Date of Approval:

Approved by: _____
Niklas Blomberg

Implementation Study

Title: "Development of Architecture for Software Containers at ELIXIR and its use by EXCELERATE use-case communities"

Scope

This project belongs to the Tools Platform with coordinated efforts with the ELIXIR Compute, Interoperability, and Training Platforms and demonstrators from the ELIXIR Use Cases around human data.

Purpose

The aim of this implementation study is to provide a stable infrastructure for unifying software containers solutions within ELIXIR. This infrastructure will provide an access point for end-users to find, generate, store, monitor, and even benchmark software containers solutions. Hardware infrastructure will be provided by an ELIXIR Node from the ELIXIR Compute Platform for software containers deployment while ELIXIR-ES will provide the backup system using EUDAT protocols and infrastructures. In the long-term this registry could become a relying service to the ELIXIR AAI allowing infrastructures to manage users accounts.

The impact of this infrastructure will be demonstrated across ELIXIR Platforms and Use Cases. Software containers are a key technology which enables the rapid deployment of software resources including workflows across a variety of systems e.g. HPC, Cloud environments, and local computers; and the connection with existing database repositories. Additionally, this technology will be used to support training activities carried out by ELIXIR, where trainers will be able to focus on the training content rather than in the technological framework of the training, during face to face or remote sessions. Such a leading role on the development of this infrastructure will greatly increase ELIXIR's visibility across many domains of life sciences and even beyond. The coordinated effort to develop this infrastructure is similar to previous efforts carried out in ELIXIR, such as the Beacon Project and Bioschemas and will also link into work taking place in the ELIXIR Compute and Interoperability Platforms in coordination with the GA4GH.

The ELIXIR Compute and Interoperability Platforms are using the container technology as the basis of a task based execution model within a life-science data analysis platform. As part of this Implementation Study, in collaboration with the ELIXIR Compute Platform, a central Bioinformatics Containers Central Service (BCCS) will be implemented, extending and reusing services like BioContainers and/or GA4GH TRS[1]. Such registry will effectively enable software containers to be registered for later use within analysis workflows.

Node	Name of PI	Role	PMs	Other Platforms	Projects				Use cases	
					1	2	3	4	5.1	5.2
EMBL-EBI	Yasset Perez Riverol	Technical Lead	6		1	1	4			
DE	Björn Grüning	Technical Co-Lead	6		1	1	2	2		
ES	Salvador Capella, Josep Ll. Gelpi, Sergi Beltran, Jordi Rambla	Use-case Lead	8		2	2			2	2
BE	Frederik Coppens	Member	3			1		2		
FR	Francois Moreews / Olivier Collin	Member	6		3		2	1		
DK	Jon Ison	Member	2			2				
EMBL-EBI	ELIXIR Compute Platform (Steven Newhouse)	Member	2	2	4					
FR	Training Platform (Victoria Dominguez)	Member		1				1		
IT+ES	Human Data (Salvador Capella, Rita Casadio, Giuseppe Profiti)	Member		3					0	(ES) 1 + (IT) 2
Hub	Tools Platform coordinator + Rafael C. Jimenez	Member	0							
		Total	33	396	11	7	8	6	2	5
Delivery	Starting from 1 January 2018 for a period of 12 months									

[1] https://www.youtube.com/watch?v=SuwONuO8LoA

Description of Work

The main goal of this implementation study is to **support existing community efforts** working on bioinformatics software deployments with traditional packages and containers. This proposal **focuses on the BioContainers** initiative. BioContainers is backed-up by a broader community and has several ELIXIR partners (e.g. Galaxy ELIXIR Working Group, BioShadock, CWL) actively contributing. BioContainers collaborates, integrates, and supports other related efforts like **BioConda**, **BioShadock**, **Bio.Tools, and OpenEBench**. The activities planned in this implementation study are relevant to all of the technical pillars and working groups within the ELIXIR Tools Platform, have a clear connection with the other ELIXIR Platforms, and can be applied to all ELIXIR Use Cases.

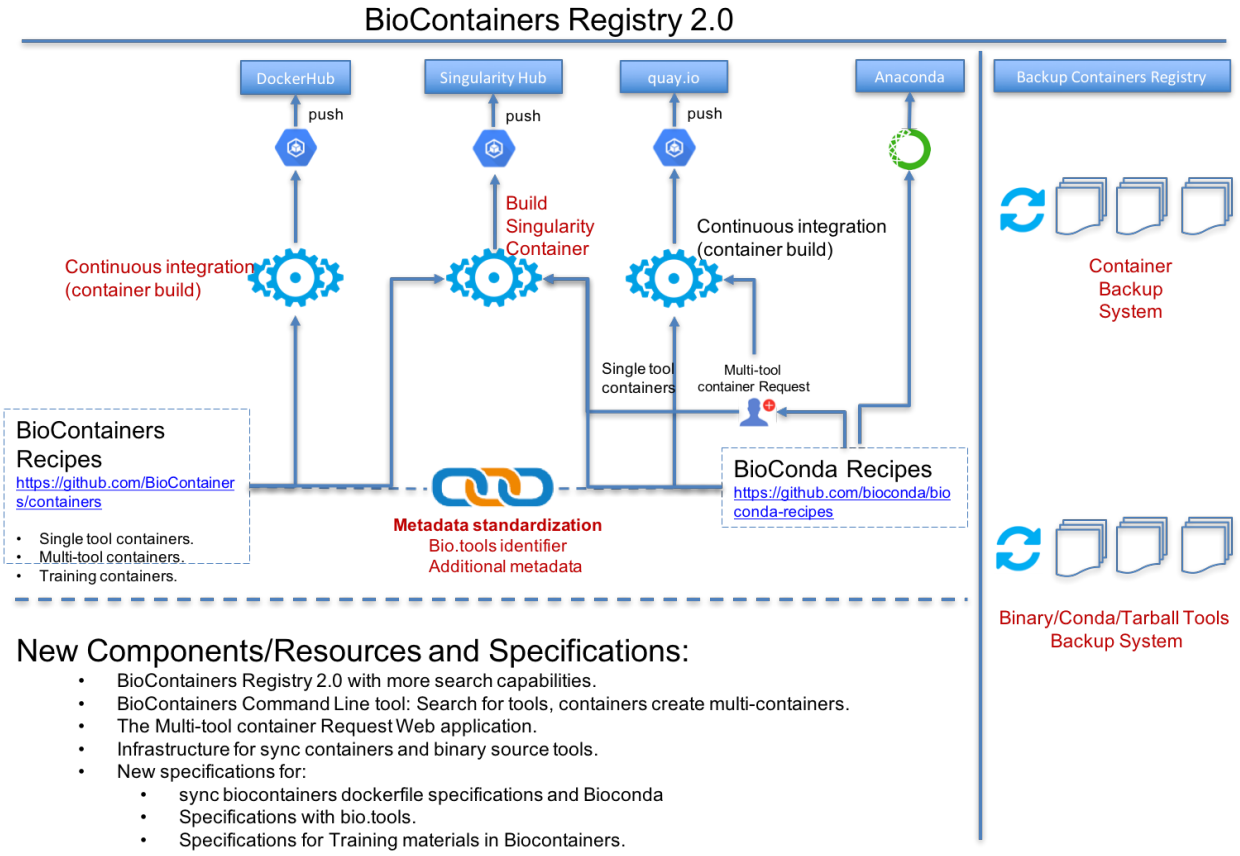

Figure 1: Proposed Architecture to be developed under the current Implementation Study.

This implementation study is put forward by the ELIXIR Software Deployment Group from ELIXIR Tools Platform. At the moment, this group is sustained by ELIXIR Nodes sharing a

common interest to work together on a harmonised strategy. The activities of the group successfully started in the first quarter of 2017 with the integration of the major community efforts on software containers driven by ELIXIR partners: BioContainers [1] and BioShadock [2]. The ELIXIR Software Deployment Group has already made available more than 2,000 bioinformatics containers (http://BioContainers.pro/registry).

This implementation study aims to push forward priorities discussed and approved by the ELIXIR Tools Platform and HoNs to support the bioinformatics developer and end-user communities in ELIXIR and beyond (figure 1). The study is conceived along four lines of development, and a validation on two selected ELIXIR Use Cases.

1. Develop the Bioinformatics Containers Central Service (BCCS) to support the use of software containers in ELIXIR.

The current release/deployment infrastructure of BioContainers is build on open-source/online services such as github, travis or quay.io. This architecture has been one of the cornerstones of BioContainers but also presents some major drawbacks: i) all containers, software packages, tarballs are stored on internet services which open the possibility for data lost at some point; ii) some of the heavy containers can not be built and/or deployed with those services making it impossible the use of some bioinformatics solutions; iii) the increasing number of software containers will soon reach the limit of repositories on some of these services. The BioContainers group will establish, with support from ELIXIR Nodes in the ELIXIR Compute Platform, the infrastructure to backup binaries, tarballs, package recipes and containers. This will drive the creation of a single logical Bioinformatics Containers Central Service (BCCS) without the replication of multiple and/or different registries across ELIXIR Nodes. This is the first step to prototype an infrastructure that will be able to continuously generate software containers, and submit them to appropriate repositories, backup any data preventing any lost and, more importantly, enable the deployment of such images across different computational clouds in a concerted effort with the ELIXIR Compute Platform and ELIXIR Nodes providing computational access to such systems via ELIXIR AAI, and potentially any other service.

ELIXIR-FR has developed a first prototype with a local CI (Continuous Integration) system that builds containers on github registry. The system is able to commit and pushes containers to remote/local registry. It also sends notifications on successful builds to a "web registry" (the user frontend) which updates the list of available containers (according to their type, version, etc.) to get an automatic and up-to-date catalog. This prototype can be used as a first study or test object for the Biocontainers group to support the work of the group, bringing back some insights

on technical and organizational issues. It will also allow the group to address the diversity of technologies available in a domain changing at a rapid pace.

Figure 1 shows the proposed infrastructure to deploy the new containers on BioContainers and make them available through BCCS.

2. BioContainers integration with bio.tools and OpenEBench.

BioContainers and bio.tools are highly complementary resources that should be integrated. Our strategy is based upon the vision of bio.tools becoming the sustainable primary archive for basic tool metadata, providing persistent references to “canonical” tool descriptions for re-use by BioContainers and other relevant resources. In addition, OpenEBench combines the bio.tools software metadata with other sources, including versions, with different software metrics which depend on the nature of each resource, e.g. web-servers vs binary packages. In the present study, we will annotate BioContainers with unique bio.tools identifiers, while information about versions, together with additional metrics, will be provided by OpenEBench. This will enable the BioContainer Registry to pull metadata from these resources for each container, thereby increasing its search capabilities. Conversely, we will implement an API in BioContainers that enables bio.tools and OpenEBench to retrieve the corresponding container metadata for each software.

3. BioContainers Registry 2.0 and Command Line tool.

During the present proposal, BioContainers will be scaled to different packaging flavours (e.g. docker containers, singularity images, conda packages, rkt containers). In order to be able to search/find the desired containers, a new version of BioContainers registry that extends the current functionalities (BioContainers.pro/registry/) should be developed. The new version should, by the use of ontologies such as EDAM[2], enable searching by keywords, software names, software versions and type of omics. In addition, the BioContainers user community will prefer to use a command line tool that enables the same search capabilities as the web interface; providing in advance the url from where the tool can be pulled (e.g. Docker pulls BioContainers/blast).

ELIXIR-FR has developed BioShaDock[3]. The next step is to achieve the integration of BioShaDock with Biocontainers with the automatic addition of bio related packages from the Debian distribution (hundreds of tools) with the scripts already written for BioShaDock.

The synchronization of the two repositories should result in a unique entry point for bioinformatics containers (single web GUI and domain).

[2] http://edamontology.org/page
[3] https://f1000research.com/articles/4-1443/v1

Achieving this goal will require distinct feature improvements related to synchronisation, redundancy treatment and standardisation.

4. BioContainers for Training and Support

On the training context, the ELIXIR training group modules have been classified into three types of trainers regarding specific needs: Research (TtR), Developers and Infrastructure operators (TtD) and Trainers (TtT).

We would like to engage the ELIXIR training community (TtD and TtT) to reuse and create containers for training purposes. To do so we propose to define a Container specification for *Training Containers* that will provide a **centralized point to download training solutions** that include: **Software** requirements to perform the training, **Presentations and Materials** and **Data**. The training containers will include more information including data and metadata needed. This work will also promote and include more *training containers* from some providers such as the **Galaxy community** and the **ELIXIR Training platform**.

Regarding the TrR community, ELIXIR-FR has created some containers in ELIXIR WP10.3 “training on genome assembly annotation”, which could be used in this project.
The Training Platform contribution to this project is defined in a separate project plan and linked to this project[4].

5. Demonstration of the use of software containers in selected ELIXIR’s Use Cases

The selected use-cases within this implementation study are led by the Human Data Use Case:

- EGA Integration
- Rare Diseases use-case. Proof-of-concept implementation of a human genomics variant calling pipeline using BioContainers and Galaxy.

Currently, BioContainers containers have been used by several platforms and frameworks [1]. The Galaxy Project has recently proposed Docker containers as a new way to solve workflow dependencies (BioContainers.pro/docs). Also, the PhenoMeNal H2020 project (related to the foreseen Metabolomics Elixir Use Case) has adopted and implemented BioContainers guidelines and deploying their containers into the BioContainers architecture. In addition, the work proposed in section 4 to deliver BioContainers for training in collaboration with the ELIXIR Training Platform, we are proposing to enable two additional ELIXIR Use Cases related to Human Data to adopt the present architecture of BioContainers.

[4] https://docs.google.com/document/d/1JRoGOVLVo_5MPKYFGGkP1dO6uTDl66obuAgVl6eTrf0/

5.1. EGA Integration.

The European Genome-phenome Archive (EGA) (https://ega-archive.org) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects. The EGA goal is to allow the best re-use of datasets deposited in its archive. Thus, the EGA is strongly interested in extending the current archival services by leveraging a service, or interrelated set of services, that could safely host tools and workflows, most likely as a set of containers, used to perform primary and/or secondary analyses. This service should keep these objects for a long period of time, equivalent to the life-cycle of the corresponding datasets.

In the current use case, the EGA would like to extend the current submission system to annotate the containerized versions of tools and workflows which have been used to analyze the data subject of submission. The current BioContainers architecture will enable EGA users to reproduce the corresponding analyses by accessing to the containers deposited in BioContainers and the data in EGA. In addition to the system extension, this proposal will select workflows from representative projects deposited at EGA, and provide them as containerized solutions within BioContainers, including the inclusion of those recipes/BioContainers ID as part of the studies metadata. As a proof of concept, selected datasets will be re-analyzed using these solutions.

5.2. Rare Diseases use-case. Proof-of-concept implementation of a human genomics variant calling pipeline using BioContainers and Galaxy.

The Rare Disease community has an increasing need to run validated and secure germline variant calling, annotation and analysis pipelines as whole exome and genome sequencing becomes standard practice in molecular diagnosis. These pipelines should be deployed, properly validated and audited for the sake of quality and reproducibility. Container-based solutions make it easier to achieve such requirements, even without a large expertise in preparing and deploying such type of pipelines. We propose a use case to make a proof-of-concept implementation of an automated variant calling workflow and validate it with a public gold-standard dataset.

The variant calling workflow will be based on the RD-Connect pipeline [3] including at least one aligner, one BAM processor and one variant caller (SNVs and InDels). The different components will be packaged in containers and made available through BioContainers. Galaxy will be the framework used to orchestrate and manage the different components. The Galaxy

workflow will also be made available through OpenEBench as part of further efforts on workflows benchmarking (project 2).

We will use a local Galaxy installation to test the workflow by analyzing the NA12878 Platinum Genome FASTQ files (https://www.illumina.com/platinumgenomes.html). The results obtained (SNVs and InDels) will be compared with the gold-standard developed by the U.S National Institute of Standards and Technology (NIST) and the Genome in a Bottle Consortium [4].

This proof-of-concept will be the foundation step towards providing a full whole exome/genome analysis pipeline for the rare-diseases community by demonstrating the feasibility of using BioContainers and an orchestrator for variant calling.

Alignment with other relevant Implementation Studies.

1. Led by the Training Platform:
    - Using clouds and VMs for bioinformatics training (Workshop as a Service), related with **Project 4** in current proposal.

2. Led by the Interoperability Platform:
    - Enabling the reuse, extension, scaling, and reproducibility of scientific workflows, related with **Project 1,2,3** in current proposal.

3. Led by the Compute Platform:
    - ELIXIR Implementation Study to provide a GA4GH Compatible Cloud Analysis Platform on ELIXIR Compute Platform (ECP) resources, related with **Project 1** in current proposal.

    - Data Movement: ELIXIR Proof of concept study on the availability of big datasets on remote compute infrastructure, related with **Project 1** in current proposal.

    - ELIXIR AAI Research Infrastructure production and proposed implementation study, related with **Project 1** in current proposal.

Milestones.

M1.1. Provision of bioinformatics containers following guidelines enabling their use in multiple architectures i.e. cloud, HPC, local computers, and different containers technologies i.e. Singularity, rkt or Docker. **Month 4. ELIXIR-FR, EMBL-EBI.**

M1.2. Enhance the current continuous deployment system to build Singularity-based containers of all available BioContainers. **Month 6. ELIXIR-DE.**

M1.3. The BCCS will implement a backup system to assist in case any of the main service providers of container images is not available. **Month 9. ELIXIR Compute Platform (EMBL-EBI).**

M1.4. The registry will be compatible with the GA4GH Tools Registry Service (TRS), so the ELIXIR Interoperability and Compute Platforms can use the registry to store and retrieve tools (encapsulated in containers) to be deployed and run within workflows. **Month 9. ELIXIR-DK.**

M2.1. Extend the BioContainers metadata (Bioconda recipes) to support the bio.tools unique tool identifier (toolID), and useful stable identifiers where desirable, such as from OpenEBench. **Month 4. EMBL-EBI.**

M3.1. Pull of metadata from bio.tools and OpenEBench APIs into BioContainers Registry for searching purposes, thereby increasing the search/find capabilities of BioContainers registry. **ELIXIR-ES. Month 6.**

M4.1. A specification for training containers, how to generate, store and find them. **ELIXIR-FR, EMBL-EBI. Month 9**

M4.2. Training and community engagement (using workshops and training) to generate more training containers. **ELIXIR-DE. Month 9**

M5.1.1. Extend the EGA submission system to annotate BioContainers used to process data being submitted. **ELIXIR-ES. Month 10.**

M5.1.2. Re-annotation of a set of exemplar submitted data to include existing BioContainers following established guidelines. **ELIXIR-ES. Month 11.**

M5.2.1. Implement a variant calling workflow using BioContainers and Galaxy. **ELIXIR-IT, ELIXIR-ES. Month 10.**

M5.2.2. Implement a RNASeq analysis workflow using BioContainers and Galaxy. **ELIXIR-ES. Month 11.**

Deliverables.

D1.1. Refine and deploy existing continuous deployment systems for the docker-based containers, connecting this system with BioContainers recipes, and pushing directly into DockerHub and similar repositories i.e. quay.io. **ELIXIR-DE. Month 3.**

D1.2. Establish a prototype Bioinformatics Containers Central Service (BCCS) that will provide a registry pointing to container images, and tarballs, binaries, and/or conda images. **ECP (EMBL-EBI). Month 9.**

D2.1. Create a mechanism (processing the BioContainers recipe metadata file) to automatically add multiple unique tool identifiers from bio.tools, and OpenEBench identifiers, into the multi-containers images. **ELIXIR-DE. Month 6.**

D2.2. Create a software component in BioContainers that reports to bio.tools and OpenEBench all the containers lacking identifiers in those resources. **ELIXIR-DE. Month 6.**

D2.3. Improve bio.tools and OpenEBench APIs for retrieving metadata for each BioContainers based on unique tools identifiers. **ELIXIR-ES, ELIXIR-DK. Month 9**.

D3.1. Progress review. EMBL-EBI, Month 6 (for contracting purposes)

D3.2. A Web interface and API that allows to search and find BioContainers. **EMBL-EBI. Month 9.**

D3.3. A Command Line tool that enables searching of BioContainers using the same capabilities as the web interface. **ELIXIR-DE. Month 9.**

D4.1. Generation of a list of containers (e.g https://quay.io/organization/galaxy) for training and push them into BioContainers registries. **ELIXIR-DE. Month 6.**

D4.2. Community engagement plan established to generate more training containers. **ELIXIR-BE. Month 9.**

D4.3. BioContainers for Training available in the TESS Portal. **ELIXIR-FR. Month 12.**

D5.1. Reanalysis of exemplar data sets deposited at EGA making use of newly associated BioContainers. **ELIXIR-ES. Month 12.**

D5.2. Test and validate implemented workflows on Galaxy with a gold-standard dataset and deposit results in OpenEBench. **ELIXIR-ES, ELIXIR-IT, Month 12.**

D5.3. Project final report delivered with outcomes and impact of the study. All participating Nodes respectively. Month 12.

Resources (in EUR)

Node	EMBL-EBI	DE	ES	BE	FR	DK	IT
PM	10	6	9	3	7	2	2
Salary	108,116	57,425	78,750	24,225	43,500	13,425	15,660
Travel (All Hands & project meetings)	3,444	1,500	10,000	1,500	4,000	5,000	500
Other costs		1,500[5]	2,000[6]		8,000[7]		
Overhead				4,759	13,875	4,606	3,915
Total per Node	**111,560**	**60,425**	**90,750**	**30,484**	**69,375**	**23,031**	**20,075**
					Grand total		405,700 EUR

Timeline

[5] ELIXIR-DE other costs: Workshops
[6] ELIXIR-ES other costs: publication fees
[7] ELIXIR-FR other costs: Train the Trainer workshop

Months/Projects	1	2	3	4	5	6	7	8	9	10	11	12
Project 1												
Project 2												
Project 3												
Project 4												
Use-Case 5.1												
Use-Case 5.2												
Workshops					Developers				ECCB	de.NBI		

Possible workshops:

Workshop	Expected Funding
Train the trainers	ELIXIR-EXCELERATE
Capacity Building of ELIXIR Nodes	ELIXIR-EXCELERATE
Tools developers workshop (expected May 2018)	ELIXIR Hub
Users workshop (ECCB - September 2018)	ELIXIR-DE
Bioconda/Biocontainers hackathon at the de.NBI annual meeting (expected Oct 2018)	ELIXIR-DE

Communication

The project outcomes will be presented in an ELIXIR webinar, communicated at the ELIXIR All Hands meeting and in an international conference.

References

[1] Leprevost, F. D. V.; Grüning, B. A.; Aflitos, S. A.; Röst, H. L.; Uszkoreit, J.; Barsnes, H.; Vaudel, M.; Moreno, P.; Gatto, L.; Weber, J.; et al. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* **2017**.

[2] Moreews, F.; Sallou, O.; Ménager, H.; Bras, Y. L.; Monjeaud, C.; Blanchet, C.; Collin, O. BioShaDock: a community driven bioinformatics shared Docker-based tools registry. *F1000Research* **2015**.
[3] Laurie, S., Fernandez-Callejo, M., Marco-Sola, S., Trotta, J.-R., Camps, J., Chacón, A., … Beltran, S. From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. *Human Mutation* **2016**.
[4] Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology* **2014**.