience in communication, dissemination and outreach within different scientific communities, including the so-called "long tail of science".

Publications, products, achievements

1. A. Albeyatti et al. “Towards a European health research and innovation cloud (HRIC)”. In: 2019 accepted in Genome Medicine.
2. M. Chauvin et al. “OpenDose: Generating reference data for Nuclear Medicine dosimetry”. In: European Journal of Nuclear Medicine and Molecular Imaging 44.S2 (Sept. 2017), pp. 119–956. DOI: 10.1007/s00259-017-3822-1.
3. D. Salas et al. “Resource-Centered Distributed Processing of Large Histopathology Images”. In: 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES). Aug. 2016, pp. 367–370. DOI: 10.1109/CSE-EUC-DCABES.2016.210.
4. S. Marcatili et al. “Model-based versus specific dosimetry in diagnostic context: Comparison of three dosimetric approaches”. In: Medical Physics 42.3 (2015), pp. 1288–1296. DOI: 10.1118/1.4907957.
5. D. Sarrut et al. “A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications”. In: Medical Physics 41.6 Part1 (2014), p. 064301. DOI: 10.1118/1.4871617.
6. I. Perseil et al. “An Efficient Modeling and Execution Framework for Complex Systems Development”. In: 2011 16th IEEE International Conference on Engineering of Complex Computer Systems. Apr. 2011, pp. 317–331. DOI: 10.1109/ICECCS.2011.38.
7. A. Divoli et al. “Effect of Patient Morphology on Dosimetric Calculations for Internal Irradiation as Assessed by Comparisons of Monte Carlo Versus Conventional Methodologies”. In: Journal of Nuclear Medicine 50.2 (2009), pp. 316–323. DOI: 10.2967/jnumed.108.056705.
8. I. Perseil and L. Pautet. “Foundations of a new software engineering method for real-time systems”. In: Innovations in Systems and Software Engineering 4.3 (Oct. 2008), pp. 195–202. ISSN: 1614-5054. DOI: 10.1007/s11334-008-0067-y

Relevant projects or activities

Inserm is the leading academic biomedical research institution in Europe with more than 13,000 publications a year; and second in the world (behind the American National Institutes of Health).

Inserm has 24 international cooperation agreements, 33 associated European laboratories (AELs) and associated international laboratories (AILs), and 183 Horizon 2020 contracts since 2014 - of which 45 were signed in 2017. 67 ERC winners have been hosted at Inserm since 2012, 13 of whom in 2017. Inserm is involved in many ESFRIs:

1. ERINHA2 (H2020)
2. ERINHA (FP7)
3. ADOPT BBMRI-ERIC (H2020)
4. BioMedBridges (FP7)
5. MRTdosimetry EMPIR (H2020)
6. MetroMRT (REG)

Inserm is also one of the funding partners of the French NGI, integrated within EGI.

Significant infrastructure

Inserm has more than 350 research units spread across France and internationally. These are supported by 13 Regional Commissions for local oversight. Scientific activities are organized around 9 “Inserm Thematic Institutes”, corresponding to the main fields of biomedical and health research.

4.1.7 QuantStack: QUANTSTACK (FR)

QuantStack was founded in 2016 by a team of developers and maintainers of key packages of the open-source scientific computing stack. QuantStack provides support and custom development services in the Jupyter and Scientific Python ecosystems. Clients and partners of QuantStack range from financial software companies to robotics startups and public research institutions. The team comprises several core developers of Jupyter subprojects and authors of popular scientific computing and visualization software used in both academic and industrial contexts.

Beyond Project Jupyter, projects developed at QuantStack include data visualization packages for Jupyter such as bqplot, ipyvolume, ipyleaflet, and ipysheet, as well as Jupyter language kernels such as xeus-cling and xeus-python, and JupyterLab extensions like te draw.io and sidecar. QuantStack is also behind the development of the xtensor framework, a high-level array computing library and C++ dataframe.

Curriculum vitae of the investigators

Sylvain Corlay (leadPI, male, 6 PM) Sylvain Corlay is the founder and CEO of QuantStack. He holds a PhD in applied mathematics from University Paris VI.

As an open source developer, Sylvain contributes to Project Jupyter in the area of interactive widgets for the notebook, and is steering committee member of the Project. He also serves as a member of the board of directors for the NumFOCUS foundation, and co-organizes the PyData Paris Meetup, a regular seminar series on open-source scientific computing.

Sylvain was one of the 15 core Jupyter developers to receive the 2017 ACM Software System Award.

Besides Jupyter, Sylvain contributes to a number of scientific computing open-source projects such as bqplot, xtensor and ipyleaflet.

Prior to founding QuantStack, Sylvain was a quant researcher at Bloomberg and an adjunct faculty member at the Courant Institute and Columbia University.

Johan Mabille (R, male, 0 PM) Johan Mabille is a scientific software developer at QuantStack specializing in high-performance computing in C++. He holds master's degree in computer science from Centrale-Supelec.

As an open source developer, Johan coauthored xtensor and xeus , and is the main author of xsimd. Prior to joining QuantStack, Johan was a quant developer at HSBC.

Martin Renou (R, male, 0 PM) Martin Renou is a Scientific Software Developer at QuantStack. Prior to joining QuantStack, Martin also worked as a Software developer at Enthought. He studied at the French Aerospace Engineering School ISAE-Supaero, with major in autonomous systems and programming.

As an open source developer, Martin has worked on a variety of projects, such as SciviJS (a JavaScript 3-D mesh visualization library) and simphony-remote (a web-service allowing to run desktop applications like Mayavi on the browser).

Passionate about 3-D rendering, Martin has also developed an open source 3-D Chess GUI based on OpenGL during his spare time.

Martin is the main author of xeus-python, and xleaflet. He is also a maintainer of ipyleaflet.

Wolf Vollprecht (R, male, 0 PM) Wolf Vollprecht is a scientific scientific software developer at QuantStack. He finished his Master in Robotics, Systems and Controls at ETH Zurich in 2017 with a specialization in AI and Deep Learning.

During his thesis work at Stanford University he was involved in developing fast machine learning algorithms on Tensorflow to anticipate human driver behavior.

His current work focuses on making xtensor faster, and more useful in the context of robotics and machine learning.

Software Engineer (R, 41 PM) We will hire a software engineer with experience working in large open-source projects. They will benefit from the mentoring of the other Jupyter contributors of the QuantStack team.

Publications, products, achievements

1. QuantStack developers participate in the continuous development of *Project Jupyter*. The team is especially active in the area of interactive widgets, as well as JupyterLab and the Jupyter Server.
2. QuantStack is the main driving force behind the *xtensor* project, a C++ tensor expression system for high-performance computing. Xtensor comes along with language bindings for Python, R, and Julia, as well as interfaces to BLAS, FFTW, and means to input and output a large number of standard file formats.
3. QuantStack also develops the *xeus* project, a framework for creating Jupyter language kernels. Xeus is used as a foundation for the C++ Jupyter kernel "xeus-cling", built upon the Cling C++ interpreter from CERN. Xeus was also adopted in Kitware's *Slicer* medical imaging software for its Jupyter integration.
4. The QuantStack team includes the authors and maintainers of some of the most popular Jupyter interactive widgets packages, including *bqplot*, a 2-D interactive plotting system, *ipyvolume*, a 3-D volume rendering package, *ipyleaflet*, a maps visualization toolkit.
5. QuantStack contributes extensively to the *conda-forge* project, a community-maintained collection of packages for scientific computing. Nearly a hundred "recipes" for conda-forge are maintained by QuantStack.
6. QuantStack developers are also behind the *vaex* data decimation engine for interactive visualization of large datasets.

Relevant projects or activities

Beyond open-source scientific computing development, QuantStack promotes scientific open source software development through the organization of events and by volunteering in non-profit organizations promoting the ecosystem.

1. QuantStack team members co-organize the regular *PyData Paris Meetup*, a free event series taking place every two to three months. After a year, the group counts over two thousand members in Paris.
2. We also support the *NumFOCUS Fondation* as volunteers as a member of the team is a member of the board of directors of the foundation.

4.1.8 UiO: University of Oslo (NO)

The University of Oslo (UiO) is Norway's oldest institution for research and higher education, with 28,000 students and 6,000 employees. UiO has 8 faculties, 2 museums and several centres. In addition, UiO has 10 Norwegian Centres of Excellence, is ranked as the world's 62nd university, and has had 5 Nobel prize laureates. UiO aims to become an international hub for the research-based integration of computing into science education and has financed a university-wide hosting service for Jupyter notebooks through JupyterHub to introduce a computational aspect to all curriculum programs in all science disciplines from bachelor to postdoctoral studies.

The University of Oslo is a Silver Partner to The Carpentries, an international successful community driven project with Instructors, Trainers, Maintainers, helpers, and supporters who share a mission to teach foundational computational and data science skills to researchers.

The Department of Geosciences of the Faculty of Mathematics and Natural Sciences is the broadest geoscience research-based teaching environment in Norway, and covers a wide range of disciplines from deep mantle processes to atmospheric sciences. It is organised in five sections and an administrative unit and supports two main strategic research initiatives:

- Land-Atmosphere Interactions in Cold Environments (LATICE)
- Interface Dynamics in Geophysical Flows (EarthFlows)

The geosciences department has several large research projects financed by The Research Council of Norway, EU and Norwegian companies.

The University of Oslo aims to manage research data according to international standards, such as the FAIR principles[9], and thereby support the development of a global research community in which research data is widely shared. Since November 2017, UiO's policy follows the "open as standard" principle in respect of access to research data [Datb].

Curriculum vitae

Anne Fouilloux (leadPI, female, 24 PM) PhD, is a highly experienced Research Software Engineer dedicated to supporting researchers towards the adoption of Open Science best practices.

With a solid background in Computer Sciences, she worked in various application fields, including environmental sciences, Intelligent Transport Systems, High-Performance computing, bio-informatics, meteorology and Geosciences.

She is currently working in the IT group of the department of Geosciences at the University of Oslo and holds a 25% at the Nordic e-Infrastructure Collaboration (NeIC) where she is involved on the Nordic Collaboration on e-Infrastructures for Earth System Modeling (NICEST) and CodeRefinery1 projects on Training and e-Infrastructure for Research Software Development.

Since 2015, Anne Fouilloux has been very active with The Carpentries, a diverse and global community of volunteers and she teaches foundational coding and data science skills to students and young researchers. She is a certified Carpentries instructor, instructor trainer and maintainer. She has volunteered to help build CarpentryCon 2020 a biannual conference for members of the global Carpentries community and people with similar interests.

She is a member of the core team of the Carpentry@UiO and is leading the studyGroup@UiO where students and researchers at the University of Oslo are committed to sharing skills, experiences, and ideas around open science, open source, code, and community in research.

Publications, products, achievements

1. Publication ready scientific reports and presentations with Jupyter notebooks, Anne Fouilloux, Research Bazaar 2019, DOI 10.5281/zenodo.2548936
2. Reproducible Research with Interactive Jupyter Dashboards, 2018, Ana Costa Conrado, Gladys Nalvarte, Benjamin Ragan-Kelley and Anne Fouilloux, Research Bazaar 2018, DOI 10.5281/zenodo.1168721
3. Working with Spatio-temporal data in Python, 2017, Anne Fouilloux, DOI 10.5281/zenodo.1165281

Relevant projects or activities

1. CodeRefinery (2016-2021):
   The goal of this project is to provide students and researchers with infrastructure and training in the necessary tools and techniques to create sustainable, modular, reusable, and reproducible software. This is a project within the Nordic e-Infrastructure Collaboration (NeIC), an organisational unit under NordForsk. NeIC is a Platinium Partner to The Carpentries.
   The result of this project is a set of software development e-infrastructure solutions, coupled with necessary technical expertise and extensive training and on-boarding activities, training material and best practices guides which together form a Nordic platform for research groups and institutes to develop a better collaboration on software and thereby to catalyze reproducible research and

[9] Findable, Accessible, Interoperable and Reusable

collaboration.
CodeRefinery training material is licensed under CC BY-SA 4.0 and code examples are OSI-approved MIT license.
The University of Olso is a CodeRefinery partner and will ensure the complementarity of the two projects thus avoiding potential fragmentation. BOSSEE will benefit from all this experience as well as the estbalished network in the Nordic Countries and beyond to fully realize the potential of BOSSEE EOSC services.

2. Nordic Collaboration on e-Infrastructures for Earth System Modeling (NICEST, 2017-2019) :
   This project aims at networking, intensifying existing collaboration, and facilitating joint work on very specific topics helping, for example, building up knowledge and competency, and harmonising certain procedures concerning e-Infrastructure topics.
3. UiOHive (2018-2019) :
   UiOHive provides a vital and novel competence at the Union of Internet of Things (IoT), Microcontroller / Hardware development, Artificial Intelligence (AI) and Machine Learning, and Data Science to enhance and strengthen collaboration between domains at the application of the aforementioned technologies. and the disciplines with competence to further develop technologies. The purpose of GEOHive is to establish a central knowledge hub, centered around individuals interested in utilizing IoT technologies, applying Artificial Intelligence and Machine Learning to data challenges, and sharing knowledge across relevant interdisciplinary domains.
4. LATICE (Land-ATmosphere Interactions in Cold Environments, 2015-2022): LATICE aims to advance the knowledge base concerning land atmosphere interactions and their role in controlling climate variability and climate change at high northern latitudes.
5. EarthFlows (Interface Dynamics in Geophysical Flows, 2015-2022): The dynamics of interface processes during flows on Earth, including the geosphere, the hydrosphere, the cryosphere, and the atmosphere, including the behavior of the complex interfaces separating ‘Fluid Earth’ from ‘Solid Earth’.
   The goal for the EarthFlows project is to provide fundamentally new understanding of the dynamics of fluid-solid interfaces for a number of important geophysical systems.

Significant infrastructure

1. Infrastructure as a Service: the University of Oslo is part of the Norwegian Cloud Infrastructure for Research and Education and provide researchers with compute and storage medium-size resources. These include multi-GPUs clusters for big data analysis. The department of Geosciences is heavily relying on this services both for teaching and research work.
2. UNINETT Sigma-2: UNINETT Sigma2 manages the national infrastructure for computational science in Norway and offers services in High Performance Computing (HPC) and Data Storage and data analysis (Research Platform as a Cloud Service). The services are organized into infrastructural activities, financed by the Research Council of Norway and the Sigma2 consortium partners, which are the universities in Oslo, Bergen, Trondheim and Tromsø.
   Services are freely available to individuals and groups involved in research and education at Norwegian universities and colleges, and other organizations and project funded with public money. Cost efficient development, procurement, coordination and operation of the national e-infrastructure for research and education is the main focus for Sigma2.
   The Department of Geosciences (University of Oslo) has been granted access to over 2 petabytes and several millions of CPU hours on the Norwegian High-Performance computers.

4.1.9 UPSud: UNIVERSITÉ PARIS-SUD (FR)

Université Paris-Sud is among the 40 top universities worldwide in the 2013 Shanghai ranking, and is one of the top two French research universities. With about 27000 students, 1800 permanent faculty and 1300 permanent research scientists from national research organisations (CNRS, Inserm, INRA, Inria), it is the largest campus in France. Since 2006, scientists from the University were awarded two Fields medals, one Nobel Prize and a number of other national and international prizes (European Inventor Award 2013, Wolf Prize 2010, Holweck Prize 2009, Japan prize 2007). Université Paris-Sud offers a wide range of qualifications, from the exact sciences to life and health sciences (including medical practice), legal sciences and economics. Research at Université Paris-Sud is an essential part of academic understanding and includes research activities with high commercial potential. Research contracts and partnership with companies make Université Paris-Sud a key actor and a major player in French research. The University is located partly on the Plateau de Saclay, the largest cluster of public and private R&D institutions in France (with ca. 16000 research staff), and is one of the core members of University Paris-Saclay – a world-class university and a world-renowned research and innovation hub.

In the context of this project, Université Paris-Sud is a member of the Open Source Thematic Group of the Systematic Paris Region Systems and ICT Cluster. It is the home of one of the largest group worldwide of developers of the open source SageMath computational mathematics system. The University also hosts a major research group in Human-Centered Computing and manages the Digiscope network of high-end visualisation platforms, which will provide critical assets to the project.

Finally, A variety of courses are delivered at Université Paris Sud using Jupyter technologies. This includes for example programming classes in C++ at lower undergraduate level (400 students per year since 2017), a series of undergraduate and graduate math courses (computer aided mathematics, computer algebra, numerical methods), or courses in physics, bio-informatics, etc. To support these courses, a JupyterHub service has been deployed in 2017 and progressively improved since, on Paris Sud's local cloud infrastructure Cloud@VD (see below), enabling students and teachers to work from anywhere and any device.

Curriculum vitae of the investigators

Michel Beaudouin-Lafon (PI, male, 2 PM) Michel Beaudouin-Lafon (PhD, Université Paris-Sud) is a Professor of Computer Science, classe exceptionnelle, at Université Paris-Sud and a senior fellow of Institut Universitaire de France. His research interests include fundamental aspects of interaction, novel interaction techniques, computer-supported cooperative work and engineering of interactive systems. He has published over 180 papers and is a member of the ACM SIGCHI Academy. He is the laureate of an ERC Advanced Grant exploring instrumental interaction and information substrates. Michel was director of LRI, the laboratory for computer science joint between Université Paris-Sud and CNRS. He now heads the Human-Centered Computing lab at LRI and chairs the Computer Science department at Université Paris-Saclay. He was Technical Program Co-chair for CHI 2013 (3500 participants), sits on the editorial boards of ACM Books and ACM TOCHI, and has served on many ACM committees. He received the ACM SIGCHI Lifetime Service Award in 2015.

Viviane Pons (PI, female, 2 PM) Maître de Conférences at the Laboratoire de Recherche en Informatique, Viviane Pons is a young researcher in Algebraic Combinatorics. She defended her thesis in 2013 and has 4 papers in international journals and 6 communications in international conferences, including a talk at PyCon US 2015. She was also invited as a keynote speaker at Pycon FR 2018. Since January 2019, she is in the editorial board of the Journal of Open Source Software. Before starting her research career, she worked for two years in industry as a Java and web developer.

She discovered SageMath during her first SageMath Days in 2010 and has since been an active user and contributor with 10 (co)authored tickets improving the support of combinatorial objects in SageMath. She is heavily involved in the promotion of SageMath, participating in SageMath Days and running SageMath introduction tutorials or SageMath presentations at various conferences. She has also been involved in devloping the project FindStat dedicated to databases in combinatorics.

Viviane is leading the very successful Community Building and Dissemination work package of the European Research Infrastructures project OpenDreamKit (2015-2019), in which 66 events (development workshops, training sessions, ...) were organized or coorganized, with more than a thousand trainees. Viviane herself organized or coorganized several of them, including two week-long workshops dedicated to women (one in 2017 and one to come in spring 2019).

Nicolas M. Thiéry (leadPI, male, 5 PM) Professor at the Laboratoire de Recherche en Informatique, Nicolas M. Thiéry is a senior researcher in Algebraic Combinatorics with 18 papers published in international journals. Among other things, he is a member of the permanent committee of FPSAC, the main international conference of the domain, a founding member of the upcoming Numfocus Europe non-profit, and a member of Work group on Free and Open Source software for the "Open Science Committee" of the French Ministry for Research. He has collaborators in the US and Canada where he cumulatively spent more than three years (Colorado School of Mines, UC Davis, Providence, Montréal), and in India. He also co-organised fourteen international workshops, in particular SageMath Days, and the semester long program on "Automorphic Forms, Combinatorial Representation Theory and Multiple Dirichlet Series" hosted in Providence (RI, USA) by the Institute for Computational and Experimental Research in Mathematics.

Algebraic combinatorics is a field at the frontier between mathematics and computer science, with heavy needs for computer exploration. Pioneer in community-developed open source software for research in this field, Thiéry founded in 2000 the Sage-Combinat software project (incarnated as MuPAD-Combinat until 2008); with 50 researchers in Europe and abroad, this project has grown under his leadership to be one of the largest organised community of Sage developers, gaining a leading position in its field, and making a major impact on one

hundred publications[10]. Along the way, he coauthored part of the proposal for NSF Sage-Combinat grant OCI-1147247, and co-organised or taught at a dozen training and dissemination actions (workshops, summer schools, etc.), in America, Africa, Europe, and India.

With 150 tickets (co)authored and as many refereed, Thiéry is himself a core SageMath developer, with contributions including key components of the SageMath infrastructure (e.g. categories), specialised research libraries (e.g. root systems), thematic tutorials, and two chapters of the book “Calcul Mathématique avec SageMath” and its English translation.

Based on this experience, and to tackle the pressing funding needs in the ecosystem of open source mathematical software, Thiéry initiated and lead the European Research Infrastructures project OpenDreamKit #676541 (2015-2019, 15 sites, 50 participants, 8M€), engaging the Jupyter project on board. This in turn increased his involvement in using, promoting, and contributing to Jupyter, for use in mathematics and education.

Software Developer (R, 21 PM) We will hire a full time experienced software developer to work on task **T4.7** and **T6.4** under the leadership of Nicolas M. Thiéry.

The fellow will have a strong software engineering and web development experience, ideally in the Python, Javascript, and/or Jupyter ecosystem. We further require good communication and team working skills, in particular to work in tight collaboration with international open-source developer communities.

Software Developer (R, 12 PM) We will hire a full time experienced software developer to work on task **T2.4** under the leadership of Michel Beaudouin-Lafon.

The fellow will have a strong software engineering and web development experience (HTML/CSS/Javascript), and ideally good knowledge of the Python/Jupyter ecosystems and/or collaboration technologies. We further require good communication and team working skills, in particular to work in tight collaboration with international open-source developer communities.

Publications, achievements

1. Leadership of the Sage-Combinat software project.
2. Coauthoring of the open source book “Calcul Mathématique avec Sage” and its English translation , the first of its kind comprehensive introduction to computational mathematics in SageMath for education.
3. Contribution of more than 500 tickets to SageMath.
4. Michel Beaudouin-Lafon, Olivier Chapuis, James Eagan, Tony Gjerlufsen, Stéphane Huot, Clemens Klokmose, Wendy Mackay, Mathieu Nancel, Emmanuel Pietriga, Clément Pillias, Romain Primet, Julie Wagner (2012). Multi-surface Interaction in the WILD Room, *IEEE Computer*, 45(4):48–56. IEEE Computer Society.
5. Klokmose, C.N., Eagan J.R., Baader, S., Mackay, M. and Beaudouin-Lafon, M. (2015) Webstrates: Shareable Dynamic Media. In *Proceedings of the 28th annual ACM symposium on User interface software and technology (UIST '15)*. ACM.

Relevant projects or activities

1. OpenDreamKit (GA No. 676541) Open Digital Research Environment Toolkit for the Advancement of Mathematics, **coordination**.
2. Hosting or coorganisation of dozens of Sage Days or Jupyter workshops (week-long training and development workshops).
3. ERC Advanced Grant ONE “Unified Principles of Interaction” (PI: Michel Beaudouin-Lafon) that develops new user interface concepts, in particular for multi-user, multi-device environments.

Significant infrastructure

UPSud hosts a local OpenStack based cloud infrastructure Cloud@VD (3500 cores / 1 Po storage) for its personnel. The participants are regular users of this infrastructure, and in close contact with its maintainers. As a continuation of the existing deployment of a JupyterHub service on this infrastructure (joint work with EP), Cloud@VD will be available to the participants as test bed for deploying Jupyter based services (see e.g. **T5.3**).

UPSud also manages the Digiscope (`http://digiscope.fr`) network of high-end visualisation platforms and hosts the WILD and WILDER platforms, two ultra-high resolution wall-sized displays with motion capture and touch input for conducting research on collaborative human-computer interaction and visualisation of large datasets.

[10] `http://sagemath.org/library-publications-combinat.html`, `http://sagemath.org/library-publications-mupad.html`

4.1.10 Silesia: UNIVERSITY OF SILESIA (PL)

The University of Silesia in Katowice was established in 1968. Now, with 12 faculties and several interdisciplinary schools and centres, over 30000 students and over 2000 academic staff the University is one of the largest in Poland. Students are educated at three educational levels: Bachelor, Master and Doctoral and their achievement are accumulated using European Credit Transfer and Accumulation System (ECTS). Located in the heart of Upper Silesia, Poland's old industrial region with distinct history and cultural identity, the university attracts many scientists and students.

The origins of the *Faculty of Mathematics, Physics and Chemistry* date back to the academic year 1968/1969 and coincide with the establishment of the University of Silesia. One of the largest university units, the faculty incorporates, as its name indicates, three separate departments: mathematics, physics and chemistry, each with several divisions and subdivisions carrying out the research and educational activities. There are over 1900 students, both full-time and part-time, educated at three educational levels: Bachelor's, Master's and Doctoral. The Faculty is entitled to grant doctoral degrees in the natural sciences. The Faculty staff consists of 243 academics who are both teachers and researchers.

In the context of this project, University of Silesia has started offering notebook based resources for teaching and research since 2011, based on SageMath system. Now it offers courses in science and programming based on Jupyter notebook as well as collaborates with local high schools in this matter.

Curriculum vitae of the investigators

Marcin Kostur (leadPI, male, 3 PM) is an assistant Professor at the Institute of Physics. He is the author of over 50 publication cited over 2000 times in the field of statistical physics, solid state physics (Josephson Junction dynamics), microfluidics and biophysics. He is experienced in application of GPU architecture to numerical simulations of stochastic processed in physics. His recent computational interests are focused at the Open Source project Sailfish – HPC implementation of Lattice Boltzmann Method on GPU. He is leader few projects including computations in the science education and e-infrastructure:

- Infrastructure for cloud-based system education: scalable implementation of Jupyter notebook system for scientific explorations, project funded by Erasmus+, Key Action 2 - "Strategic Partnership", (budget: €160k, 2017-2019)
- Computing in high school science education - iCSE4schools, project funded by Erasmus+, Key Action 2 - "Strategic Partnerships", (budget: €263k, 2014-2017)
- "Computers in Science Education: iCSE" http://icse.us.edu.pl (budget: €1m, funded by EFS, 2011-2014)
- PAAD (Platform for Analysis and Archiving of Data) project funded by POIG program for 2014-2015 with a total budget of €4m. The task coordinator"Interactive HPC services for science".

Jerzy Łuczka (PI, male, 5 PM) Prof. Dr. Jerzy Łuczka (http://zft.us.edu.pl/luczka) is a full professor of physics at the University of Silesia (Katowice, Poland) and the Head of the the Department of Theoretical Physics.

He published more than 150 papers in journals which have been cited almost 3000 times.

He is an Editor of European Physical Journal B, Chairman of the Statistical and Nonlinear Physics Division (European Physical Society), Fellow of the Institute of Physics (United Kingdom) and Outstanding Referee (American Physical Society). He was Co-director of the NATO Advanced Research Workshop "Stochastic Systems. From randomness to complexity", 2002, Erice (Italy) and Member of the Steering Committee of the program : "Stochastic Dynamics: Fundamentals and applications" (European Science Foundation), 2003-2008. He received the DAAD research fellowship (Forschungsaufenthalte für Hochschullehrer und Wissenschaftler) 1995, 2009 and 20012. He was a leader of several Polish and two German-Polish grants. He has collaborators in Germany, Italy and Spain. He has also co-organised international conferences.

Łuczka's research interests lie in areas of stochastic processes in physics, quantum open systems, transport phenomena, physical fundamentals of quantum information. He has teaching experience with SageMath in physics, biophysics and econophysics.

Research Engineer (R, 16 PM) We will hire a part time researcher with strong programming skills to work on task **T4.4** under leadership of Marcin Kostur. The fellow will have a strong knowledge of GPU computing as well as 3d data visualisation. We further require good communication and team working skills, in particular to work in tight collaboration with international open-source developer communities.

Publications, products, achievements

1. Leadership on development K3D-jupyter project which is an 3d visualisation Juypyter widget, (https://github.com/K3D-tools/K3D-jupyter)
2. Leadership on development Sailfish-cfd which is an GPU implementation of the lattice Boltzmann method. (https://github.com/sailfish-team/sailfish)[JK14]
3. Marcin Kostur has received the Award of the Minister of Science and Higher Education for implementing "Computers in Science Education" programme.
4. The project Computing in high school science education - iCSE4schools, has received an award of Foundation for the Development of the Education System.

Relevant projects or activities

1. OpenDreamKit (GA No. 676541) Open Digital Research Environment Toolkit for the Advancement of Mathematics, (site leader)
2. Infrastructure for cloud-based system education: scalable implementation of Jupyter notebook system for scientific explorations, project funded by Erasmus+, Key Action 2 - “Strategic Partnership”, (budget: €160k, 2017-2019)
3. Computing in high school science education - iCSE4schools, project funded by Erasmus+, Key Action 2 - “Strategic Partnerships”, (budget: €263k, 2014-2017)
4. “Computers in Science Education: iCSE” http://icse.us.edu.pl (budget: €1m, funded by EFS, 2011-2014)
5. 2011-2014 - iCSE (innovative Computing in Science Education) - € 1m grant from European Social Fund, incorporating computational perspective in teaching of mathematics, physics and chemistry using cloud based SageMath system and Python language.
6. 2014-30.11.2015 PAAD (Platform for data analysis and archiving) € 3.8m, funded is mostly HPC centre for research with interactive access based on web based notebook UI.
7. 2014-30.11.2015 CNS: Centre of Applied Science, Infrastructure grant includes € 0.5m funding for small HPC and cloud infrastructure for education.

Significant infrastructure

The University of Silesia has finished or currently implements ESF grants totaling to about € 120m for infrastructure, laboratories, and computing centers. New HPC centres created as a part of PAAD and CNS projects provide necessary hardware for development and implementation of cloud-based research and teaching. In particular Silesia hosts a local cloud infrastructure for education available for students of the Faculty of Mathematics, Physics, and Chemistry. It contains 320 cores system and 8 GPU and provides hosting to various instances of Jupyterhub. It is independently availavle a small heterogenous HPC cluster dedicated to research, containing GPU, high-memory nodes and Xeon Phi.

4.1.11 WildTree: WILD TREE TECH (CH)

Wild Tree Tech GmbH is a limited liability company under Swiss law established in 2017. Its business is built on three pillars: custom data driven software products, hosted JupyterHub services and training courses in machine-learning techniques.

Clients include NGOs, companies, hospitals, universities and UN organisations from Switzerland, France and the USA.

Wild Tree Tech employees co-create, co-lead and contribute to international open-source projects used by thousands people who use computers for teaching and data-science. The Binder Project creates, advances, and promotes open technology that makes it easy for people to connect their data science communications, educational materials, and scientific work with computational environments where their work can be run and shared with others. The Binder project operates a public infrastructure at https://mybinder.org. Wild Tree Tech contributes resources for operation and maintenance of this free public service.

Curriculum vitae

Tim Head (leadPI, male, 12 PM) Tim Head is the founder and CEO of Wild Tree Tech GmbH. He holds a PhD in High Energy Particle Physics from The University of Manchester.

He is a project lead on the Binder Project which forms a central part of this proposal. Tim sets the direction of the project, works on sustainability and governance issues, and growing the community. He is closely familiar with key parts of the code and is part of the team that operates mybinder.org a public infrastructure serving over 2.5million reproducible research environments to users from around the world in 2018.

As an open-source developer he created the scikit-optimize library implementing algorithms to tune hyper-parameters of artificial intelligence algorithms.

He created and organises the PyData meetup in Zurich, Swizterland, a regular seminar series on open-source scientific computing.

Before founding Wild Tree Tech he worked as a CERN research fellow on the LHCb experiment, one of the four major experiments at the Large Hadron Collider. He also worked at the Ecole Polytechnique Fédérale de Lausanne as a research associate.

Software Engineer (R, 24 PM) We will hire a software engineer with experience working in distributed teams and open-source projects. They will have experience in using Kubernetes and Jupyter.

Publications, products, achievements

1. Hub Hero, JupyterHubs for workshops, lecture courses and institutions. Harness the power of Jupyter notebooks for classes allowing teachers to teach interactively without needing tech support.

Relevant projects or activities

1. Binder

Significant infrastructure

1. mybinder.org, Wild Tree Tech helps operate a BinderHub available to the public that allows anyone to turn a Git repository into a collection of interactive Jupyter notebooks. This service is available for free and was used to launch over 2.5 million notebooks in 2018 alone.

4.2 Third parties involved in the project (including use of third party resources)

Only participants with third parties involved in the project are listed below.

4.2.1 CNRS-Observatoire astronomique de Strasbourg

Does the participant plan to subcontract certain tasks (please note that core tasks of the project should not be sub-contracted)	No
If yes, please describe and justify the tasks to be subcontracted	
Does the participant envisage that part of its work is performed by linked third parties	Yes
***The University of Strasbourg** (UNISTRA) is one of the largest universities in France, with over 46000 students and over 4000 researchers. The University also offers access to 25 modern languages, multinational diplomas, jointly supervised doctorates, upholding renowned international postgraduate schools and student exchange agreements. The scale of research activity at Strasbourg is substantial, involving 10 doctoral schools and 73 research units and 6 research federations covering a broad range of disciplines. UNISTRA is linked to CNRS-ObAS via a signed Convention de site. Dr. Sebastien DERRIERE and Mr. Thomas BOCH are employed by Universite de Strasbourg and work in the Centre de Donnees astronomique de Strasbourg (CDS) within The Strasbourg Observatory (ObAS), a Joint Research Unit of CNRS and Universite de Strasbourg.*	
Does the participant envisage the use of contributions in kind provided by third parties (Articles 11 and 12 of the General Model Grant Agreement)	No
If yes, please describe the third party and their contributions	
Does the participant envisage that part of the work is performed by International Partners (Article 14a of the General Model Grant Agreement)?	No
If yes, please describe the International Partner(s) and their contributions	

4.2.2 Uniwersytet Slaski

Does the participant plan to subcontract certain tasks (please note that core tasks of the project should not be sub-contracted)	Yes
***Uniwersytet Slaski** have much experience in academic research based on 3d visualisation software for fluid dynamics. However, the expertise in computer graphics, especially WebGL, is not enough at the Department of Mathematics, Physics and Chemistry. Instead of building such an expertise, it is financially more efficient to specify and outsource the programming task to professionals. The subcontracting will cover work done in the development od K3D-jupyter package, and it will be performed under supervision of leadPI.*	
Does the participant envisage that part of its work is performed by linked third parties	No
If yes, please describe the third party, the link of the participant to the third party, and describe and justify the foreseen tasks to be performed by the third party	
Does the participant envisage the use of contributions in kind provided by third parties (Articles 11 and 12 of the General Model Grant Agreement)	No
If yes, please describe the third party and their contributions	
Does the participant envisage that part of the work is performed by International Partners (Article 14a of the General Model Grant Agreement)?	No
If yes, please describe the International Partner(s) and their contributions	

5 Ethics and Security

5.1 Ethics

5.1.1 Ethics framework and relevant legislation

All activities of the BOSSEE project will conform to National, EC and International legislation as listed and described below:

- The Charter of Fundamental Rights of the EU.
- The European Convention for the Protection of Human Rights and Fundamental Freedoms.
- The European Charter for Researchers and the Code of Conduct for the Recruitment of Researchers
- The Data Protection Directive (95/46/EC) of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
- The European General Data Protection Regulation (GDPR)
- The Directive on Privacy and Electronic Communications (2002/58/EC) as well as the new ePrivacy directive.
- The Directive on the Re-use of Public Sector Information (2003/98/EC) as well as the new revised version.

5.1.2 Protection of personal data

The aim of BOSSEE is to improve the accessibility, interactivity, and reproducibility of computational research in the EOSC. The handling and protection of personal data must therefore be carefully considered. For this reason, the following activities are foreseen:

- Appointment of a Data Protection Officer (DPO) for the project. The DPO will be responsible for overseeing data protection strategy and implementation to ensure compliance with ethics and legal requirements, particularly focusing on GDPR provisions.
- For organisations that must appoint a DPO under the GDPR: Involvement of the data protection officer (DPO).
- For all other organisations: Details of the data protection policy for the project (i.e. project-specific, not general).
- Elaboration of a Data Management Plan (D1.2, D1.4), which will include, but will not be limited to, details of procedures for data collection, anonymisation, storage, protection, retention, destruction, and re-use.
- Providing details of the security measures to prevent unauthorised access to personal data.
- Anonymisation/Pseudoanonymisation in case network traffic needs to be stored for processing. This will include not only replacement of IP addresses, but also replacement of HTTP requests, since these also may contain data which can be associated with individuals.
- Informing about details of the data transfers (type of data transferred and country to which it is transferred ? for both EU and non-EU countries).

5.2 Security

The BOSSE project does NOT involve any of the following:

- activities or results raising security issues: NO
- 'EU-classified information' as background or results: NO

References

[Acm] *ACM Software System Award*. 2017. URL: https://awards.acm.org/software-system.

[BP17] D. Baron and D. Poznanski. "The weirdest SDSS galaxies: results from an outlier detection algorithm". In: *Monthly Notices of the Royal Astronomical Society* 465.4 (2017), pp. 4530–4555. DOI: 10.1093/mnras/stw3021.

[Bqp] *bqplot*. URL: https://github.com/bloomberg/bqplot.

[BT18] O. Benassy and N. M. Thiéry. *Exploratory support for semantic-aware interactive widgets on mathematical objects*. Aug. 2018. URL: https://github.com/OpenDreamKit/OpenDreamKit/raw/master/WP4/D4.16/report-final.pdf.

[Cha+17] M. Chauvin et al. "A collaborative effort to produce reference dosimetric data with Monte Carlo simulation software". In: *Physica Medica* 42 (2017), pp. 32–33. DOI: 10.1016/j.ejmp.2017.09.081.

[CM17] S. Corlay and J. Mabille. "Xeus: A framework for writing native Jupyter kernels". In: *Jupytercon*. 2017. URL: https://conferences.oreilly.com/jupyter/jup-ny-2017/public/schedule/detail/60038.

[CO+18a] D. Cortés-Ortuño et al. *Data set for: Proposal for a micromagnetic standard problem for materials with Dzyaloshinskii-Moriya interaction*. Zenodo doi:10.5281/zenodo.1174311. GitHub: https://github.com/fangohr/paper-supplement-standard-problem-dmi. Apr. 2018. DOI: 10.5281/zenodo.1174311.

[CO+18b] D. Cortés-Ortuño et al. "Proposal for a micromagnetic standard problem for materials with Dzyaloshinskii–Moriya interaction". In: *New Journal of Physics* 20.11 (2018), p. 113015. DOI: 10.1088/1367-2630/aaea1c.

[Coc] *CoCalc*. URL: https://cocalc.com/.

[Col] *Colaboratory*. URL: https://colab.research.google.com/.

[Data] *Scientific Data Policy of European X-Ray Free-Electron Laser Facility GmbH*. 2017. URL: https://www.xfel.eu/users/experiment_support/policies/scientific_data_policy/index_eng.html.

[Datb] *University of Oslo Policies and guidelines for research data management*. 2017. URL: https://www.uio.no/english/for-employees/support/research/research-data-management/policies-and-guidelines/.

[Eri+17] T. A. Erickson, B. Granger, J. Grout, and S. Corlay. *Interacting with Petabytes of Earth Science Data using Jupyter Notebooks, IPython Widgets and Google Earth Engine*. 2017.

[Eux] *European XFEL Scientific Data Policy*. 2017. URL: https://www.xfel.eu/users/experiment_support/policies/scientific_data_policy/index_eng.html.

[For+18] J. Forde, T. Head, C. Holdgraf, Y. Panda, G. Nalvarte, B. Ragan-Kelley, and E. Sundell. "Reproducible research environments with repo2docker". In: *ICML 2018 RML Proceedings*. 2018.

[Gry] *Gryd*. URL: https://gryd.us.

[Ham16] J. B. Hamrick. "Creating and Grading IPython/Jupyter Notebook Assignments with NbGrader". In: *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. SIGCSE '16. Memphis, Tennessee, USA: ACM, 2016, pp. 242–242. ISBN: 978-1-4503-3685-7. DOI: 10.1145/2839509.2850507.

[Han18] M. Hansen. *How We Got Published in The New York Times*. 2018. URL: https://journalism.columbia.edu/nyt-twitter-story.

[Hir16] T. Hirst. "The Rise of Transparent Data Journalism – The BuzzFeed Tennis Match Fixing Data Analysis Notebook". In: (2016). URL: https://blog.ouseful.info/2016/01/18/the-rise-of-transparent-data-journalism-the-buzzfeed-tennis-match-fixing-data-analysis-notebook/.

[HRA18] J. Hamman, M. Rocklin, and R. Abernathy. *Pangeo: A Big-data Ecosystem for Scalable Earth System Science*. 2018.

[Hug+14] A. Hughes, Z. Liu, M. Raftari, and M. E. Reeves. "A workflow for characterizing nanoparticle monolayers for biosensors: Machine learning on real and artificial SEM images". In: *PeerJ PrePrints* 2 (Dec. 2014), e671v2. ISSN: 2167-9843. DOI: 10.7287/peerj.preprints.671v2.

[Ipya] *Interactive Python (IPython)*. http://ipython.org. 2019.

[Ipyb] *ipyvolume*. URL: https://github.com/maartenbreddels/ipyvolume.

[Ipyc] *ipywidgets*. URL: https://github.com/jupyter-widgets/ipywidgets.

[Ivo] *IVOA*. URL: http://ivoa.net/.

[JCDF17] M. Juric, D. Ciardi, and G. Dubois-Felsmann. *LSST Science Platform Vision Document*. 2017. URL: https://ls.st/LSE-319.

[JK14] M. Januszewski and M. Kostur. "Sailfish: A flexible multi-GPU implementation of the lattice Boltzmann method". In: *Computer Physics Communications* 185.9 (2014), pp. 2350–2368.

[Jup] *Project Jupyter*. http://jupyter.org. 2019.

[K3d] *K3D*. URL: https://github.com/K3D-tools/K3D-jupyter.

[Klu+16] T. Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Stand Alone* 0.Positioning and Power in Academic Publishing: Players, Agents and Agendas (2016), 87–90. ISSN: 0000-0000. DOI: 10.3233/978-1-61499-649-1-87.

[Lak15a] B. Laken. *The cosmic ray flux and the Indian Summer Monsoon*. 2015. DOI: 10.6084/m9.figshare.1299413.v3.

[Lak15b] B. A. Laken. *Reply to 'Influence of cosmic ray variability on the monsoon rainfall and temperature': a false-positive in the field of solar-terrestrial research*. 2015. eprint: arXiv:1502.00505.

[Lat] *LA Times Data Desk Jupyter Notebooks*. 2018. URL: https://github.com/datadesk/notebooks.

[Lig] *Gravitational Wave Open Science Center*. URL: https://www.gw-openscience.org/about/.

[Myb] *MyBinder.org Events Archive*. URL: https://archive.analytics.mybinder.org.

[Nbd] *nbdime*. URL: https://nbdime.readthedocs.io/en/latest/.

[Nbs] *Nbsphinx*. URL: https://nbsphinx.readthedocs.io/.

[Nbv] *nbval*. URL: https://github.com/computationalmodelling/nbval.

[Okp] *OK*. URL: https://okpy.org/.

[Pan] *Photon and Neutron Open Science Cloud: European project (financed by the INFRAEOSC-04 call) for making FAIR data a reality in 6 ESFRI Research Infrastructures*. 2018. URL: https://panosc-eu.github.io.

[Par19] P. Parente. *Estimate of Public Jupyter Notebooks on GitHub: Latest Report*. 2019. URL: https://github.com/parente/nbestimate.

[Per18] J. M. Perkel. "Why Jupyter is data scientists' computational notebook of choice". In: *Nature* 563.7729 (2018), pp. 145–146. DOI: 10.1038/d41586-018-07196-1.

[PG15] F. Pérez and B. Granger. *Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science*. 2015. URL: http://archive.ipython.org/JupyterGrantNarrative-2015.pdf.

[Soi+18] P. Soille, A. Burger, D. D. Marchi, P. Kempeneers, D. Rodriguez, V. Syrris, and V. Vasilev. "A versatile data-intensive computing platform for information retrieval from big geospatial data". In: *Future Generation Computer Systems* 81 (2018), pp. 30–40. DOI: 10.1016/j.future.2017.11.007.

[Vas+12] V. Vassilev, P. Canal, A. Naumann, and P. Russo. "Cling – The New Interactive Interpreter for ROOT 6". In: *Journal of Physics: Conference Series* 396.5 (May 2012), p. 052071. DOI: 10.1088/1742-6596/396/5/052071.

[Voi] *Voila*. URL: https://github.com/QuantStack/voila.

[WM16] Z Wang and A Ma'ayan. "An open RNA-Seq data analysis pipeline tutorial with an example of reprocessing data from a recent Zika virus study [version 1; referees: 3 approved]". In: *F1000Research* 5.1574 (2016). DOI: 10.12688/f1000research.9110.1.

[Jup+18] Project Jupyter et al. "Binder 2.0 - Reproducible, interactive, sharable environments for science at scale". In: *Proceedings of the 17th Python in Science Conference*. Ed. by Fatih Akici, David Lippa, Dillon Niederhut, and M. Pacer. 2018, pp. 113 –120. DOI: 10.25080/Majora-4af1f417-011.