EU Brazil Cloud Connect
EU Brazil Cloud Computing for Science

EUBrazilCC

EU-Brazil Cloud infrastructure Connecting federated resources for Scientific Advancement

D5.1 Report on Use Case Requirements

Contract number:	FP7-614048 / CNPq 490115/2013-6
Start Date of Project:	1 October 2013
Duration of Project:	28 months
Work Package:	WP5 – Use Cases
Due Date:	M9 – 30/06/2014
Submission Date:	23/06/2014
Partner Responsible for the Deliverable:	Universitat Politècnica de València – UPV
Dissemination Level:	PU – Public
Nature:	R – Report
Author(s):	Erik Torres, Ignacio Blanquer – UPV
Reviewer(s):	Fabrizio Gagliardi – BSC, José Luis Vivas – UFCG

SEVENTH FRAMEWORK PROGRAMME

Ciência e Tecnologia
Ministério da Ciência e Tecnologia

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Change Log

Version	Date	Description	Author(s)
v1.0	30/04/2014	Initial table of contents	Erik Torres, Ignacio Blanquer – UPV
v2.0	05/05/2014	First draft	Erik Torres, Ignacio Blanquer – UPV
v3.0	21/05/2014	Questionnaires section added	Erik Torres, Ignacio Blanquer – UPV, Jacek Cala – UNEW
v4.0	30/05/2014	Requirements & preliminary prototypes sections added	Erik Torres, Ignacio Blanquer – UPV, Sandro Fiore, Giovanni Aloisio – CMCC, Israel Cruz – ISCIII, Mariano Vázquez – BSC, Antonio Tadeu – LNCC, Daniele Lezzi – BSC, Harry Seijmonsbergen – UvA
v5.0	03/06/2014	Circulation to WP5 list	Erik Torres, Ignacio Blanquer – UPV
v6.0	09/06/2014	Include feedback from 2nd technical meeting	Erik Torres, Ignacio Blanquer – UPV
v7.0	16/06/2014	Include comments after final complete WP5 review	Erik Torres, Ignacio Blanquer – UPV, Sandro Fiore, Giovanni Aloisio – CMCC, Jacek Cala – UNEW, Israel Cruz – ISCIII, Mariano Vázquez – BSC, Harry Seijmonsbergen – UvA, Iana Fufino – UFCG
v8.0	23/06/2014	Include comments after final complete QA review	Erik Torres, Ignacio Blanquer – UPV
v9.0	27/06/2014	Include comments after final complete PMB review	Erik Torres, Ignacio Blanquer – UPV

Document Review

Review	Version	Date	Reviewers	Comments
Draft	v2.0	05/05/2014	Fabrizio Gagliardi – BSC, José Luis Vivas – UFCG	Review of the ToC
QA	v8.0	23/06/2014	Fabrizio Gagliardi – BSC, José Luis Vivas – UFCG	Review of the deliverable
Final	v9.0	27/06/2014	PMB	Final version edits

Table of contents

List of figures

List of tables

Disclaimer

EUBrazilCloudConnect EU-Brazil Cloud infrastructure Connecting federated resources for Scientific Advancement (2013-2015) (hereinafter “EUBrazilCC”) is a Small or medium-scale focused research project (STREP) funded by the European Commission under the Cooperation Programme, Framework Programme Seven (FP7) Objective FP7-ICT-2013.10.2-EU-Brazil Research and Development cooperation, and the National Council for Scientific and Technological Development of Brazil (CNPq) of the Brazilian Ministry of Science and Technology (MCT) under the corresponding matching Brazilian Call for proposals MCT/CNPq 013/2012.

This document contains information on core activities, findings, and outcomes of EUBrazilCC project, and in some instances, distinguished experts forming part of the project’s External Expert Committee. Any references to content in both website content and documents should clearly indicate the authors, source, organization and date of publication.

The document has been produced with the co-funding of the European Commission and the National Council for Scientific and Technological Development of Brazil. The content of this publication is the sole responsibility of the EUBrazilCC Consortium and its experts and cannot be considered to reflect the views of neither the European Commission nor the National Council for Scientific and Technological Development of Brazil.

Executive Summary

The purpose of the Report on Use Case Requirements is to provide a comprehensive list of functional and non-functional requirements related to the three use cases of EUBrazilCC.

This report will serve as a guide to all partners, establishing the guidelines and procedures to be followed in the project for the analysis and management of requirements, ensuring that the partners adopt the same practices, conventions and procedures to best meet the needs of the use cases.

This document will be complemented by the deliverable D5.2 - Report on Use cases implementation at first year; D5.3 - Final report on Use cases implementation; and D5.4 - Validation report on the EUBrazilCC platform.

This report is considered to be a live document during the project lifetime and it is open to be modified according to the possible upcoming necessities.

1. EUBrazilCC use cases overview

EUBrazilCC aims at creating an intercontinental federated e-infrastructure for scientific usage. This e-infrastructure will join resources from different frameworks, like private clouds, supercomputing and opportunistic desktop resources to offer the community high-level scientific gateways and programming models.

The overarching objective of EUBrazilCC is to drive cooperation between Europe and Brazil by strengthening the scientific and knowledge-based society as key to sustainable and equitable socioeconomic development. The core of this collaboration consists of the three selected scientific use cases, which will require the collaboration between Brazil and Europe in the provision of data, services and expertise. None of the three use cases could feasibly be taken forward without the cooperation between Brazilian and European entities and without the availability of computing and data resources.

However, contributing to the technological and scientific advancement in the particular field of study of each use case is only one part of a much broader goal, in which other factors, such as sustainability and interoperability, will play a major role for the success of EUBrazilCC. Beyond the expected advances, the three use cases will provide a practical validation scenario for the EUBrazilCC infrastructure.

Moreover, the EUBrazilCC federated e-infrastructure has the added value of bringing similar scientific communities from across the Atlantic and helping them to learn how to work together. Since the communities need to use different IT environments and need to use all computing power and data processing capability they can access, integrating different computing infrastructures and tools becomes a need. The EUBrazilCC federated e-infrastructure aims at fulfilling this need, also contributing to close the gap between cloud service providers and scientific users. EUBrazilCC will build on the experience gained from the previous highly successful projects: VENUS-C[1] –co-funded by the European Commission under Framework Programme 7 (2007-2013) Research infrastructures projects– and EUBrazilOpenBio [2] –funded under the Objective FP7-ICT-2011-EU-Brazil Research and Development cooperation and by the Brazilian Minister of Science Technology and Innovation (MCTI) - National Council for Scientific and Technological Development (CNPq). Four partners of EUBrazilCC participated in the research project VENUS-C, where 27 applications from several scientific user communities were adapted to a cloud computing environment that runs on resources provided exclusively by European donors. This experience was later applied to the development of the EUBrazilOpenBio hybrid data infrastructure that integrated and federated existing European and Brazilian infrastructures and resources for the study of biodiversity. EUBrazilCC is the logical step towards the consolidation of the intercontinental federated e-infrastructure that was initiated with EUBrazilOpenBio and the generalization of the programming models that were developed in VENUS-C, allowing us to extend and enhance the services we are able to offer to our scientific communities.

[1] http://www.venus-c.eu/
[2] http://www.eubrazilopenbio.eu/

1.1. UC1 – Leishmaniasis Virtual Laboratory (LVL)

Leishmaniasis is one of the world's most neglected diseases, affecting largely the poorest of the poor, mainly in developing countries; 350 million people are considered at risk of contracting leishmaniasis, and 1.5-2 million new cases occur every year. Main endemic areas include the Indian subcontinent, Latin America, North and East Africa, and the Euro-Mediterranean basin. It is caused by protozoan parasites of the *Leishmania* genus, and transmitted by an insect vector, the sand flies. The disease is spreading because its control is exacerbated by three escalating risk factors: human-made and environmental changes; immune status (essentially because of *Leishmania*/HIV co-infection); treatment failure and drug resistance.

More effective control of neglected tropical diseases like Leishmaniasis is vital to achieve poverty reduction and spur social-economic development without the need to wait for countries to fully develop and for living conditions to improve over a potentially long period of time.

1.1.1. Better understanding of the etiology of the disease

There is today an insufficient clarity on the interaction of leishmaniasis risk factors due to scarcity of investigation. The intersection between ecology and epidemiology has been explored in a few studies on host-parasites-diseases scenarios. Furthermore, the recent northward and southward spread of leishmaniasis may be related to environmental and/or climate changes, and is exemplified by the recent emergence of canine and human visceral leishmaniasis in the province of Misiones (Argentina), currently the southernmost focus of the disease in Latin America; the northward spread of the disease in European affects both endemic areas such as Italy and Spain and other previously non-endemic areas in Europe.

The correct identification of the etiological agent could be crucial for the prognostic of the disease since different species or strains can produce different clinical outcomes. Also, treatments or even diagnostic tools could be inefficient regarding different species or genetic populations. Molecular methods like Multilocus Sequence Analysis (MLSA) and Multilocus Microsatellite Typing (MLMT) have been applied to the identification of *Leishmania* species and strains, and to assess its population structure. Next-Generation Sequencing (NGS) has also been used to understand the biology of *Leishmania* species.

In nature, the interaction *Leishmania*-sand fly seems to be species specific, particularly in the Neotropic where biodiversity of both groups is greater. Changes on environments have an important impact on vector species populations and the correct determination of species is crucial for entomological surveillance. Also, the genetic structure of vector populations is important for control strategies management.

EUBrazilCC will devote efforts to create a database of molecular markers of *Leishmania* and sand flies to contribute to a better surveillance and knowledge of leishmaniasis.

1.1.2. The Leishmaniasis Virtual Laboratory

EUBrazilCC will adapt GIS tools and process pipelines for molecular data in the platform through the integrated programming framework, leveraging from the previous experience in EUBrazilOpenBio and

VENUS-C projects. This integrated data will enable the development of a Leishmaniasis Virtual Laboratory (LVL) similar to the Virtual Vector Laboratory for the Chagas disease in the Americas[3], developed by the Biodiversity Institute of the University of Kansas in collaboration with Oswaldo Cruz Institute (Fiocruz) and other partners in the region. Having a LVL for leishmaniasis will enable public health workers and researchers to access and supply relevant and in-depth information or data on the parasite and vector responsible for this disease. This will also be an automated tool for species determination given a more precise and gaugeable identification of parasites and vectors of Leishmaniasis. However, such a facility will require the integration of distributed data (geographical, biomolecular and clinical) and the availability of computing resources for executing the proper processing pipelines.

The LVL use case needs to integrate data from four main sources: CLIOC (Coleção de Leishmania do Instituto Oswaldo Cruz, COLFLEB (Coleção de Flebotomíneos do Instituto Oswaldo Cruz), Collection of Leishmania of the WHO-Collaborating Centre for Leishmaniasis ISCIII, and speciesLink, the species occurrence database. Once integrated, those databases will provide the necessary data to run several processing pipelines to join classical data from collections (geographical coordinates of species occurrence) with data from molecular typing (Barcoding, MLSA), and to explore those data using GIS tools in order to generate an atlas of parasite/vector increasing the knowledge on their species and genetic populations distribution.

1.2. UC2 – Heart Simulation

Cardiovascular diseases have a huge impact on population, particularly people low and middle incomes. Since the early 1950s, there have been increasing efforts to develop computational models and simulation-based techniques in order to assess physiological and pathophysiological conditions accounting for the multiple time scales and levels of spatial organization present in the cardiovascular system (CVS). Applications such as diagnosis, treatment and surgical planning have benefited enormously from these complementary tools. Simulating a heartbeat is a complex, multi-scale problem. This means that many scales are coupled, covering different orders of magnitude from descriptions of electrical propagation, cells arrangement into a spatial description, generally known as myofiber orientation, up to the geometry of the cardiac chambers.

EUBrazilCC leverages the integration of heterogeneous supercomputing and virtualized infrastructures with the orchestration –through the components integrated in the platform– of two simulation codes from Brazil and Spain, addressing two complementary problems in cardiovascular modelling.

1.2.1. Cardiac Electromechanical Modelling

Recently, electro-mechanical cardiac simulations have become of frequent use to understand the emergent properties of complex, multi-scale systems tightly interconnected in the heart. Electromechanical simulations have proved useful, for example, for predicting the effects of cardiac resynchronization therapy, patient-specific applications and cardiac growth. Computational models of electrophysiology have been used to understand arrhythmias that have been acquired with age, drug

[3] Costa et al.: "Distributional potential of the *Triatoma brasiliensis* species complex at present and under scenarios of future climate conditions". Parasites & Vectors 2014, 7:238

use, or are of genetic nature. The most current research is also able to identify the need of morphological details in the anatomical models used as a substrate for arrhythmic episodes.

In EUBrazilCC, the goal of this use case is to analyse the output sensitivity to different fibre fields and initial conditions. EUBrazilCC uses for this purpose the Alya Red System, a simulation tool that has been fully developed by the Barcelona Supercomputing Centre (BSC), from the numerical methods up to the parallel implementation, including mesh-generation and visualization. The Alya Red System is built on top of the Alya System, which is a Computational Mechanics code specially designed for running with high efficiency standards in large-scale supercomputing facilities, capable of solving different physics in a coupled way (fluids, solids, electrical activity, species concentration, free surface, combustion, embedded bodies, etc.). The goal of the Alya Red System is to develop a Cardiac Computational Model at organ level, and to simulate fluid-electro-mechanical coupling.

1.2.2. One-Dimensional Arterial Blood Flow Modelling

The scientific community recognizes that an integrative approach to the modelling of the cardiovascular system (CVS) would need to be able to model and accurately simulate the interactions between phenomena taking place at different time and spatial scales, going from genes expression to the whole functioning of the organism. From pioneering works studying the basic theoretical ingredients through the early developments of topological descriptions of the CVS, subsequent improvements and alternatives, and finally reaching incontestable in-vitro validations and in-vivo verifications, one-dimensional models have had a prolific existence, and are currently established as a more interesting tool to gain insights into the most diverse aspects of the systemic circulation.

The Anatomically-Detailed Arterial Network (ADAN) model, developed at the Laboratório Nacional de Computação Científica (LNCC), starts from anatomical data and physiological concepts in order to perform cutting-edge cardiovascular research supported by the modelling of physical phenomena and simulation-based techniques. The ADAN model includes, in the definition of the vascular topology, most of the arteries which are acknowledged in the medical and anatomical literature for an average male. This requires taking approximately 1,585 arteries into account.

1.2.3. Creating the most complex cardiovascular model

In EUBrazilCC, this use case will couple the Alya Red heart model with the ADAN model in order to deliver a novel model of the blood flow circulation in the cardiovascular system, making it possible to widen the range of cardiovascular scenarios that the model is able to address. In this sense, it is possible to study the effect of wave propagation back into the heart to analyse, for instance, the impact of aortic regurgitation on hypertrophy in the cardiac muscle, or the consequences of arterial stiffening on heart efficiency. Likewise, changing parameters in the heart functioning allows us to understand how they affect the pressure pulse conformation in detail. In any case, the coupled model naturally integrates phenomena taking place at such vascular entities (heart, systemic arteries).

The integrated Alya+ADAN simulator will leverage the EUBrazilCC federated platform to compute high-resolution models, which provides an invaluable added value to these simulations, enabling researchers to increase the level of detail of the models without the need of dealing with the underlying technology-related complexity.

1.3. UC3 – Biodiversity and Climate Change

It is necessary to gain a better understanding of the mutual interaction between climate change and biodiversity dynamics through joint actions aimed at integrating data at a global scale and with the involvement of experts on both sides of the Atlantic. Current knowledge gaps imply that key parameters on the impact of the biodiversity system on the climate system (and reverse) are missing and currently entered as assumptions in climate models. Another approach would require a holistic methodology to identify and unravel patterns and processes in the system-system interactions. This methodological approach assumes the availability of both overview and detailed data sets with observations and measurements of the systems components, together with advanced analytical and modelling software. It also requires computational capacity to run the demanding workflows on huge data sets.

On biodiversity & climate change, two Brazilian areas are objects of the case study: The Semiarid region and the Ducke Reserve (Amazon Forest). The Semiarid region of Brazil is a large drought and desertification prone region, with 1 million square km and highly populated. It is one of the main targets of the considerable Brazilian efforts on poverty reduction. Water resource scarcity, land degradation and desertification are driven by climate variability and poverty (overgrazing and deforestation for energy use), which put major pressures on natural resources. Appropriate use of satellite images in their full potential would substantially help a best understanding about the main land use changes drivers: human activities and climate changes. Satellite images derived from several sensors will be processed by specific algorithms to produce estimates of energy balance and evapotranspiration of water to the atmosphere. These estimates, combined with the analysis of historical time-series, allow the detection of changes in the terrestrial plant systems and they will be used to discriminate the influences from human occupation and those from climate variability and/or change on energy fluxes and land cover. The algorithms have to be calibrated and validated using ground-based data. Thus, a large multiple source set of satellite and ground data has to be processed and comparatively analysed. Differences in temporal and spatial resolution among the satellites' images and between them and ground data, as well as differences in the time series extension and algorithms features and parameterization, introduce further processing challenges. Identification of a basic set of climate change indicators will be cross-related with biodiversity indicators data and plant species occurrences data. It can improve the understanding about climate change and biodiversity dynamics of terrestrial plant systems through multi-level and multi-dimensional analysis. Early warning systems (EWS) have been developed and applied for both environmental forecasting and monitoring in the past two decades, which helped to cope with the recurrent droughts and mitigate their effects. However, those early warning systems still make limited use of remotely sensed data. Appropriate use of satellite images in their full potential would substantially increase the benefits provided by the current EWS in the region.

The second case study is the Adolfo Ducke Forest. It is a 10,000 ha protected area on the outskirts of Manaus (AM). The reserve sits at the intersection of two major drainage areas, the Amazon River" and the Negro River. The reserve is made up of research plots designed to study the biota of the regions, which may serve as a basis for biodiversity surveys in other areas of the Amazon region, and to study the impacts of fragmentation. Biodiversity and its interaction with climate change is only sparsely monitored, mostly on the basis of field observations at a local scale. Remote sensing technologies can be applied at scales ranging from local to global extent. Various technologies, such as LiDAR have shown to provide spectral signals that offer great opportunities to describe the 3D structure of the vegetation and the patterns that underlie the earth surface. Hyper spectral images provide information on the physical (soil water) and chemical (nutrients) quality of the land surface and biochemical properties The methods used in this use case intend to extract 3D vegetation patterns from LiDAR elevation data and hyperspectral imagery (Remote sensing data) of the vegetation. The combination of 3D-structure and quality of the vegetation can thereupon be used as a proxy for biodiversity and habitat suitability. The inversion of remote sensing data to 3D vegetation structures is still at its infancy because of the high throughput computation needed. Also Object Based Image Analysis (OBIA) and classification, including segmentation and machine learning, generates the need for extensive computational power.

This approach has the potential to help identify forest patches and vegetation structures (and the associated biodiversity) that have an important role in mitigating the negative impacts of climate change. The workflow for this use case requires the processing of data from various origins, computing of models and structures, and the visualization of results. EUBrazilCC is well-placed to implement a combined workflow for the study of climate change that gathers experience from two major European initiatives for biodiversity and climate change research: 1) the European e-Science infrastructure for biodiversity and ecosystem research (LifeWatch), which is led in The Netherlands by the University of Amsterdam (UvA); and 2) the European Network for Earth System Modelling (ENES), through the Euro-Mediterranean Center on Climate Change (CMCC), with the Ecological Niche Modelling tools and data provided by the Centro de Referência em Informação Ambiental (CRIA).

1.3.1. Driving cloud adoption to tackle biodiversity & climate change

In EUBrazilCC, applications and services will be adapted to provide a framework for the study of the impact of biodiversity and climate change. The cooperation from Brazilian and European centres is key, as expertise in biodiversity modelling (e.g. openModeller) and data (e.g. speciesLink) comes from the Brazilian side (CRIA), while the expertise on climate change data analysis, together with the access to the ESGF federated data archive and additional remote sensing data sources, comes from both the European (CMCC) and Brazilian (UFCG) side. In EUBrazilCC, this use case uses the tools and frameworks provided to run these services on the federated resources. It is important to remark that the extent of the project requires the selection of one or two study areas with available data. The climate datasets will be used to infer a set of climate indicators (e.g. surface temperature) for the selected areas. Such indicators will be properly identified together with the analytics workflows in order to compute them. In this regard a comprehensive analysis regarding multiple scenarios from the CMIP5 data archive will be taken into consideration to provide climate indicators projections, on the target areas, over the next century. Analytics workflows for the computation of the climate indicators will be implemented thanks to COMPSs and PDAS. The former will provide the workflow support and coordination, and the latter will be responsible for running massive data reduction operators on large amount of climate data (terabytes order). The results will be stored into a clearinghouse system, a multidimensional database (e.g. data warehouse) that will act as a cache for the entire set of indicators.

The biodiversity workflow will include several steps like (1) filtering out the areas with human induced changes; (2) analysing vegetation structure and physical/chemical quality of soil surface; (3) modelling ecological niche to compute (potential) species distributions/abundance; and (4) modelling and identifying key forest systems to buffer climate change.

2. EUBrazilCC infrastructure overview

EUBrazilCC leverages a set of components for the use of supercomputing, private cloud and cloud opportunistic resources in desktops. EUBrazilCC will expose these resources through programming frameworks and scientific gateways, easing the adaptation and deployment of the applications that use data and computing resources in both sides of the Atlantic. The integration of the components takes into account existing standards to maximize interoperability with provisioning systems and existing infrastructures. Tools for transparent access to supercomputers such as CSGrid are combined with execution frameworks such as COMPSs in order to deploy applications on multiple private and public cloud resources and with complex workflow managers that run on top of different infrastructures through e-Science Central. EUBrazilCC includes the parallel data analysis service (PDAS) for big data analytics, the fogbow middleware to federate opportunistic resources and the mc2 framework for scientific gateways.

EUBrazilCC applications involve complex workflows and access to huge datasets. Thanks to the collaboration between European and Brazilian research centres the project will access different data sources in Europe and Brazil to perform the three selected use cases.

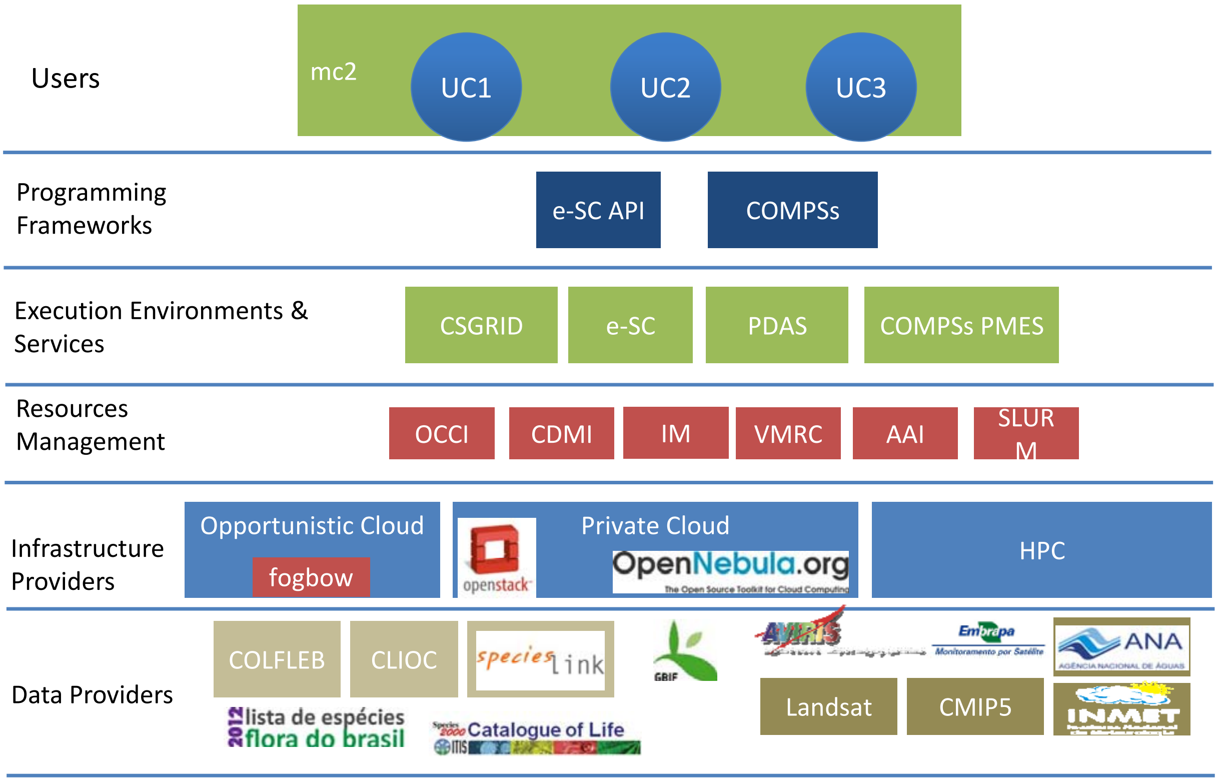

Figure 1 – EUBrazilCC architecture

Figure 1 shows a diagram of the EUBrazilCC general architecture, which is currently under development. The components of the architecture are layered in the diagram following the logical order of resource and service provisioning: the top side of the figure represents the services that constitute the highest layer of the infrastructure, which is closest to the end-users –represented in the figure by the use cases–, while the lower layers represent the internal services and resources that are intended to support the implementation of the use cases. Both e-Science Central (e-SC) and COMPSs provide programmatic access through APIs to the workflow management and execution functionalities that are provided by these tools. Those APIs are part of the programming frameworks layer in Figure 1 and are available to the end-user applications. At the same time, both e-Science Central and COMPSs provide the necessary bridges to the execution environments like CSGrid and PDAS that leverage on the available resource management mechanisms –which can be systems like IM+VMRC or cloud standards such as OCCI– to allocate computing and storage resources from the different EUBrazilCC infrastructure providers. The lowest layer in Figure 1 consists of the data sources that are integrated in the infrastructure via different protocols and mechanisms that are specific to the data provider.

It's expected that more details of the infrastructure become available in the next months, specifically with regard to the relationships among the components, their interactions, and other dynamic aspects of the infrastructure.

3. Procedure for the analysis of requirements

The use cases were presented to the consortium in the first plenary meeting. After the overall discussion of each use case, the partners continued with small, focus group discussions. Before the focus groups were held, the guidelines and the action plan for the analysis of requirements were circulated to the partners and a brief presentation was given by the WP5 coordinators to introduce the plan.

The action plan includes activities that would be completed in 4 phases (elicitation, analysis, specification and validation), shown in Table 1, along the project life cycle. All these activities were scheduled after the WP5 session of the first plenary meeting. Table 2 presents a checklist of 12 tasks to be completed in order to fully address the requirements of the use cases.

Activity	Requires	Produces
Elicitation		
The expected user classes (actors) and other stakeholders are identified from the scenarios described by users.		Report on use case communities (Sections 4.1.1, 4.2.1 and 4.3.1).
Users and developers reach a common understanding of the tasks and goals of the use case actors.		
Users describe the environment in which they usually perform their work.		Inventory of applications and data sources (Sections 4.1, 4.2 and 4.3).
User's quality expectations are discussed.		
Analysis		
High-level requirements are decomposed into an appropriate level of detail that can be addressed by the developers.		Initial list of requirements (Wiki[4]).
New functional requirements are generated from the analysis of the information.		

[4] http://eubrazilcc-rm.i3m.upv.es/projects/eubrazilcc/wiki/Requirements/history

Quality attributes are transformed into non-functional requirements (e.g. usability, performance, security, interoperability, scalability).		
Requirements are allocated to infrastructure functions and components.	WP3 and WP4 advance in the definition of the infrastructure.	
Implementation priorities are defined.		Implementation plan (Sections 6.1, 6.2 and 6.3).
Requirements that cannot be addressed in the project are rejected.		List of use case requirements (Sections 4.1.2, 4.1.3, 4.2.2, 4.2.3, 4.3.2 and 4.3.3).
Specification		
Requirements are transformed into technical documents and diagrams.		D5.1 – Report on Use Cases Requirements.
Validation		
Documented requirements are reviewed and accepted.	D5.1 is approved.	
Early prototypes are developed and validated in the infrastructure.	Integration tests are successfully passed.	D5.2 – Report on Use Cases impl. at 1st year.
Users confirm that use case implementation meets their needs.	User acceptance tests are successfully passed.	D5.3 – Final report on Use Cases impl.

Table 1 – Phases of the analysis of use case requirements

№	Action	Date
1	Create the elicitation plan, schedule meetings and demonstrations in the calendar of the project.	PM1
2	Create a Wiki page with the initial list of requirements.	PM3
3	Verify that all user classes are represented.	PM6
4	Verify that all requirements align with the overall project's vision and scope.	PM6
5	Verify the level of detail of the requirements.	PM6
6	Add priorities to requirements.	PM6
7	Allocate requirements to infrastructure.	PM6
8	Model the requirements.	PM6-9
9	Develop early prototypes, include APIs and user interfaces.	PM9-13
10	Evolve the infrastructure to support the use cases.	PM9-13
11	Develop validation tests.	PM9-16
12	Validate use case requirements.	PM22-28

Table 2 – Requirements analysis checklist

This methodology was adapted from the approach described by Wiegers et al.[5], (2013). The rest of this report covers the analysis procedure to the phase of specification. Validation will be documented in the subsequent deliverable D5.2 and D5.3.

For the elicitation phase, two questionnaires were defined and filled-in by the WP5 partners. The questionnaires are described in the next sub-sections. The results are described in Section 4. Section 5 presents the consolidated analysis of the use case requirements, defining a list of general requirements. This information was carefully analysed and discussed in the second technical meeting that took place in June 4-6, in Barcelona, leading to the first release of use case requirements, which are described in this document.

3.1. Questionnaire for the identification of use case requirements

This questionnaire addressed six major technical points and an additional asking for a description of examples relevant to the use case analysis. This information has been used as a basis for the extraction of use case requirements and the consolidated analysis of the requirements.

The questionnaire had seven points which are described below, including the objective for each one. The answers and the analysis are included in Section 4.

1. Describe your type of application.

This refers basically the type of interaction that you expect from your users. For example, your application is executed in a command line, it is used through a web interface, or it is an executable program with its own graphical interface.

2. Describe your data and your data sources.

First, you should list the data sources that your application will use. Then, you should complete this information by describing: 1) the nature of the data (is the data publicly available on the Web or, on the contrary, is there a licence required to access the data?); 2) what is the expected data access volume required by a single execution (in your maximum and average cases); 3) what protocols are available to access the data (e.g. ftp, http)?; and 4) in which format is the data stored? In case that you use any domain-specific data, please, provide as many details as possible about the format, including links to existing standards or definitions and software libraries that can be used to manipulate your data.

3. Describe your execution model.

Define the CPU time and memory required by a single execution in your maximum and average cases. Does your application involve any parallelization strategy (high-performance or high-throughput computing, multiple cores, multiple processing workflows)? How many concurrent executions do you expect to have? How many concurrent users?

[5] Software Requirements, (2013) Third Edition (ISBN 9780735679665), by Karl Wiegers and Joy Beatty.

4. Describe your network traffic.

Define the amount of data that will be transferred over the network (bandwidth consumed) in your maximum and average cases. Note that you don't need to measure the amount of data transmitted and received by your application. Instead, you can report here the size of your data and the expected frequency of download and upload operations.

5. Describe your security constraints on accessing data and computing.

Do your data include sensitive, personal information that require any special security measures? For example, working with patients' data requires privacy protection measures. Similar protection against data breaches could be applied also to computing resources, for example, by enforcing strong isolation between virtual machines, re-deploying applications instead of reusing the previously allocated ones, etc.

6. Describe your constraints and preferences regarding programing languages, applications and platforms.

Provide references to the tools your application uses. Are they freely available for download? Include information about licenses. Give some examples of how you normally use the tools. Explain your reasons for using or not using a particular framework or programing language to implement your application.

7. Provide examples of existing applications or projects that illustrate your needs.

Include links to the documentation of the applications and to the website of the projects. Describe the positive and negative features of the example applications that you provide.

3.2. Questionnaire for the identification of user communities

This questionnaire addressed four major points for the description of the target users and an additional point for the identification of experts in the field of each use case who can contribute with the analysis of the requirements. This information has been used to identify the target communities, as well as the user classes.

The questionnaire had five points which are described below, including the objective for each one. The answers and the analysis are included in Section 4.

1. Define your user classes and representative users.

From your application's point of view, what are the different user classes that you will support? Describe the roles played by each group that you define. An example of such a role is: a person who runs data processing, or a person who validates the data entered to the system, finds errors and repairs them. Note that a single user may play multiple roles.

2. Describe your users from a quantitative point of view.

How many users are in your community? How many of them are expected to use your application?

3. Which are the most representative IT tools of your domain?

Mention any Information Technology (IT) tool that is extensively used in your community (Web portals, online databases). For example, PubMed[6] and GenBank[7] are widely used in Bioinformatics.

4. What is the relation of your community with the new generation of Internet technologies and applications?

Is your community actively participating in the development of Internet-based applications? If so, list the most representative projects, the projects that are more close to you and to your application. Describe the experience of your community in the use of Cloud computing and other distributed computing infrastructures, such as Grid computing infrastructures.

5. Define your requirements decision-makers.

In your opinion, who can mediate over disagreements which may arise with regard to the requirements and ultimately decide upon using one approach over the other?

3.3. Requirements management tools

Requirements evolved from the user stories that were presented to the technical teams in meetings and teleconferences to the models presented in this document. However, this process will require a number of iterations over the cycle of gathering, analysing, improving and disseminating information about the requirements to maintain a good coherence between the requirements and the use case implementations. The following tools are available in the project to manage the life-cycle of the requirements:

- Requirements changes are documented through the issue tracking system of the project. This tool keeps information about the progress of the activity:
    - http://eubrazilcc-rm.i3m.upv.es/projects/eubrazilcc/issues
- Raw notes acquired during the elicitation phase are stored in the document management system. These notes include the questionnaires and other user-edited information that is maintained in its original form for consultation at any moment of the project:
    - http://eubrazilcc-rm.i3m.upv.es/projects/eubrazilcc/dmsf

3.4. Quality attributes

The process to validate a requirement depends on the type of goals users have set for that particular requirement. Functional requirements are usually validated by users themselves, who verify whether

[6] http://www.ncbi.nlm.nih.gov/pubmed/
[7] http://www.ncbi.nlm.nih.gov/genbank/

the implementation works as they expect. However, non-functional requirements are more difficult to validate since they usually refer to characteristics of the system that are prone to subjective opinion, such as how easy it is to use the system or how secure the system is. Therefore, the procedure for the analysis of requirements defines the attributes that describe the quality of the system, how they are measured in the project and what information is needed to write the validation tests that verify whether or not the requirement is met.

The following quality attributes will be considered in the project:

Availability

Availability measures the time during which the system's services are available for use and fully functional. In some cases, the system should recognize and prioritize the execution of critical tasks among all others. Availability will be measured in the project as the ratio of up-time to total operating time. An alternative way of measuring the availability will consider the ratio of the system uptime to the operational-total time that the system is expected to be up.

When availability is crucial, the use case must define a subset of the system functionality whose availability is most critical and the time periods for which availability is imperative. The latter is very important because users might be distributed throughout several time zones; at least Brazil (UTC-4, UTC-3 and UTC-2) and Europe (UTC, UTC+1 and UTC+2).

Installation cost

Installation cost measures how easy it is to correctly install, uninstall, and reinstall the system on the appropriate device or platform. It addresses the initial installation of the system in a plain operating-system, update of the installation to a new system version, installation of additional components or updates and complete uninstallation of the system. Installation cost will be measured in the project as the number of manual steps needed during installation and the mean time an untrained user needs to perform a successful initial installation of the system.

Interoperability

Interoperability measures how readily the system can exchange data and services with other systems. It will be measured as a feature that is present or absent in the system. For example, to verify that a system can be connected and exchange data with public research infrastructures, such as EGI, the system must be able to exchange authentication credentials with the remote system (VOMS proxy certificates in an EGI-like system), execute operations remotely and receive and process response messages sent from the remote system.

Performance

Performance measures the responsiveness of the system to user actions. It includes several indicators, such as response time and throughput. In general, performance will be measured in representative cases and the values of the performance indicators that are relevant to each case will be studied.

Security

Security measures the capability of the system to block unauthorized access to functions or data. This includes measures to protect the data that is stored in the system, transmitted over a network or processed in the system. In some cases, systems must be protected also against denial-of-service and other cyber-attacks. Security quantification will include a report on the features that are present or absent in the system, as well as the extent to which they are implemented (security criteria can be fulfilled only partially, depending on the specific needs of each use case). For example, it can include verification of compliance with security standards, use of appropriate encryption mechanisms and use

of mechanisms to protect the system against known vulnerabilities, such as Cross-Origin Resource Sharing (CORS) and Cross-Site Request Forgery (CSRF).
The use case must define the authorization or privilege levels and user access controls, and describe the security constraints for access of data and computing resources.

Scalability
Scalability measures the ability of the system to grow to accommodate more users, requests, data, network traffic and other services without compromising performance or correctness. It will be measured as the degradation of service performance due to an increase in the number of concurrent requests sent to the system.
When scalability is required, the use case must define the minimum acceptable performance criteria that must be satisfied regardless of the number of concurrent requests.

Usability
Usability measures the effort required to operate a system. In general, in the project we will evaluate that user interfaces must respond quickly to user commands and must have to be simple to use so that any member of the target community would be able to use them. In particular, usability will be measured as the maximum delay or wait time when performing a task, the maximum number of mouse clicks, keystrokes, touch-screen gestures required to accomplish a task, and the average time needed for a specific type of user to complete a particular task correctly.
When usability is crucial, the use case must include examples or guides demonstrating how to properly use the tools they will provide and will secure an appropriate number of users to participate in the usability study.

4. Basic identification of use case requirements

Each use case was first analysed separately to understand technical details, identify the actors involved and their roles. This initial analysis only involves the use case team members and WP5 coordinators. This section presents the information gathered per use case. The answers to the questionnaires are summarized in two tables, one with the results of the technical questionnaire and other with the answers to the community questionnaire. The analysis of this information led to the identification of functional and non-functional requirements, which are presented at the end of the use case subsection.

4.1. UC1 requirements

The LVL will provide a number of online collaborative and research tools for advancing the current knowledge about leishmaniasis. These tools must allow researchers to share the produced knowledge in an effective way. Data sources, computing systems to execute experiments, etc., must be seamless integrated in the LVL. From the point of view of an end-user, there must be only one entry point to the LVL and inputs and outputs should be consistent across all the LVL operations.
This imposes several challenges to the design of the LVL that are addressed in the following sections. Table 3 shows the information gathered from UC1 users in the elicitation phase.

Scope	User-provided information
User interface	Web-based access.
Data access	The LVL will use user-provided data and public data sources (GenBank, PubMed), and will produce data that must be protected from unauthorized access. It is expected that the data access volume required by a single execution does not exceed 100MB, although this value may be exceeded in specific uncommon cases. Standard data formats and access protocols are widely used and there exist open-source software libraries that can be used to access and manipulate the data.
Execution model	Experiments are modelled and executed as repeatable workflows. Users select a workflow template and customize input and output parameters.
Network traffic	Moderate network use is expected.
Security constraints	User data must be protected from unauthorized access. A user can grant access to other user or group of users.
External tools	None.
Relevant examples	Multilocus Sequence Typing project[8]. **Pros:** provides access to high-quality data. **Cons:** analysis tool is sometimes unavailable.

Table 3 – UC1 information gathered in the requirements elicitation survey

4.1.1. Target communities

The potential users of the LVL include researchers, international organizations, pharmaceutical companies and public sector bodies, such as government departments and public health authorities. The size of this community can be estimated from the number of attendants to the World Congress on Leishmaniasis, which is one of the most important scientific conferences on leishmaniasis and is held every 4 years. The last edition of this congress, celebrated on May 2013 in Brazil, received 555 researchers and 822 students from 50 different countries. Among the supporters there were leading pharmaceutical companies like Bayer and Sanofi, the World Health Organization (WHO), and the Drugs for Neglected Diseases initiative (DNDi), which is a leading (non-profit) drug research and development organization.

This community has a vast experience in the use of online databases and research tools, such as GenBank, PubMed and the Eukaryotic Pathogen Database Resources Centre (EuPathDB), which is an added value for the project, since this experience can be helpful to model, design and validate the use case.

4.1.1.1. User classes and representative users

[8] http://www.mlst.net

Scope	User-provided information	Representative users
Curator	A person who assesses the quality of the information stored with the LVL, and supports user activities by creating and maintaining pipeline templates. They will interact directly with the development team to solve the possible issues that might arise from the use of the system.	Fiocruz, ISCIII.
Researcher	A person who uses the LVL to search for information, customizes and executes pipelines in the LVL. They can upload custom data to the LVL, granting access permissions to other researcher or groups of them.	Fiocruz, ISCIII.
Public health professional	A person who uses the LVL to create reports of the data stored with the LVL for surveillance purposes. They usually will be interested in a specific geographic area, rather than the global information. Occasionally, they will execute a pipeline in the LVL, although this is not the expected behaviour.	Public health authorities.

Table 4 – UC1 user classes and representative users

4.1.2. Functional Requirements

- The LVL **must** allow the submission of new sequences: a user uploads a new biological sequence of *Leishmania* or sand fly and selects characteristics from a list of predefined characteristics. A new record is created in the LVL collection. The status of this new record is pending for approval. Users with curator role are notified.
- The LVL **must** allow sequence curation and versioning: a curator user lists the sequences pending from approval, edits the information of the record and updates the status of the record (accepted, rejected). When the edition introduces substantial changes in the sequence, a new version **must** be assigned and the status of the previous record **must** be updated to deprecate the previous version. Sequence submitter is notified in all cases.
- The LVL **must** control the access to the records in the LVL collection: sequence submitter decides who can access to the sequence (private or public access).
- The LVL **must** get sequences from *Leishmania* and sand fly collections: the system **must** be able to retrieve sequence records from the LVL collection as well as from existing collections and present them in an integrated manner to the user. The system **must** support at least interaction with ISCIII-WHO-CCL collection, COLFLEB and CLIOC. At least the following filters **must** be supported to search and display the information: country, clinical form, HIV-status, year of isolation/collection.
- The LVL **must** get new publications from PubMed and new sequences from GenBank: the system **must** periodically inspect these resources for new records on Leishmaniasis. Users with administrator role are notified when new records are found.
- The LVL **must** provide assisted data import: user with administrator role imports references from PubMed and GenBank. The system **must** at least assist in identifying the following fields: PubMed identifier, sequence accession number, geographical coordinates and/or country/region.

- The LVL **must** get DNA sequences from biological databases: the system **must** at least download sequences from GenBank, taking an accession number as input.
- The LVL **must** store and make available molecular analysis workflows: the system **must** at least store a file with the workflow definition, a description of the protocol and the information about the authors and the different versions of the workflow.
- The LVL **must** allow executing molecular workflows and access to the results: a user adjusts the parameters of a workflow and submits it for execution in a computing back-end, receiving a ticket from the system. Execution status as well as results can be retrieved with the ticket. These saved workflow parameters **must** be made available to a group of users for reuse and edition.
- The LVL **must** provide permanent references: LVL collection records, workflow definitions, experiments (workflow parameters) and results **must** be referable even from external sources (e.g. publications, scientific papers).
- The LVL **must** get occurrence records from biodiversity collections: the system **must** support at least interaction with speciesLink and COLFLEB.
- The LVL **must** allow geospatial analysis and visualization: the system **must** at least support search for and visualization in a map of PubMed references and GenBank sequences. At least the following operations **must** be supported: clustering & aggregation of geo-referenced data by area (country, shaped area) and simultaneous visualization of unrelated data (references, sequences, occurrence points).
- The LVL **must** allow users to comment and review the published information: a user opens a new thread in the social media tool (to be defined: wall, blog, forum) and writes a post about an item (e.g. collection record, workflow). Other users can subscribe to the thread to receive updates and provide their own input.
- The LVL **must** provide optimal viewing experience in mobile phones, tablets and computer monitors: the user interface **must** at least render with sufficient quality in the most common resolutions used in these devices.

4.1.3. Non-functional Requirements

It is likely that some components of the LVL will be installed or replicated outside of the EUBrazilCC infrastructure. For example, the ISCIII will deploy their own copy of the database to develop and test new pipelines without affecting LVL users. Therefore, there **must** be a documented procedure to correctly install the LVL and make it work in other operating environments.

The new pipelines will be integrated into the LVL, once they are fully developed and tested. Therefore, there **must** be a documented mechanism to extend and enhance the LVL with new pipelines.

A validation test will be written to validate these two requirements by measuring the **installation cost** of the LVL. A successful implementation will facilitate the automatic installation of the LVL with only three manual steps: a user (1) provides the version of the LVL or the URL address to download the LVL installer; (2) specifies the local directory under which to install the LVL; and (3) configures the local installation to match the user needs.

LVL is expected to be used as a tool for surveillance in rural areas and in developing countries. Therefore, the LVL **must** provide optimal loading time even in poor Internet connection conditions. In field conditions, it is probable that users will use a portable device to connect to the Internet. Therefore, the user interface **must** use device battery efficiently, disabling any unnecessary animation and other features that may consume battery. This will be validated in a **usability** study involving real users working with the LVL in field conditions.

Scalability is also critical for the LVL, in particular the ability to allocate additional computing resources off-premises to execute a large number of molecular analysis workflows in parallel, which is a common scenario in an outbreak of leishmaniasis. Similarly, a minimum **availability** will be required in these cases of high workload and molecular analysis services **must** be preserved by implementing additional measures, such as disabling non-critical functions, when the LVL is used in such conditions. **Interoperability** is also a key for accessing additional computing resources, such as the EGI federated cloud[9].

4.2. UC2 requirements

Use Case 2 will deliver an integrated environment for blood flow and heart simulation. This environment will leverage on the existing user interfaces of Alya Red System and ADAN to create and submit experiments for execution in High-performance computing (HPC) systems and cloud environments. In a typical experiment, a 3D model of a computational heart and a 1D model of the arterial network run coupled together. Two different scenarios are foreseen: direct solution and parameters estimation. In the first case, large-scale cases are solved. In the second one, a large number of problems are simulated on a coarse level to explore the space of certain parameters using optimization software such as Dakota[10].

After computing the Alya+ADAN coupled model, the ParaView open-source visualization tool will be used to gather the results from the storage resources and to combine them in a common image that is presented to the user.

An additional piece of software will be developed in the project to couple Alya Red with ADAN. This software will connect the input and output of both modelling tools to create an integrated model.

Table 5 shows the information gathered from UC2 users in the elicitation phase.

Scope	User-provided information
User interface	Experiments will be configured and submitted to the simulator through a Web-based or command-line interface. Simulation results will be visualized with the ParaView[11] open-source visualization tool.
Data access	Alya Red input is about 50MB in the average case. Most of this space is occupied by the geometry of the heart, which is provided by the user (for example, medical imaging of the heart), and the rest of the inputs are parameters taken from the literature. The same geometry can be reused in several experiments. ADAN input is about 3MB. Model output is approximately 4GB for Alya Red and 800MB for ADAN. In both cases, input data is provided in custom plain text file format. Output models are stored in HDF5 or EnSight[12] format.
Execution model	Both Alya Red and ADAN require HPC resources to reduce the model computation time. Alya Red uses MPI for parallelism. For a large-scale problem, it requires 2 hours to complete one cardiac cycle (500 are required in the average case) in BSC supercomputer Marenostrum (2,000 cores, 8TB of RAM memory). ADAN requires 10 minutes to complete one cardiac cycle in 1 core, 1GB of RAM processor.

[9] http://www.egi.eu/infrastructure/cloud/
[10] http://dakota.sandia.gov/software.html
[11] http://www.paraview.org/
[12] http://www.ceisoftware.com/

	The estimated execution time to compute an integrated model that couples Alya Red and ADAN is about 6 hours. Model visualization requires a GPU cluster. Visualization of computational heart requires 1 hour in 4096 CUDA cores, while visualization of arterial network model requires 1 hour in 1024 CUDA cores. In the case of the parameter estimation, coarser meshes are used. Then, Alya needs no more than ten cores, running the simulation in a few minutes. Dakota[10] the open-source tool for optimization and uncertainty analysis, provides the parameter's estimation schemes.
Network traffic	About a few tens of MB will be transmitted over the network between the coupling software, the modelling tools and the optimization software. However, this amount of data will be transmitted in many small packages. In the case of visualization, a fast network connection will be needed between the server and the client to remotely display the model.
Security constraints	Privacy protection is not needed at a first stage when animal models are used (e.g. rabbit heart). However, in the future, when patient data will be used to compute the models, some special security measures will be needed to guarantee patient privacy.
External tools	Alya Red runs on HPC facilities that support MPI. Both Alya Red and ADAN are tested in Linux environment. They are programmed in Fortran 90/95 and compiled with the Intel Fortran compiler. ADAN requires the open-source library for scientific computation PETSc[13], and the coupling software requires open-source numeric library GSL[14]. For visualization, ParaView[11] open-source visualization tool is required. Dakota[10] is the open-source optimization tool.
Relevant examples	None.

Table 5 – UC2 information gathered in the requirements elicitation survey

4.2.1. Target communities

The potential users of the integrated environment for blood flow and heart simulation are cardiologists and researchers who are dedicated to better understanding cardiovascular diseases. Although this is a very large community, the tools developed in the UC2 will be used by a small group of about 10 researchers before they can be used for cardiovascular research.

One of the main characteristics of the group of researchers who will use the UC2 is their strong relationship to IT. Therefore, UC2 will release a very simple user interface for the command line intended for advanced users. More comprehensive user interfaces will be developed in the future as part of the Hemodynamic Modelling Laboratory (HeMoLab)[15], which is an online tool for the simulation of the human cardiovascular system developed by LNCC, and the Computer Applications in Science and Engineering (CASE), which is a BigData visualization tool developed by BSC.

[13] http://www.mcs.anl.gov/petsc/

[14] http://www.gnu.org/software/gsl/

[15] http://hemolab.lncc.br/principal.php

4.2.1.1. User classes and representative users

Scope	User-provided information	Representative users
Developer	The team of developers who will develop the coupling software, and will extends Alya Red and ADAN in order to support additional features that are needed to address the integrated model.	BSC and LNCC.
Advanced finite element method (FEM) modeller	The team of developers who will use the integrated environment for blood flow and heart simulation to create models and to validate the output models.	BSC and LNCC.

Table 6 – UC2 user classes and representative users

4.2.2. Functional Requirements

- The integrated environment **must** allow running parameter sweeping studies for the Alya-ADAN coupled case, using a third-party code for driving the optimization process, Dakota being the preferred one.
- The integrated environment for blood flow and heart simulation **must** allow running Alya Red and ADAN in parallel. Alya and ADAN **should** simulate a coupled problem either with both codes in the same place or in different places, therefore using different computing facilities to create heart and blood flow models.
- The integrated environment **must** allow running parameter sweeping studies to pre-calibrate both Alya Red and ADAN with smaller settings before running large simulations in the HPC facilities.
- The optimizing tool **must** be capable of starting/stopping the simulation tools and retrieve their outputs through the Internet.
- When running in different places, the coupling software **must** synchronize Alya Red with ADAN executions through an Internet-capable communication protocol that **must** fulfil the restrictions imposed locally by the provider to access the HPC facilities.
- Model visualization **must** allow integrating both blood flow and heart models in the same view.
- Model visualization **must** allow integrating models stored in different (possibly distant) servers in the same view.
- The user interface **should** integrate the necessary tools to submit operations to both Alya Red and ADAN.

4.2.3. Non-functional Requirements

The simulation of the complete cardiovascular system is a computationally-intensive problem that demands the use of parallel computing. However, the problem of creating a 3D model of a computational heart or a 1D model of an arterial network involves many simulations that can be computed separately. Therefore, the underlying computing back-end **must** allow multiple instances of a model simulation (Alya Red or ADAN) to run in parallel in the same computing element (processor, server, etc.).

The visualization of the models produced in the UC2 requires combining information about the heart and the arterial network. However, this information is likely to be stored in distant servers, since in the project there are two different providers of HPC resources: BSC in Europe and LNCC in Brazil. Therefore, the visualization tool **should** be able to load the models from the remote sites in an optimal manner, providing the best possible user experience. This requirement combines **performance** and **usability**, since the goal is to load and visualize the model in the user environment causing the least possible delay, while providing the image quality needed to analyse the model.

Finally, medical images from patients **must** be protected from disclosure, enabling additional **security** measures to ensure data privacy both in the storage and in the transmission systems.

4.3. UC3 requirements

The UC3 will deliver a scientific gateway that integrates tools to study how climate change affects the biodiversity dynamics of terrestrial plants. Workflows will be produced that combine the analysis of data acquired with different technologies, such as LiDAR, hyper-spectral imagery, satellite images and ground level sensors, with meteorological and biodiversity data to study the impact of climate change in regions with high interest for biodiversity conservation, such as the Brazilian Amazon and the semi-arid & Caatinga regions in Brazil. The analysis of remote sensing images will provide 3D information concerning the structure of the vegetation, such as the biomass distribution within the forest canopy and forest gap density patterns, which should improve biodiversity indicators such as the energy balance and evapotranspiration.

Synergy of hyper-spectral imagery and LiDAR data can substantially improve 3D vegetation structure information, especially for environmental parameter extraction and biodiversity mapping/species definition. Therefore, imaging spectroscopy sensor data such as AVIRIS[16] will be used in those regions where ground level information is absent or incomplete.

The scientific gateway will connect biodiversity researchers to the workflows, allowing them to customize their experiments using different parameters and input datasets. Also, the gateway will provide tools to visualize output datasets and to perform spatiotemporal analysis.

The expected results of the UC3 include the analysis of historical time-series to discriminate the influences of human occupation and natural causes of biodiversity loss, such as climate variability, changes on energy fluxes and land cover.

Scope	User-provided information
User	Web-based access for end-users through a scientific gateway.

[16] http://aviris.jpl.nasa.gov/data/

interface	
Data access	Through the gateway, this use case will provide results originated from different data sources. 1) **Meteorological data** from land surface monitoring stations is publicly available (requires registration) from government agencies such as the Brazilian National Institute of Meteorology (INMET), Brazilian National Water Agency (ANA) and National Environmental Data Systems (Sinda). This data is provided as CSV files of a few KB size. Other data will be used, such as SRTM elevation data[17], which is stored in .hgt files (about 10MB per file) or other raster formats. 2) A second data source is the **hyper-spectral (satellite-based) imagery**. Such data contains a multitude of bands (e.g. 224 grid layers), whereby each cell stores particular absorption/reflection characteristics of the earth surface. Orbital data is publicly available from international agencies, such as the United States Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA). In this regard, Landsat represents the world's longest continuously acquired collection of space-based moderate-resolution land remote sensing data. Four decades of imagery provides a unique resource for those who work in agriculture, geology, forestry, regional planning, education, mapping, and global change research. Landsat images are also invaluable for emergency response and disaster relief. The Landsat project is an integral part of the Remote Sensing Missions component of the USGS Land Remote Sensing (LRS) Program. Remote sensing satellite data are stored as raster images mostly using the standard GeoTIFF format. A single unstaged image of Landsat satellite is about 400MB, while the size of pre-staged images is about 10MB. Analysis will require over a hundred images to cover a single region. Expected temporal and spatial resolutions are: LANDSAT - 30x30m (pixel) and 16 days image frequency; MODIS - 250 m (bands 1–2) 500 m (bands 3–7) 1000 m (bands 8–36) and 1 day image frequency. 3) Another relevant data source is represented by **LiDAR data**, which is provided as: (i) raw data, holding x, y, z point; and (ii) processed data, such as point clouds or interpolated raster data. The raw LiDAR data hold all original unfiltered data point clouds, either in unclassified or in classified format. Unclassified point clouds have not been classified into first/last returns. Classified point clouds contain classifications of ground (last return) and non-ground returns (first and other than last returns). Processed LiDAR data are based on the raw LiDAR data and have often been interpolated to grids. Such grids describe the topography of the terrain, that is: Digital Terrain Models (DTMs), or the surface of the vegetation, that is: Digital Surface Model (DSM). For the prototype application discrete LiDAR Data is available from an area near Manaus, Brazil, from the EMBRAPA (Brazilian Enterprise for Agricultural Research[18]) repository. LiDAR data is available as open source data or restricted licensed data, which requires signing an agreement to use licensed data for research/educational purposes. Download protocols to download LiDAR data from the NOAA'S Digital Coast services centre[19] are also well documented[20]. Usually – and similar to remote sensing imagery, LiDAR data can be downloaded via ftp and internet connections, using map

[17] http://www2.jpl.nasa.gov/srtm/
[18] https://www.embrapa.br/
[19] http://csc.noaa.gov/digitalcoast/
[20] http://www.academia.edu/4912612/Protocols_for_Downloading_and_Processing_LiDAR

	interfaces in which areas or metadata of data can be entered as search. The LiDAR data used in the project is accessible via secured protocol from the EMBRAPA repository[21] by entering user name/password. LiDAR-based data (point clouds) is delivered by the distributor usually in the uncompressed binary .las format or in compressed .laz format. LAS format is a published standard file format for the interchange of LiDAR data. It maintains specific information related to LiDAR data. It is a way for vendors and clients to interchange data and maintain all information specific to that data. E.g. ArcGIS 10.1 supports LAS versions 1.0, 1.1, 1.2 and 1.3. Older LiDAR data were provided in ASCII format. Processed, interpolated datasets are grids that are stored in .tiff format or any raster-based format that can be handled by GIS and/or remote sensing software such as ERDAS Imagine (.img), ENVI, ArcGIS (grid) and modelling software such as MATLAB. 1 LiDAR tile occupies up to 150 MB, multiple tiles may reach >10GB. The free dataset of the Netherlands (AHN2) covers 4TB of compressed raw LiDAR data. Brazilian country wide LiDAR data is expected to occupy several times this amount in the future. 4) **Biodiversity data** will be provided principally by the speciesLink network[22], the List of Species of the Brazilian Flora[23] and the Global Biodiversity Information Facility (GBIF). speciesLink and GBIF data is publicly available on the Web, while the list of plants of Brazil (version 2012) will be provided by CRIA. speciesLink provides web-services for researchers to access the data using the TAPIR protocol[24]. The data is provided in the Darwin Core file format[25], which is a standard to facilitate the exchange of species occurrence data. The GBIF occurrence web-service provides a range of filters for selecting occurrence records. The currently supported data formats include Darwin Core and KML (for use with Google Earth). The information to access biodiversity resources is maintained in the Biodiversity Catalogue[26]. With regard to the availability of biodiversity data in the selected regions, from the ~18,000 species of flowering plants that occur in the Amazon biome, 4,250 are endemic species and 331 of them have at least one occurrence point within the Adolfo Ducke Forest Reserve. Similarly, 9,526 species of plants were registered as potential targets for the analysis of the semi-arid region. 5) In terms of **climate data**, the CMIP5 Federated Data Archive will represent a comprehensive and relevant “data source”. The CMIP5 federated data archive collects 61 global climate models from 29 different modelling groups with a total amount of about 2 petabytes of datasets that can be accessed via any one of distributed data nodes of the Earth System Grid Federation, the official site for CMIP5 outputs. Specifically, CMIP5 promotes a standard set of model simulations in order to (i) evaluate how realistic the models are in simulating the recent past; (ii) provide projections of future climate change; and (iii) understand some of the factors responsible for differences in model projections including quantifying some key feedbacks such as those involving clouds and the carbon cycle. CMIP5 also provides a large number of complex models running at high resolution, with a complete representation of external forces and different types of scenario. In the EUBrazilCC project, climate change indicators on the targeted areas will be evaluated starting from

[21] http://repositorio.cnpm.embrapa.br/sl_download/JAM_A02_2013_LiDAR/
[22] http://splink.cria.org.br/
[23] http://floradobrasil.jbrj.gov.br/
[24] http://www.tdwg.org/dav/subgroups/tapir/1.0/docs/tdwg_tapir_specification_2010-05-05.htm
[25] http://www.tdwg.org/activities/darwincore
[26] https://www.biodiversitycatalogue.org

	the data available through this federated data archive.
Execution model	CPU use depends on the resolution of the LiDAR data (point density per square meter). No parallel optimization is currently used. Approximately 50 concurrent users are expected. Processing one single satellite image using one processor will cost several hours, but only 200MB of RAM memory are required. This process can run in parallel, although no parallel algorithm is currently available. The pipeline on remote sensing data (e.g Landsat) that calculates the evapotranspiration needs to be executed every 16 days (when new data arrives) and takes approximately 700 CPU hours. This is highly parallelizable and can run on a small private cloud in a few hours. However, our rough estimation is that running this procedure for the last 30 years of available data will require approximately 500,000 CPU hours, which will require a large-scale infrastructure. In addition, some processing will need to be done with PDAS to extract statistical information from the data generated by this background pipeline. The computation of analytics workflows will strongly depend on the input size and the associated workflow complexity (from a single operator to tens of operators in the same workflow experiment). In this regard, the PDAS will be exploited in two different modes: (i) batch, for analysing input data and extract a distilled set of information to be stored in clearinghouse systems for further exploitation (as expected for the climate indicators); and (ii) online, for end users interactive data analytics tasks from the scientific gateway (as expected for climate and remote sensing data).
Network traffic	User Internet connection is a potential bottleneck for performance, especially low bandwidths. However, the proposed approach relies on “server-side” processing & analytics, where users are not expected to retrieve large amounts of data (e.g. raw data) from the gateway. Only the final results of an analysis (e.g., images, maps, reports and summaries, typically megabytes or even kilobytes) will be downloaded. Such an approach will drastically limit the traffic of data over the network, making possible the service and data exploitation even from low-bandwidth connections.
Security constraints	User registration is required for accounting purposes.
External tools	Geo-processing algorithms from GDAL[27] will be used. Testing of data will make use of commercial GIS (ArcGIS) and remote sensing software (ERDAS) or open GIS software (Grass) and Lastools, which are common tools used in LiDAR pre-processing, processing and visualization, and can be run as extensions of ArcGIS software or as a stand-alone executable program[28]. The BEAM open-source toolbox will be used to process remote sensing raster data[29]. openModeller[30] will be used to carry out ecological niche modelling experiments, specifically the EUBrazilOpenBio openModeller Web Service 2.0[31] that relies on cloud resources offered by the EGI federated cloud and other openModeller providers, such as CRIA. The algorithms for LiDAR point cloud processing are currently in MATLAB scripting and in ArcGIS.

[27] http://www.gdal.org/

[28] http://rapidlasso.com/lastools/

[29] http://www.brockmann-consult.de/cms/web/beam/

[30] http://openmodeller.sourceforge.net/

[31] http://enm.eubrazilopenbio.d4science.org/om/omws2

	In terms of data analysis, the PDAS will incorporate new core functions (array-based primitives) from a number of scientific numerical libraries or command line tools (GSL, PetsC, etc.).
Relevant examples	LiDAR online[32] provides an example of platform to market LiDAR, GIS data on the Internet in an intuitive user interface. Such an approach provides an intuitive map-based search window for the end-user, and easy to find products under the Geo-Services tab > forestry. However, it is difficult to add data or find metadata of input products, and tutorials are unclear and not intuitive, especially for non-specialists. NASA LiDAR Access System (NLAS) provides an example of the future role of an application to serve world-wide products[33]. NDVI changes (Esri)[34] allows time series analysis with a user-friendly interface. However, these products for analysis are limited: multi-temporal analysis cannot be customized, the user interface does not allow downloading the product and overlapping layers as vector type files. Series View (LAF/INPE)[35] allows downloading graphs and specific-location data. However, it provides a small number of products for analysis and it does not provide the images used in the application. The Brazilian Semiarid NDVI Viewer (INSA/UFAL)[36] gives the possibility of running temporal data analysis of socioeconomic data. On the other hand it provides a low number of land surface information: it only uses one vegetation index and does not allow crossing of socioeconomic and environmental information. World Evapotranspiration Web Viewer (Esri)[37] provides an important environmental variable (evapotranspiration) for all land surfaces. However, the information provided is summarized at watershed scale. It does not allow the user to obtain more detailed information on higher resolutions. The information on evapotranspiration refers to annual totals, being thus of limited applicability for environmental studies. It is not possible to download the data used in the application. Concerning climate data, the most relevant production-level tools for data analysis are Climate Data Operators (CDO) and NetCDF Operators (NCO). In both cases they are client-side, sequential and disk-based. These are key limiting factors to tackle (near) real-time data analytics. Other software products like Live Acess Server (LAS) have a server-side support in terms of data visualization, but they rely nevertheless on a sequential analytics back-end.

Table 7 – UC3 information gathered in the requirements elicitation survey

4.3.1. Target communities

The user community for LiDAR and GIS technologies in general is rapidly growing upon increased availability of data. Examples of users are governmental agencies (both research & education) and

[32] http://www.lidar-online.com/tools/maps/
[33] http://www.opentopography.org/index.php/about/nlas
[34] http://changematters.esri.com/compare
[35] https://www.dsr.inpe.br/laf/series/index.php
[36] http://www.insa.gov.br/ndvi
[37] http://www.arcgis.com/apps/OnePane/main/index.html?appid=48268eba0f414713be00f75ac3289bb4

commercial companies. Non-scientific users come from the field of forestry, earth science, biology, environmental conservation, and others.

The LiDAR geo-community consists of hundreds of research institutions and companies and many thousands of individual and/or group users, and is growing. The number of scientific institutions potentially interested to use the applications will be >10%. If the gateway is extended to global datasets, its use and citation potential could grow rapidly.

In general, the LiDAR community is actively playing part in Internet-based applications, and there are several successful experiences on the use of online databases and research tools, such as the LiDAR online data and geoservices[38], the National Dutch LiDAR data AHN2[39], and the NASA LiDAR Access System (NLAS)[40].

The gateway will be also useful for other professionals of many fields, such as hydrologists, agriculturalists, climatologists, environmentalists and remote sensing professionals, who will use the services provided in the gateway for their labour. The gateway will help them in making policies to mitigate land degradation and desertification, and to develop conservation projects. More than 100 users with this profile are estimated to become frequent users of the gateway.

4.3.1.1. User classes and representative users

Scope	User-provided information	Representative users
Biodiversity researcher	A person who uses the scientific gateway to execute experiments and to analyse the results produced.	CRIA
Other professionals	Governmental agencies, private companies and non-governmental organizations (NGOs) that are interested in the use and diffusion of the results.	Food and Agriculture Organization (FAO) of the United Nations (UN)
Developer	The team of developers who will extends the geo-processing toolset, prepare newly available data for use in the gateway and adapt the tools for the infrastructure.	CMCC, UvA

Table 8 – UC3 user classes and representative users

4.3.2. Functional Requirements

- The gateway **must** integrate satellite images data sources, at least from Landsat and AVIRIS.
- The gateway **must** integrate meteorological/climate data sources, at least from INMET, ANA and Sinda.
- The gateway **must** integrate biodiversity data, at least from speciesLink, the list of plants of Brazil and GBIF.

[38] http://www.lidar-online.com/tools/maps/
[39] https://www.pdok.nl/
[40] http://www.opentopography.org/index.php/about/nlas

- Metadata **must** be included into the gateway to describe the area covered, the year of acquisition of the data, the type/format of the data and any other technical specification that is necessary. Metadata **should** link up to existing standards and directives (e.g. INSPIRE initiative[41]).
- The gateway **must** allow the processing of satellite images series using the SEBAL algorithm.
- The gateway **must** provide geo-processing tools/algorithms to calculate 3D vegetation products for geographical regions.
- The gateway **must** store 3D vegetation products. Examples of such products are DTMs, DSMs, CHMs, tree top locations, tree top heights, biomass and related forest products.
- The gateway **must** allow the processing of large-scale dataset series, at least the processing of the Brazilian Semiarid region (evapotranspiration series over 30 years and current day onward).
- The gateway **must** support the user analysing the output of the SEBAL workflow in terms of (i) time series analysis, (ii) data reduction (e.g. by aggregation), (iii) data subsetting, and (iv) data transformation.
- The gateway **must** present and visualise the key output of the use case in different ways: maps, graphs, tables, and comparative charts. Export of results **should** be also supported (CSV, JPEG formats).
- The gateway **must** provide access and visualization support regarding the indicators stored into the clearinghouse system both for the climate and biodiversity part.
- Download of aggregated results and products regarding satellite images series will be also supported.
- The user interface **must** facilitate the end-user to select the data sources, temporal and spatial scales and output format for historical time-series analysis.
- The user interface **must** facilitate the end-user to select an area of interest (e.g. the Duc Reserve near Manaus, Brazil), for that area the available data (e.g. LiDAR) **must** be retrieved from the gateway and the end-user **must** have the possibility through the web-interface to select derived 3D vegetation products for that area.
- The user interface **must** make available the 3D vegetation products for the end-user. Users **must** be able to visualise/retrieve the processed products from the LiDAR data via the web-interface.
- The gateway **should** also allow programmatic access to the workflow processing facilities, providing a Web service API to access the operations and retrieve results.
- The gateway **should** provide extensible interfaces to the workflow processing facilities. At least, it **should** provide a mechanism to add new data sources and algorithms in the future. Upon newly available (LiDAR/hyper-spectral) data, new derived products **could** be added to the infrastructure and made available to the end-user via the web-interface.
- Besides user classes, the gateway **should** distinguish the following groups of users with different access and their roles:
    - Administrator or system manager: full control access over all users, user groups and data.
    - Data administrator: managing database.

[41] http://inspire.ec.europa.eu/

- Scientific administrator: managing geo-processing tools, ensuring data accuracy.
- End-users and possible sub groups of end-users (research/education/individual).

4.3.3. Non-functional Requirements

The analysis of remote sensor data is a data-intensive problem. The UC3 will develop a pilot to analyse specific regions of Brazil, such as the Duc Reserve. Future developments of the gateway **must** be able to host global products from satellite sensors. This requirement targets the **scalability** of the gateway to accommodate larger data analysis that covers a wide area, such as the entire territory of Brazil, and also to support the analysis of historical time series datasets. To validate this requirement a larger dataset (scene) **must** be created from real data or simulations, which resembles the real data volume expected in a wide analysis.

The user interface needs displaying many different data using maps and tables. This **should** be approached by prototyping and iterating over the user interface several times before the final version of the gateway is released to the community. This requirement targets the **usability** of the gateway.

A distinctive feature of the UC3 is that end-users are mostly interested in the final products rather than on raw data. Therefore, the access to/visualization of products from the web-interface **must** be fast. This requirement targets the **performance** of the gateway to post-process and display the products.

5. Consolidated analysis and abstraction of general requirements

This section presents a consolidated analysis of the use case requirements. From this analysis, a set of general functional requirements have been identified to guide the design of the EUBrazilCC infrastructure. These requirements will be prioritised and all of them are expected to be enforced in the final infrastructure. This document sets the starting point for the discussion and common understanding of the infrastructure needs and capabilities.

The requirements have been classified into three types: requirements for data access, requirements for execution, and requirements for security and logging.

All the use cases will contribute to the objective of creating a federation of heterogeneous infrastructures for computing and data resources. UC1 and, to a lesser degree UC3, will contribute to the objective of providing intensive computing using federated computing resources. UC1 and UC3 will execute complex workflows on a wide geographic area using the EUBrazilCC federated infrastructure. Finally, UC3 and, to a lesser degree UC1, will process massive amounts of data.

5.1. General requirements for data access

Access to large-scale data storage and data processing facilities is the most important requirement of the project. As open-access databases continue to grow in scale and more information becomes available, scientists are finding more ways in which this information can be exploited to develop the new generation of scientific tools and applications.

For example, UC1 leverages on the increasing availability of DNA sequence data to develop new pipelines for the study of leishmaniasis propagation. These pipelines will use information filtered by geospatial proximity. DNA sequence databases are beginning to realize the importance of georeferencing their sequences, and major providers, such as GenBank, are beginning to annotate their entries with the location where the DNA sample was collected. However, this information is still insufficient (less than 5% of phlebotome DNA sequences are geo-referenced in GenBank). Therefore, UC1 will rely on other publicly available information, such as PubMed references, to geo-reference DNA sequences. However, this will require analysing a huge amount of data that grows and evolves every day.

UC3 faces an even bigger problem, since ground-level sensors and satellite images are becoming cheaper and widespread, and the number of available data is exponentially growing, at the same time that new high-resolution acquisition methods are also emerging. Although currently this information is not available for several of the most biodiversity-rich areas, it's expected that more information become available in the next years.

The comprehensive analysis of the use case requirements yielded the following global requirements for data access:

- The infrastructure **must** support the integration of external data from existing data sources. This integration **must** be complemented with methods for referencing the data in their original locations, and to pre-process and annotate the data with additional information. Metadata standards **must** be used when available to annotate the data. Also, automatic synchronization with original data sources **must** be addressed (updating the infrastructure with the latest releases of the data), considering the individual needs of each case, which range from simply discovering and downloading new data when it becomes available, to running complex data pre-pre-processing before storing the data in the infrastructure.
- The infrastructure **must** store the data processing products, taking the necessary steps to ensure data persistence and data protection, when necessary.
- The infrastructure **must** provide access to authorized applications to access and process the data, supporting the application data processing model.
- The infrastructure **must** facilitate the end-user to access the data, providing the most appropriate protocols and data formats to enable developers with the necessary means to build usable user interfaces.
- User Internet connection is a potential bottleneck for performance, especially low bandwidths as expected in field conditions. Therefore, the infrastructure **should** facilitate the access to the data even in poor Internet connections.

5.2. General requirements for execution

Computing requirements are very heterogeneous in the project. UC1 and UC3 are data-intensive Web applications, with well-defined workflows that can run concurrently using the computing resources available in the infrastructure. UC1 can have workload peaks due to a disease outbreak, while UC3 is more predictable with regard to the use of computing resources. However, the integration of new sensor data sources into the infrastructure can causes an overhead in the services of the infrastructure.

To cope with this, advanced resource provisioning mechanisms should be provided with the infrastructure to enable applications to adapt to capacity and performance requirements as the volume of available data and workload changes over time. Those mechanisms must be supported widely in the infrastructure, allowing both the programming frameworks (like e-SC and COMPSs) and the execution environments (like CSGrid and PDAS) to negotiate with the resource providers regarding the capacities required to fulfil the application needs. Therefore, there will be necessary to modify the existing systems in order to interact with the services responsible for managing the infrastructure resources, for example, to create new virtual infrastructures on-demand or to release unneeded resources.

On the other hand, UC2 targets a computationally-intensive problem that demands the use of HPC resources both to run model simulations and to achieve data visualization. Therefore, UC2 requires specialized facilities, such as supercomputers and clusters of computers. Access to HPC resources generally requires to adapt the application to run on a specific HPC resource, compiling the source code for the target processors and adjusting network parameters to proper values. Nevertheless, the added value of running the UC2 models on a federated platform is that new computing resources, apart from those required to create the high-resolution models, can be allocated for optimizing the input parameters, potentially reducing the time required to compute a high-resolution model using an HPC facility. This experience could be helpful in improving other HPC applications.

This analysis has reached several requirements on the need of computing resources:

- The infrastructure **must** support two different execution models: (1) High-throughput computing (HTC); and (2) High-performance computing (HPC).
- The infrastructure **must** be self-adapting in order to accommodate the workload peaks that can appear in HTC applications.
- The infrastructure **must** support the execution of big data workflows, where the input data and the products can be large (in the order of tens of GBs) and they can be stored on servers geographically distant from each other.

5.3. General requirements for security and logging

In general, there are no strong requirements for security, since the data will be made available to the community openly. UC1 and UC3 will require user registration for accounting purposes and to support group-level policies to access the data.

On the other hand, UC2 will require privacy protection when patient data is used to compute the models. However, models will be computed using HPC resources, so this requirement has to be met by the HPC provider.

Major security requirements are expected from other project activities that deal with interoperability. This fact might cause the use cases to adopt a specific security mechanism in order to successfully interoperate with other major cloud computing provider, such as the EGI federated cloud.

- The infrastructure **must** support end-user authentication for access control and accounting purposes.
- The infrastructure **must** support end-user authorization for accessing the data and the applications deployed with the infrastructure.

- The infrastructure **must** support group-level policies to access the data.
- Data privacy protection **must** be ensured in the HPC resources where the UC2 will run.

6. Early prototypes

Although the main activities of implementation of the use cases will begin in the coming months, some progress has been made in the design of the architectures in this phase of the project. The three use cases heavily rely on the development of early prototypes to analyse their requirements. Each prototype provides a different level of detail, but they all share the common objective of providing the use case team with a joint vision of the final product.

The prototypes will be developed in the next months and will be released in PM9, which is the main objective of the milestone MS51.

6.1. Leishmaniasis Virtual Laboratory

Figure 2 shows the architecture of the LVL prototype. The UC1 team has focussed on identifying the different components of the LVL, as well as the technologies and standards that will support the operations of the LVL. Virtualization is essential to facilitate the installation of the LVL in other operating environments, as well as to improve LVL availability and performance.

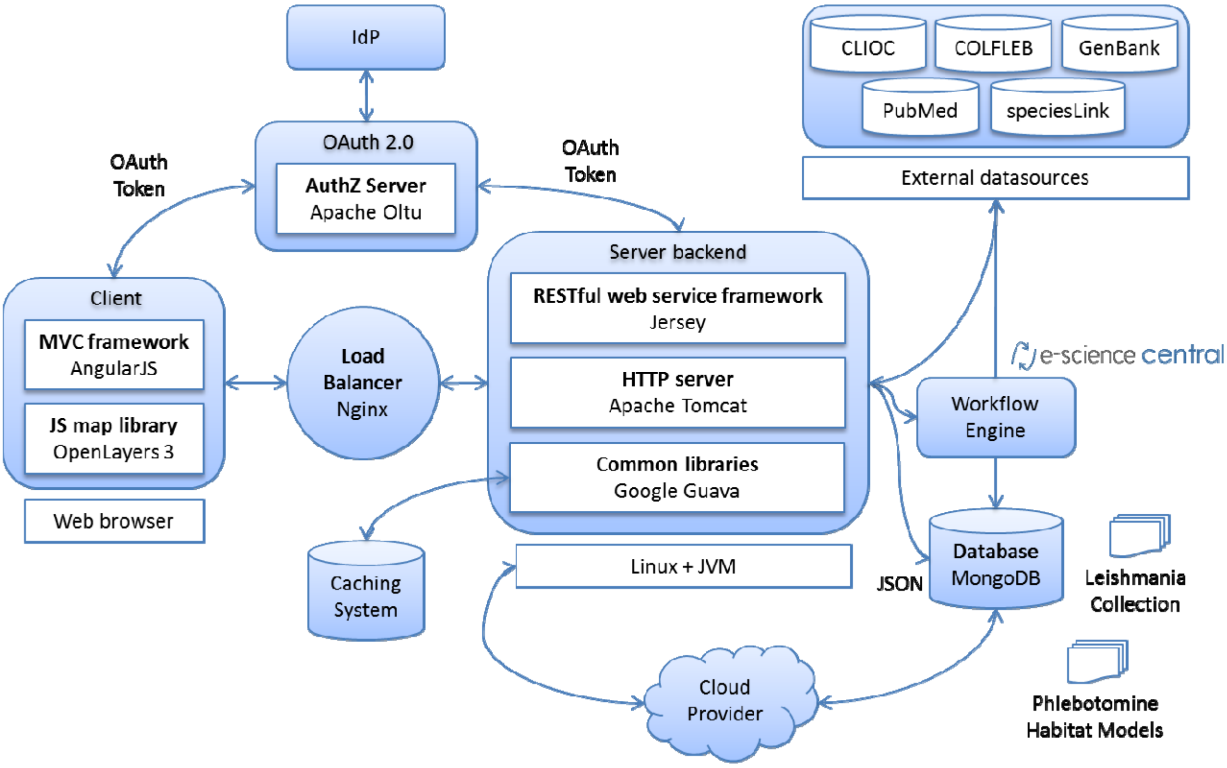

Figure 2 – UC1 prototype architecture

The LVL team has identified the following components in the proposed architecture:

- An authorization (AuthZ) server will provide the necessary authorization and access control mechanisms to access the data and the operations in the LVL. This server will rely on the standard OAuth 2.0[42,43], specifically in the open-source implementation provided by Apache Oltu[44]. The AuthZ server will rely on the identity provider (IdP) deployed in the EUBrazilCC infrastructure to establish end-user identity.
- The core of the LVL is the database that stores the *Leishmania* collection and other related information, such as the phlebotomine habitat models. This database will be stored in the EUBrazilCC infrastructure, in principle using a MongoDB[45] database optimized for high-throughput access from the workflows. This database will be designed to facilitate deployment and replication across a cloud computing infrastructure.
- E-science central will support the workflow execution.
- A LVL server back-end will orchestrate all the LVL operations, exposing a RESTful interface. This server will include an HTTP server and a caching system. It will be designed to deploy on top of a Linux-based operating system with a minimum number of software requirements, and will preferably depend only on the JVM. This will facilitate the deployment of several instances of the LVL server across a cloud computing infrastructure. Each instance will be autonomous and will operate independently of the other instances. A workload balancer will distribute the requests among the available replicas of the LVL server.
- A browser-based client completes the LVL. This client will rely on HTML5 and JavaScript to connect end-users to the LVL server. A responsive web design will be applied with the help of a MVC JavaScript framework to support mobile devices and desktops. OpenLayers 3[46] will be used to render and display maps in the client side. This open-source technology is preferred over other proprietary map APIs, such as Google Maps, to permit the use of the LVL to institutions that restrict the use of proprietary software due to privacy or other legal concerns.

6.2. Integrated environment for blood flow and heart simulation

Figure 3 shows the architecture of the integrated environment for blood flow and heart simulation. The main challenge of the UC2 is the coupling of Alya Red and ADAN modellers, transmitting information over the network and across organization boundaries, while using HPC resources that are available in the infrastructure.

[42] http://oauth.net/2/
[43] http://tools.ietf.org/html/rfc6749
[44] http://oltu.apache.org/
[45] http://www.mongodb.org/
[46] http://ol3js.org/

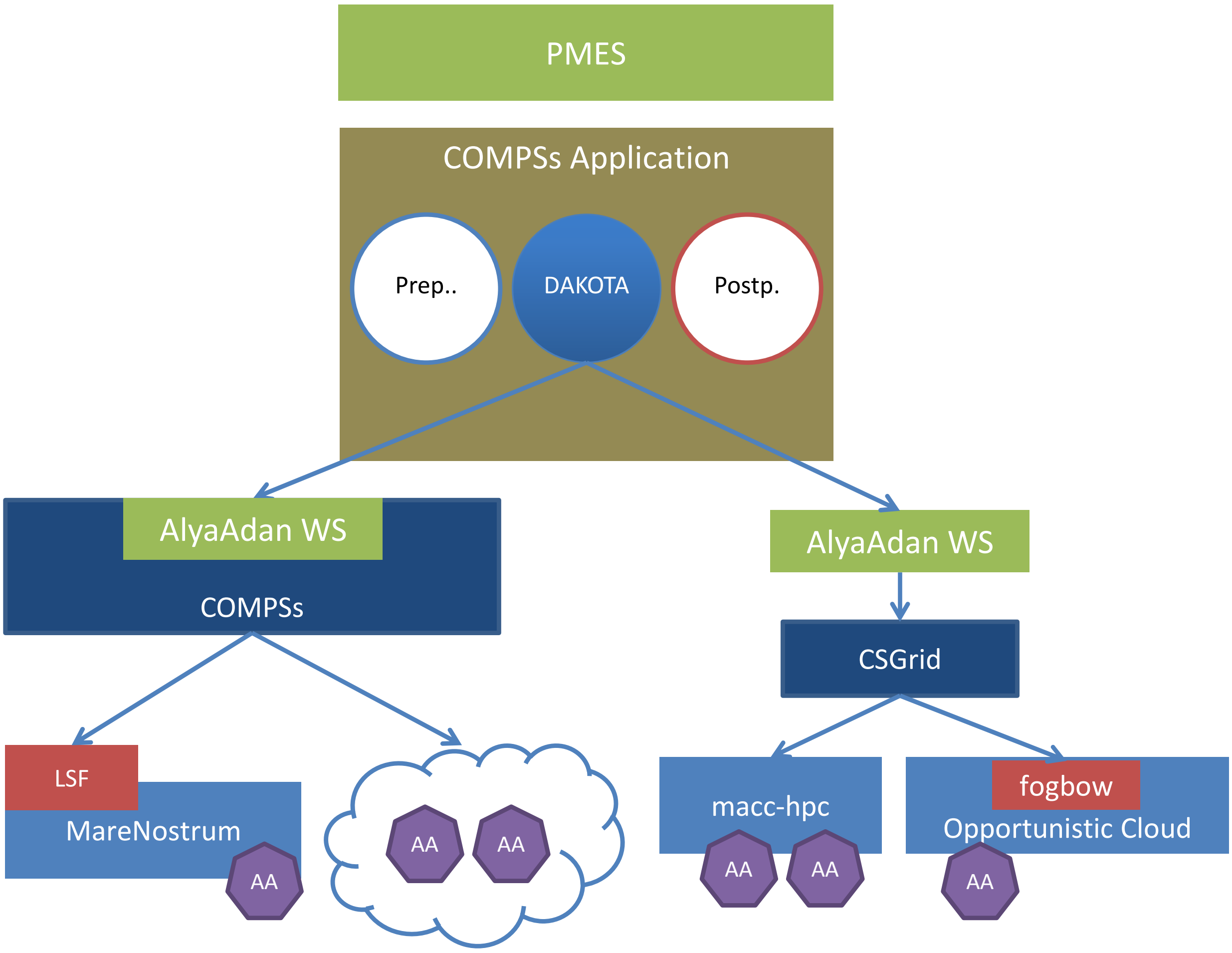

Figure 3 – UC2 prototype architecture

User provides input parameters through mc2 (not shown in Figure 3), which dispatches COMPSs application through the PMES providing it with the input parameters. The COMPSs application is composed of: 1) a pre-processing phase to prepare the data for the modelling software; 2) a DAKOTA task that generates the parametric execution requests; and 3) a post-processing phase where other additional tasks are performed with the output models, for example, to prepare the models for visualization. The executions of Alya+ADAN (AA) are performed through a web service that can be implemented with COMPSs or another enactor –that has not yet been developed–, which interacts with CSGrid. This web service will be deployed in the EUBrazilCC infrastructure that provides the execution on a given back-end. The same service can be used to schedule large-scale coupled runs of Alya+ADAN.

6.3. Scientific gateway to study climate change effect on biodiversity

Figure 4 shows the architecture of the UC3 prototype. The major challenge in UC3 is to provide fast access to a huge amount of data. Therefore, the focus of the prototype is to facilitate the design of the data access services.

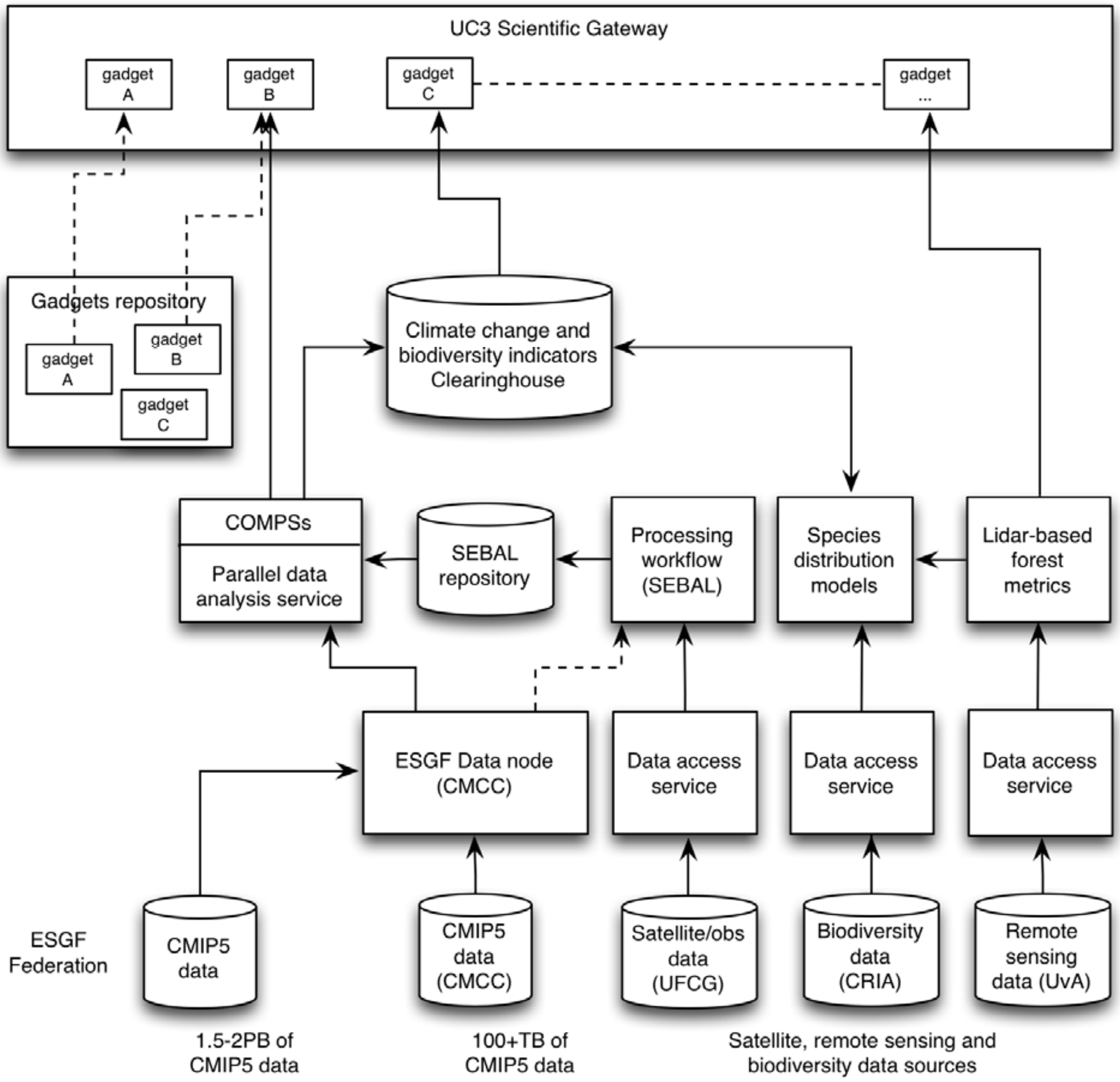

Figure 4 – UC3 prototype architecture

The bottom side of Figure 4 shows the different external data sources that will be integrated in the UC3. They will be accessed from the workflows and processed with the appropriate tool (e.g. PDAS, EUBrazilOpenBio ENM service). The execution of these workflows will be supported by the infrastructure, and the products will be stored into the clearinghouse database.

The top side of Figure 4 shows the Web interface of the scientific gateway, which will be developed as a collection of user interface components (gadgets) that will be integrated in a Web portal. Gadgets will connect users to the products stored in the infrastructure, and will allow them to configure and run processing and analytics workflows.

7. Final considerations

This section presents additional information that arises from the analysis of the use cases that are not strictly speaking part of the requirements, but are important for the implementation of the use cases.

7.1. Software licenses

Software developed in the project that implements the use cases will be released under the European Union Public Licence (EUPL)[47]. In case any of the partners cannot could not use the EUPL, then the software will be released under a dual license consisting of the EUPL and the Apache License 2.0[48].

7.2. Version control

The EUBrazilCC organization was registered in GitHub to facilitate community involvement in the development and validation of the use cases: https://github.com/eubrazilcc. Each use case will create a repository within this organization for managing and distributing its own source code.

8. ACRONYMS AND ABBREVIATIONS

Acronym	Definition
AOB	Any Other Business
CA	Consortium Agreement
CNPq	National Council for Scientific and Technological Development of Brazil
CooA	Coordination Agreement
DoW	Description of Work
e-SC	e-Science Central
EC	European Commission
EEC	External Expert Committee
EWS	Early Warning Systems
GA	EC Grant Agreement
GIS	Geographic Information System

[47] https://joinup.ec.europa.eu/software/page/eupl
[48] http://www.apache.org/licenses/LICENSE-2.0.html

KPI	Key Performance Indicator
LiDAR	Light Detection And Ranging
LVL	Leishmaniasis Virtual Laboratory
MCT	Ministry of Science and Technology
NGS	Next-Generation Sequencing
OBIA	Object Based Image Analysis
PEB	Project Executive Board
PM	Person Month
PMB	Project Management Board
TSC	Technical Steering Committee
UC	User Committee
WHO	World Health Organization
WP	Work Package
WPL	Work Package Leader