Task Area 1:
Developing workflows and tools for data management

Workshop | PAHN-PaN
August 22-23 2019, DESY

Christoph Wissing (DESY), Andreas Heiss (KIT)

Task Area 1 from the LoI

Task area 1 ”Developing workflows and tools for data management”

This task area defines and develops data handling standards and data processing workflows which are as generic and interoperableas possible while respecting theFAIRdata principles. In this context, data can be either raw experimental data or data thatcomprise all information on the experimental apparatus or codes in theoretical calculations/simulations that were used togenerate data. High-level services necessary to implement the data processing workflows will be selected and (further) developed.This includes existing data management software like dCache, XRootD, Dynafed, RUCIO, or IRODS, as well as transfer serviceslike FTS3. Workflows will be based on the services of the task area ”FAIRdata management infrastructures and open data”,which will build the foundation of the distributed computing and data management environment. There is an increasing demand for the ability to utilize a spectrum of resources including HPCs and cloud systems. In addition, upcoming workflow tools need to support specialised hardware architectures like GPUs and FPGAs. In particular, the following items will be pursued: workflows and middleware of specific and generic data access methods including authentification and authorisation, datasecurity and access rights; workflows and middleware to generate standardised (cross-disciplinary) meta-data; user-transparent inclusion of heterogenous, opportunistic and long-term IT resources into data processing workflows; workflows and middleware to support the definition of application-specific machine learning architectures. It is important to us that this work is carried out in accordance with the international activities and collaborations in the research field. Furthermore, services, standards and solutions developed in this task area also need to fit into global structures such as the European Open Science Cloud (EOSC).

Task Area 1 from the LoI: summarized

- Define and develop data handling and data processing workflows.
  - As genereic and interoperable as possible.
  - Take FAIR principles into account.
- Workflows implementations based on
  - selected existing data management services (dCache, xrootd, RUCIO, IRODS, FTS3)
  - services of TA2 ”FAIRdata management infrastructures and open data”
- Provide techniques to make use of various types of resources and special hardware:
  - HPC, cloud / GPUs, FPGAs, application-specific machine learning architectures, …
- Use results of international R&D in this research field and make sure that everythings fits into developing distributed infrastructures (e.g. EOSC).

Task Area 1 proposed topics / contributions in short (from questionnaire)

- RWTH Aachen
  - Inter- and long-term operability
- Bonn
  - Exploiting heterogeneous resources in a transparent operation with focus on lattice and HPC
- DESY
  - Development of automated data movement depending on forseen data usage, access pattern or durability to appropriate media
  - AAI: implement AARC recommendations for their stoarage solutions
  - Integration of endpoints to modern and autoscaling infrastructures in experiment frameworks
- Erlangen
  - Creating standardised metadata and ccordinating this process with intern. astroparticle physics
- Göttingen
  - Exploitation and integration of heterogeneous resources and cloud resources

Task Area 1 proposed topics / contributions in short (from questionnaire)

- GSI
  - Developing data management workflows and tools
- FZ Jülich
  - Software development (middleware, user tools) for the management of LQCD data, ensuring reproducibility (data lineage) and ease of storage and access, management of the local LQCD data repository.
- KIT
  - Work on analysis workflows and utilization of HPC and cloud resources
  - Work on data access methods in distributed environments including opportunistic resources
- Mainz
  - Development focusing on co-existence between different data management software like dCache, XRootD, iRODS, … and opportunistic scheduling.
- Münster
  - Connection to EOSC project CS3MESH4EOSC, utilization of invenio RDM techology.

Task Area 1 proposed topics / contributions in short (from questionnaire)

- Regensburg
  - Develop and implement workflows for long-term archiving of research data.

- Wuppertal
  - Development of monitoring for containerized jobs including standardized interfaces. Development should be based on existing monitoring tools for containers. The goal is a (itself containerized) software suite, which allows to validate the successful execution of containers as well as error detection when running previously archived containers.
  - Tutorials for data management and container utilization (is actually Cross Cutting C).
  - Development of data management software for storage, sharing & archival (reproducibility) of primary and secondary LQCD data.

FTE planned by Co-Applicants in PAHN-PaN

Institute Co-Applicant E-Mail	Task Area 1	Task Area 2	Task Area 3	Task Area 4	Cross Topic A	Cross Topic B	Cross Topic C	Governance	Total	FTE Contribution	Comment
Aachen Alexander.Schmidt@physik.rwth-aachen.de	1		2		x	x	x		3	2	
Bielefeld tarsch@physik.uni-bielefeld.de		1	1						2	2	
Bonn liechti@physik.uni-bonn.de	1.5	1	1		x				3.5	2.5	
Darmstadt soyaal@ikp.tu-darmstadt.de		0.5	0.5				x		1	0.2	
DESY Thomas.schoerner@desy.de	2	1.5	1.5	?	x	x	x	1	6	2	Possibly another FTE in TA4
Dortmund kevin.kroeninger@cern.ch			2.5	1	x		x		3.5	3.5	
Erlangen uli.katz@physik.uni-erlangen.de	0.25	0.25	0.5						1	0.5	Distribution over TA is a guess!
FIAS redelbach@compeng.uni-frankfurt.de				1			x		1	0.5	
Freiburg markus.schumacher@physik.uni-freiburg.de	0.5		0.5						1	0.5	
Goettingen arnulf.quadt@cern.ch	1		0.5				x		1.5	1	
GSI K.Schwarz@gsi.de	1	1	1		x	x		1	4	1	
Hamburg gregor.kasieczka@uni-hamburg.de			2		x		x		2	1	Actually 1.5. TA1, 0.5 CT-A
Heidelberg plehn@uni-heidelberg.de			0.75				x		0.75	2	FTE request might increase
Jülich – FZJ s.krieg@fz-juelich.de	0.5	0.5	1			x	x			1	TA assignment not 100% clear yet
KIT andreas.haungs@kit.edu	2	2			x	x	x	1	5	2	
Köln smayer@ikp.uni-koeln.de	0.75	0.75			x				1.5	1	TA assignment not spelled out
LMU Thomas.kuhr@lmu.de	0.5		2		x				2.5	1	0.5 FTE in TA1 is actually CT-A
Mainz buescher@uni-mainz.de	1	1	1.5	1.5					5	5	
Münster hmg@uni-muenster.de	0.5	0.5				x			1	0.5	
Regensburg gunnar.bali@ur.de	0.5		1						1.5	0.9	
TUM nora.brambilla@ph.tum.de			1						1	1	
Wuppertal heinemz@uni-wuppertal.de	1		1						2		
(Participants not included)											Can request budget for them
Sum	14	10	21.25	3.5				3	**51.75**	**15.4**	Total

Stand: 21.8.2019