Spoke 3 - WP4 - Big Data Management, Storage and Archiving

Work Plan

Cristina Knapic

Objectives and Methodologies

- **<u>Objective 2</u>. Big data processing and visualization, via adopting innovative approaches (e.g. Artificial Intelligence, inference via Bayesian statistics) for the analysis of large and complex data volumes and for their exploration (e.g. in-situ visualization), capable of efficiently exploiting HPC solutions.**

- **<u>Objective 3</u>. High Performance storage, Big Data management, and archiving applying the Open Science principles and implementing them in the Big Data Archives. Experimenting and adopting novel technologies and computational approaches for fast and scalable I/O**

- ACO-S will promote the FAIRness of the research outputs and services across research communities involved in the project. WP0 and WP4 will ensure that the FAIR principles are fully endorsed and implemented in the Spoke activity. In particular, WP4 will evaluate the services proposed in the technical WPs and propose guidelines to improve the FAIRness of the data and research output of the project

Tasks description

Overview:

WP4 will analyze, explore, standardize and store data of different collections, characterizing them in the appropriate way in order to facilitate the respect of the FAIR principles and enable users to benefit from an innovative storage and archiving platform. A distributed archive infrastructure with hot and cold storages and proper access tools will be implemented or customized by existing ones, respecting the interoperability directives (RDA, IVOA, ...). The requirements of the storage in terms of identification of data collection characteristics (data models, data dimension, accessibility methods, cold or hot storage etc...) will be analyzed, defining and implementing suitable technical solutions.

Tasks description

T4.1 Data management, Standardization and Interoperability This task is intended to gather all the information related to collect, organize, analyze, preserve, curate and share data. It will be documented in light of the analysis of relevant use cases. The general information related to the data collection content will be exposed using the most current and updated standards available in the international consesses like RDA, IVOA, Open Access, so they will be interoperable with similar resources in order to reach the data FAIRness.

Strong interaction with WP0 (DMP), WP1 (meta descriptors definition), WP2 (meta descriptors definition), WP3 (access for data visualization including possibility to transfer data to computational clusters).

Tasks description

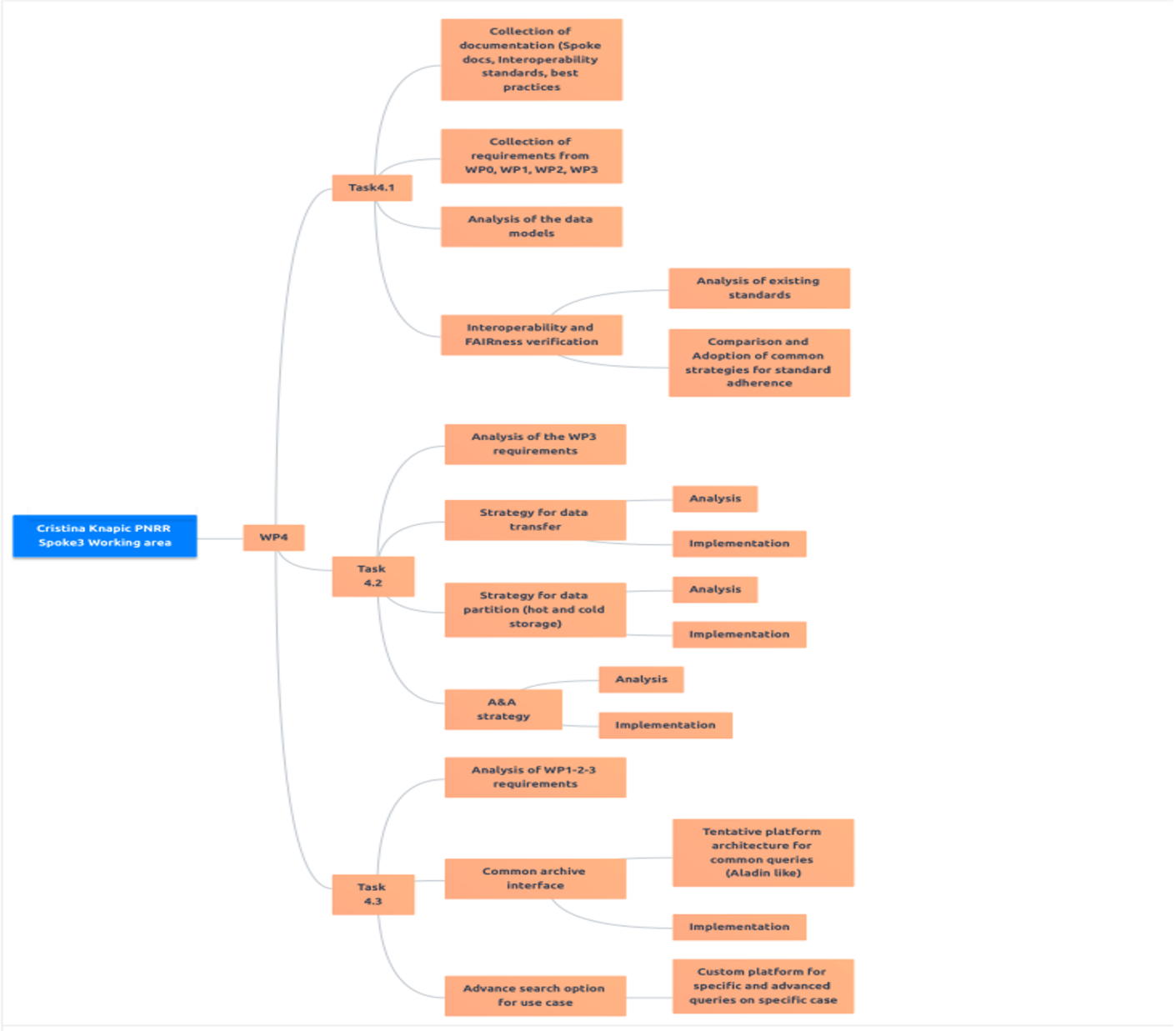
T4.2 Local and Distributed Long and Short term Storage Optimization This task will be carried on in synergy with WP3 and will prototype the protocol stack for data transfer both externally (from data providers) and internally (from the storage space to the data-intensive computation area), by using as study implementation, the existing hardware infrastructures. Transfer protocols and data management systems will be tested and configured to optimize the band occupancy and parallel file systems will be tested and configured to guarantee writing on device and computing performances.

Strong interaction with WP3 (data transfer to and from computational clusters) and WP6

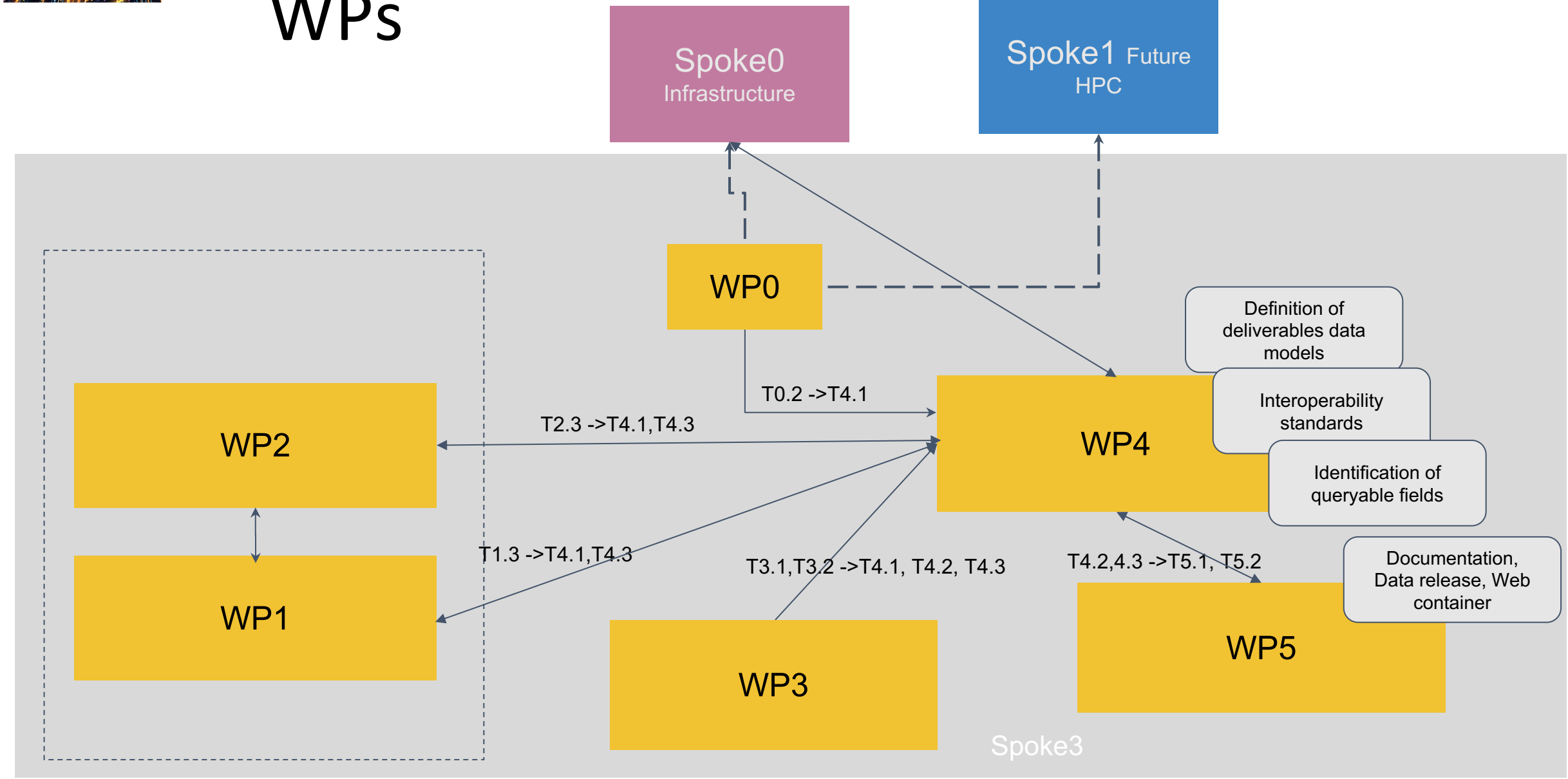
Tasks description

T4.3 Archive and Repository Definition and Implementation This task is related to the software integration of the Distributed Data Archive Framework and the access tools already implemented or customized by the project. It will compose a framework where users can perform data storage, retrieval and exploitation. The User Space (storage part) implementation will be fully VO compliant. A web interface will give access to remote resources that will adopt appropriate libraries and tools for data exploitation. The access to the storage resources will allow authentication of users for private data and will be performed with standard and well consolidated authentication and authorization (A&A) tools, fully compliant with the Single Sign On paradigm.

Strong interaction with WP0 (policy on data), WP2-3 (definition of investigable fields..), WP6 (for deployment).


Cristina Knapic PNRR Spoke3 Working area
WP4
Task4.1
Collection of documentation (Spoke docs, Interoperability standards, best practices
Collection of requirements from WP0, WP1, WP2, WP3
Analysis of the data models
Interoperability and FAIRness verification
Analysis of existing standards
Comparison and Adoption of common strategies for standard adherence
Task 4.2
Analysis of the WP3 requirements
Strategy for data transfer
Analysis
Implementation
Strategy for data partition (hot and cold storage)
Analysis
Implementation
A&A strategy
Analysis
Implementation
Task 4.3
Analysis of WP1-2-3 requirements
Common archive interface
Tentative platform architecture for common queries (Aladin like)
Implementation
Advance search option for use case
Custom platform for specific and advanced queries on specific case

WPs


Spoke0
Infrastructure
Spoke1 Future
HPC
WP0
WP2
WP1
WP3
WP4
WP5
T0.2 ->T4.1
T2.3 ->T4.1,T4.3
T1.3 ->T4.1,T4.3
T3.1,T3.2 ->T4.1, T4.2, T4.3
T4.2,4.3 ->T5.1, T5.2
Definition of deliverables data models
Interoperability standards
Identification of queryable fields
Documentation, Data release, Web container
Spoke3

Partner Efforts

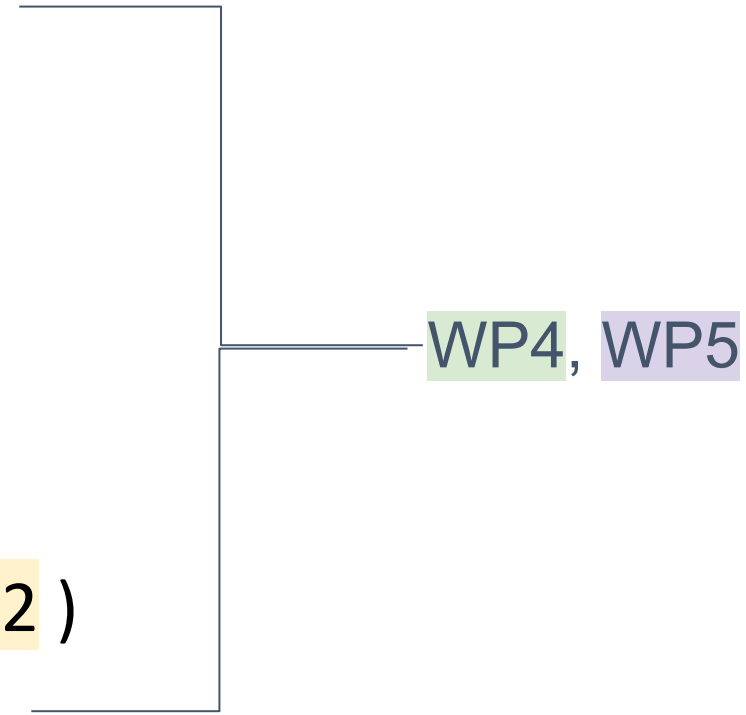
PARTNER		TASK	Persona di riferimento	PM Anno 1	PM Anno 2	PM Anno 3	Tot
INAF		T4.1- T4.2 - T4.3	Deborah Busonero	6	11	6	23
		T4.2- T4.3	TD				
UNITO		T4.1-T4.3	Andrea Mignone	0.3	0.3	0.3	0.9
UNITS		T4.1	Francesco Longo	0,75	0,75	0,75	2,25
INAF	OACagliari	T4.1,T4.2,T4.3	Andrea Possenti	1	1	1	3
INAF	OACagliari	T4.1,T4.2,T4.3	TD (resp Possenti)	0-6	0-12	0-6	0/24
INAF	OACagliari	T4.1,T4.2,T4.3	PhD (resp Possenti)	0-12	0-12	0-12	0/36
INAF	OATs	T4.3	TD (resp. Knapic)	6	12	12	30
INFN	PG	T4.3	Sara Cutini	1	1	1	3
INFN	FE	T4.3	Martina Gerbino	0	1	1	2
ROMTOV		T4.1, T4.2, T4.3	Tecnologo	2	2	2	6
Unipol SAI - Leithà		T4.3	--	6	6	6	18
Totale				23,05/41,05	35,05/59,05	30,05/48,05	88,15/148,15

Deliverables

D19	Identified Technologies integration into existing archives (M24.6)	Identified Technologies integration into existing archives report	WP4	INAF	R	PU	M24
D20	Data Transfer protocols description (M24.7)	Data Transfer protocols description document	WP4	INAF	R	PU	M24
D21	Distributed Data Archiving infrastructure description document (M24.8)	Distributed Data Archiving infrastructure description document (INAF)	WP4	INAF	R	PU	M24
D28	Final WP4 activity report (M36.5)	Final WP4 activity report and public release of all WP products: infrastructures, codes and documentation (INAF, INFN, UniTS, UniTO, UniTOV)	WP4	INAF	R	PU	M36

Working groups (proposal)

1. Eulerian Codes (WP1, WP2)
2. Lagrangian Codes (WP1, WP2)
3. Time series (WP1, WP2, WP3)
4. Feature extraction (WP3)
5. Bayesian inference(WP2 , WP3)
6. Deep learning (WP1,WP3)
7. Visualization (WP3)
8. Data-reduction & imaging (WP1, WP2)
9. Semi-numerical codes (WP1, WP2)
10. Web-tools (WP4, WP5)
11. Platforms (WP4, WP5)
12. Data-model (WP1, WP2, WP3 ,WP4, WP5)



WP4 Organization

- Definition of Task leaders will be done depending on partners efforts
- Participation to WGs is suggested also by personal effort offered (0,3 PM per year can only be a consultant and a reporter from other WPs as already stated);
- Data models WG can be lead by one of the key person in the WP4 but candidatures are welcome!
- Several work will be done in strict conjunction with WP5…

Conclusions

- identification of affordable tasks is still in progress;
- a draft work plan of the whole project has to be don by the end of the year;
- by the end of November we should have a WP effort table updated and effective;
- risks have to be mitigated: currently we can rely on 18 PM on the first year, 29 PM on the second and 24 PM on the third for developments…

Reference documents