INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.*P0*: SELECTED CASE STUDY FROM *TRUFA*

(TRANSCRIPTOMES USER-FRIENDLY ANALYSIS)

INPUT TO EU DELIVERABLE: D 2.1

Document identifier:	INDIGO-WP2-D2.1-ANNEX-1P0-TRUFA-V10
Date:	**16/06/2015**
Activity:	**WP2**
Lead Partner:	**EGI.eu**
Document Status:	**DRAFT**
Dissemination Level:	**CONFIDENTIAL (INTERNAL)**
Document Link:	

Abstract

This report summarizes the findings of T2.1 and T2.2 **for partner CSIC** along the first three months of the project. It is an integrated document including a general description of the research communities involved and the selected Case Studies proposed, in order to prepare deliverable D2.1, where the requirements captured will be prioritized and grouped by technical areas (Cloud, HPC, Grid, Data management) etc. The report includes an analysis of DMP (Data Management Plans) and data lifecycle documentation aiming to identify synergies and gaps among different communities.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/Activity	Date
From	Fernando Aguilar, CSIC	**CSIC P0**/WP2	16-jun 2015
Reviewed by	**Moderators:** P.Solagna, F.Aguilar, J.Marco **Internal Reviewers:** <<To be completed by project office on submission to PMB>>		
Approved by	**PMB** <<To be completed by project office (no submission)>>		

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	5-may-2015	First draft, v01	J.Marco, F.Aguilar CSIC
9	15-june-2015	Draft revised	F.Aguilar, CSIC
10	16-june-2015	Draft to be circulated for internal review, v10	F.Aguilar, CSIC
11	20-june-2015	Comments included, version for release v11	P.Solagna, EGI.eu

TABLE OF CONTENTS

INTRODUCTION AND CONVENTIONS

PLEASE, READ CAREFULLY BEFORE COMPLETING THE ANNEX:

*This Annex is an example of compilation of the information needed to support adequately a **Case Study** of interest in a Research Community. Each partner in INDIGO WP2 is expected to provide such information along the first three months of the project (i.e. by June 2015), and it will be used to compile Deliverable D2.1 on Initial Requirements from Research Communities.*

There will be around 10 Annexes, for example Annex 1.P1 for partner 1 in WP2 (i.e. UPV), will cover Case Studies from EuroBioImaging research community.

The initial version will be discussed with INDIGO Architectural team to agree on a list of requirements.

Some relevant definitions:

***A Case Study** is an implementation of a research method involving an up-close, in-depth, and detailed examination of a subject of study (the case), as well as its related contextual conditions.*

We should focus on Case Studies that are representative both of the research challenge and complexity but also of the possibilities offered by INDIGO-DataCloud solutions on it!

*The Case Study will be based on a set of User Stories, i.e. how the researcher describes the steps to solve each part of the problem addressed. **User Stories** are the starting point of **Use Cases**, where they are transformed into a description using software engineering terms (like the actors, scenario, preconditions, etc). Use Cases are useful to capture the Requirements that will be handled by the INDIGO software developed in JRA workpackages, and tracked by the Backlog system from the OpenProject tool.*

The User Stories are built by interacting with the users, and a good way is to do it in three steps (CCC): Card, Conversation and Confirmation[1]*.*

Use Cases can benefit from tools like "mock-up" systems where the user can describe virtually the set of actions that implement the User Story (i.e. by clicking or similar on a graphical tool).

Different parts of this document should be completed with the help/input of different people:

RESEARCH MANAGERS

-Section 1, SUMMARY, is to be reviewed/agreed with them as much as possible

RESEARCHERS

*-Section 2, INTRODUCTION is designed to be filled with direct input from (senior) researchers describing the interest of the application, and written in such a way that it can be included in related technical papers. It is likely that such introduction is already available for some communities (for example, for several research communities in WP2 like DARIAH, CTA,EMSO, Structural Biology, one may start from the **Compendium of e-Infrastructure requirements for the digital ERA**[2] **from EGI***

APPLICATION DEVELOPERS AND INTEGRATORS WITHIN THE RESEARCH COMMUNITIES

-Sections 3, 4, 5, 6: should be discussed from their technical point of view (including data management as much as possible).

MIDDLEWARE DEVELOPERS AND E-INFRASTRUCTURE MANAGERS

-Sections 7, 8: should be discussed with them

[1] For a nice intro, see: https://whyarerequirementssohard.wordpress.com/2013/10/08/when-to-use-user-stories-use-cases-and-ieee-830-part-1/ , and also https://whyarerequirementssohard.wordpress.com/2015/02/12/how-do-we-write-good-user-stories/ etc.

[2] *https://documents.egi.eu/public/ShowDocument?docid=2480*

The logical order to fill the sections is: 2,3,4,5,6,1,7,8. Sections 1 and 8 will go into deliverable D2.1.

Other conventions and instructions for this document:

As this document/template is to be reused, the convention to use it as a questionnaire is that:

1) -text in italics provides its structure and questions,

2) -input/content should be written using normal text, replacing **<input here>**

Also the following conventions are used to identify the purpose of some parts of the questionnaire:

Bold text in blue corresponds to indications/suggestions to complete the questionnaire

Bold text in dark red marks technical issues particularly relevant that should be carefully considered for further analysis of requirements

Text in red indicates pending issues or ad-hoc warnings to the reader

0 EXECUTIVE SUMMARY ON THE CASE STUDY

Summarize the research community applications/plans/priorities (max length 2 pages).
To be completed after section 2 and reviewed later. Supervision by a senior researcher is required.

1.1 Identification

- *Community Name*: **LifeWatch**
- *Institution/partner representing the community in INDIGO:* **IFCA-CSIC**
- *Main contact person*: **Fernando Aguilar**
- *Contact email:* **aguilarf@ifca.unican.es**
- *Specific Title for the Case Study:* **Transcriptome User-Friendly Analysis**

1.2 Brief description of the Case Study and associated research challenge

Please include also a brief description of the community regarding this Case Study: partners collaborating, legal framework, related projects, etc.

Describe the research/scientific challenge that the community is addressing in the Case Study

Application of next-generation sequencing (NGS) methods for transcriptome analysis (RNA-seq) has become increasingly accessible in recent years and are of great interest to many biological disciplines including, eg, evolutionary biology, ecology, biomedicine, and computational biology. Although virtually any research group can now obtain RNA-seq data, only a few have the bioinformatics knowledge and computation facilities required for transcriptome analysis.

Here, we discuss the Case Study around TRUFA (TRanscriptome User-Friendly Analysis), an open informatics platform offering a web-based interface that generates the outputs commonly used in de novo RNA-seq analysis and comparative transcriptomics[3]. TRUFA provides a comprehensive service that allows performing dynamically raw read cleaning, transcript assembly, annotation, and expression quantification. Due to the computationally intensive nature of such analyses, TRUFA is highly parallelized and benefits from accessing high-performance computing resources. The complete TRUFA pipeline was validated using four previously published transcriptomic data sets. TRUFA's results for the example datasets showed globally similar results when comparing with the original studies, and performed particularly better when analyzing the green tea dataset. The platform permits analyzing RNA-seq data in a fast, robust, and user-friendly manner. Currently accounts on TRUFA are provided freely upon request at https://trufa.ifca.es. TRUFA has been developed by IFCA in collaboration with MNCN (Spanish Natural Science Museum, also in CSIC). Access to the web portal is available under subscription for the research community.

[3] See Kornobis E, Cabellos L, Aguilar F, Frías-López C, Rozas J, Marco J, Zardoya R. (2015) **TRUFA: A User-Friendly Web Server for de novo RNA-seq Analysis Using Cluster Computing.** *Evol Bioinform Online* 11:97-104.

1.3 Expectations in the framework of the INDIGO-DataCloud project

What do you think could be your main objectives to be achieved within the INDIGO project in relation to this Case Study?

The main expectation for TRUFA is to get the capabilities to deploy the different components in a flexible and scalable way in the Cloud framework.

Currently, TRUFA has three different layers: web interface, job scheduler and computational layer. These two last layers are deployed in our Supercomputer, Altamira, so resources are not unlimited and the scalability of the framework is limited by corresponding TRUFA quotas.

So that, the expectation is to substitute limited layers by a scalable solution that allows managing a growing number of users. For instance, one solution could be to substitute these scheduler and supercomputer-execution layers by cloud oriented layers: an orchestrator that manages the deployment of Virtual Machines or Dockers within an e-infrastructure like EGI FedCloud. Furthermore, the web portal could be replaced by a SaaS solution, in particular including the currently deployed web-based file manager. Potentially additional post-processing from a python or R interface could be also interesting.

For data storage and management we need also a cloud solution that provides users not only with capacity but also with a web based or desktop application to manage files: check output, add output as input, etc.

A corresponding user authentication system solution is also needed.

1.4 Expected results and derived impact

Describe the research results and impact associated to this Case Study.

TRUFA being a recent product, it is not easy to estimate its potential impact. The experience with local users is that around 1-2 Million hours and 10 TB are only starting numbers to produce a few papers in the area. After the publication of a recent comment on TRUFA in the rna-seq blog[4] we have started to receive a significant number of requests from many different sites in the world (including Brazil, China, US, Argentina, India and of course Europe).

As NGS techniques are increasingly popular and useful, it is expected that if TRUFA is found useful the impact could be quite relevant in the short future, given that scalability is provided.

1.5 References useful to understand the Case Study

Include previous reports, articles, and also presentations describing the Case Study

Kornobis, Cabellos, Aguilar, Frias-Lopez, Rozas, Marco & Zardoya (2015). TRUFA: A user-friendly web server for de novo RNA-seq analysis using cluster computing. Evolutionary Bioinformatics. 11:97-104. http://www.la-press.com/article.php?article_id=4857

The user guide and an intro video can be found in the web portal, http://trufa.ifca.es

[4] http://www.rna-seqblog.com/trufa-a-user-friendly-web-server-for-de-novo-rna-seq-analysis-using-cluster-computing/

2 INTRODUCTION TO THE RESEARCH CASE STUDY

Summarize the Case Study from the point of view of the researchers (max length 3 pages + table). Input by the research team in the community addressing the Case Study is required.

2.1 Presentation of the Case Study

Describe the Case Study from the research point of view

Since the introduction of the RNA-seq methodology around 2006, studies based on whole transcriptomes of both model and non-model species have been flourishing. RNA-seq data are widely used for discovering novel transcripts and splice variants, finding candidate genes, or comparing differential gene expression patterns. The applications of this technology in many fields are vast, including researches on, eg, splicing signatures of breast cancer, host–pathogen interactions, the evolution of the frog immunome, the plasticity of butterfly wing patterns, the study of conotoxin diversity in Conus tribblei and the optimization of trimming parameters for de novo assemblies.

Despite the tremendous decrease in sequencing costs, which allows virtually any laboratory to obtain RNA-seq data, transcriptome analyses are still challenging and remain the main bottleneck for the widespread use of this technology. User-friendly applications are scarce and the post-analysis of generated sequence data demands appropriate bioinformatics know-how and suitable computing infrastructures.

When a reference genome is available, which is normally the case for model system species, a reference-guided assembly is preferable to a de novo assembly. However, an increasing number of RNA-seq studies are performed on non-model organisms with no available reference genome for read mapping (particularly those studies focused on comparative transcriptomics above the species level), and thus require a de novo assembly approach. Moreover, when a reference genome is available, combining both de novo and reference-based approaches can lead to better assemblies. Analysis pipelines encompassing de novo assemblies are varied, and generally include steps such as cleaning and assembly of the reads, annotation of transcripts, and gene expression quantification. A variety of software programs have been developed to perform different steps of the RNA-seq analysis, but most of them are computationally intensive. The vast majority of these programs run solely with command lines. Processing the data to connect one step to the next in RNA-seq pipelines can be cumbersome in many instances, mainly due to the variety of output formats produced and the postprocessing needed to accept them further as input. Moreover, as soon as a large computing effort is required, interactive execution is usually not feasible and an interface with the underlying batch systems used in clusters or supercomputers is needed.

2.2 Description of the research community including the different roles

Please include a description of the scientific and technical profiles, and detail their institutions

Describe the research community specifically involved in this Case Study

The community is composed by developers and users:

- TRUFA development team: composed by IT and bioinformatics stuff.
- Users (researchers): different background and expertise: PhD students, senior researchers, etc. Biology, bioinformatics, medical informatics, etc.

2.3 Current Status and Plan for this Case Study

Please indicate if the Case Study is already implemented or if it is at design phase.

Describe the status of the Case Study and its short/mid term evolution expected

TRUFA web portal and underlying components are fully operational and are used by a growing number of users. The web portal is installed in a server placed at IFCA. The different software needed to complete the TRUFA pipeline is installed in Altamira Supercomputer, placed at IFCA also and is manage by a scheduler system that handle the different jobs generated by the web portal.

The evolution expected for short/mid term is TRUFA "cloudification", which means to substitute the different TRUFA components by new cloud components: SaaS web portal, file management and storage workspace, orchestration system to manage all different steps in TRUFA and orchestrated computational components (dockers or virtual machines with needed software). Also we need an authentication system that allows TRUFA managers to handle the growing number of users.

2.4 Identification of the KEY Scientific and Technological (S/T) requirements

Please try to identify what are the requirements that could make a difference on this Case Study (thanks to using INDIGO solutions in the future) and that are not solved by now.

Indicate which are the KEY S/T requirements from your point of view

- Orchestration of different cloud-based steps in the workflow.
- Distributed storage system. File management.
- User management.

2.5 General description of e-Infrastructure use

Please indicate if the current solution is already using an e-Infrastructure (like GEANT, EGI, PRACE, EUDAT, a Cloud provider, etc.) and if so what middleware is used. If relevant, detail which centres support it and what level of resources are used (in terms of million-hours of CPU, Terabytes of storage, network bandwidth, etc.) from the point of view of the research community.

Detail e-Infrastructure resources being used or planned to be used.

TRUFA is a very recent released tool that currently is exploited by around 30 users. The computing layer is currently deployed in Altamira Supercomputer, so it is using HPC for processing and it is managed by a batch system. The modular design of TRUFA allows different types of deployment so web server does not need a particular batch system or computing infrastructure for working.

Regarding storage, the average use of disk needed is around 15GB per use case but it is variable.

2.6 Description of stakeholders and potential exploitation

Please summarize the potential stakeholders (public, private, international, etc.) and relate them with the exploitation possibilities. Provide also a realistic input to table on KPI.

Describe the exploitation plans related to this Case Study

- Public potential international TRUFA users from research community in many different fields and range (students, senior researchers).
- Private companies, eg, pharmaceutical industry.

Please indicate (as realistic as possible) the expected impact for each topic in the following table:

Area	***Impact Description***	***KPI Values***
Access	*Increased access and usage of e-Infrastructures by scientific communities, simplifying the "embracing" of e-Science.*	• *Number of ESFRI or similar initiatives adopting advanced middleware solutions* ESFRIs: **2 (LifeWatch, LTER)** • *Number of production sites supporting the software* **Currently 1 (IFCA), next LW sites**
Usability	*More direct access to state-of-the art resources, reduction of the learning curve. It should include analysis platforms like R-Studio, PROOF, and Octave/Matlab, Mathematica, or Web/Portal workflows like Galaxy.* *Use of virtualized GPU or interconnection (containers).* *Implementation of elastic scheduling on IaaS platforms.*	• *Number of production sites running INDIGO-based solutions to provide virtual access to GPUs or low latency interconnections* **GPUs are not used (yet)** • *Number/List of production sites providing support for Cloud elastic scheduling* **as above** • *Number of popular applications used by the user communities directly integrated with the project products:* **TRUFA integrates different sub applications, it is likely that R-studio could be integrated** • *Number of research communities using the developed Science Gateway and Mobile Apps:* **2** • *Research Communities external to INDIGO using the software products*: **Unknown (until an analysis of users is done)**
Impact on Policy	*Policy impact depends on the successful generation and dissemination of relevant knowledge that can be used for policy formulation at the EU, or national level.*	• *Number of contributions to roadmaps, discussion papers:* **1**
Visibility	*Visibility of the project among scientists, technology providers and resource managers at high level.*	• *Number of press releases issued*: **2 per year** • *Number of download of software from repository per year:* **does not apply** • *List of potential events/conferences/workshops:* **2 per year** • *Number of domain exhibitions attended* **2 per year** • *Number of communities and stakeholders contacted* **2 worldwide**
Knowledge Impact	*Knowledge impact creation: The impact on knowledge creation and dissemination of knowledge generated in the project depends on a high level of activity in dissemination to the proper groups.*	• *Number of journal publications:* **Undefined yet (5-10 as target)** • *Number of conference papers and presentations*: **idem**

Table 1 Key Performance Indicators (KPI) associated to different areas. Add in this table how your community would contribute to the KPIs. Note: this table will NOT be included in the deliverable.

3 TECHNICAL DESCRIPTION OF THE CASE STUDY

Describe the Case Study from the point of view of developers (4 pages max.)
Assemble it using preferably an AGILE scheme based on User Stories.

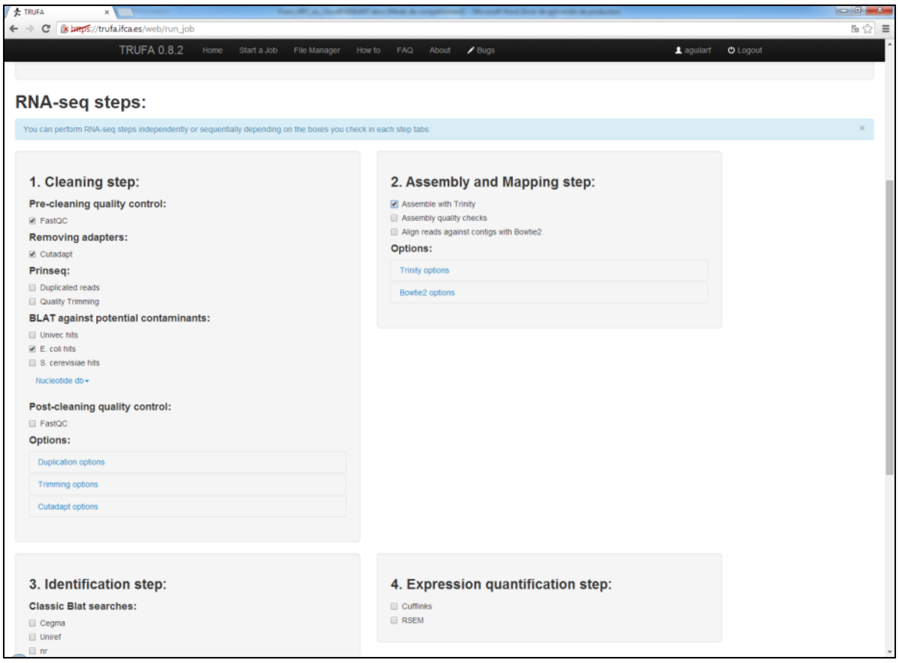
3.1 Case Study general description assembled from User Stories

Please describe here globally the Case Study. If possible use as input "generic" User Stories built according to the scheme: short-description (that fits in a "card") + longer description (after "conversation" with the research community). Provide links to presentations in different workshops describing the Case Study when available. Include schemes as necessary.

Describe the Case Study showing the different actors and the basic components (data, computing resources, network resources, workflow, etc.). Reference relevant documentation.

TRUFA is a system composed by different modules that work together in different layers. The top layer is the web server that provides the user an interface where they can follow different steps in order to set up the configuration needed to run an analysis. This web server is based both on a python and apache server, so the resources needed are not so many (just 1 GB of memory and 1 core). Also this web allows users to learn about TRUFA and how it works (using How To and FAQs), change user password and report bugs through a feedback system.

The second layer of TRUFA is the pipeline itself. It is composed of a driver script that takes the configuration sent by the web layer and spawns several jobs executed in the third layer. The following snapshot shows the different options that user can choose to configure the analysis including input files and steps that must be taken.



RNA-seq steps:

You can perform RNA-seq steps independently or sequentially depending on the boxes you check in each step tabs:

1. Cleaning step:

Pre-cleaning quality control:
☑ FastQC

Removing adapters:
☑ Cutadapt

Prinseq:
☐ Duplicated reads
☐ Quality Trimming

BLAT against potential contaminants:
☐ Univec hits
☑ E. coli hits
☐ S. cerevisiae hits
Nucleotide db ▾

Post-cleaning quality control:
☐ FastQC

Options:
Duplication options
Trimming options
Cutadapt options

2. Assembly and Mapping step:
☑ Assemble with Trinity
☐ Assembly quality checks
☐ Align reads against contigs with Bowtie2

Options:
Trinity options
Bowtie2 options

3. Identification step:

Classic Blat searches:
☐ Cegma
☐ Uniref
☐ nr

4. Expression quantification step:
☐ Cufflinks
☐ RSEM

Figure 1. TRUFA's configuration form

When user clicks "Start" button, the pipeline script checks the entire form getting chosen options and set up a file that is sent to Altamira supercomputer, in this case, as a job file. This script handles a number of different actions like the unzipping of input files, addition of commands depending on the options chosen by the user, etc.

Once the job is launched we can find the last TRUFA layer that is the process of the analysis itself. We currently take these steps in Altamira supercomputer. Users can also check the execution status (job pending, running, finished); so that when the job is finished they can get the output.

The following list represents the different software used by TRUFA, according to the number of cores that every program can manage.

Software	*Number of CPUs*	*Type*
B2G	*16*	*identify*
BLASTP	*16*	*assembly/mapping*
BLAT	*64*	*cleaning, identify*
BOWTIE	*16*	*assembly/mapping*
CEGMA	*16*	*assembly/mapping*
CUFFLINKS	*16*	*expression*
CUTADAPT	*16*	*cleaning*
FASTQC	*2*	*cleaning*
HMMER	*96*	*identify*
INTERPROSCAN	*64*	*identify*
MPIBLAST	*96*	*identify*
PRINSEQ	*16*	*cleaning*
RSEM	*1*	*expression*
SAMTOOLS	*1*	*assembly/mapping*
TRINITY_RNA_SEQ	*16*	*assembly/mapping*

For getting and managing analysis outputs, we have used a tool called FileManager (simogeo, https://github.com/simogeo/Filemanager). This tool provides users a web based file manager system where they can handle with the files that have been generated during the process. The source code have been modified in order to add certain functionalities like see html outputs as web pages or add output files as input files.

For user managing, we currently create a local user in the web server and they use a generic user once they want to run an analysis (this is transparent for the user) in our supercomputer Altamira.

3.2 User categories and roles

Describe in more detail the different user categories in the Case Study and their roles, considering in particular potential issues (on authorization, identification, access, etc.)

- TRUFA developers: web portal administration access with root privileges, software and middleware configuration permissions. Developers need permissions to set up new steps in the pipeline, configuration of software within the workflow, etc.
- TRUFA Users: web portal access (as users). Currently they need to ask for access and a local user account is created for them. Once they log into the web portal they can configure and launch analysis.

3.3 General description of datasets/information used

List the main datasets and information services used (details will be provided in next section)

Input: Currently, the input data accepted by TRUFA includes Illumina read files and/or reads already assembled into contigs. Read files should be in FASTQ format and can be uploaded as gzip compressed files (reducing uploading times). Reads from the NCBI SRA databases can be used but should be first formatted into FASTQ format using, eg, the SRA toolkit. Already assembled contigs should be uploaded as FASTA files. Other FASTA files and HMM profiles can be uploaded as well for custom blast-like and protein profile-based transcript annotation steps, respectively. Thus far, no data size limitation is set.

Output: TRUFA generates a large amount of output information from the different programs used in the customized pipeline. Briefly, a user should be able to download FastQC html reports, FASTQ files with cleaned reads (without duplicated reads and/or trimmed), Trinity-assembled transcripts (FASTA), read alignments against the transcripts (BAM files), GO annotations (.txt and.dat files which can be imported into the Blast2GO java application), and read counts (text files providing read counts and TPM). Various statistics are computed at each step and are reported in text files, such as the percentage of duplicated/trimmed reads, CEGMA completeness report, assembly sequence composition, percentage of mapped reads, and read count distributions.

3.4 Identification of the different Use Cases and related Services

Identify initial Use Cases based on User Stories, and describe related (central/distributed) Services

There is essentially one general use case that can be useful as an example of the way users handle TRUFA:

A) TRUFA user wants to perform a RNA-seq analysis in order to study of conotoxin diversity in Conus tribblei.

See section 6 for further details on Use Cases.

From the point of view of "services" definition, the following ones have been identified.

-Data management

TRUFA Users need to manage data that is used to perform the different analysis. On one hand, users need to handle input data: upload own data, add data/files from external databases directly, transform files between formats, etc. On the other hand, users need also to manage output files: move data within the workspace, check image and text files directly from the browser, tag output as input files or download output files. Having a "dropbox"-like web based file manager could be very useful.

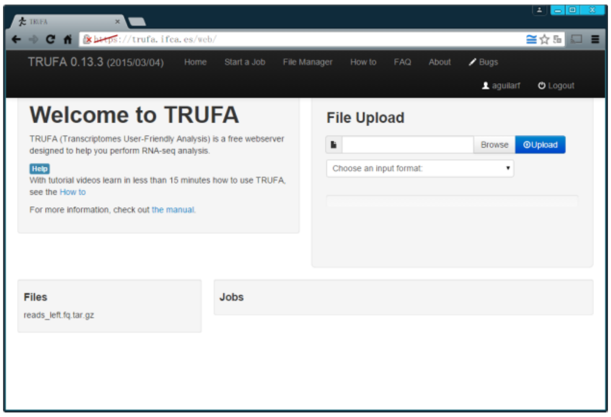
-Intensive Computing

The different steps of the pipeline that can be selected using TRUFA have several different computing requirements. Currently these steps are performed in Altamira that provides up to 96 cores per step.

3.5 Description of the Case Study in terms of Workflows

Summarize the different Workflows within the Case Study, and in particular Dataflows. Include the interaction between Services.

1. User access to TRUFA web portal.



Welcome to TRUFA

TRUFA (Transcriptomes User-Friendly Analysis) is a free webserver designed to help you perform RNA-seq analysis.

With tutorial videos learn in less than 15 minutes how to use TRUFA, see the How to

For more information, check out the manual.

File Upload

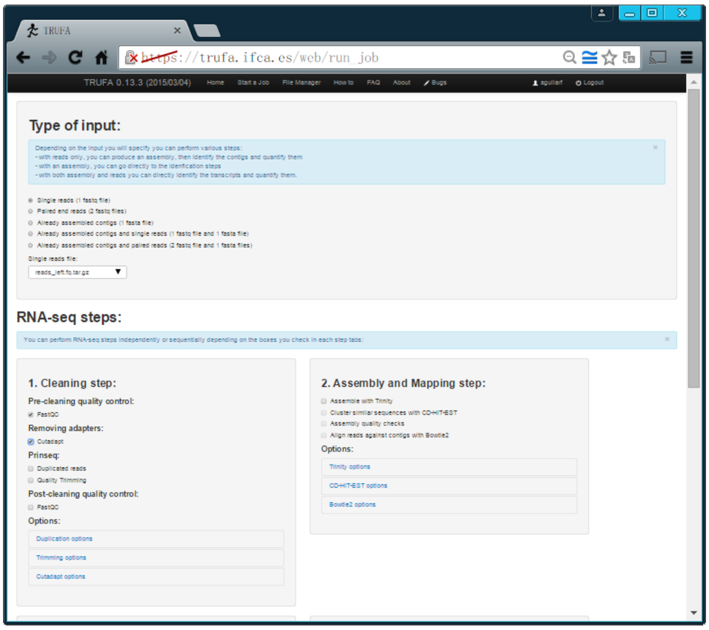
Choose an input format:

Files

reads_left.fq.tar.gz

Jobs

2. User prepares all the input files required to perform an analysis. This can be done uploading files or using online file manager. Data can also come from external sources like NCBI SRA database.
3. Files are uploaded to a storage space owned by the user. Currently, this space is located at IFCA and distributed through GPFS. This way, all the files can be used both from web portal and from Altamira.
4. User starts to fill a form that have to include input files and all the steps required for the analysis.



Type of input:

Depending on the input you will specify you can perform various steps:
- with reads only, you can produce an assembly, then identify the contigs and quantify them
- with an assembly, you can go directly to the idenfication steps
- with both assembly and reads you can directly identify the transcripts and quantify them.

● Single reads (1 fastq file)
○ Paired end reads (2 fastq files)
○ Already assembled contigs (1 fasta file)
○ Already assembled contigs and single reads (1 fastq file and 1 fasta file)
○ Already assembled contigs and paired reads (2 fastq file and 1 fasta files)

Single reads file: reads_left.fq.tar.gz

RNA-seq steps:

You can perform RNA-seq steps independently or sequentially depending on the boxes you check in each step tabs:

1. Cleaning step:

Pre-cleaning quality control:
☑ FastQC

Removing adapters:
☑ Cutadapt

Prinseq:
□ Duplicated reads
□ Quality Trimming

Post-cleaning quality control:
□ FastQC

Options:

2. Assembly and Mapping step:

□ Assemble with Trinity
□ Cluster similar sequences with CD-HIT-EST
□ Assembly quality checks
□ Align reads against contigs with Bowtie2

Options:

5. User clicks on “Start” button and a script located in the server configures the list of jobs that have to be runned in Altamira supercomputer. Some of these jobs are dependant between them, so dependencies has to be managed.
6. Altamira process all the different steps taking input data from the user GPFS storage space and put generated data also there. That way user will be able to manage all that files using the web portal and the online file manager.
7. User can check job status using the web portal. Once all the steps have finished, user can get the outputs: check image and text files directly from the web server, download other type of files or tag output as input files.

3.6 Deployment scenario and relevance of Network/Storage/HTC/HPC

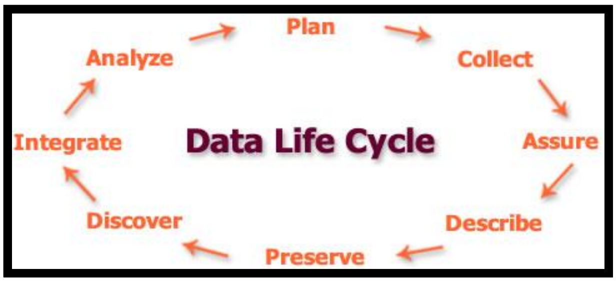
Indicate the current deployment framework (cluster, Grid, Cloud, Supercomputer, public or private) and the relevance for the different Use Cases of the access to those resources.

TRUFA workflow is currently deployed in Altamira Supercomputer where all the different software needed is installed, so that HPC is currently needed. Different steps need high requirements in terms of CPU (up to 96 cores) and memory that is why this kind of infrastructures is needed to perform that complex pipeline. However, in order to increase the scalability, we think that more flexible resources as cloud computing can substitute the current TRUFA deployment.

4 DATA LIFE CYCLE

INDIGO-DataCloud is a DATA oriented project. So the details provided in this complex section are KEY to the project. Please try to be as complete as possible with the relevant information.

Using the DataONE scheme, shown below, the different stages in the data life cycle are considered under the perspective of preparation of a DMP (Data Management Plan) following the recommendations of the UK DCC and H2020 guidelines.



BEFORE FILLING NEXT SECTIONS, CONSIDER CONSULTING:

https://www.dataone.org/all-best-practices-download-pdf and https://dmponline.dcc.ac.uk/

4.1 Data Management Plan (DMP) for this Case Study

According to EU H2020 indications[5], following UK DCC tool indications

[5] *In Horizon 2020 a limited pilot action on open access to research data will be implemented. Projects participating in the Open Research Data Pilot will be required to develop a Data Management Plan (DMP), in which they will*

4.1.1 Identification of the DMP

Plan identification: <*Code, ID*> **<input here>**

Associated grants: <Funded Projects, other grants> **<input here>**

Principal Researcher: **<input here>**

DMP Manager: **<input here>**

Description: **<input here>**

4.1.2 DMP at initial stage (to be prepared before data collection)

The DMP should address the points below on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be produced.

For each data set provide:

Description of the data that will be generated or collected; indicate its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

Data set reference and name **<input here>**

Data set description **<input here>**

Standards and metadata **<input here>**

Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created (see also below).

Connection to Instrumentation,

Sensors, Metadata, Calibration, etc (pending definitive form, see next sections)

<input here>

Vocabularies and Ontologies

Are they relevant? Internal vocabularies related to the specific fields. RDA groups.

specify what data will be open. Other projects are invited to submit a Data Management Plan if relevant for their planned research. The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project. The first version of the DMP is expected to be delivered within the first 6 months of the project. More elaborated versions of the DMP can be delivered at later stages of the project. The DMP would need to be updated at least by the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start. The templates provided for each phase are based on the annexes provided in the Guidelines on Data Management in Horizon 2020 (v.1.0, 11 December 2013).

(pending definitive form, see next sections)

<input here>

Data Capture Methods

Outline how the data will be collected / generated and which community data standards (if any) will be used at this stage. Indicate how the data will be organised during the project, mentioning for example naming conventions, version control and folder structures. Consistent, well-ordered research data will be easier for the research team to find, understand and reuse.

- *How will the data be created?* **<input here>**
- *What standards or methodologies will you use?* **<input here>**
- *How will you structure and name your folders and files?* **<input here>**
- *How will you ensure that different versions of a dataset are easily identifiable?* **<input here>**

Metadata

Metadata should be created to describe the data and aid discovery. Consider how you will capture this information and where it will be recorded e.g. in a database with links to each item, in a 'readme' text file, in file headers etc. Researchers are strongly encouraged to use community standards to describe and structure data, where these are in place. The UK Data Curation Center offers a catalogue of disciplinary metadata standards.

- *How will you capture / create the metadata?* **<input here>**

- *Can any of this information be created automatically?* **<input here>**
- *What metadata standards will you use and why?* **<input here>**

Data sharing

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

<input here>

Method for Data Sharing

Consider where, how, and to whom the data should be made available. Will you share data via a data repository, handle data requests directly or use another mechanism? The methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data. Mention earlier examples to show a track record of effective data sharing.

- *How will you make the data available to others?* **<input here>**
- *With whom will you share the data, and under what conditions?* **<input here>**

Restrictions on Sharing

Outline any expected difficulties in data sharing, along with causes and possible measures to overcome these. Restrictions to data sharing may be due to participant confidentiality, consent agreements or IPR. Strategies to limit restrictions may include: anonymising or aggregating data; gaining participant consent for data sharing; gaining copyright permissions; and agreeing a limited embargo period.

• *Are any restrictions on data sharing required? e.g. limits on who can use the data, when and for what purpose.* **<input here>**

• *What restrictions are needed and why?* **<input here>**

• *What action will you take to overcome or minimise restrictions?* **<input here>**

Data Repository

Most research funders recommend the use of established data repositories, community databases and related initiatives to aid data preservation, sharing and reuse. An international list of data repositories is available via Databib or Re3data.

• *Where (i.e. in which repository) will the data be deposited?* **<input here>**

Archiving and preservation (including storage and backup)

Questions to consider before answering:

•What is the long-term preservation plan for the dataset? e.g. deposit in a data repository

•Will additional resources be needed to prepare data for deposit or meet charges from data repositories?

Researchers should consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

• What additional resources are needed to deliver your plan?

• Is additional specialist expertise (or training for existing staff) required?

• Do you have sufficient storage and equipment or do you need to cost in more?

• Will charges be applied by data repositories?

• Have you costed in time and effort to prepare the data for sharing / preservation?

Carefully consider any resources needed to deliver the plan. Where dedicated resources are needed, these should be outlined and justified. Outline any relevant technical expertise, support and training that is likely to be required and how it will be acquired. Provide details and justification for any hardware or software which will be purchased or additional storage and backup costs that may be charged by IT services. Funding should be included to cover any charges applied by data repositories, for example to handle data of exceptional size or complexity. Also remember to cost in time and effort to prepare data for deposit and ensure it is adequately documented to enable reuse. If you are not depositing in a data repository, ensure you have appropriate resources and systems in place to share and preserve the data.

Describe the procedures that will be put in place for long-term preservation of the data.

<input here>

Indicate how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered. **<input here>**

4.1.3 DMP at final stage (to be ready when data is available)

*SCIENTIFIC RESEARCH DATA SHOULD BE EASILY **DISCOVERABLE***

Questions to consider:

- ***How will potential users find out about your data?***
- ***Will you provide metadata online to aid discovery and reuse?***

Guidance: Indicate how potential new users can find out about your data and identify whether they could be suitable for their research purposes. For example, you may provide basic discovery metadata online (i.e. the title, author, subjects, keywords and publisher).

*Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. **Digital Object Identifier**)?* **<input here>**

SCIENTIFIC RESEARCH DATA SHOULD BE **ACCESIBLE**

Questions to consider:

- ***Who owns the data?***
- ***How will the data be licensed for reuse?***
- ***If you are using third-party data, how do the permissions you have been granted affect licensing?***
- ***Will data sharing be postponed / restricted e.g. to seek patents?***

State who will own the copyright and IPR of any new data that you will generate. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. If purchasing or reusing existing data sources, consider how the permissions granted to you affect licensing decisions. Outline any restrictions needed on data sharing e.g. to protect proprietary or patentable data. See the DCC guide: How to license research data.

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses? (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc) **<input here>**

*SCIENTIFIC RESEARCH DATA SHOULD BE **ASSESSABLE** AND INTELLIGIBLE*

- ***What metadata, documentation or other supporting material should accompany the data for it to be interpreted correctly?***
- ***What information needs to be retained to enable the data to be read and interpreted in the future?***

Describe the types of documentation that will accompany the data to provide secondary users with any necessary details to prevent misuse, misinterpretation or confusion. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, any assumptions made, the format and file type of the data.

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review?, e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgments can be made about their reliability and the competence of those who created them **<input here>**

***USABLE** BEYOND THE ORIGINAL PURPOSE FOR WHICH IT WAS COLLECTED*

• What is the long-term preservation plan for the dataset? e.g. deposit in a data repository

• Will additional resources be needed to prepare data for deposit or meet charges from data repositories?

Researchers should consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

Guidance on Metadata:

• How will you capture / create the metadata?

• Can any of this information be created automatically?

• What metadata standards will you use and why?

Metadata should be created to describe the data and aid discovery. Consider how you will capture this information and where it will be recorded e.g. in a database with links to each item, in a 'readme' text file, in file headers etc.

Researchers are strongly encouraged to use community standards to describe and structure data, where these are in place. The DCC offers a catalogue of disciplinary metadata standards.

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data? e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists? **<input here>**

INTEROPERABLE** TO SPECIFIC QUALITY **STANDARDS

• What format will your data be in?

• Why have you chosen to use particular formats?

• Do the chosen formats and software enable sharing and long-term validity of data?

Outline and justify your choice of format e.g. SPSS, Open Document Format, tab-delimited format, MS Excel. Decisions may be based on staff expertise, a preference for open formats, the standards accepted by data centres or widespread usage within a given community. Using standardised and interchangeable or open lossless data formats ensures the long-term usability of data?

See the UKDS Guidance on recommended formats

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc?, e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins

<input here>

4.2 Data Levels, Data Acquisition, Data Curation, Data Ingestion

4.2.1 General description of data levels

Indicate if the DATASETS are organized into different levels (LEVEL-0, 1, 2, 3,4) and if so what are the relevant definitions and how DOI are provided. **<input here>**

4.2.2 Collection/Acquisition

Gathering RAW data

Specify how do you gather/collect your data (e.g. sensors, observations, satellites, etc.)? **<input here>**

How do you pre-process, transfer and store your RAW data? **<input here>**

From RAW Data to Calibrated Data

Describe the processes applied for Data Calibration, Validation, Filtering, etc. **<input here>**

4.2.3 Access to external data

Describe the identification and access to External Data **<input here>**

Indicate if there is a procedure for validation of External Data **<input here>**

4.2.4 Data curation

Specify any automatic check applied, like completing series, detecting outlier **<input here>**

Describe manual quality checks **<input here>**

Are there quality flags applied to the data? **<input here>**

4.2.5 Data ingestion / integration

Describe transformations applied to data taking into account ontologies/metadata. Indicate also if there is any "harmonization procedure" (to share/integrate data) and how linking internal and external data is made if relevant. **<input here>**

4.2.6 Further data processing

Describe, if relevant, the different additional processing steps (and the associated software and resources) applied to the (collected/curated) datasets to provide a "final" dataset collection that can be used in the analysis **<input here>**

4.3 Analysis

4.3.1 Basic analysis and standard analysis suites

Describe usual examples of basic analysis in the Case Study **<input here>**

Specify if software packages/tools like MATLAB, R-Studio, iPython,etc. are used **<input here>**

4.3.2 Data analytics and Big Data

Describe relevant examples of advanced analysis in the Case Study (like for example application of neural networks, series analysis, etc.) **<input here>**

Specify the resources and additional software required **<input here>**

Identify analysis challenges that can be classified as "Big Data" **<input here>**
List Big Data driven workflows **<input here>**

4.3.3 Data visualization and interactive analysis

Indicate the need for data and analysis results visualization **<input here>**

Indicate how visualization is made and if interactivity/steering is needed **<input here>**

Specify the User Interfaces (web, desktop, mobile, etc.) **<input here>**

4.4 Data Publication

Describe the information flow from the analysis to the publication **<input here>**

Indicate the requirements from publishers/editors to access data, and how it is made available (open data?) **<input here>**

5 INTENSIVE CPU REQUIREMENTS

Describe the Simulation/Modelling requirements in this Case Study. Please identify also any other intensive CPU mainly activity as required.

5.1 General description of simulation/modelling needs

Describe the different models used (including references) **<input here>**

*Indicate the type and quantity of simulations needed in the Case Study, and how they are incorporated in the general workflow of the solution***<input here>**

5.2 Technical description of simulation/modelling software

For each simulation package:

Identify the simulation software **<input here>**

Provide a link to its documentation, and describe its maturity and support level **<input here>**

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) **<input here>**

Tag the simulation software as HTC or HPC **<input here>**

*List the input files required for execution and how to access them***<input here>**

Describe the output files and how they will be stored **<input here>**

Reference an existing installation and performance indicators **<input here>**

Specify if the simulation software is parallelized (or could be adapted) **<input here>**

Specify if the simulation software can exploit GPUs **<input here>**

Specify how the simulation software exploits multicore systems **<input here>**

Specify if parametric runs are required **<input here>**

Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) **<input here>**

5.3 *Simulation Workflows*

Describe if there are workflows combining several (HTC/HPC) simulations or simulations and data processing

TRUFA is a workflow itself.

6 DETAILED USE CASES FOR RELEVANT USER STORIES

This section tries to put the focus on the preparation of detailed Use Cases starting from User Stories most relevant to the Case Study considered.

6.1 *Identification of relevant User Stories*

Examples of relevant User Stories linked to roles like for example Final User, Data Curator, etc.

List User Stories based on data collection, curation, processing, analysis, simulation, etc, that are considered most relevant for the Case Study being analyzed

TRUFA user wants to perform a RNA-seq analysis to study of conotoxin diversity in a Conus.

A.1) New user applies for a new TRUFA account.

A.2) TRUFA team checks new user identity and create a new account. User is notified.

A.3) User access to the portal, where temporal password can be changed.

A.4) User prepares the input files for the analysis that can be uploaded (files owned by user) or linked from external databases (like NCBI SRA).

A.5) User fills the form selecting input files to analyse and different steps to be performed.

A.6) User launches the analysis.

A.7) TRUFA pipeline script set up the workflow to be performed in Altamira supercomputer. List of jobs and dependencies between them are established.

A.8) Jobs are processed in Altamira.

A.9) User gets information about the status of the jobs from the web server.

A.10) Once jobs have finished, user can access to file manager to handle the output files: download, check images and text files directly in the web server, delete, move, tag as input files, etc.

For each relevant User Story:

Draft a basic card **<input here>**

Provide details from conversation with the researchers' teams

The input is based on the definition of the workflow and also on the TRUFA user guide, that defines the basic interaction between users and TRUFA.

Draft as a Use Case **<input here>**

Analyze tools to support the definition of the Use Case (like mockups). Integrate in the analysis the requirements on user interfaces (like the use of mobile resources, under different flavours, access through web interfaces, etc.)

TRUFA has a web based interface that allows users to perform different actions: upload input files, define the pipeline (form), and manage files (web based file manager).

Describe the way to extract requirements and define acceptance criteria **<input here>**

Include if possible an example of support for Big Data driven workflows for e-Science, with requirements for scientific workflows management, under a "Workflow as a Service" model, where the proper workflow engines will be selected according to user needs and requirements.

In such case please describe the scenario for Big Data analysis, and assure that the Use Case considers which levels of workflow engines are needed (e.g., "coarse gran", which targeting distributed (loosely coupled) experiments, through workflow orchestration across heterogeneous set of services; "fine grain", which targeting high performance (tightly coupled) data analysis through workflows orchestration on big data analytics frameworks)

7 INFRASTRUCTURE TECHNICAL REQUIREMENTS

Describe the Case Study from the point of view of the required e-infrastructure support.

INDIGO Data-Cloud will support the use of heterogeneous resources.

7.1 Current e-Infrastructures Resources

Start from the current use of e-infrastructures.

7.1.1 Networking

Describe the current connectivity

Altamira nodes are connected by FDR IB

Describe the key requirements (availability, bandwidth, latency, privacy, etc) Availability, privacy

Specify any current issue (like last mile, or access from commercial, etc) **<input here>**

7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the current use of each of these type of resources: size and usage

Up to 96 cores per step, memory depends on step, but usually => 2GB.

Indicate if there is any mode of "orchestration" between them

Batch system to manage jobs.

7.1.3 Storage

Describe the current resources used

Currently there are 10TB per all users (around 30). Distributed between components/layers(GPFS).

Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator)

7.2 Short-Midterm Plans regarding e-Infrastructure use

Plans for next year (2016) and in 5 years (2020).

7.2.1 Networking

Describe the proposed connectivity **Connectivity**

Describe new/old key requirements (availability, bandwidth, latency, QoS, private networking, etc) **<input here>**

Specify any potential solution/technique (for example SDN) **<input here>**

7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the evolution expected: which infrastructures, total "size" and usage **<input here>**

Detail potential "orchestration" solutions **<input here>**

7.2.3 Storage

Describe the resources required **<input here>**

Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) **<input here>**

7.2.4 SPECIFIC QUESTIONS REGARDING USE OF EGI.eu (FROM EGI DOC 2478)

Sample questions to capture details of a support case

These questions can help case supporters interview the case submitter and the NGIs to refine the technical details of the case and ultimately to move towards a suitable technical setup. These questions aim at understanding the user's need, the technical and other requirements/constrains of the case, and the impact that a solution would bring to the scientific community. These questions provide only guidance – Ticket owners can use other questions or even other methods to identify details of their support case(s).

- *What does the user/community want to achieve? (What's the user story?)*
- *For who does the case request resources for? (CPU/storage capacity, SW tools, consultant time, etc.) For a group? For a project? For a collaboration? Etc.*
- *What is the size of the group that would benefit from these resources, and where these people are? (which country, institute)*
- *Approximately how much compute and storage capacity and for how long time is needed? (may be irrelevant if the activity is for example assessment of an EGI technology)*

- *Does the user need access to an existing allocation (→ join existing VO), or does he/she needs a new allocation? (→ create a new VO)*
- *What is the scientific discipline?*
- *Which institute does the contact work for (or those he/she represents)?*
- *Does the case include preferences on specific tools and technologies to use?*
  - *For example: grid access to HTC clusters with gLite; Cloud access to OpenStack sites; Access to clusters via standard interdafaces; Access to image analysis tools via Web portal*
- *Does the user have preferences on specific resource providers? (e.g. in certain countries, regions or sites)*
- *Does the user (or those he/she represents) have access to a Certification Authority? (to obtain an EGI certificate)*
- *Does the user (or those he/she represent) have the resources, time and skills to manage an EGI VO?*
- *Which NGIs are interested in supporting this case? (Question to the NGIs)*

7.3 On Monitoring (and Accounting)

Please outline any requirements for monitoring of the platforms and the applications.

If you have specific tools already in use, please outline them.

Please also specify monitoring, metrics at different levels: system, performance, availability, network QoS, website, security, etc.

Currently we just have metrics regarding the general use of TRUFA in terms of computing time and storage used per user. We would need specific accounting for user: CPU time, storage space, etc.

7.4 On AAI

(From EGI, revise and check with WP4/5/6)

Describe the current AAI status of your community/research infrastructure

• *Does your community/research infrastructure already use AAI solutions?*

TRUFA only currently use local authentication system, with a local database.

• *Can you describe the solutions you have adopted highlighting as applicable: Technology adopted (e.g. X509, SAML Shibboleth,...), Identity Providers (IdP) federations integrated (e.g. eduGAIN) or approximate number of individual IdPs integrated, Solution for homeless users (users without an insitutional IdP), Solutions to handle user attributes* **<input here>**

Describe the potential needs and expectations from an AAI integration in the ***services and platforms provided by INDIGO***

- *Type of IdP to be integrated (e.g. institutional IdP part of national federations and eduGAIN or non federated, social media credentials, dedicated research community catch-all IdP, ...)*

TRUFA would need to integrate some institutional IdP like eduGAIN or certificates. One solution could be use certificates within a certain Virtual Organization....

- *Preferred authentication technology, and requirements for support of multiple technology and credential translation services (e.g. SAML -> X509 translation)* eduGAIN/certificates
- *Community level authorization/attribute based authorization to support different authorization levels for the users*

  Only two types of users: developers and users. For web access only one type of authorization is needed.
- *Web access and/or non-web access*

  Web portal access
- *Need for delegation (e.g. execute complex workflows on behalf of the user)*

  Delegation for infrastructure beyond the web portal (HPC, cloud).
- *Support for different level of assurance credentials, and need to use the information about users with lower level of assurance credentials to limit their capability* **<input here>**
- *Requirements for high level of assurance credentials (e.g. to access confidential/sensitive data)* **<input here>**

7.5 On HPC

Describe any specific issue related to the use of supercomputers.

TRUFA currently uses HPC but can be substituted by other infrastructures like cloud.

7.6 Initial short/summary list for “test” applications (task 2.3)

Software used	*Software/applications/services required, configuration, dependencies (Describe the software/applications/services name, version, configuration, and dependencies needed to run the application, indicating origin and requirements.)* **Web portal: python, python based web server, web-based file manager. Batch system: SLURM (can be substituted thanks to TRUFA modular architecture).** **Pipeline: Altamira management software, B2G BLASTP BLAT BOWTIE CEGMA CUFFLINKS CUTADAPT FASTQC HMMER INTERPROSCAN MPIBLAST PRINSEQ RSEM SAMTOOLS TRINITY_RNA_SEQ**
Operating system requirements	**Any Linux distribution. Currently web portal is in Ubuntu and Altamira uses Scientific Linux 6**
Run libraries requirements	*Run API/libraries requirements (e.g., Java, C++, Python, etc.)* **Web server: Python, etc.** **Pipeline: those required by different software.**
CPU requirements (multithread,MPI,“wholenode”)	**Up to 96 parallelized cores, MPI.**
Memory requirements	**Memory requirements depends on the step, but usually ~ 2GB minimum for any analysis.**
Network requirements	**Enough for handling few GB of data in a reasonable time.**
Disk space requirements (permanent, temporal)	*Include the requirements for data transferring (upload and download of data objects: files, directories, metadata, VM/container images, etc.)* **50-100GB per user**
External data access requirements	**External databases like NCBI**
Typical processing time	**Depending on the number of steps. Up to 3 days in current setup**
Other requirements	*Requirements for data synchronization* *Requirements for data publication* *Requirements for depositing data to archives and referring them* *Requirements for mobile application components for data storage and access* *Requirements for data encryption and integrity control-related functionality* **<input here>**
Other comments	**<input here>**
Relevant references or URLs	**http://trufa.ifca.es**

8 CONNECTION WITH INDIGO SOLUTIONS

<To be filled by INDIGO JRA >

8.1 IaaS / WP4

8.2 PaaS / WP5

8.3 SaaS / WP6

8.4 Other connections

9 FORMAL LIST OF REQUIREMENTS

<this will be further edited within WP2>

10 REFERENCES

R 1	
R 2	
R 3	
R 4	
R 5	