H2020-FETHPC-1-2014 ANTAREX-671623

# AutoTuning and Adaptivity approach for Energy efficient eXascale HPC systems

http://www.ANTAREX-project.eu/

# Deliverable D1.2: Data Management Plan V1.7

European Commission

Horizon 2020
European Union funding
for Research & Innovation

| | | | |
|---|---|---|---|
| **Deliverable Title:** | Data Management Plan | | |
| **Lead beneficiary:** | POLIMI (Italy) | | |
| **Keywords:** | Publica Data, Private Data, Data Management Plan | | |
| **Author(s):** | Giovanni Agosta (POLIMI), Andrea Bartolini (ETHZ) Andrea Beccari (DOMPE), Luca Benini (ETHZ), João Bispo (UPORTO), João Cardoso (UPORTO), Radim Cmar (SYGIC), Carlo Cavazzoni (CINECA), Jan Martinovic (IT4I), Gianluca Palermo (POLIMI), Pedro Pinto (UPORTO), Erven Rohou (INRIA), Nico Sanna (CINECA), Katerina Slaninova (IT4I), Cristina Silvano (POLIMI); | | |
| **Reviewer(s):** | João Cardoso (UPORTO), Cristina Silvano (POLIMI). | | |
| **WP:** | WP1 | **Task:** | T1.2 |
| **Nature:** | Report | **Dissemination level:** | Public |
| **Identifier:** | D1.2 | **Version:** | V1.7 |
| **Delivery due date:** | March 1st, 2016 | **Actual submission date:** | March 11th, 2016 |

| | |
|---|---|
| **Executive Summary:** | Based on the activities carried out in **Task 1.2**, this deliverable describes the **Data Management Plan - DMP (D1.2)** for the ANTAREX project according to the **Guidelines on Data Management in H2020** (Version 2.0 dated 30/10/2015) and **Guidelines on Open Access to Scientific Publications and Research Data in H2020** (Version 2.0 dated 30/10/2015). The **DMP** specifies what data will be generated in the project and what data will be exploited and/or shared/made accessible for verification and reuse and how this data will be maintained. The **DMP** will evolve during the project lifetime to present the project's progresses in terms of data management. Updated versions of **D1.2** are planned to be released at **M18** and finally at the end of the project (**M36**).<br>In this way, ANTAREX project will become eligible for the **Pilot Action on Open Access to Research Data** as stated in H2020.<br>The described policy reflects the **ANTAREX Consortium Agreement** (signed by all partners in December 2015) regarding data management and it is consistent with the exploitation and protection of results. |

| | |
|---|---|
| **Approved and issued by the Project Coordinator:**<br>C Silvano | **Date:** March 11th, 2016 |

**Project Coordinator**: Prof. Dr. Cristina SILVANO – Politecnico di Milano
**e-mail**: silvano@elet.polimi.it - **Phone:** +39-02-2399-3692- **Fax:** +39-02-2399-3411

# Table of Contents

# 1 Data Management Plan

## 1.1 Summary

Data Management Plans (DMPs) are introduced in the Horizon 2020 Work Programmes:

> *A further new element in Horizon 2020 is the use of Data Management Plans (DMPs) detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved. The use of a Data Management Plan is required for projects participating in the Open Research Data Pilot. Other projects are invited to submit a Data Management Plan if relevant for their planned research.*

The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the applicants with regard to all the datasets that will be generated by the project. The DMP is not a fixed document, but evolves during the lifespan of the project.

This document describes the Data Management Plan - DMP (D1.2) for the ANTAREX project, generated according to the Guidelines on Data Management in H2020 (Version 2.0 dated 30/10/2015) and Guidelines on Open Access to Scientific Publications and Research Data in H2020 (Version 2.0 dated 30/10/2015). According to the ANTAREX DoW, the ANTAREX DMP is planned to be issued at M06 as D1.2, while updated versions of D1.2 are expected to be released at M18 and finally at the end of the project (M36).

In this way, ANTAREX project will become eligible for the **Pilot Action on Open Access to Research Data** as stated in H2020.

> *A detailed description and scope of the Open Research Data Pilot requirements is provided on the Participants Portal (Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020). Projects taking part in the Pilot on Open Research Data are required to provide a first version of the DMP as an early deliverable within the first six months of the project. Projects participating in the pilot as well as projects who submit a DMP on a voluntary basis because it is relevant to their research should ensure that this deliverable is mentioned in the proposal. Since DMPs are expected to mature during the project, more developed versions of the plan can be included as additional deliverables at later stages. The purpose of the DMP is to support the data management life cycle for all data that will be collected, processed or generated by the project.*

A DMP as a document outlining how research data will be handled during a research project, and after it is completed, is very important in all aspects for projects participating in the Horizon 2020 Open Research Data Pilot as well as almost any other research project. Especially where the project participates in the Pilot it should always include clear descriptions and rationale for the access regimes that are foreseen for collected data sets. This principle is further clarified in the following paragraph of the Model Grant Agreement:

> *As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex I, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.*

## 1.2 *Public Data Management Policies*

### 1.2.1 Open Access Infrastructure for Research in Europe OpenAIRE

OpenAIRE[1] is an initiative that aims to promote open scholarship and substantially improve the discoverability and reusability of research publications and data. The initiative brings together professionals from research libraries, open scholarship organisations, national e-Infrastructure and data experts, IT and legal researchers, showcasing the truly collaborative nature of this pan-European endeavour.

**Project details:**

| | |
|---|---|
| Project n°: | 643410 |
| Project type: | Research and Innovation |
| Start date: | 01/01/2015 |
| Duration: | 42 months |
| Total budget: | 13 132 500 € (4 mi are targeted towards the FP7 post grant gold OA pilot) |
| Funding from the EC: | 13 000 000 € |

A network of people, represented by the National Open Access Desks (NOADs), organises activities to collect H2020 project outputs, and supports research data management. Backing this vast outreach, is the OpenAIRE platform, the technical infrastructure that is vital for pulling together and interconnecting the large-scale collections of research outputs across Europe. The aim of the project is to create workflows and services on top of this valuable repository content, which will enable an interoperable network of repositories (via the adoption of common guidelines), and easy upload into an all-purpose repository (via Zenodo).

OpenAIRE2020 assists in monitoring H2020 research outputs and should be key infrastructure for reporting H2020's scientific publications as it will be loosely coupled to the EC's IT backend systems as stated in the project description. The EC's Research Data Pilot is supported through European-wide outreach for best research data management practices and Zenodo, which will provide long-tail data storage. Other activities include: collaboration with national funders to reinforce the infrastructure's research analytic services; an APC Gold OA pilot for FP7 publications with collaboration from LIBER; novel methods of review and scientific publishing with the involvement of hypotheses.org; a study and a pilot on scientific indicators related to open access with CWTS's assistance; legal studies to investigate data privacy issues relevant to the Open Data Pilot; international alignment with related networks elsewhere with the involvement of COAR.

***Zenodo***

Zenodo[2] is developed by CERN under the EU FP7 project OpenAIREplus (grant agreement no. 283595). The repository is open to all research outputs from all fields of science regardless of funding source. Given that Zenodo was launched within an EU funded project, the knowledge bases were first filled with EU project codes, but they are keen to extend this to other funders. Zenodo is free for the long tail of Science. In order to offer services to the more resource hungry research, they have a

[1] https://www.openaire.eu

[2] http://www.zenodo.org

ceiling to the free slice and offer paid for slices above, according to the business model developed within the sustainability plan.

Zenodo allows to create own collections for communities and to accept or reject uploads submitted to it. It can be used for example for workshops or other activities.

***Content***

All research outputs from all fields of science are welcome. In the upload form it can be chosen between types of files: publications (book, book section, conference paper, journal article, patent, preprint, report, thesis, technical note, working paper, etc.), posters, presentations, datasets, images (figures, plots, drawings, diagrams, photos), software, videos/audio and interactive materials such as lessons. Zenodo assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citeable. Further information is in Terms of Use and Policies.

***Size limits***

Zenodo currently accepts files up to 2GB (several 2GB files per upload); there is no size limit on communities. However, they don't want to turn away larger use cases. The current infrastructure has been tested with 10GB files, so possibly they can raise the file size limit per community or for the whole of Zenodo if needed. Larger files are allowed on demand. Since they target the long-tail of science, they want public user uploads to always be free.

***Data safety***

The data is stored in CERN Data Center. Both data files and metadata are kept in multiple online replicas and are backed up to tape every night. CERN has considerable knowledge and experience in building and operating large scale digital repositories and a commitment to maintain this data centre to collect and store 100s of PBs of LHC data as it grows over the next 20 years. In the highly unlikely event that Zenodo will have to close operations, they guarantee that they will migrate all content to other suitable repositories, and since all uploads have DOIs, all citations and links to Zenodo resources (such as data) will not be affected.

***Open and closed uploads***

Zenodo is a strong supporter of open data in all its forms (meaning data that anyone is free to use, reuse, and redistribute) and takes an incentives approach to encourage depositing under an open license. They therefore only display Open Access uploads on the front-page. Closed Access upload is still discoverable through search queries, its DOI, and any community collections where it is included.

Since there isn't a unique way of licensing openly and nor a consensus on the practice of adding attribution restrictions, they accept data under a variety of licenses in order to be inclusive. However, they take an active lead in signaling the extra benefits of the most open licenses, in terms of visibility and credit, and offer additional services and upload quotas on such data to encourage using them. This follows naturally from the publications policy of the OpenAIRE initiative, which has been supporting Open Access throughout, but since it aims to gather all European Commission/European Research Area research results, it allows submission of material that is not yet Open Access.

***Future funding for Zenodo***

Zenodo was launched within the OpenAIREplus project as part of a Europe-wide research infrastructure. OpenAIREplus deliver a sustainability plan for this infrastructure with an eye towards future Horizon 2020 projects and is thus one of our possible funding sources. Another possible source of funding is CERN itself. CERN hosts and develops several large services, such as CERN Document

Server and INSPIRE-HEP, which run the same software as Zenodo. Additionally, CERN is familiar with preserving large research datasets because of managing the Large Hadron Collider data archive of 100 petabytes.

*Information of this section was collected from official OpenAIRE and Zenodo web sites.*

### 1.2.2 Benchmarks

Although the two use cases provided by the partners will guide the research we will do during ANTAREX, we plan to further test the developed methodologies, techniques and tool flows using **open source benchmarks** (e.g., Table 1). We will make available the configurations needed to execute the benchmarks, as well as the obtained results and the information needed to reproduce the experiments (e.g., execution times, memory accesses, profiling and characteristics of the machines where the tests run).

**Table 1. Set of possible benchmarks to be used to validate and test ANTAREX**

| Benchmark | Type | URL |
|---|---|---|
| CORAL | HPC | asc.llnl.gov/CORAL-benchmarks |
| HPL | HPC | icl.eecs.utk.edu/hpl |
| HPCG | HPC | www.hpcg-benchmark.org |
| Green Graph 500 | HPC | green.graph500.org |
| ASC | HPC | www.lanl.gov/projects/codesign/proxy-apps/index.php |
| NAS | HPC | www.nas.nasa.gov/publications/npb.html |
| HPCC | HPC | icl.cs.utk.edu/hpcc |
| BSC | HPC | pm.bsc.es/projects/bar |
| PARSEC | HPC | parsec.cs.princeton.edu |
| San Diego Vision | Vision | parallel.ucsd.edu/vision |
| PaRMAT | Graph | github.com/farkhor/PaRMAT |
| Stanford SNAP | Graph | snap.stanford.edu/data |

## *1.3 Private Data Management Policies*

This Section describes the facilities and policies to be used for ANTAREX Project by each partner manage private data. For the two industrial partners, DOMPE’ and SYGIC, the private data will be managed by the two supercomputing centers, CINECA and IT4I respectively, according to the next sections.

The **Primary Sygic contact,** Radim Cmar, is the physical person responsible for the ANTAREX project for Sygic and for approving other Sygic user accesses to the project. He is also the representative for Sygic for the data management process.

The **Primary Dompe' contact,** Andrea Beccari, is the physical person responsible for the ANTAREX project for Dompe' and for approving other Dompe' user accesses to the project. He is also the representative for Dompe' for the data management process.

### 1.3.1 IT4I Data Management Policies

#### *Human roles and administration process*

**IT4Innovations System Administrators** are full-time internal employees of IT4Innovations, department of Supercomputing Services. The system administrators are responsible for safe and efficient operation of the computer hardware installed at IT4Innovations. Administrators have signed a confidentiality agreement.

User access to IT4Innovations supercomputing services is based on projects, membership in a project provides access to the granted computing resources (accounted in corehours consumed). There will be one common project for ANTAREX.

The project will have one **Primary Investigator,** a physical person, who will be responsible for the project, and is responsible for approving other users access to the project. At the beginning of the project, Primary Investigator will appoint one Company Representative for each company involved in the project.

**Company Representatives** will be responsible for approving access to **Private Storage Areas** belonging to their company. Private Storage Areas are designated for storing sensitive private data. Granting access permissions to a Private Storage area must be always authorized by the respective Company Representative AND Primary Investigator.

**Users** are physical persons participating in the project. Membership of users to ANTAREX project is authorized by Primary Investigator. Users can log in to IT4Innovations compute cluster, consume computing time and access shared project storage areas. Their access to Private Storage Areas is limited by permissions granted by Company Representatives.

User data in general can be accessed by:

1. IT4Innovations System Administrators
2. The user, who created them (i.e. the UNIX owner)
3. Other users, to whom the user has granted permission *and at the same time* have access to the particular Private Storage Area (in the case of data stored in the Private Storage Area) granted via the “Process of granting of access permissions” process.

### *Process of granting of access permissions*

All communication with participating parties is in the manner of signed email messages, digitally signed by a cryptographic certificate issued by a trusted Certification Authority. All requests for administrative tasks must be sent to IT4Innovations HelpDesk. All communication with HelpDesk is archived and can be later reviewed.

Access permissions for files and folder within the standard storage areas (HOME, SCRATCH) can be changed directly by the owner of the file/folder by respective Linux system commands. The user can request HelpDesk for assistance on how to set the permissions.

Access to Private Storage Areas is governed by the following process:

1. A request for access to Private Storage Area for given user is sent to IT4Innovations HelpDesk via a signed email message by a user participating in the project.
2. HelpDesk verifies the identity of the user by validating the cryptographic signature of the message.
3. HelpDesk sends a digitally signed message with request of approval to the respective Company Representative and to the Primary Investigator.
4. Both the Company Representative and the Primary Investigator must reply with a digitally signed message with explicit approval of the access to the requested Private Storage Area.
5. System administrator at HelpDesk grants the requested access permission to the user.

Company representative or Primary Investigator can also send a request to HelpDesk to revoke access permission for a user.

### *Data storage areas*

There are four types of relevant storage areas: **HOME, SCRATCH**, **BACKUP and PRIVATE. HOME, SCRATCH and BACKUP** are standard storage areas provided to all users of IT4Innovations supercomputing resources (file permissions apply). **HOME** storage is designed for long-term storage of data and is archived on the tape library **- BACKUP**. **SCRATCH** is a fast storage for short- or mid-term data, with no backups. **PRIVATE** storages are dedicated storages for sensitive data, stored outside the standard storage areas.

### *HOME storage*

HOME is implemented as a two-tier storage. First tier is disk array and the second tier is a NL-SAS disk array together with a partition of T950B tape library. Migration between the two tiers is provided by SGI DMF software. DMF creates two copies of data migrated to the second tier: one to NL-SAS drives and the second on LTO6 tapes for backup.

HOME is realized on CXFS file system by SGI. Access to this file system on the cluster is provided by three CXFS Edge servers and a pair of DMF/CXFS Metadata servers, which export the file system via NFS protocol.

Each user has a designated home directory on the HOME file system at /home/username, where username is login name given to the user. By default, the permissions of the home directory are set to 750, and thus it is not accessible by other users.

### *SCRATCH storage*

SCRATCH is running on parallel Lustre filesystem with fast access. SCRATCH filesystem is divided into two areas: WORK and TEMP.

1. WORK filesystem. Users may create subdirectories and files in directories **/scratch/work/user/username** and **/scratch/work/project/projectid.** The /scratch/work/user/username is private to user, much like the home directory. The /scratch/work/project/projectid is accessible to all users involved in project projectid.
2. TEMP area. In this area, files that are not accessed for more than 90 days will be automatically deleted. Users may freely create directories in this area, and are fully responsible for setting correct access permissions of the directories.

### *PRIVATE storage*

In order to provide additional level of security of sensitive data, we will setup dedicated storage areas for each company participating in the project. PRIVATE storage areas will be setup in a separate storage and will be not accessible to regular IT4Innovation users. IT4Innovations can additionally provide encryption of PRIVATE storage; the particular solution will be discussed with regards to security and performance considerations.

### *BACKUP storage*

Contents of HOME storage are automatically backed up to tape library. There is a minimal period of retention, but no maximal, so we cannot guarantee time when the backups are removed from the tapes.

### *PRIVATE BACKUP storage*

It is possible to setup dedicated backups of PRIVATE storage. In this case, unlike with the regular BACKUP, we can guarantee secure removal of data archived in PRIVATE BACKUP.

### *Data access*

### *Physical security*

All data storage is placed in a single room, which is physically separated from the rest of the building, has a single entry door and no windows. Entry to the room is secured by electromechanical locks controlled by access cards with PINs and non-stop alarm system. The room is connected to CCTV system monitored at reception with 20 cameras, recording and backup. Reception of the building has 24/7 human presence and external security guard during night. Reception has a panic button to call a security agency.

### *Remote access and electronic security*

All external access to IT4I resources is provided only through encrypted data channels (SSH, SFTP, SCP and Cisco VPN)

Control of permissions on the operating system level is done via standard Linux facilities – classical UNIX permissions (read, write, execute granted for user, group or others) and Extended ACL mechanism (for a more fine-grained control of permissions to specific users and groups). PRIVATE storage will have another level of security that will not allow mounting the storage to non-authorized persons.

### *Data lifecycle*

1. **Transfer of data to IT4Innovations:** User transfers data from his facility to IT4Innovations only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
2. **Data within IT4Innovations:** Once the data are at IT4Innovations data storage, access permissions apply.

3. **Transfer of data from IT4Innovations:** User transfers data from to facility from IT4Innovations only via safely encrypted and authenticated channels (SFTP, SCP). Users are strongly advised not to initiate unencrypted data transfer channels (such as HTTP or FTP) to remote machines.
4. **Removal of data:** On SCRATCH file system, the files are immediately removed upon user request. However, the HOME system has a tape backup, and the copies are kept for indefinite time. We advise not to use HOME storage if you do not wish to keep copies of your data on tapes. PRIVATE storage will be securely deleted upon request or when the project ends.

***Data in a computational job lifecycle***

When a user wants to perform a computational job on the supercomputer the following procedure is applied:

1. User submits a request for computational resources to the job scheduler
2. When the resources become available, the nodes are allocated exclusively for the requesting user and no other user can login during the duration of the computational job. The job is running with same permissions to data as the user who submitted it.
3. After the job finishes, all user processes are terminated and all user data is removed from local disks (including ramdisks).
4. After the cleanup is done, the nodes can be allocated to another user, no data from the previous user are retained on the nodes.

All Salomon computational nodes are diskless and cannot retain any data.

There is a special SMP server UV1 accessible via separate job queue, which has different behavior from regular computational nodes: it has a local hard drive installed and multiple users may access it simultaneously.

### 1.3.2 CINECA Data Management Policies

***Human roles and administration process***

**CINECA HPC System Administrators** are full-time internal employees of CINECA, department of DSET (System&Technology Dept). The system administrators are responsible for safe and efficient operation of the HPC computer hardware installed at CINECA. Administrators have signed a confidentiality agreement.

User access to CINECA supercomputing services is based on personal Username/password information (for system access) and Projects (for resource allocation).

Membership in a project provides access to the granted computing resources (accounted in core-hours consumed in the batch mode interactive use is not accounted) as well as to a private storage area ($WORK) reserved to the members of the project.

Projects are hierarchically grouped into “root entities”, even if each single sub-project is completely autonomous in terms of PI, budget, private storage area and collaborators.

There will be several sub-projects for ANTAREX, one for each Company involved, all of them grouped into a single root project “Antrx_”.

Each sub-project will have one **Principal Investigator,** a physical person representative for the corresponding Company, who will be responsible for the project, and is responsible for approving other users access to the project.

The collaborators of each sub-project will have exclusive access to the WORK area, a p**rivate Storage Areas** associated to the project itself. The WORK area is designated for storing sensitive private data. It is a permanent area maintained for the full duration of the project.

**Users** are physical persons participating in the project. Users must register to the CINECA Database of Users (UserDB) following the normal CINECA Policy for users. They will be given a "personal username" and password that will permit the access to CINECA supercomputing platforms.

General users will become members of the ANTAREX project only when they will be associated to one or more ANTAREX sub-projects by one of the Principal Investigators. Only at this point, users shall be allowed to log into the compute cluster, consume computing resources and access the project private storage areas.

Several data areas are available on our systems:

1. Personal storage areas (HOME and SCRATCH): each user owns such areas on the system
2. Project private storage area (WORK): each project owns such area opened to all (and only) project collaborators
3. Data Resources (DRES): private data areas owned by a physical person (DRES owner) who can share it with collaborators or even projects (all collaborators of the project)

User data in general can be accessed by:

1. System Administrators and help-desk consultants
2. The user, who created them (i.e. the UNIX owner)
3. Other users, to whom the user has granted permission for personal data areas
4. *Other collaborators of the same project, to whom the user has granted permission, for the WORK* or DRES Private Storage Area.

### *Process of granting of access permissions*

All communication with participating parties is in the manner of signed email messages, digitally signed by a cryptographic certificate issued by a trusted Certification Authority.

All requests for administrative tasks must be sent to Cineca HelpDesk (superc@cineca.it). All communication with HelpDesk is archived in a Trouble Ticketing system and can be later reviewed.

Access permissions for files and folder within the personal storage areas (HOME, SCRATCH) can be changed directly by the owner of the file/folder by respective Linux system commands. The user can request HelpDesk for assistance on how to set the permissions.

Access to Private Storage Areas is exclusively reserved to the collaborators of the sub-project. In order to access it the user must be included among the project collaborators by the PI of the project. The PI is also allowed to remove collaborators from its project.

### *Data storage areas*

There are several types of relevant storage areas: **HOME, SCRATCH**, **TAPE** (user oriented), **WORK** and **DRES** (project oriented).

**HOME, SCRATCH and TAPE** are standard storage areas provided to all users of supercomputing resources (file permissions apply). **HOME** storage is designed for long-term storage of data and is archived on the tape library (a disk quota applies); **SCRATCH** is a fast storage for short- or mid-term data, with no backups and periodic data cleaning (no disk quota). **TAPE** storages are dedicated to personal archiving to the tape library (disk quota applies).

**WORK** is a storage area for sensitive data, provided for each project, disk quota applies, only project collaborator can access it, data are preserved for the full duration of the project.

**DRES** is similar to WORK, but provided only on specific request and can be associated to multiple projects.

All storage areas in the CINECA HPC environment are managed by GPFS (General Parallel File System). The Tape library is connected to the data storage by the LTFS technology.

### *HOME storage*

Each user has a designated home directory on the HOME file system at /<host>/userexternal/<username>, where <host> is the system name (GALILEO, FERMI or PICO) and <username> is login name given to the user. By default, the permissions of the home directory are set to 700, and thus it is not accessible by other users. The user is however free to open the permissions giving access to others to its own files.

There is a disk quota on this filesystem of 50 GB that can be extended on request. The filesystem is daily saved to Magnetic tapes by backup. Data here are preserved as long as the user is defined on the system.

### *SCRATCH storage*

SCRATCH is given to each user, though the $CINECA_SCRATCH environmental variable. No quota applies to this filesystem and the occupancy is regularly checked by HD staff not to overcome a given threshold. By default the permission are set to 755, that is, open in read access to all. The user is however free to modify the permissions closing the access. In this area a cleaning procedure is active, deleting all files that are not accessed for more than 30 days.

### *TAPE storage*

This area is given to a user on request and is reachable thought the $TAPE environment variable. Data stored here migrates automatically to magnetic tapes thanks to the LTFS system. A default quota of 1TB applies, even if this limit can be increased on request. Data here are preserved as long as the user is defined on the system.

### *WORK storage*

This area is given to each project active on the system and is reachable via the $WORK environment variable. If the user participates to more than one project he will be entitled to more than one WORK area; he will choose among them using a specific command (chprj – Change Project). A default quota of 1 TB applies, but the value can be increased on request. Access here is strictly reserved to project's collaborators and it is not possible to open this area to others. Data here are preserved as long as the project is defined on the system.

### *DRES storage*

This area can be created only on request and is stored on the gss (GPFS Storage System) disks. It is owned by a user (DRES owner) and it is characterized by a quota, a validity and a type (FS – normal Filesystem; ARCH – tape storage; REPO – iRods based repository).

This area is reachable from all HPC systems in CINECA (at least from the login nodes) and can be linked to one or more projects. In this case all collaborators of the projects are entitled to access the storage area. Data here are preserved as long as the DRES itself is defined on the system.

### *Data access*

#### *Physical security*

All data storage is placed in a single room, one of the two machine rooms of CINECA. Entry to the room is secured by electromechanical locks controlled by access cards with PINs and non-stop alarm system. The room is connected to CCTV system monitored at reception with dozens of cameras, recording and backup.

Reception of the building has 24/7 human presence, staff during working hours and external security guards during nights and week-ends.

#### *Remote access and electronic security*

All external access to cineca resources is provided only through encrypted data channels (SSH, SFTP, SCP and Cisco VPN)

Control of permissions on the operating system level is done via standard Linux facilities – classical UNIX permissions (read, write, execute granted for user, group or others) and Extended ACL mechanism (for a more fine-grained control of permissions to specific users and groups).

### *Data lifecycle*

1. **Transfer of data to CINECA**
   User transfers data from his facility to CINECA only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.

2. **Data within CINECA**
   Once the data are at CINECA data storage, access permissions apply.

3. **Transfer of data from CINECA**
   User transfers data from **CINECA** to local facility only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.

4. **Removal of data**
   Normally the files are immediately removed upon user request. However, the HOME system has a tape backup, and the copies are kept for indefinite time.

   Data on HOME and TAPE have a life cycle related with the life of the Username (removed one year after the removal of the Username).

   Data on SCATCH will be preserved only one month in not used on a daily bases

   Data on WORK follows the life of the project (removed six months after the conclusion of the project).

   Data on DRES follows the life of the DRES itself (removed six months after the conclusion of the DRES).

### *Data in a computational job lifecycle*

When a user wants to perform a computational job on the supercomputer the following procedure is applied:

- User submits a request for computational resources to the job scheduler, specifying the project to be accounted for.

- When the resources become available, the cores are allocated exclusively for the requesting user. Other jobs con share the nodes, if they are not requested in an exclusive way. The job is running with same permissions to data as the user who submitted it.
- The job should only use the gpfs storage filesystems. Even when local disks are present, they are not guaranteed.
- After the job finishes, all user processes are terminated and the resources can be allocated to another job, no control about data from the previous user written on local disks.

### 1.3.3 POLIMI Data Management Policies

***Human roles and administration process***

The **Project Coordinator**, Prof. Cristina Silvano, is the physical person responsible for the ANTAREX project and for approving other users access to the project. The Project Coordinator is also the **Representative** for POLIMI for the data management process.

**Users** are physical persons participating in the project. Membership of users to ANTAREX project is authorized by Project Coordinator. Users can log in to the computer hardware dedicated to the ANTAREX project at POLIMI and access the shared project storage areas. Access to POLIMI resources is available to POLIMI users, as well as to users from other parties upon request from the party Representative, and following authorization by the Project Coordinator.

**System Administrators** are members of the POLIMI staff involved in the ANTAREX project, since the computer hardware resources used for the ANTAREX project at POLIMI are dedicated, and not shared with general POLIMI scientific or IT personnel.

User data in general can be accessed by:

- The user who created them (i.e., the UNIX owner)
- System Administrators
- Other users who have been granted permission by the owner

***Process of granting of access permissions***

Access permission requests for POLIMI resources should be sent via registered mail signed by the party Representative. Such communication will be archived for the duration of the project plus 5 years.

Access permissions for files and folder within the standard storage areas (HOME) can be changed directly by the owner of the file/folder by respective Linux system commands.

***Data storage areas***

***HOME storage***

HOME is implemented via two 2TB SATA Western Digital Black disks in RAID-1 mirror.

Each user has a designated home directory on the HOME file system at /home/username, where username is login name given to the user. By default, the permissions of the home directory are set to 700, and thus it is not accessible by other users.

***SWAP storage***

SWAP storage is implemented via a 120 GB Samsung 150 Evo solid state disk, mounted as a Linux swap partition.

***Data access***

***Physical security***

All data storage is placed in a single cabinet, in a room physically separated from the rest of the building. Entry to the room is secured by electromechanical locks controlled by access cards. An alarm system is active when no personnel is present.

***Remote access and electronic security***

All external access to POLIMI ANTAREX resources is provided only through encrypted data channels (SSH, SFTP, SCP).

Control of permissions on the operating system level is done via standard Linux facilities – classical UNIX permissions (read, write, execute granted for user, group or others) and Extended ACL mechanism (for a more fine-grained control of permissions to specific users and groups).

***Data lifecycle***

- Transfer of data: User transfers data from his facility to POLIMI only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
- Data within POLIMI: Once the data are at POLIMI data storage, access permissions apply.
- Transfer of data from POLIMI: User transfers data from to facility from POLIMI only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
- Removal of data: On HOME file system, the files are immediately removed upon user request, or after 2-years from the project end. The SWAP storage is managed as a standard swap partition, and no long-term storage takes place there.

### 1.3.4 UPORTO Data Management Policies

***Human roles and administration process***

The **Project Coordinator**, Prof. Cristina Silvano, is the physical person responsible for the ANTAREX project and for approving other users access to the project. The **Representative** for UPORTO for the data management process is Dr. João Bispo.

**Users** are physical persons. Membership of users to ANTAREX project is authorized by Project Coordinator. Users can log in to services hosted at UPORTO and access the shared project storage areas. Access to ANTAREX resources is available upon request from the party Representative, and following authorization by the Project Coordinator.

**System Administrators** are part of the ANTAREX consortium.

User data in general can be accessed by:

- The user who created them (i.e., the account owner)
- System Administrators
- Other users who have been granted permission by the owner

***Services provided by UPORTO***

***ANTAREX OwnCloud***

OwnCloud is a self-hosted Dropbox-like solution for private file storage. It is used in the project as a repository to store files related to the project (e.g., reports, publications, dissemination materials). We use a free version of OwnCloud as the repository server. It is an open platform which can be accessed through a web interface or a sync client (available for desktop and mobile platforms). Members of the ANTAREX Consortium can access the repository files using accounts, previously created by a system administrator. It is possible to create public links to individual files of the repository, which can later be used to share files publicly in the website.

### *ANTAREX Wiki*

We setup a self-hosted wiki in order to facilitate the communication of knowledge between the members as well as to aid in a multitude of collaborative tasks. This wiki is based on the Detritus release of DokuWiki. The wiki is closed to the general public, meaning that even reading the wiki is not possible for someone that is not logged in. In order to keep the wiki private, new user accounts are created on demand by the system administrators. The wiki provides a way to discuss subjects and to work in a collaborative way in some topics.

### *ANTAREX Website*

The ANTAREX website is hosted externally, by the Portuguese company AMEN, which is part of the European company DADA S.p.A. The hosting service for the website also supports the mailing lists and official project emails. The hosting is done over a Linux, using Apache as the HTTP server. The code of the website is being developed in a private Git repository hosted by BitBucket, which is responsible for maintaining a backup of the data, ensuring the integrity of the website.

Having the website hosted externally is more secure, since we avoid possible attack vectors related with website hosting. Since all data published in the website is public, there is no problem in hosting it externally. All public documents are accessed through links hosted by the self-hosted OwnCloud repository and are not stored in the website.

### *Data access*

### *Physical security*

All data storage is hosted on virtual machines provided by UPORTO. The physical machines are placed in dedicated rooms and entry to the room is secured by electromechanical locks controlled by access cards. Users do not have access to these rooms.

### *Remote access and electronic security*

All external access to ANTAREX resources hosted by UPORTO is provided through secure data channels (e.g., HTTPS).

### *Process of granting of access permissions*

Access permissions for files and folder within the repository and the wiki is controlled by system administrators.

### *Data lifecycle*

- **Transfer of data:** User transfers data from his facility to UPORTO via safely encrypted and authenticated channels (HTTPS).
- **Data within UPORTO:**Once the data is at UPORTO repository/wiki, access permissions apply.
- **Transfer of data from UPORTO:** User transfers data to facility from UPORTO via safely encrypted and authenticated channels (HTTPS).
- **Removal of data:** The virtual machine is included in a system of daily backups to hard-disk and bi-weekly backups to tapes, to ensure the integrity of data. They will be maintained for

at least 3 years after the end of the project. The website host and domain will be available for two years after the end of the project. After the end of the project, the website will be moved to a machine at UPORTO.

### 1.3.5 ETHZ Data Management Policies

***Human roles and administration process***

The **Project Coordinator**, Prof. Cristina Silvano, is the physical person responsible for the ANTAREX project and for approving other users access to the project. The **Representative** for ETHZ for the data management process is Prof. Luca Benini.

Users are physical persons participating in the project. Membership of users to ANTAREX project is authorized by Project Coordinator. Users can log in to the computer hardware dedicated to the ANTAREX project at ETHZ and access the shared project storage areas. Access to ETHZ resources is available to ETHZ users, as well as to users from other parties upon request from the party Representative, and following authorization by the Project Coordinator.

**System Administrators** are members of the ETH Zurich and Integrated System Laboratory staff.

User data in general can be accessed by:

- The user who created them (i.e., the account owner)
- System Administrators
- Other users who have been granted permission by the owner

***Data storage areas***

***HOME storage***

HOME is stored remotely in a shared data center of ETH Zurich in a physically separated machine ZFS and backup regularly on RAID-6 and tape.

Each user has a designated home directory on the HOME file system at /home/username, where username is login name given to the user. By default, the permissions of the home directory are set to 755. Thus are visible from other users at institute level.

***SCRATCH storage***

Scratch storage is local on user's workstations and shared servers using SSD disks.

***Data access***

***Physical security***

All data storage as well as servers and user's workstations are part of a virtual private network (VPN) at institute level.

***Remote access and electronic security***

All external access to ETHZ ANTAREX resources is provided only through encrypted data channels (SSH, SFTP, SCP).

Control of permissions on the operating system level is done via standard Linux facilities – classical UNIX permissions (read, write, execute granted for user, group or others) and Extended ACL mechanism (for a more fine-grained control of permissions to specific users and groups).

***Data lifecycle***

1. **Transfer of data** User transfers data from his facility to ETHZ only via safely encrypted and

authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.

2. **Data within ETHZ:** Once the data is at ETHZ data storage, access permissions apply.
3. **Transfer of data from ETHZ:** User transfers data from to facility from ETHZ only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
4. **Removal of data:** On HOME file system, the files are immediately removed upon user request. At project end data are archived and preserved. The SCRATCH storage is managed as a standard scratch partition, and no long-term storage takes place there.

### 1.3.6 INRIA Data Management Policies

***Human roles and administration process***

The **Project Coordinator**, Prof. Cristina Silvano, is the physical person responsible for the ANTAREX project and for approving other users access to the project. The **Representative** for Inria for the data management process is Dr. Erven Rohou.

**Users** are physical persons. Membership of users to ANTAREX project is authorized by Project Coordinator. Users can log in to the computer hardware at Inria and access the shared project storage areas. Access to ANTAREX resources is available to Inria users, as well as to users from other parties upon request from the party Representative, and following authorization by the Project Coordinator.

**System Administrators** are members of the Inria staff.

User data in general can be accessed by:

- The user who created them (i.e., the UNIX owner)
- System Administrators
- Other users who have been granted permission by the owner

***Data storage areas***

***Inria Forge***

Inria Forge is a service offered to facilitate the scientific collaborations of people working at Inria. It offers easy access to revision control systems, mailing lists, bug tracking, message boards/forums, task management, site hosting, permanent file archival, full backups, and total web-based administration. The objective is to provide everyone working at the institute with an infrastructure for their scientific collaborations with internal and/or external partners.

***HOME storage***

HOME is implemented as a shared Network File System (NFS), mounted from user machines. Users do not have admin privilege on the machines where a NFS volume is mounted.

Each user has a designated home directory on the HOME file system at /udd/username, where username is login name given to the user. By default, the permissions of the home directory can be set to 700, and thus not accessible by other users.

***Data access***

***Physical security***

All data storage is placed in dedicated rooms physically separated from the rest of the building. Entry to the room is secured by electromechanical locks controlled by access cards. Users do not have access to these rooms, only System Administrators do. Inria Rennes has 24/7 on-site security

### *Remote access and electronic security*

All external access to Inria ANTAREX resources is provided only through encrypted data channels (SSH, SFTP, SCP).

### *Process of granting of access permissions*

Access permission requests for Inria resources should be sent via registered mail signed by the party Representative. Such communication will be archived for the duration of the project plus 5 years.

Access permissions for files and folder within the standard storage areas (HOME) can be changed directly by the owner of the file/folder by respective Linux system commands.

Control of permissions on the operating system level is done via standard Linux facilities – classical UNIX permissions (read, write, execute granted for user, group or others). Access to the data in the Inria Forge is based on Extended ACL mechanism (for a more fine-grained control of permissions to specific users and groups).

### *Data lifecycle*

- **Transfer of data:** User transfers data from his facility to Inria only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
- **Data within Inria:** Once the data are at Inria data storage, access permissions apply.
- **Transfer of data from Inria:** User transfers data  to facility from Inria only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
- **Removal of data:** On HOME file system, the files are immediately removed upon user request, or after 2 years from the project end.

## 1.4 Data Management Plan Template

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.

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | Identifier for the data set to be produced<br>DOI |
| 2 | **Data set description** | Description of the data that will be generated or collected, 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. |
| 3 | **Standards and metadata** | Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created. |
| 4 | **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.).<br>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). |
| 5 | **Archiving and preservation (including storage and backup)** | Description of the procedures that will be put in place for long-term preservation of the data. Indication of 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. |

### 1.4.1 Partner: IT4I Data Table 1

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **Graph500 benchmark results**<br>DOI from service defined in Section Public Data Management Policies. |
| 2 | **Data set description** | Graph 500 is an HPC benchmark, which emphasizes the speed of memory access instead of the speed of arithmetical operations like other widely used benchmarks such as Top 500. The main idea behind Graph 500 is to measure the number of traversed edges per second (TEPS) using the Breadth First Search (BFS) algorithm on artificially generated graph. During the testing TEPS and time are collected in 64 runs. The resulting data set then contains information about the problem size and aggregated performance results from all 64 runs.<br>The result will be used for assessing effectivity and usability of ANTAREX technologies developed within WP2 and WP3. |
| 3 | **Standards and metadata** | The detailed description of Graph 500 output standard can be found at http://www.graph500.org/specifications |
| 4 | **Data sharing** | Data sharing will follow rules of selected service defined in Section Public Data Management Policies. |
| 5 | **Archiving and preservation (including storage and backup)** | Archiving and preservation will follow rules of selected service defined in Section Public Data Management Policies. |

### 1.4.2 Partner: IT4I Data Table 2

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **Benchmark dataset for betweenness centrality**<br>DOI from service defined in Section Public Data Management Policies. |
| 2 | **Data set description** | Betweeness centrality is a measure of graph vertices indicating how well is a particular graph node connected to other nodes. It is useful for determining important nodes of a network. Importance of a node depends not only on its degree but also on weight of its adjacent edges. The edges can be weighted by various values such as distance, average speed, type of the road, etc. Removal of these nodes would result in severe degradation of flow throughput in the network. We will use the computed betweenness for traffic routing optimization on road networks. The result will be used for assessing efectivity and usability of the ANTAREX technologies developed within WP2 and WP3.<br>Input data of the benchmark will be collected from OpenStreetMap data and preprocessed to suit the needs of the benchmark. Several graphs will be obtained, each having different properties (graph size, node density, etc.).<br>Output of the benchmark will consist of values of given performance metrics and will be stored for evaluation. The gathered performance metrics can serve as baseline for future improvements and optimizations of the developed toolset. |
| 3 | **Standards and metadata** | The OpenStreetMap data are obtained from volunteers contributing to the project in form of a results of their own geographical surveys. The data are managed by non-profit organization OpenStreetMap Foundation based in UK. The OpenStreetMap data are available under ODC Open Database Lincense<br>(http://opendatacommons.org/licenses/odbl/1.0/).<br>We will use publicly available export of the map data in the form of a binary file encoded in the Protocol Buffers binary format<br>(http://planet.openstreetmap.org). |
| 4 | **Data sharing** | Data sharing will follow rules of selected service defined in Section Public Data Management Policies. |
| 5 | **Archiving and preservation (including storage and backup)** | Archiving and preservation will follow rules of selected service defined in Section Public Data Management Policies. |

### 1.4.3 Partner: IT4I Data Table 3

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **Benchmark of Time dependent routing algorithm**<br>DOI from service defined in Section Public Data Management Policies. |
| 2 | **Data set description** | Time dependent routing is an extension of a standard vehicle routing task. In ordinary routing problem, the routes are determined from a static unweighted graph representation of a road network based on given points of origin and destination. The resulting route is always the same between two given points. The route determined by the time dependent algorithm for the same two origin and destination points can vary in time. For example, it is more beneficial for some days of the week to take a detour from the standard route to avoid the morning commute and minimize the risk of possible delays.<br>The time dependent algorithm works with routes extracted from graph representation of the road network where the edges hold additional metadata about the road network throughput and state for a given timeframe.<br>The input dataset for the benchmark will contain pre-defined set of routes computed for a given set of simulated pairs of origin and destination points and generated speed profiles. The original algorithm will be optimized by ANTAREX technologies developed within WP2 and WP3 and executed multiple times under different conditions. Various metrics of effectivity and profiling data will be collected during each run of the algorithm and stored for future the analysis. |
| 3 | **Standards and metadata** | Time dependent routing algorithm is described in the following publications:<br>Tomis R., Rapant L., Martinovič, J., Slaninová K. & Vondrák I., Probabilistic Time-Dependent Travel Time Computation using Monte Carlo Simulation, accepted to HPCSE 2015.<br>Tomis, R., Martinovič, J., Slaninová, K., Rapant, L., & Vondrák, I., Time-Dependent Route Planning for the Highways in the Czech Republic. In Lecture Notes in Computer Science, 9339, pp. 145-153, 2015. |
| 4 | **Data sharing** | Data sharing will follow rules of selected service defined in Section Public Data Management Policies. |
| 5 | **Archiving and preservation (including storage and backup)** | Archiving and preservation will follow rules of selected service defined in Section Public Data Management Policies. |

### 1.4.4 Partner: IT4I. Data Table 4

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **Data used and created within UC2**<br>DOI from service defined in Section Public Data Management Policies. |
| 2 | **Data set description** | Due to the private nature of UC2, the data set description will be included into private deliverables of UC2. |
| 3 | **Standards and metadata** | Due to the private nature of UC2, the description of standards and metadata will be included into private deliverables of UC2. |
| 4 | **Data sharing** | Selected data will be privately available to selected ANTAREX participants. |
| 5 | **Archiving and preservation (including storage and backup)** | All the data collections created, maintained and processed within UC2 by IT4I and Sygic will be preserved, stored, and maintained following the rules defined in Section IT4I Data Management Policies. |

### 1.4.6 Partner: CINECA. Data Table 5

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **GALILEO-HPL:** Galileo HPL benchmark for Top500.<br>DOI from OpenAIRE/Zenodo service. |
| 2 | **Data set description** | This dataset collect the data stored during the procedure of evaluation of the Galileo machine at CINECA in order to classify it for Top500 list.<br>Dataset also include a report summarizing the results of benchmarks (STREAM for single node memory assessment and HPL for HPC parallel performance) carried out in May 2015. |
| 3 | **Standards and metadata** | Dataset is made up of ASCII files (Unix format) assembled as a tar gzipped archive.<br>Full metadata description are provided within the standard dataset creation in OpenAIRE/Zenodo service.<br>**Keywords:** Galileo; CINECA; TOP500; HPL; STREAM; HPC; benchmarks. |
| 4 | **Data sharing** | GALILEO-HPL dataset is and will be PUBLIC.<br>Access is guaranteed by OpenAIRE/Zenodo service and is widely open to public without any restriction.<br>GALILEO-HPL dataset is provided through the OpenAIRE/Zenodo web interface to end-user and no additional software is necessary for its dissemination and sharing.<br>GALILEO-HPL dataset is indexed within OpenAIRE and exposed to external end-user via standard OpenAIRE retrieval tools like those available within the Zenodo software. |
| 5 | **Archiving and preservation (including storage and backup)** | Storage persistence in OpenAIRE/Zenodo service is guaranteed for unlimited time. |

### 1.4.7 Partner: CINECA. Data Table 6

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **GALILEO-HPCG:** Galileo HPCG benchmark for assessing Galileo machine performance on hybrid configuration.<br>DOI from OpenAIRE/Zenodo service. |
| 2 | **Data set description** | This dataset collect the data stored during the procedure of evaluation of the Galileo machine at CINECA when using Xeon Phi and K80 GPU as numerical coprocessors.<br>Dataset also include a report summarizing the results of benchmarks carried out in May 2015. |
| 3 | **Standards and metadata** | Dataset is made up of ASCII files (Unix format) assembled as a tar gzipped archive.<br>Full metadata description are provided within the standard dataset creation in OpenAIRE/Zenodo service.<br>**Keywords:** Galileo; CINECA; HPCG; XeonPhi; K80GPU; HPC; benchmarks. |
| 4 | **Data sharing** | GALILEO-HPCG dataset is and will be PUBLIC.<br>Access is guaranteed by OpenAIRE/Zenodo service and is widely open to public without any restriction.<br>GALILEO-HPCG dataset is provided through the OpenAIRE/Zenodo web interface to end-user and no additional software is necessary for its dissemination and sharing.<br>GALILEO-HPCG dataset is indexed within OpenAIRE and exposed to external end-user via standard OpenAIRE retrieval tools like those available within the Zenodo software. |
| 5 | **Archiving and preservation (including storage and backup)** | Storage persistence in OpenAIRE/Zenodo service is guaranteed for unlimited time. |

### 1.4.8 Partner: CINECA. Data Table 7

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **LiGen-DOCK:** Dataset of protein receptors and ligands inputs and corresponding docking results.<br>PID from EUDAT B2SHARE service |
| 2 | **Data set description** | This dataset collect the data stored during the procedure of evaluation of UC1 LiGen-DOCK mini-app on the Galileo machine at CINECA.<br>Dataset is made out of a comprehensive input set of protein receptors taken from the Protein Data Bank (PDB) and the largest set of ligand's chemical structures from commercial catalogs like, i.e., Sigma-Aldrich and/or Enamine(1,2).<br>LiGen-Dock dataset will also include the output of the performance evaluation of the UC1 mini-app on performing ligand-receptor docking workflow in various computational scenarios(2).<br>Dataset also include a report summarizing the results of LiGen-DOCK benchmarks. |
| 3 | **Standards and metadata** | Dataset is made up of ASCII files (Unix format) assembled as a tar gzipped archive.<br>Full metadata description are provided within the standard dataset creation in EUDAT B2SHARE.<br>**Keywords:** Galileo; CINECA; LiGen; Docking; PDB; Sigma-Aldrich; Enamine; benchmarks. |
| 4 | **Data sharing** | LiGen-DOCK dataset is and will be PUBLIC.<br>Access is guaranteed by OpenAIRE/Zenodo service and is widely open to public without any restriction.<br>LiGen-DOCK dataset is provided through the OpenAIRE/Zenodo web interface to end-user and no additional software is necessary for its dissemination and sharing.<br>LiGen-DOCK dataset is indexed within OpenAIRE and exposed to external end-user via standard OpenAIRE retrieval tools like those available within the Zenodo software. |
| 5 | **Archiving and preservation (including storage and backup)** | Storage persistence in OpenAIRE/Zenodo service is guaranteed for unlimited time. |

(1) It is expected to select a subset from PDB made out of tens of protein receptors and 1 to 10 million of ligands chemical structures.

(2) Final dimension of dataset will be defined at the second revision of this deliverable at M18.

### 1.4.9 Partner: UPORTO. Data Table 8

| No. | Item | Description |
|---|---|---|
| 1 | **Data set reference and name** | **ANTAREX-DSL:** DSL transformations. DOI from OpenAIRE/Zenodo service. |
| 2 | **Data set description** | Collection of DSL codes used to adapt the set of applications that can be made publicly available, together with the corresponding application code and the transformed code after applying the DSL codes. This dataset represents the output of the first part of the ANTAREX proposed tool-flow, and shall cover the two use cases of the proposal and tested benchmarks. This dataset can be useful as an example of how we are specifying the runtime adaptation and non-functional requirements in the DSL, and the resulting code. The DSL compiler shall be made available (possibly as a web interface) and the dataset will allow any person to validate the results from the DSL transformations and to evaluate and try the DSL compiler. |
| 3 | **Standards and metadata** | Dataset is made up of ASCII files assembled as a zipped archive.<br>Full metadata description are provided within the standard dataset creation in OpenAIRE/Zenodo service.<br>**Keywords:** LARA; DSL; benchmarks. |
| 4 | **Data sharing** | ANTAREX-DSL dataset is and will be PUBLIC.<br>Access is guaranteed by OpenAIRE/Zenodo service and is widely open to public without any restriction.<br>ANTAREX-DSL dataset is provided through the OpenAIRE/Zenodo web interface to end-user and no additional software is necessary for its dissemination and sharing.<br>ANTAREX-DSL dataset is indexed within OpenAIRE and exposed to external end-user via standard OpenAIRE retrieval tools like those available within the Zenodo software. |
| 5 | **Archiving and preservation (including storage and backup)** | Storage persistence in OpenAIRE/Zenodo service is guaranteed for unlimited time. |