

H2020-FETHPC-1-2014 ANTAREX-671623



**AutoTuning and Adaptivity approach for Energy
efficient eXascale HPC systems**

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


**Deliverable D6.10:
Open Research Data Pilot Report**



European
Commission

Horizon 2020
European Union funding
for Research & Innovation

Deliverable Title:	Open Research Data Pilot Report		
Lead beneficiary:	IT4I (Czech Republic)		
Keywords:	ORDP, Public Data, Private Data, Data Collections, Benchmarks		
Author(s):	Katerina Slaninova (IT4I), Jan Martinovic (IT4I), João M.P. Cardoso (UPORTO), Carlo Cavazzoni (CINECA), Federico Ficarelli (CINECA), Andrea Beccari (DOMPE)		
Reviewer(s):	Loic Besnard (CNRS), Cristina Silvano (POLIMI)		
WP:	WP6	Task:	T6.2
Nature:	ORDP report	Dissemination level:	Public
Identifier:	D6.10	Version:	V1.4
Delivery due date:	November 31st, 2018	Actual submission date:	January 10th, 2019

Executive Summary:	<p>ANTAREX project is a part of Open Research Data Pilot in Horizon 2020. Deliverable D1.2 Data Management Plan as one of the activities in Task 1.2 in M06 was defined to specify what data will be generated during 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 deliverable was prepared according to the Guidelines on Data Management in H2020 and Guidelines on Open Access to Scientific Publications and Research Data in H2020. 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.</p> <p>This deliverable D6.10 is the Open Research Data Pilot report type. It consists of the list of data collections and benchmarks made during the projects. The data collections and benchmarks are available in public long-term repositories to allow the long-term access for the verification and their reuse. After the review in M18, it was recommended to create this new deliverable Open Research Data Pilot Report instead of the updated Data Management Plan.</p>
Approved and issued by the Project Coordinator:	Date: January 10th, 2019
	

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	Summary.....	4
2	Outputs from the Project	5
2.1	Partner: IT4I – D1	6
2.2	Partner: IT4I – D2	8
2.3	Partner: IT4I – D3	9
2.4	Partner: IT4I – D4	11
2.5	Partner: IT4I – D5	13
2.6	Partner: IT4I – D6	14
2.7	Partner: SYGIC – D7	15
2.8	Partner: UPORTO – D8	16
2.9	Partner: UPORTO – D9	17
2.10	Partner: UPORTO – D10	18
2.11	Partner: CINECA – D11	19
2.12	Partner: CINECA – D12	20
2.13	Partner: CINECA – D13	21
2.14	Partner: ETHZ – D14	22

1 Summary

This new public deliverable has been added to comply ANTAREX project with the Grant Agreement option article 29.3 for Open Research Data Pilot in Horizon 2020. Deliverable D1.2 Data Management Plan as one of the activities in Task 1.2 in M06 was defined to specify what data will be generated during 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 deliverable was prepared according to the Guidelines on Data Management in H2020 and Guidelines on Open Access to Scientific Publications and Research Data in H2020. 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. After the review in M18, it was recommended to create this new deliverable Open Research Data Pilot Report instead of the updated Data Management Plan.

The deliverable is Open Research Data Pilot type (ORDP). It consists of the list of data collections and benchmarks made during the project. The data collections and benchmarks are available in public long-term repositories to allow the long-term access for the verification and their reuse. A detailed description of the repositories, data management policies of the project partners, and data management maintenance was described in D1.2.

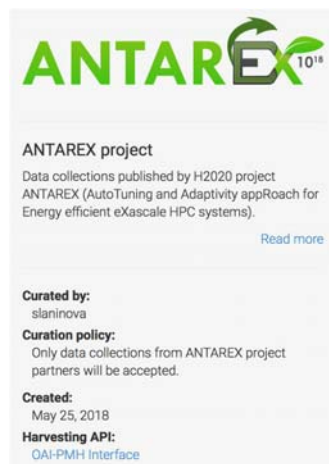
2 Outputs from the Project

The following tables include list of data collections and benchmarks from the ANTAREX project available within Open Research Data Pilot. Each table includes the partner responsible for the output, data set name and description, licence, and the link of the repository the data set is available. The outputs have information how to use/repeat experiments, often with the links to relevant conference, workshop, and journal publications with the description of the ANTAREX tools and experiments published and made available in open access repositories complying with the guidelines of OpenAIRE, the platform funded by EC.

The ANTAREX project made available more outputs than originally planned in D1.2 Data Management Plan. The collections and data benchmarks are archived in Zenodo¹, the long-term repository opened for research outputs developed under the EU FP7 project OpenAIREplus. The ANTAREX community was created at Zenodo repository to link all the data collections and benchmarks to this project. Where appropriate, repositories like EUDAT² or GitHub³ were used (data collection exceeds the limit of Zenodo, or the data collection is closely connected with a software developed).

Also other institutional repositories of Politecnico di Milano (research catalogue RE.PUBLIC@POLIMI⁴) and of IT4Innovations, VSB - Technical University of Ostrava (digital repository DSPACE⁵) are used to make available appropriate documentation or relevant publications. A detailed description of publications relevant to the ANTAREX project is available at D6.6 Dissemination of the Foreground Knowledge: Activities and Plans (Dissemination Plan 3).

The code developed within the ANTAREX project is stored in code repositories according to IPR and its availability, for example GitHub or private code repositories of the partners. A detailed description of code developed during the ANTAREX project is available at D6.9 Final Plan to Use the Foreground Knowledge (Exploitation Plan 3).



¹ ZENODO - <http://www.zenodo.org>

² EUDAT - <https://eudat.eu>

³ GitHub - <https://github.com>

⁴ RE.PUBLIC@POLIMI - <https://www.polimi.it/en/scientific-research/research-at-the-politecnico/publications-and-open-access/>

⁵ DSPACE - <http://dspace.vsb.cz/?locale-attribute=en>

2.1 Partner: IT4I – D1

No.	Item	Description
1	Data set reference and name	D1: Benchmark dataset for betweenness centrality DOI: 10.5281/zenodo.1290209
2	Data set description	<p>Betweenness 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. Outputs from betweenness centrality were used for traffic routing optimization on road networks. The result was used for assessing effectivity and usability of the ANTAREX technologies developed within WP2 and WP3.</p> <p>Input data of the benchmark were collected from OpenStreetMap data and preprocessed to suit the needs of the benchmark. Several graphs were obtained, each having different properties (graph size, node density, etc.).</p> <p>Output of the benchmark consists 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.</p> <p>The dataset includes road network graphs for betweenness centrality algorithm. Weighted graph representation of a road network in selected regions. Derived from Open Street Map https://www.openstreetmap.org. The dataset can be used as input for the betweenness centrality algorithm implemented here: https://github.com/It4innovations/Betweenness.</p>
3	Standards and metadata	<p>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 License (http://opendatacommons.org/licenses/odbl/1.0/).</p> <p>We used publicly available export of the map data in the form of a binary file encoded in the Protocol Buffers binary format (http://planet.openstreetmap.org).</p>
4	Data sharing	<p>Data sharing follows rules of selected service defined in Section Public Data Management Policies.</p> <p>Licence: Creative commons</p>

5	Archiving and preservation (including storage and backup)	Archiving and preservation follow the rules of selected service defined in Section Public Data Management Policies. Data set is available at ZENODO: https://zenodo.org/record/1290209#.XAD1Qy2E4W4
---	--	---

2.2 Partner: IT4I – D2

No.	Item	Description
1	Data set reference and name	D2: Graph500 benchmark results
2	Data set description	<p>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.</p> <p>The result will be used for assessing effectivity and usability of ANTAREX technologies developed within WP2 and WP3.</p>
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. Licence: Free
5	Archiving and preservation (including storage and backup)	Archiving and preservation will follow rules of selected service defined in Section Public Data Management Policies. Link to repository: https://github.com/It4innovations/Graph500

2.3 Partner: IT4I – D3

No.	Item	Description
1	Data set reference and name	D3: Simulated probabilistic speed profiles for selected routes in Prague (Benchmark for Probabilistic time dependent routing algorithm). DOI: 10.5281/zenodo.2275647
2	Data set description	<p>Simulated probabilistic speed profiles for 1608 road segments in Prague, Czech Republic.</p> <p>This dataset contains simulated speed profiles for selected road segments in Prague, Czech Republic. The profiles describe changes in average driving speed on a particular segment over a given time interval discretised to a time steps of a fixed size. Number of the time steps is stored in the "IntervalsPerSegment" attribute of the root group. Size of a single time step in seconds is stored in the "TimeStep" attribute. Total length of time interval covered by the profiles is computed as IntervalsPerSegment x TimeStep.</p> <p>The file contains two types of profiles. The first type are plain speed profiles which provide just average driving speed assigned to each time step - dataset "speed_profile". The second type are probabilistic speed profiles which contain discrete probability distribution of the driving speed for particular segment and given time step.</p> <p>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.</p> <p>The input dataset for the benchmark contains pre-defined set of routes computed for a given set of simulated pairs of origin and destination points and generated speed profiles.</p>
3	Standards and metadata	<p>Time dependent routing algorithm is described in the following publications:</p> <p>Tomis, R., Rapant, L., Martinovič, J., Slaninová, K., Vondrák, I.: Probabilistic time-dependent travel time computation using Monte Carlo simulation, (2016) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9611, pp. 161-170</p> <p>Golasowski, M., Tomis, R., Martinovič, J., Slaninová, K., & Rapant, L.: Performance Evaluation of Probabilistic Time-Dependent Travel Time Computation, (2016) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),</p>

		9842 LNCS, pp. 377-388. Tomis, R., Martinovič, J., Slaninová, K., Rapant, L., & Vondrák, I. (2015). Time-Dependent Route Planning for the Highways in the Czech Republic. In Computer Information Systems and Industrial Management (pp. 145-153). Springer International Publishing.
4	Data sharing	Data sharing follows rules of selected service defined in Section Public Data Management Policies. Licence: Open Data Commons Attribution
5	Archiving and preservation (including storage and backup)	Archiving and preservation follow rules of selected service defined in Section Public Data Management Policies. Zenodo: https://zenodo.org/record/2275647#.XBkD6C2E4W4

2.4 Partner: IT4I – D4

No.	Item	Description
1	Data set reference and name	D4: Traffic network routing index for the Czech Republic DOI: 10.5281/zenodo.2275557
2	Data set description	<p>The time dependent routing 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.</p> <p>The input dataset for the benchmark contains pre-defined set of routes computed for a given set of simulated pairs of origin and destination points and generated speed profiles.</p> <p><i>Graph representation of road network of the Czech Republic</i> This dataset contains graph of the entire Czech road network. The data are derived from the Open Street Map project and stored in a HDF5 file. The file contains graph topology and metadata for edges and vertices.</p> <p>Spatial index is also included for the purpose of point snapping and other spatial queries. The spatial index is stored in SQLite file with SpatiaLite extension.</p> <p><i>Routing Index in HDF5</i> The graph topology is stored in the following way. Nodes have assigned a row index in the edges dataset which points to an outbound edge plus a number of the subsequent edges which are also output to this node. The edge metadata are stored in the EdgeData dataset to avoid redundancy. NodeMap dataset provides a convenient way to query nodes based on their unique identifiers.</p>
3	Standards and metadata	<p>Time-dependent routing is described in the following publications:</p> <p>Tomis, R., Rapant, L., Martinovič, J., Slaninová, K., Vondrák, I.: Probabilistic time-dependent travel time computation using Monte Carlo simulation, (2016) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9611, pp. 161-170</p> <p>Golasowski, M., Tomis, R., Martinovič, J., Slaninová, K., & Rapant, L.: Performance Evaluation of Probabilistic Time-Dependent Travel Time Computation, (2016) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9842 LNCS, pp. 377-388.</p>
4	Data sharing	Data sharing follows rules of selected service defined in Section Public Data Management Policies.

		Licence: Open Data Commons Attribution
5	Archiving and preservation (including storage and backup)	Archiving and preservation follow rules of selected service defined in Section Public Data Management Policies. Zenodo: https://zenodo.org/record/2275557#.XBlrDy2E4W4

2.5 Partner: IT4I – D5

No.	Item	Description
1	Data set reference and name	D5: Floating Car Data Collection for Processing and Benchmarking DOI: 10.5281/zenodo.2250119
2	Data set description	The floating car data (FCD representing movement of cars with their position in time) is produced by the traffic simulator software (further referred to as Simulator) published in (Ptošek, 2018) and can be used as an input for data processing and benchmarking. The dataset contains FCD of various quality levels based on the routing graph of the Czech Republic derived from Open Street Map openstreetmap.org .
3	Standards and metadata	Floating car data collection for processing and benchmarking is described in the following publications: Ptošek, V., Ševčík, J., Martinovič, J., Slaninová, K., Rapant, L., Cmar, R. Real time traffic simulator for self-adaptive navigation system validation (2018). In Proceedings of 30th European Modeling and Simulation Symposium, EMSS 2018, pp. 274-283. A. El Abbous and N. Samanta. A modeling of GPS error distributions, In proceedings of 2017 European Navigation Conference (ENC), 2017.
4	Data sharing	Data sharing follows rules of selected service defined in Section Public Data Management Policies. Licence: Creative Commons
5	Archiving and preservation (including storage and backup)	Archiving and preservation follow rules of selected service defined in Section Public Data Management Policies. Zenodo: https://zenodo.org/record/2250119#.XBPEIC2E4W4

2.6 Partner: IT4I – D6

No.	Item	Description
1	Data set reference and name	D6: Data used and created within UC2
2	Data set description	Due to the private nature of UC2, the data set description is included into private deliverables of UC2.
3	Standards and metadata	Due to the private nature of UC2, the description of standards and metadata is included into private deliverables of UC2.
4	Data sharing	Selected data is 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 is preserved, stored, and maintained following the rules defined in Section IT4I Data Management Policies. IT4I private repository: https://code.it4i.cz

2.7 Partner: SYGIC – D7

No.	Item	Description
1	Data set reference and name	D7: Sygic FCD Stream
2	Data set description	Stream of floating car data from Sygic navigated vehicles in Slovakia, Czech Republic, Austria
3	Standards and metadata	Due to the private nature of UC2, the description of standards and metadata is included into private deliverables of UC2.
4	Data sharing	Selected data is 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 is preserved, stored, and maintained following the rules defined in Section IT4I Data Management Policies. Online ftp stream: adaswftp.vsb.cz

2.8 Partner: UPORTO – D8

No.	Item	Description
1	Data set reference and name	D8: Data used and created within UC2
2	Data set description	Due to the private nature of UC2, the data set description is included into private deliverables of UC2.
3	Standards and metadata	Due to the private nature of UC2, the description of standards and metadata is included into private deliverables of UC2.
4	Data sharing	Selected data is privately available to selected ANTAREX participants.
5	Archiving and preservation (including storage and backup)	All the data collections created, maintained and processed within UC2 is preserved, stored, and maintained following the rules defined in Section IT4I Data Management Policies. UPORTO private repository: https://code.it4i.cz/routing/AntarexIT4I/tree/feup

2.9 Partner: UPORTO – D9

No.	Item	Description
1	Data set reference and name	D9: Code output by AutoPar-Clava using the benchmark codes in the repositories PolyBench and NAS as input DOI: 10.5281/zenodo.1889368
2	Data set description	Code automatically parallelized using AutoPar-Clava and targeting OpenMP
3	Standards and metadata	Dataset is made up of ASCII files assembled as a zipped archive. Full metadata description is provided within the standard dataset creation in OpenAIRE/Zenodo service. Keywords: LARA; DSL; benchmarks; Auto-Parallelization.
4	Data sharing	Data sharing follows rules of selected service defined in Section Public Data Management Policies. Licence: Creative Commons
5	Archiving and preservation (including storage and backup)	Archiving and preservation follow rules of selected service defined in Section Public Data Management Policies. UPORTO repository: https://github.com/specs-feup/specs-lara/tree/master/ANTAREX/AutoPar Zenodo: https://zenodo.org/record/1889368

2.10 Partner: UPORTO – D10

No.	Item	Description
1	Data set reference and name	D10: ANTAREX-DSL: DSL usage examples
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 covers 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 Clava compiler is 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 Clava compiler.
3	Standards and metadata	Dataset is made up of ASCII files with codes and source code and jar file for the Clava compiler. Keywords: Clava, LARA; DSL; benchmarks.
4	Data sharing	ANTAREX-DSL dataset is public. Access is guaranteed by GitHub access and Clava webpage open to public without any restriction. Demo version of the Clava compiler is available via a web interface to end-user and no additional software is necessary for its dissemination and sharing. License: MIT for the LARA examples, and Apache 2.0 for Clava.
5	Archiving and preservation (including storage and backup)	Storage persistence according to services' availability of GitHub. GitHub: <ul style="list-style-type: none"> • DSL examples: https://github.com/specs-feup/specs-lara/tree/master/ANTAREX • Clava jar: http://specs.fe.up.pt/tools/clava.jar • Clava source code: https://github.com/specs-feup/clava • Clava interface web: http://specs.fe.up.pt/tools/clava

2.11 Partner: CINECA – D11

No.	Item	Description
1	Data set reference and name	D11: Marconi supercomputer KNL partition benchmark (HPL and single node stream) DOI: 10.5281/zenodo.2426600
2	Data set description	This dataset collects the data stored during the procedure of evaluation of the Marconi KNL machine at CINECA in order to classify it for Top500 list. Dataset also includes a report summarizing the results of benchmarks (STREAM for single node memory assessment and HPL for HPC parallel performance) carried out in Nov. 2016.
3	Standards and metadata	Dataset is made up of ASCII files (Unix format) assembled as a tar gzipped archive. Full metadata description are provided within the standard dataset creation in OpenAIRE/Zenodo service. Keywords: Galileo; CINECA; TOP500; HPL; STREAM; HPC; benchmarks.
4	Data sharing	Marconi-HPL dataset is and will be PUBLIC. Access is guaranteed by OpenAIRE/Zenodo service and is widely open to public without any restriction. Marconi-HPL dataset is provided through the OpenAIRE/Zenodo web interface to end-user and no additional software is necessary for its dissemination and sharing. Marconi-HPL dataset is indexed within OpenAIRE and exposed to external end-user via standard OpenAIRE retrieval tools like those available within the Zenodo software. Licence: Creative Commons Attribution
5	Archiving and preservation (including storage and backup)	Storage persistence in OpenAIRE/Zenodo service is guaranteed for unlimited time. Zenodo: https://zenodo.org/record/2426600#.XBqc9y2E4W4

2.12 Partner: CINECA – D12

No.	Item	Description
1	Data set reference and name	D12: MARCONI KLN supercomputer HPCG benchmark: MARCONI HPCG benchmark for assessing Galileo machine performance on hybrid configuration. DOI: 10.5281/zenodo.2427207
2	Data set description	This dataset collects High Performance Conjugate Gradient (HPCG) benchmark results of Marconi KNL (Intel Xeon PHI partition) supercomputer, carried out in May 2017.
3	Standards and metadata	Dataset contains an ASCII report of the benchmark result a JPEG table and a JPEG graph. Full metadata description are provided within the standard dataset creation in OpenAIRE/Zenodo service. Keywords: Marconi, Xeon PHI; CINECA; HPCG; XeonPhi; HPC; benchmarks.
4	Data sharing	Marconi-HPCG dataset is and will be PUBLIC. Access is guaranteed by OpenAIRE/Zenodo service and is widely open to public without any restriction. Marconi-HPCG dataset is provided through the OpenAIRE/Zenodo web interface to end-user and no additional software is necessary for its dissemination and sharing. Marconi-HPCG dataset is indexed within OpenAIRE and exposed to external end-user via standard OpenAIRE retrieval tools like those available within the Zenodo software. Licence: Creative Commons Attribution
5	Archiving and preservation (including storage and backup)	Storage persistence in OpenAIRE/Zenodo service is guaranteed for unlimited time. Zenodo: https://zenodo.org/record/2427207#.XBqaPy2E4W4

2.13 Partner: CINECA – D13

No.	Item	Description
1	Data set reference and name	<p>D13: GeoDock: Dataset of ligand inputs and corresponding geometrical docking results. EUDAT B2SHARE PID: 11100/53fd0ff6-100d-11e9-a612-e41f13eb41b2</p> <p>EUDAT B2SHARE direct handle: https://hdl.handle.net/11100/53fd0ff6-100d-11e9-a612-e41f13eb41b2?noredirect</p>
2	Data set description	<p>This dataset collects the data stored during the procedure of evaluation of UC1 GeoDock mini-app on the A2 partition of Marconi system at CINECA.</p> <p>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 generated from commercial catalogs like, i.e., Sigma-Aldrich and/or Enamine. GeoDock 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.</p> <p>Dataset also includes a report summarizing the results of GeoDock benchmarks.</p>
3	Standards and metadata	<p>Dataset is composed of:</p> <ul style="list-style-type: none"> • ligands molecular structure in standard mol2 format, as individual bzip2 compressed ASCII files; • ligands molecular structure preprocessed and converted in the GeoDock high-throughput, HDF5-based format as described in D4.4; • geometrical docking results produced during benchmark campaigns in GeoDock high-throughput, HDF5-base format as described in D4.4; <p>Full metadata description are provided within the standard dataset creation in EUDAT B2SHARE.</p> <p>Keywords: Marconi; CINECA; GeoDock; LiGen; Docking; PDB; HDF5; Sigma-Aldrich; Enamine; benchmarks.</p>
4	Data sharing	<p>Full set of data is privately available to ANTAREX participants and available upon request.</p> <p>A significant subset (more than 4.5 million of ligands) is publicly available using the EUDAT handle reported above.</p>
5	Archiving and preservation (including storage and backup)	<p>Storage persistence on EUDAT B2SHARE is granted by EUDAT free of charge for unlimited time, thanks to an in-kind contribution of CINECA.</p>

2.14 Partner: ETHZ – D14

No.	Item	Description
1	Data set reference and name	D14: Antarex HPC Fault Dataset DOI: 10.5281/zenodo.1453949
2	Data set description	The ANTAREX dataset contains trace data collected from the homonymous experimental HPC system located at ETH Zurich while it was subjected to fault injection, for the purpose of conducting machine learning-based fault detection studies for HPC systems. Acquiring our own dataset was made necessary by the fact that commercial HPC system operators are very reluctant to share trace data containing information about faults in their systems.
3	Standards and metadata	For a detailed analysis on the structure and features of the ANTAREX dataset, please refer to the research paper "Online Fault Classification in HPC System through Machine Learning", by Netti et al. Additional details can be found in the research paper "FINJ: a Fault Injection Tool for HPC System" by Netti et al., whereas all source code can be found on the GitHub repository of the FINJ tool.
4	Data sharing	Data sharing follows rules of selected service defined in Section Public Data Management Policies. Licence: Creative Commons
5	Archiving and preservation (including storage and backup)	Archiving and preservation follow rules of selected service defined in Section Public Data Management Policies. Zenodo: https://zenodo.org/record/1453949#.XBPTfi2E6BY