Use Case Application: Galaxy

Contact person(s):

Name	Email	Node
Björn Grüning	bjoern.gruening@gmail.com	Germany
Frederik Coppens	frederik.coppens@ugent.vib.be	Belgium
Gildas Le Corguillé	lecorguille@sb-roscoff.fr	France

Contributors:

Name	Email	Node
Gabriella Rustici Krzysztof Poterlowicz Ajit Singh Carole Goble Ralf Weber	gr231@cam.ac.uk k.poterlowicz1@bradford.ac.uk ajit.singh@rothamsted.ac.uk carole.goble@manchester.ac.uk r.j.weber@bham.ac.uk	UK
Christophe Antoniewski Anthony Bretaudeau Christophe Caron Victoria Dominguez del Angel Olivier Inizan Fabien Mareuil Hervé Ménager Julien Seiler	christophe.antoniewski@upmc.fr anthony.bretaudeau@inra.fr christophe.caron@inra.fr victoria.dominguez@france-bioinformatique.fr olivier.inizan@inra.fr fabien.mareuil@pasteur.fr herve.menager@pasteur.fr seilerj@igbmc.fr	France
Alessandro Cestaro Federico Zambelli Marco Antonio Tangaro	alessandro.cestaro@fmach.it f.zambelli@ibiom.cnr.it ma.tangaro@ibiom.cnr.it	Italy
Rolf Backofen Uwe Scholz Bérénice Batut	backofen@informatik.uni-freiburg.de scholz@ipk-gatersleben.de berenice.batut@gmail.com	Germany
Kjell Petersen Abdulrahman Azab Sveinung Gundersen	Kjell.Petersen@uib.no abdulrahman.azab@usit.uio.no sveinung.gundersen@medisin.uio.no	Norway
Oswaldo Trelles Esteban Pérez-Wohlfeil	ortrelles@uma.es estebanpw@uma.es	Spain
Leon Mei Saskia Hiltemann Andrew Stubbs David van Zessen Youri Hoogstrate Chao Zhang Celia van Gelder	H.Mei@lumc.nl s.hiltemann@erasmusmc.nl a.stubbs@erasmusmc.nl d.vanzessen@erasmusmc.nl y.hoogstrate@erasmusmc.nl c.zhang@vu.nl celia.van.gelder@dtls.nl	Netherlands
Hans-Rudolf Hotz	hrh@fmi.ch	Switzerland

Daniel Sobral	dsobral@igc.gulbenkian.pt	Portugal
Pavel Fibich Petr Novak	pavel.fibich@cesnet.cz petr@umbr.cas.cz	Czech Republic
Brane Leskošek Jure Dimec Aleš Papič Andrej Kastrin	brane.leskosek@mf.uni-lj.si jure.dimec@mf.uni-lj.si ales.papic@mf.uni-lj.si andrej.kastrin@mf.uni-lj.si	Slovenia
Martin Reczko Artemis Hatzigeorgiou	reczko@fleming.gr hatzigeorgiou@fleming.gr	Greece
Fiona Mary Roche Karsten Hokamp	FMROCHE@tcd.ie KAHOKAMP@tcd.ie	Ireland

SECTION 1 - User community

Galaxy is a workflow management system that 1) provides support for reproducible science, 2) facilitates sharing of data and results and 3) removes the need for users to compile and install tools. Galaxy offers a user-interface, through a web browser, in which virtually any command line tool can be integrated. This is done by defining the inputs, outputs and parameters in a wrapper script. As analyses usually consist of multiple steps, tools can be composed in **workflows**, which facilitates the processing of multiple samples and reproduction of analyses. Galaxy is available as a world-wide free-to-use online portal, following open-source policy development and can be freely downloaded for a local installation.

The Galaxy workflow system is extensively used as part of national infrastructures in several ELIXIR Nodes. Galaxy itself is considered an integral part of bioinformatics infrastructure by many bioinformatics researchers and core facility groups because it enables simplified access to data and analysis tools under a single “intuitive” interface. There are over 97 public Galaxy servers worldwide and roughly 50 public instances in Europe, with many more private installations, and over 2400 virtual Galaxy instances have been launched on the Amazon cloud commercial service in 2015 alone. Around these servers, a strong European community of users, administrators and developers has been formed. Four of the annual Galaxy community conferences took place in Europe so far (Wageningen 2011, Oslo 2013, Norwich 2015 and Montpellier 2017), attracting over 200 participants each. The broad usage of Galaxy can also be appreciated in the F1000 Galaxy channel and the CiteULike Galaxy group, containing 4851 articles at the time of writing.

Education and training is an integral part of the Galaxy community. The Galaxy Training Network (GTN) are working since several years with Goblet and the ELIXIR Training Platform to enhance and deliver first-class training to the Scientific community - targeting not only scientists but also developers and admins. The ELIXIR Galaxy community has already provided training to thousands of researchers.
One of the recent efforts is to develop open, peer-reviewed and reusable training material in the spirit of Software/Data Carpentry. On http://training.galaxyproject.org, 63 tutorials are listed and we are welcoming everyone to contribute training material that can be used on public Galaxy instances, in Genomics Virtual Lab, ELIXIR clouds or via Docker on every personal computer.

The Galaxy Working Group was established in 2015 to monitor and foster the use of Galaxy in ELIXIR. Through a survey, it was determined that Galaxy is broadly used across ELIXIR Nodes and serves a sizeable community. This WG has organized five workshops in 2016 and 2017, and has already planned one for 2018. ELIXIR has sponsored the Galaxy Community Conferences in Norwich (2015) and Montpellier (2017) and members of the WG were involved in the organisation of both, as well as the one in Indiana, US (2016). This poster presenting the activities in ELIXIR from the latest Galaxy Community Conference highlights the focus areas of the Nodes regarding Galaxy.

SECTION 2 - Roadmap

Objectives

Coming three years

- Continue building the Galaxy community, become the focal point of the EU Galaxy Community
- Establish European equivalents to Galaxy Main server (usegalaxy.org)
- Integrate data access to facilitate efficient/simple retrieval
- Facilitate identifying and sharing tools, workflows and data
- Expand the portfolio of training material

Longer term

- Establish robust mechanism for data access & retrieval
- An international federation of Galaxy servers

Roadmap

A European network of Galaxy communities

Galaxy has always been very strong in Genomics for analysing High Throughput Sequencing data. However, in recent years, several sub-communities have formed around specific analysis task, among which several existing or prospective ELIXIR Use Cases.
The Galaxy Genome Annotation group aims at developing tools, workflows and software components to allow **genome annotation** within Galaxy. These developments enable automatic genome annotation in Galaxy and make Galaxy an 'orchestrator' for on-demand deployment of reference genomic databases. ELIXIR-DE and ELIXIR-FR are actively contributing to development in this project, and already use this architecture to host 15 insect and plant genomes on public reference databases, and new genomes are planned (~50 marine genomes, ~20 insect and plant genomes).
Two prominent examples for non-sequencing data are metabolomics and proteomics, where strong sub-groups have formed. In Metabolomics, the collaborative portal **Workflow4Metabolomics** (W4M) is dedicated to all aspects of handling metabolomic data. The W4M Galaxy instance hosts ~800 users and provides all W4M tools (publicly in the Galaxy ToolShed and as Docker container). W4M is part of the **Metabolomics Use Case** application. On the other hand, the **Galaxy-P** project deals with all types of proteomics data and the connection to OmicsDI/PRIDE is an ongoing effort with ELIXIR/EBI. Multiple European Galaxy instances are deploying Proteomics tools through Galaxy and initial training material was added. This European Galaxy community is part of the ELIXIR **Proteomics Use Case** application.
Next to the sub-communities highlighted here, there are many more: Plant Research, RNA bioinformatics, metagenomics, epigenetics, etc. The variety of communities shows that we have a very active and broad user field in Galaxy, and that there is added value of coordination at a higher level. The subcommunities we highlighted have already activities within ELIXIR, our aim is to combine efforts related to Galaxy, set up common analysis standards, related workflows and trainings.
With this application, we aim to foster interactions between data specific Galaxy communities, to set up common standards and quality control and to organize and resolve competing user request and systems requirements.

Visualisation in Galaxy

In data intensive analyses, visualisation is an important but still underutilized approach to e.g. evaluate quality of data. Galaxy already provides a pluggable framework for visualisations and interactive environments. To extend the available visualization tools, a partnership with BioJS was established 3 years ago. We want to strengthen this connection in the future to get more high-quality

visualisations into Galaxy to support researchers by producing publication ready and reproducible figures. While there is interest and efforts from many nodes (BE, DE, FR, …), no common approach has been developed so far. For that reason, we plan to develop as a consortium a common strategy for extending the visualization capabilities in Galaxy.

Galaxy cloud infrastructure across Europe

The growing number of data generated, users and communities require increasing amounts of computing power. Several members of ELIXIR have national science cloud initiatives, with a possible extension to the European Open Science Cloud. To help establish ELIXIR as the hub for life-science data, our first goal is to integrate easy authentication via ELIXIR AAI in the Galaxy codebase to be useable on different public servers. Furthermore, we want to facilitate the usage of Galaxy across the different ELIXIR clouds, e.g. by using CloudLaunch as a single entry point for users.
Some nodes already offer a centralized instance (IFB, France; de.NBI, Germany) or are building one (VSC, Belgium; The Netherlands). Italy is working towards a PaaS Galaxy with its INDIGO-Datacloud use case. During 2018, in collaboration with the Galaxy Core Team (usegalaxy.org), usegalaxy.fr and usegalaxy.eu (de.NBI) will be set up, providing European equivalents to the Galaxy Main server. Also the Australian Galaxy community is considering a usegalaxy.au initiative.
Our aim is to facilitate access to a broad portfolio of analysis workflows for European researchers, regardless of the specific compute resource used. This will require a close collaboration with the **Compute Platform** in ELIXIR.

Data access & integration in Galaxy

Getting data from public databases into a Galaxy instance is for most of the analyses the first essential step. Identifying files and their URLs and uploading these files in a computational environment is not straightforward for users with limited technical skills. This is exacerbated with datasets composed of numerous files distributed over distinct databases.
We aim to develop an interface to allow retrieval of data based on commonly used identifiers. This will enable researchers to use the familiar and well developed user interface of e.g. UniProt or BioSamples to stream data directly into Galaxy to make it as easy as possible to obtain data for scientists. A proof-of-concept has been developed at the 2017 Galaxy Community Conference hackathon. We aim in the future to cover **Core Data Resources** such as ENA, ArrayExpress, PRIDE or UniProt, and also more specialized databases such as Brenda, Silva, or RNACentral.
To optimally address the issue of data access and integration in Galaxy implies the standardization and the automation of data transfer, thus providing a solution beyond Galaxy. ELIXIR is ideally placed with all the life-science data providers in Europe to achieve this. Similarly, we aim to facilitate adding new genomes (and indices, annotations,...) for which we are developing, in cooperation with usegalaxy.org, a shared storage of common reference data across Galaxy instances. To ensure robustness, this needs to be integrated with repositories, in which ELIXIR should play a pivotal role.

Bringing Tools and Data together

Currently, a data-to-tools approach is prevalent which involves copying of a large volume of data to the compute environment. To avoid this, we propose a tools-to-data approach based on virtualization such as Docker, Singularity or rkt. To achieve this, the Galaxy community will be working together with the BioContainers project, which is part of the **Tools Platform**. At the time of writing, Galaxy has gained complete BioContainers integration, meaning that tools and workflows in Galaxy can run in isolated BioContainers. Given the rapid pace of change in life sciences, maintenance, update and extension of the BioContainers is required to keep the resource relevant and up-to-date.
On the medium to long term time scale, we aim at updating accessibility of tools and data to the next level, allowing users to easily combine public and private storage and compute cloud services. In this context, estimation of required resources is needed, which can be addressed by building on the

Galactic Radio Telescope project. There are also potential integrations with the benchmarking initiatives of the Tools Platform.

Training

Education and training is an integral part of the Galaxy community and all partners have expressed interest in the Training component. One of the recent efforts, supported by ELIXIR organized workshops (e.g. in Cambridge), is an open, peer-reviewed and reusable training material repository in the spirit of Software/Data Carpentry. To date, approximately one year after starting this initiative, http://training.galaxyproject.org contains 63 tutorials, providing material on usage and development of Galaxy. The material covers already 8 different biological domains (proteomics, metagenomics, …). This repository is open for everyone to use and contribute. Galaxy training material is already listed in TeSS and we will add support for BioSchemas in the next year. It is also planned to implement Galaxy tutorials in widely used e-Learning platforms, which is headed by the Slovenian Node. Further plans focus on facilitating usage of these materials, and the provided workflows, including more information about where this training can be run, capable instances, trainers, needed tools and virtual Galaxy images based on e.g. Docker. This will be done in collaboration with the Training, Tools and Compute platform.

FAIR and Galaxy

One cannot think of scientific data management without considering the FAIR initiative. Galaxy is devoted since more than 10 years to the goals now colloquially known as FAIR: capturing tools, versions and parameters in workflows, a ToolShed as tools registry, open source development, etc. Some of these features imply admin intervention e.g. for installing tools. To allow end users to identify (public) Galaxy instances containing the tools they need for an analysis, GalaxyCat has been developed. The GalaxyCat is an online catalog that lists all the tools available on various Galaxy instances and thus allows through a simple web interface to quickly find on which instances a tool is usable.

Galaxy already supports EDAM ontologies for data, data types and tools. All deployed tools in a Galaxy instance can be registered in bio.tools via a joined project between France, Germany, Norway, Denmark and UK called ReGaTE. Moreover, the Galaxy community developed plans with the bio.tools developers to integrate BioConda, BioContainers and the Galaxy ToolShed more tightly into the ELIXIR registry.

ErasmusMC is implementing myFAIR analysis, a secure and easy-to-use “end-to-end” FAIR data point and analysis architecture that is applicable to any translational or clinical research project. The aim is to extend the myFAIR analysis approach into a cloud based service using a scalable INDIGO-DataCloud service with built-in state of the art data security features. Also, a new version of Galaxy-EGA connector (GEGA) will be implemented within the myFAIR framework that allows users to integrate data from EGA and other sources into a single workflow in Galaxy, built on a prototype application developed to access EGA from Galaxy (2016 Human Data Implementation Study).

Workflows are a central concept in Galaxy. The Common Workflow Language (CWL) is gaining traction internationally. The **Interoperability platform** is preparing an Implementation Study based on CWL, where integration with Galaxy is within the scope of the proposal. For instance, the **Interoperability platform** is planning to annotate Galaxy objects (histories, workflows, etc.) as standardized ResearchObject to facilitate sharing.

For all these projects, which fit within the remit of ELIXIR, a Galaxy Use Case will facilitate the communication, coordination and dissemination across Nodes. We want to build on these projects that are being driven by the Nodes and improve FAIRness of Galaxy. Specific goals are to improve 1) findability of tools and Galaxy instances where these can be run, 2) usage of (access controlled) data in analysis workflows and 3) sharing and re-use of workflows.

Conclusion

This application shows that there are a lot of Galaxy communities and initiatives being organized across Europe in life sciences. An ELIXIR Galaxy Use Case has a natural role to ensure that we optimally use the expertise in our Nodes, and disseminate, re-use and integrate all knowledge and initiatives in ELIXIR.