Expression of interest for EGI/EUDAT/INDIGO-DataCloud call for thematic services, EINFRA-12 (A). Deadline for submission: 27 January 2017

Contact information

Name of applicant: Hans Fangohr
e-mail address: fangohr@soton.ac.uk
Organization: This application is sent under the "umbrella" of the E-Infra9 project OpenDreamKit. We are proposing a series of services around the Jupyter (previously IPython) project which can be grouped as a Work Package itself or as tasks according to what the future E-INFRA12 project consortium will prefer. The legal bodies that would enter the consortium if this Expression of interest is accepted are:

- University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom
- Simula Research Laboratory AS, Martin Linges VEI 17, Snaroya 1367, Norway

Service and activity descriptions

Service overview	
Thematic Service name	Jupyter e-Infrastructure
Service description	Jupyter notebooks are a popular tool for sharing computational workflows, combining code with narrative description and results, including graphical output and interactive elements. The Jupyter project includes a number of services for sharing and running notebooks. See http://jupyter.org/ for more information.
Service provider	Software is developed by the Jupyter project, under the umbrella of the NumFOCUS foundation. This proposal is to integrate software from the Jupyter project as services running on EU infrastructure.
Service catalogue	-
Value	Enables trivial deployment and sharing of Jupyter notebooks; fosters dissemination and reproducible research, and enables and accelerates collaborative compute and data-centric research.
Current TRL level[1], acceptance criteria and validation/verification results	The Jupyter software components are TRL-8 level, with instances already deployed, and millions of users of the Jupyter notebook. Some will require customization or some specific development for integration in the EGI/EUDAT/INDIGO infrastructure.

1 Technology Readiness Level: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf

Access policy	Wide access
Terms of use	Will be developed as part of the integration work.
User groups and scientific disciplines served	All research disciplines that require computation or data analysis. The service will address three main use cases:- • Users who wish to reproduce and build upon published computational workflows. • Users who require frictionless access to greater computational resources; facilitating migration from Desktop to cloud computing. • Educators who wish to provide a user-friendly, stable training environment for any subject involving computation or data.
Service business model	While some fraction of the service provision would be reserved to enable free use, especially for research and education purposes, users would also have the option to pay for guaranteed computational resources. In practice, we envisage a model where users choose at the point of access to wait for available free resources or pay for immediate, dedicated resources. We anticipate that the service could have considerable value to private business interests in addition to academic users.

Service architecture
Define the service by describing its components. A service is usually composed of different service components that enable or enhance the service. A service component is a logical part of a service that provides a function enabling or enhancing a service. Although a service component underlies one or more services, it usually does not create value for a customer alone and is therefore not a service by itself. Examples of service component are software, and services that are provided or could be provided by e-Infrastructures. For example:
- EGI https://www.egi.eu/services/
- EUDAT https://eudat.eu/services-support
- INDIGO https://www.indigo-datacloud.eu/service-component

Service components

Name of component	Functional description, applicable standards and needed resource capacity (if applicable) e.g. CPU Time, storage capacity etc.	Provider If already appointed
EC Binder	Builds a container for a GitHub repository containing Jupyter notebooks and an environment description (such as a Dockerfile), then starts a cloud server where the user may interact with the notebooks. Binder encourages the dissemination of methods and reproducible research. The current	

Name of component	Functional description, applicable standards and needed resource capacity (if applicable) e.g. CPU Time, storage capacity etc.	Provider If already appointed
	Binder service (http://mybinder.org/) is overloaded by the demand, proving that it has identified just the right service for a critical need. We hope to provide a similar service to the EC community on a larger scale.	
JupyterHub	Gives authenticated users access to private, persistent Jupyter notebook servers. See https://jupyterhub.readthedocs.io/ for more details. JupyterHub allows centralised management of Jupyter notebook servers for a group of users. For instance, a lecturer may run a JupyterHub instance for students taking their course, removing the need for students to install software locally. We plan to provide JupyterHub as a service for EU researchers and educators, leveraging centralised e-Infrastructure, and to evaluate the access patterns and infrastructure components which best support it at this scale.	

Service integration with generic e-Infrastructures
If applicable. Define the proposed technical integration/service enhancement activities for this service that are proposed to be funded in the project. For example software integration activities that concern general e-Infrastructure capabilities, like those provided by EGI, EUDAT, INDIGO-DataCloud and other software.

Integration activity and concerned service components	The basis of our proposal involves enabling large-scale access to this technology for EU researchers and educators (possibly beyond) by *deploying, hosting, and maintaining* Jupyter based web services like JupyterHub, tmpnb, or Binder. These would be integrated with scalable data storage, computational processing resources, and common authentication mechanisms. On top of this, we wish to develop mechanisms for sharing and archiving large data sets relevant to the computational narratives stored in notebooks.
Overall necessary effort (Person	60 Person-Months, provisionally costed at €585k including travel & equipment budget. We expect this to run for 36

Months) and timeline	months, with resources divided between the two sites.
List of requested service components	Compute, storage, data, authentication.
Infrastructure integration *This activity addresses challenge 2 of EINFRA-12 (A): "seamless operation of highly scalable and agile data and computing platforms and services dedicated to analytics including hardware and software components, database, compilers, analytics software, supported to easy user entry points for the community of users". By filling this section the applicant commits to make the service discoverable in a central service catalogue and available for access by international user communities*.	
Description of infrastructure integration activities relevant to the proposed thematic service (to be planned in the project)	We will deploy Jupyter services including JupyterHub, tmpnb and Binder on European e-Infrastructure, integrating with existing resources and authentication mechanisms. The Jupyter codebase and existing multi-user orchestration machinery already developed and deployed are available as an in-kind contribution to support this project.
Training *to develop human capital and generate innovation by fostering adoption by new user communities. Training activities requested in this project must be specific to this call and to the thematic services in scope in your expression of interest. They must not duplicate training activities already funded by other initiatives and projects.*	
Description of training activities relevant to the proposed thematic service (to be planned in the project)	We will provide online documentation and tutorial materials to facilitate self-directed learning. Tutorials and workshops will also train researchers and educators how to use the services most effectively, and we will develop materials from these to allow other experienced users to conduct training sessions, extending the impact of the training activities beyond sites that the project participants can attend in person. Where facilities allow, talks and tutorials will be recorded and made freely available online to reach a wider audience. The existing body of documentation, tutorial materials and talks developed by the Jupyter project forms an in-kind contribution.

Relevance to EINFRA-12 (A) challenges

EINFRA-12 (A) challenge (**remove those that are not addressed by your activity**)	*Specify your contribution to the challenges highlighted by the e-INFRA-12 (a) call, providing whenever possible concrete examples and key performance indicators.*
1. The operation of a federated European	Integrating Jupyter notebooks into the proposed infrastructure would facilitate access to IT resources for

data and distributed computing infrastructure for research and education communities will optimise the access to IT equipment and services	relatively small computational tasks with rapid feedback to the user, in addition to the large batch jobs which are commonly run on shared compute resources. In particular, these technologies encourage open sharing of computational research methods, and reproducing and building on published work. Key Performance Indicator (KPI): number of users accessing the services.
2. All European researchers and educators are in equal footing to access essential resources	Providing these services on pan-European infrastructure could enable any European institution to, for instance, teach a course using hosted Jupyter notebooks, whereas the cost of such an activity may otherwise limit it to richer institutions. As part of this proposal, we would like to work on internationalisation within Jupyter, and provide a framework for people to contribute to translating the interface into their own language. KPI: number of available languages at the conclusion of the project.
3. Partnerships with industrial and private partners	The Jupyter project already works with a number of industry partners, including Google, Microsoft, Bloomberg and O'Reilly. Therefore we are used to collaborating with the private sector and understand their requirements of disclosure, intellectual property and data security. This proposal would build on these relationships. We could also exploit an ongoing collaboration with Logilab within the OpenDreamKit project to explore the use of Simulagora to promote non-interactive, offline collaboration on batch jobs. Simulagora is a web-based platform developed by Logilab, which allows users to run large scientific computations on on-demand cloud computating resources. It has run in production at https://www.simulagora.com for two years, and is used at large companies such as EDF and SNCF, proving its scalability and reliability. KPI: number of industry partners involved in the design and use of the system.
4. Train people in research and academic organisations	We will provide training in using the service through online tutorials and documentation, through video walk throughs and recorded presentations, through attendance and delivery of talks, tutorials and workshops at conferences that are attended by the communities (for example from PyCon, SciPy, Supercomputing, ...), through engagement and direct work with organisations that are likely to act as multipliers of the knowledge, such as the software carpentry and Southampton's Centre for Doctoral Training in Next Generation Computational Modelling. The aim of the training will be to empower future users to adapt the services to their needs. KPI: tutorials taking place, number of users attending and their demographic mix.
5. Avoid the locking-in	All of the software involved is not only open source, but

to particular hardware or software platforms	built around open protocols and file formats, allowing for easy interoperation. In particular, support for multiple programming languages is modular, relying on 'kernels' which speak a common protocol. Kernels for a wide range of languages have already been developed outside the Jupyter project, demonstrating the effectiveness of this modularity. KPI: available tools that integrate with the Jupyter ecosystem or make use of notebook files.
6. More scientific communities will use storage and computing infrastructures with state-of-the-art services	By presenting code in a more appealing format, Jupyter is helping to expand the use of programming beyond traditionally numerical fields, to become more commonplace among biologists and social scientists, among others. Connecting this interface to large scale computational resources would allow a diverse set of communities to tackle interesting problems. KPI: numbers attending training from disciplines less associated with computing; may be backed up by surveying samples of the user population online.
7. The open nature of the infrastructure will allow scientists, educators and students to improve the service quality	All of the Jupyter code is developed in the open, using the popular GitHub code sharing site. The Jupyter community regularly receives improvements from people using the software. KPI: increase in contributors from European countries.
8. Increase the incentives for scientific discovery and collaboration across disciplinary and geographical boundaries. It will further develop the European economic innovation capacity and provide stability to the e-infrastructure.	Jupyter notebooks provide hugely effective communication of all details of a computing or data centric study, as each notebook is executable and can (in principle) repeat its study by being re-executed. As such, each Jupyter Notebook directly supports collaboration of groups beyond geographic boundaries where person-to-person meetings to exchange details of a calculation are difficult to arrange. The notebook increases research effectiveness, in quantity and quality, and thus naturally accelerates economic innovation across many disciplines: academic and commercial communities from high-energy research to the financial sector have embraced the Notebook as their tool of choice. The emergence of the notebook as the de-facto standard computing environment is likely to contribute to stabilising the eco-system of computational tools; an indication for this demand and emerging standardisation is that github.com has already started to provide rendering of notebook files. KPI: qualitative survey sent to a sample of the end users near the end of the project.

Information on innovation, dissemination and exploitation

The global impact of the services presented above will be an increase in the number of researchers having access to and using the deployed services. The open source model of the Jupyter project tools will help to reach out as many researchers as possible and thus will contribute to the goals

of the European Research Area (ERA).

Innovation

The services proposed in this document are at TRL (Technology readiness Level) 8, meaning they are a complete and qualified system. If developers have pushed these services to that TRL level, they now need integrate these in an operational environment that is adequate to what end-users need. It is indeed the end-user pulling the innovation process to bring a technology from a TRL 8 to a TRL 9. Jupyter project developers can already count on the supervision of the OpenDreamKit advisory board, composed of representatives from different disciplines and sectors. However the expertise that EGI/EUDAT/INDIGO could provide in the frame of the future EINFRA-12 (A) consortium would be a welcome additional support.

But most importantly, the services that are proposed here will enable end-users to innovate. Jupyter notebooks provide a step change in efficiency in carrying out computational studies (be it based on data or computation) through full integration of the following steps into a single document: assumptions, code/data, results, post-processing, analysis, visualisation, interpretation and conclusions. This cuts down on the time required to carry out a full computational study: previously, all of the above steps had to be carried out using distinct tools and environments, and eventually put together in a report manually. Such studies are a core activity in research in most fields, including the development of research roadmaps, exploration of adventurous ideas and systematic evaluation of computational technologies with lower TRLs.

The widespread accessibility of Jupyter Notebooks, as proposed here, fosters innovation through lowering the effort of exploring innovative ideas and technologies. Also, the affordability and the performance gains of the open source services we propose will support the research and teaching work of academics.

Dissemination

Jupyter notebooks are an evolutionary step up from traditional academic papers. Whereas traditional papers merely advertise how computational research was performed, Jupyter notebooks allow the reader to completely reproduce, critique and build upon it on their own computational infrastructure. Combining a narrative description of the research along with executable code and data, they form complete research objects that fully encapsulate the research. As such, they are an unrivaled vehicle for the dissemination of computational and data-centic methods.

Providing services to host and share notebooks will improve dissemination at all scales, making it easier for anyone interested in a result to re-run the analysis, and inspect every single step of the work (which is often not fully documented in traditional academic publications).

In order to foster the dissemination of academic and teaching work within the ERA, we must plan to disseminate the provided services so that they can be widely used. As already described, all the services in question will remain open source. Thus, any institution, academic, student or citizen will be able to freely benefit from the Jupyter project services. The first targeted area remains the ERA, however the developed services can be used worldwide, provided the language barrier is overcome with the necessary work on the user interface translations.

Trainings, online courses, conferences, workshops will be organised, as described above, along with the services of the future EGI/EUDAT/INDIGO infrastructure. We are also willing to publish joint publications along with the future infrastructure

partners in software-oriented journals such as the Journal of Open Research Software (JORS), Transactions on Mathematical Software (TOMS), Journal of Software for Algebra and Geometry (JSAG), or the Journal of Open Source Software (JOSS).

Exploitation

The Jupyter Notebook accelerates research and impact in two ways: the quantity of studies we can carry out using notebooks is greater than without them (see section 'Innovation'), thus providing significant additional value to any computational and data-centric research activity. In addition to this quantitative improvement (more research done per investment), there is also a significant improvement in the quality of research and dissemination: as the published notebooks contain every step of the computational study, they can be fully investigated and exploited by other groups, for example by taking a published notebook as the starting point of a new study, not having to spend any time to reproduce published results at the beginning of a new project.

Availability of the access to the notebook thus improves the rate and quality of research that can be carried out, and is a methodological improvement to benefit research in all areas, leading to more effective exploitation of research investment.