INDIGO - DataCloud

RIA-653549

INDIGO WP2: Definition of Support t
Research Communities
Report on Case Studies

(aka D2.1 INITIAL REQUIREMENTS FROM RESEARCH COMMUN

Ignacio Blanquer, UPV

Peter Solagna & Yin Chen, EGI.eu

Fernando Aguilar & Jesus Marco, IFCA-CSIC

Thanks to:

Miguel, Alexandre, Antonio, Max, Laura, Manuela, Davide, Sandro, Federico, Cristina, Riccardo, Angelo, Eva...

INDIGO-DataCloud is co-founded by the Horizon 2020Framework Programme

INDIGO: The Research Communities

A work package (WP2, NA) represents the interest of Research Communities to assure that their requirements will be satisfied by the project outcomes, by providing feedback and participating in the revision of the services deployed.

Keep the focus also on big data research use and management through a dedicated task oriented to track the different needs at the **data life-cycle**, following the reference models used by the different Research Communities.

NDIGO: The IMPACT on Research Communities

MPLIFIED IMPACT TABLE ECTED OBJECTIVES versus REQUESTS/ POTENTIAL PACT FOR COMMUNITIES Development of the INDIGO Platform based on open tware without restrictions on the e-Infrastructure	Life Sciences	Physical Sciences & Astronomy	Social Sciences & Humanities	Environm Sciences
earch Communities & Initiatives , including ESFRIs	ELIXIR INSTRUCT/ WeNMR EuroBioImaging	CTA LBT WLCG	DARIAH DCH-RP	EMSO LIFEWATCH ENES
mples of Applications	HADDOCK GROMACS AMBER GALAXY	MIDAS, IRAF, IDL, Geant4 ROOT/PROOF Geant4	Fedora Digital Libraries	Delft3D R-Studio TRUFA MATLAB
gn and development of a Platform providing advanced users and munity developers a powerful and modern environment levelopment work. This includes programming and scripting tools, composition of custom applications and software deployment	RELEVANT	**CRITICAL**	RELEVANT	**CRITICAL**
eloping a framework to enable the transparent execution on remote frastructures of existing popular applications like MATLAB / OCTAVE, T, MATHEMATICA, or R-STUDIO.	RELEVANT	**CRITICAL**	MINOR	**CRITICAL**
ide the services and tools needed to enable a secure composition of ices from multiple providers in support of scientific applications.	**CRITICAL**	**CRITICAL**	RELEVANT	RELEVAN
elop and implement a solution that is able to deploy in a transparent powerful way both services and applications in a distributed and rogeneous environment made by several different infrastructures Grid and Federated Cloud, IaaS Cloud, Helix Nebula, HPC clusters)	**CRITICAL**	RELEVANT	MINOR	RELEVAN
elop the capability in the PaaS to provide unified data access despite graphical location of data, including APIs access, based on existing dards, or virtually mount like a POSIX device to worker node, cloud al machines, personal computer etc.	**CRITICAL**	RELEVANT	**CRITICAL**	RELEVAN

.1: Initial Requirements from Research Communities

INDIGO - DataC

om DoW:

*D2.1 summarizes the findings of T2.1 (**Research Communities Requirements**) and T2.2(**Defining support to Research Data**) along the first three months of the project, providing **input to JRA activities**.*

*The report will be an integrated document including a general description of the researc communities involved and of the **use cases and workflows** proposed and will express **equirements captured**, **prioritized** and grouped by technical areas (Cloud, HPC, Grid, Da Management, etc.).*

n particular, the analysis of Data Management Plans (DMPs) and data lifecycle documentation aiming to identify both synergies and gaps among the different communities will be provided.

Methodology for D2.1

From DoW:

- *Analyze the use cases proposed by the communities participating to the consortium. Capture the requirements for efficiently running the applications and workflows on Cloud, Grid or HPC infrastructures.*

Caution: Impedance mismatch, ICT experts vs. Researchers

- *Capture requirements generated by user communities not part of the project (such as the EGI Federated Cloud users), which are relevant for the outputs of the project*
- *Liaise with the INFRADEV-4 projects to enable synergies between the projects, and interoperability between the INDIGO outputs and the VRE to be deployed by the E-INFRA-9 projects.*

Methodology for D2.1 (cont'd)

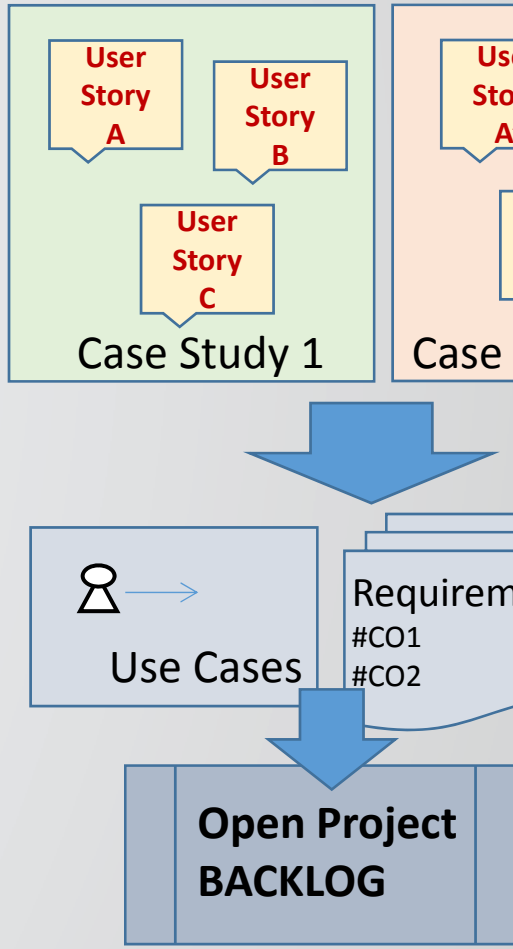
highly criticized — and refined — in several iterations

A template was **designed** to gather information from communities

- Based on **Case Studies**

  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.

- Focus on Case Studies that are representative both of the research challenge and complexity but also of the possibilities offered by INDIGO-Data Cloud solutions on it.
- A Case Study is (ideally) 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 template serves as a structured framework with guiding questions concerned by INDIGO development workpackages.



Methodology for D2.1 (cont'd)

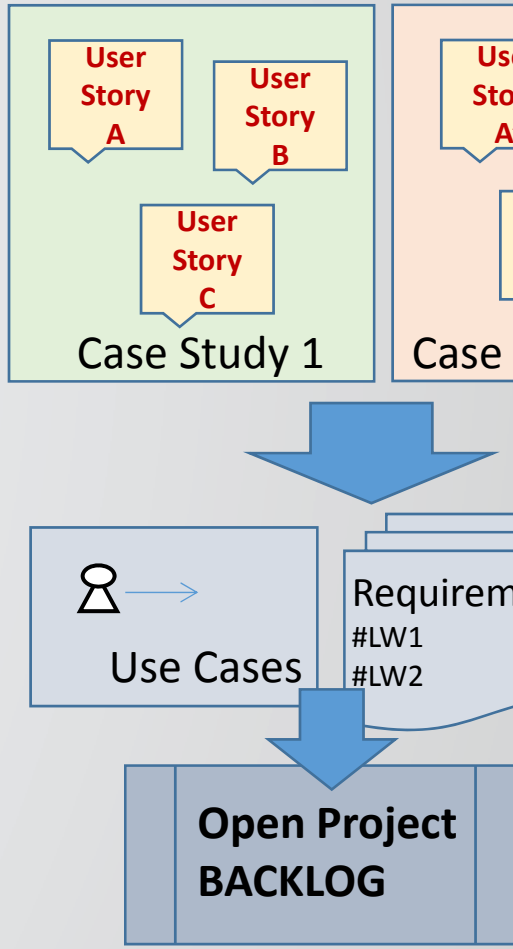
highly criticized — and refined — in several iterations

A template was **designed** to gather information from communities

- Based on **Case Studies**

  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.

- Focus on Case Studies that are representative both of the research challenge and complexity but also of the possibilities offered by INDIGO-Data Cloud solutions on it.
- A Case Study is (ideally) 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 template serves as a structured framework with guiding questions concerned by INDIGO development workpackages.



Methodology for D2.1 (cont'd)

emplate sections:

General Introduction

Technical description of the case study.

- Going into User Stories and, when possible, into Use Cases

Data Life Cycle (including DMP).

Computing intensive processes

- including simulation/modeling

Detailed use cases for relevant user stories.

Infrastructure technical requirements.

- including AAI and (service) monitoring aspects

Connection with INDIGO solutions

Executive summary

https://grid.ifca.es/wiki/INDIGO/WP2/I

Annexe Template

1st Iteration

- INDIGO-WP2-D2.1-ANNEX-1P0-V7-CMCC_v1.6.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7-EUB_v3.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7_CIRMMP.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7_INGV_v2.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7_elixir_ita.pdf
- INDIGO-WP2-D2.1-ANNEX-1P10_v1-UUtrecht.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-TRUFA-V7.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7.1INAF_LBT.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7-EGI-Fedcloud-1st.pdf
- INDIGO-WP2-D2.1-ANNEX-DARIAH.pdf
- INDIGO-WP2-D2_1-ANNEX-1P0-V7-2_ICCU_(2)_0_1.doc

2nd Iteration

Thanks to all!

- INDIGO-WP2-D2.1-ANNEX-1P0-V7-CMCC_v2.3.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7-EUB_v4.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7_CIRMMP_v2.pdf
- INDIGO-WP2-D2.1-ANNEX-1P10-Utrecht_v3.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-ALGAE-BLOOM-V10.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-TRUFA-V10.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7_elixir_ita_v2.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7_INGV_v4.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-V7_INAF-CTA.pdf
- INDIGO-WP2-D2.1-ANNEX-1P0-LBT-V7.1.pdf
- INDIGO-WP2-D2 1-ANNEX-1P0-V7-2_ICCU_0_2.doc

Methodology for D2.1 (cont'd)

ext: Analysis of the Annexes to identify requirements

Performed by Ignacio, Peter, Yin

Produced large EXCEL table, several entries per Case Study

- Community, Req#, Req. Descr., Rank (Mandatory/Convenient/Optional), Current, Gaps, Solution…)

Community	Req #	Requirement	Requirement Type (Computing / Storage / PaaS	Rank (Mandatory / Convenient / Optional)	Current workflow/solution	Gaps	Proposed improvement	Potential solution for INDIGO (User community point of view)	Potential solution for INDIGO (JRA point of view)	Comments
ENES - CMCC	ENES#7	Isolation of deployments	Computing	Convenient	Currently users share the infrastructure.	Unavailable feature	Need for minimising side-effects and Ophidia deployments are tailored to the reference data.	Deployment on containers and VMs provides the isolation.		See ENES#1, ENES#8
	ENES#8	Execution across multiple centres.	Computing / PaaS Service	Mandatory	Not provided	Unavailable feature	Interesting when exhausting resource capabilities of one deployment or when combining the processing of different data sets that are deployed on different Data Analytics infrastructures.	Task T5.3 in INDIGO deals with the geographic scheduling of workloads, however, this may not be sufficient given the interactive nature of the process. Surely changes are needed at application level and coherent global authz management could help. Metascheduling.		
		to reduce time-to-			Based on data download	Server-side approach not	It should be easy to deploy a self-configurable and auto-scalable Data	Combination of TOSCA specification, software		See ENES#1, ENES#2,

ext: identification of common requirements

Performed by Ignacio

Produced single EXCEL table

Methodology for D2.1 (cont'd)

INDIGO - DataC

#	Requirement	Requirement	Rank	Proposed improvement	Potential solution for INDIGO (User	Pot	EuB	Lif	ELI	Had	CIR	Fe	DA	INA	CMC	C
	Deployment of Interface SaaS	Computing / PaaS	Mandatory	A mechanism to facilitate the deployment of a customised Haddock	The portal could be instantiated by means of a set of containers and/or specific base		M	C	M	M	M			C	C	M
	Deployment of Customized computing back-ends as batch queues	Computing / PaaS	Mandatory	Each instance may have an independent software configuration,	A devops tool integrated with the deployment service to install and configure		M	M	M	M	M	C		C	M	C
	Deployment of user-specific software	Computing / PaaS	Mandatory	Manual installation may be cumbersome for large-scale application involving many computing	Ability of a user to easily construct a software installation and configuration specification (e.g. TOSCA) for their own		M		M						C	
	Automatic elasticity of computing batch queues	Computing / PaaS	Mandatory	When moving to the cloud, users should be provided with the exact	Monitoring services may be integrated with the deployment, which will trigger the		M	M	M	M	M			C	M	
	Terminal access to the resources.	Computing / PaaS service	Mandatory	This feature must be linked to the AAI	This will require ssh ports to be open and direct access to the VMs. The massive		M		M	M	M				M	
	Privileged access	Computing / PaaS service	Mandatory	This feature must be linked to the AAI	A single special user in the "sudo" group.		C		M	M	M				M	
	Execution of workflows	Computing / PaaS	Mandatory	Processing done on the cloud where the outputs of the processing are	Workflow engine can be deployed as any other application. Back-end could be a		M		C	C		O			M	M
	Provenance information	Computing / PaaS Service	Convenient	Very important for revision of papers and project proposals.	Repository of data and software that could be deployed or inspected on demand.		C									
	Cloud bursting	Computing / PaaS Service	Mandatory	Supplementing the computing capacity with special instances	Automatic contextualization and configuration will enhance the		C	C	C	M					M	
	Data-aware scheduling	Computing / PaaS Service	Convenient	Currently storage and computing are highly coupled.	This will affect the scheduling. Moving computing to data. Maybe the use of				C			C			M	
	Provisioning of efficient Big Data Analysis solutions exploiting server-side and declarative approaches	Computing / Storage / PaaS Service	Mandatory		Currently it uses a hierarchical set of databases that are coordinated through distributed memory parallel computing										M	
	Execution across multiple centres.	Computing / PaaS Service	Mandatory	Interesting when exhausting resource capabilities of one deployment or when	Task T5.3 in INDIGO deals with the geographic scheduling of workloads,										M	
	On-line processing of data	PaaS	Mandatory	Special management of post-processing jobs that could be sent to	Despite that this may look similar to any other processing, two aspects need to be		C	M			M			M	M	
	Special hw configuration - MPI, multicore, GPGPU	Compute / PaaS	Mandatory	More flexibility in the way the requirements are defined and the	Three main issues must be analysed here (not all for the User Cases selected)ç: 1) The		C	C			M				M	

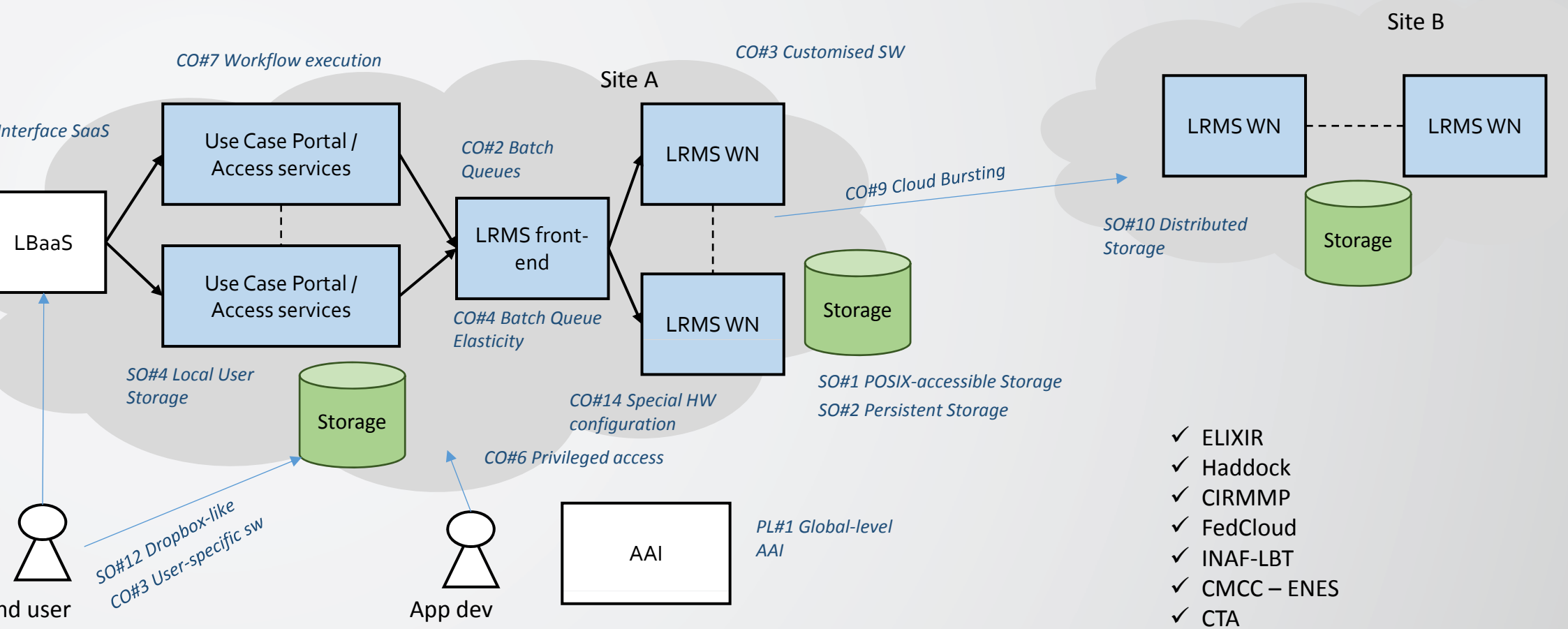
Req #	Requirement	Requirement	Rank	Proposed improvement	Potential solution for INDIGO (User community	Pot	EuB	Lif	ELI	Had	CIR	Fe	DA	INA	CMC	CT	AL	ING
SO#1	Shared storage accessible like a POSIX filesystem	Storage / PaaS Service	Mandatory	Limited storage and no scalability	Data volumes that can be mounted (R/W) on multiple VMs using an efficient protocol. Block-based storage will offer a		M	M	M	M				M			M	
SO#2	Persistent data storage	Storage	Mandatory		Disk storage in the VMs must be persistent even if the VM is undeployed, and only removed if explicitly requested.		M	M	M		M		M	M	M	M	M	M
SO#3	Long-term availability of results	Storage	Mandatory	External, long-term, self-maintained storage.	Interoperability with other infrastructures.		C				M		M					
SO#4	Local user storage	Storage / PaaS Service	Mandatory	Separate individual volumes will increase scalability and privacy.	Individual storages deployed as R&W volumes.		M	M	M					M		C		
SO#5	Availability of reference data	Storage / PaaS Service	Mandatory		A shared, read-only volume should be available with all the reference data.				M	M	M		M	M		M		
SO#6	Interoperability with IS-ENES/ESGF	Storage / PaaS Service	Mandatory	No improvement, keeping this feature	Basic data access functionality through ESGF protocols (HTTP, OPeNDAP) associated to metadata catalogues (Thredds); User authentication based on OpenID federation; and Solr search and discovery service.										M			
SO#7	Metadata management / Database as a Service	Storage / PaaS Service	Convenient		Metadata services as part of the storage services		C							C	C	C	M	M
SO#8	Share data capabilities	Storage/IaaS service	Convenient	Block storage with added NFS-like capability of multiple access.	One Data Storage solution							M			C			C
SO#9	Data replication	PaaS	Mandatory	Hide the data topology to the user, data federation, data replication capabilities	OneData used to federate community repositories, and allow an easy access to the datasets, and to replicate the data where necessary, based on community parameters.							M		M				
SO#10	Distributed storage	Storage / PaaS service	Mandatory	Cloud or grid based solutions have not proven to be efficient yet.	Cloud back-end will facilitate the deployment on a wider range of infrastructures.								M	M	M	M		M
SO#11	Dropbox-like storage	Storage / PaaS service	Convenient	Facilitate interaction with users in uploading and downloading files	Client tools for accessing storage from desktop systems.			C						C			M	

Req #	Requirement	Requirement	Rank	Proposed improvement	Potential solution for INDIGO (User community point of	Pot	EuB	Lif	ELI	Had	CIR	Fe	DA	INA	CMC	CT	AL	ING
SO#9	Data replication	PaaS	Mandatory	Hide the data topology to the user, data federation, data replication capabilities	OneData used to federate community repositories, and allow an easy access to the datasets, and to replicate the data where necessary, based on community parameters.							M		M				
SO#10	Distributed storage	Storage / PaaS service	Mandatory	Cloud or grid based solutions have not proven to be efficient yet.	Cloud back-end will facilitate the deployment on a wider range of infrastructures.								M	M	M	M		M
SO#11	Dropbox-like storage	Storage / PaaS service	Convenient	Facilitate interaction with users in uploading and downloading files	Client tools for accessing storage from desktop systems.			C						C			M	
PL#1	Global-level AAI	PaaS	Mandatory	Centralized mechanism to define general authorisation policies will give scalability and a coherent mechanism.	Use a repository of credentials and authorisation tokens that could provide a coherent global mechanism. Use of a centralised credential system and the management of users and tenants, such as OpenStack Keystone.		M	M	M	M	M		M	M	M	M		M
PL#2	On-line access to data	Computing / Storage / PaaS	Mandatory	Interactive access to the VMIs to avoid downloading huge amounts of data for consolidated inspection of results	This will have to main impacts. If VMs are not be provided of public IPs, reverse tunnelling or any other solution must ensure that the ports used in the interactive access are provided (VNC-like). In any case, firewall rules must enable this kind of traffic.		C							M	M		M	M
PL#3	Network configuration	IaaS	Optional	Extend current standard interfaces to support network configuration, such as VPN-aaS, Firewall-aaS								O						
PL#4	Monitoring and operation	PaaS	Convenient	Keep functionality	Monitoring of resources is competence of the infrastructure provider. Monitoring of the services need to be analysed.					C	C							

User Community Computing Portal Service (~SaaS)

- A user community has an application (or set of them) that can be accessed through a portal and requires a batch queue as back-end.
- Unpredictable workload and user access profile.
- The application consists on two main parts: the portal / scientific Gateway and the processing working nodes
  - Working nodes should scale-up and down according to the workload.
  - Cloud-bursting to external infrastructures may be requested.
  - Portal services should also adapt to workload.
  - Users can access reference data and provide their own local data.
- Requested by the use cases from:
  - ELIXIR, Haddock, CIRMMP, FedCloud, DARIAH, INAF-LBT, CMCC – ENES, CTA, ALGAE – BLOSSOM, INGV - MOIST

User Community Computing Portal Service (~SaaS)

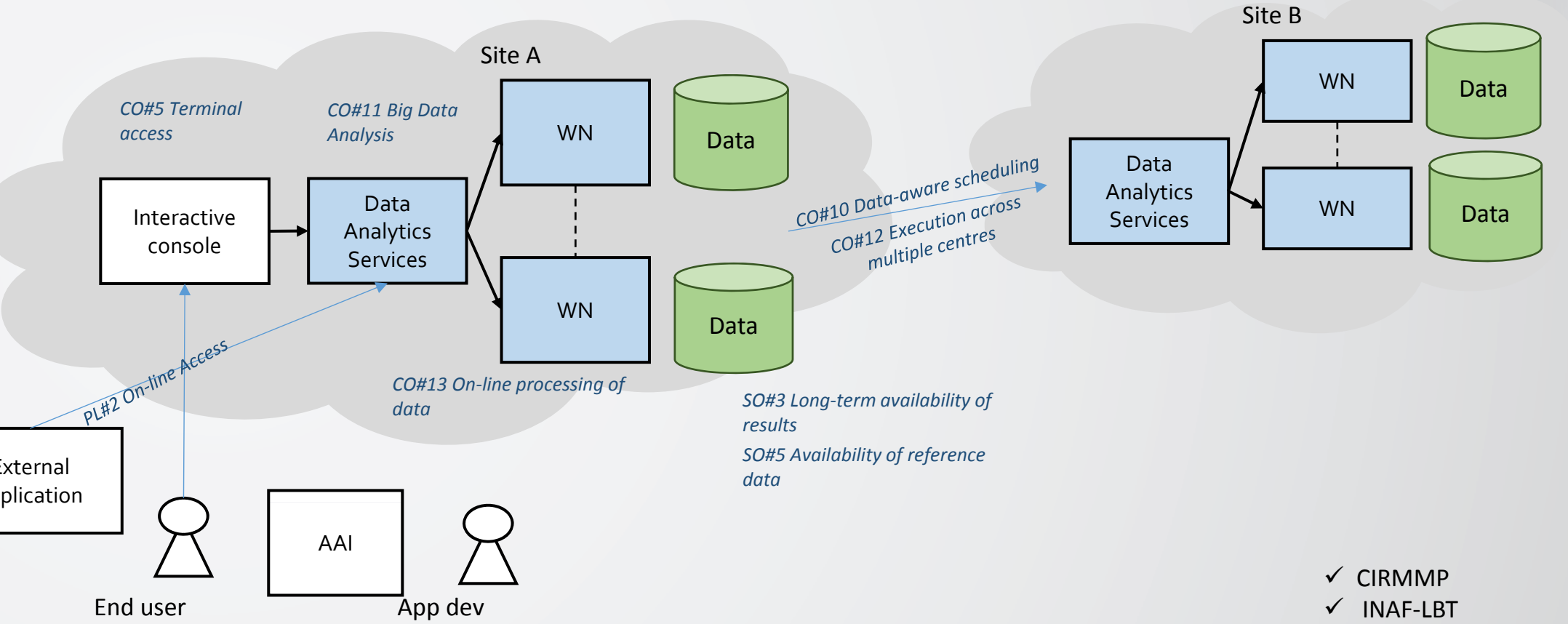


- ✓ ELIXIR
- ✓ Haddock
- ✓ CIRMMP
- ✓ FedCloud
- ✓ INAF-LBT
- ✓ CMCC – ENES
- ✓ CTA
- ✓ ALGAE – BLOSSOM

Data Analysis Service (~PaaS)

- A user community has a coordinated set of data repositories and software services to access, process and inspect them
  - Processing is interactive, requiring accessing a console deployed on data's premises.
- The application consists on a console / Scientific Gateway that interacts with the data
  - E.g "R", Python, OPHIDIA
  - It can be a complementary scenario from the previous one.
  - It can expose programmatic services.
- Requested by use cases from:
  - CIRMMP, INAF-LBT, CMCC – ENES, ALGAE – BLOSSOM, INGV – MOIST.

Data Analysis Service (~PaaS)



(Other requirements from previous case also apply)

- ✓ CIRMMP
- ✓ INAF-LBT
- ✓ CMCC – ENES
- ✓ ALGAE – BLOSSOM
- ✓ INGV - MOIST

Comments

- These are the **Initial** Case Studies
  - More Mature? High Impact? Too wide?
  - Candidates for pilots, tests (see D2.3), **training** (?)
- Evolution is expected
  - From iteration with INDIGO JRA
  - From co-evolution with VRE and INFRADEV…
  - **From impact from e-infrastructure**
- **Analysis yet to be extended for DMP and DATA topics**
  - Will evolve in Task 2.2 into D2.4, then into D2.7 (Ingestion)
- *KEY IDEA PROPOSED: identify for each Research Community*

  *a **(>50% devoted)** "champion" to guarantee that the Annex will be made real*

How are we going to Agile-interact???

- Example (from WP6 initial comments), include:
  - *the needs of the scientific workflows -WfMS*

  (as stated as an example in ENES#4, EB#3,CO#7 etc)
  - *the needs of Science Gateways*

  (as stated for example in ENES#15, EL#1, ENES#13, LBT#6)
  - *the needs of mobile apps –*

  (that are missing (besides ENES#15) since was not captured in first table)

YES WE ARE COMPLETELY OPEN AND ENCOURAGE INTERACTION/REVISION,
BUT FINAL WORD IS (MAINLY) ON RESEARCH COMMUNITIES (OUR CUSTOMERS)
(and some include ICT experts and previous/competing solutions)