ENVRI^{plus} DELIVERABLE

D7.1 Interoperable data processing for environmental RI projects: system design

WORK PACKAGE 7 – Data processing and analysis

LEADING BENEFICIARY: National Research Council of Italy (CNR)

Author(s):	Beneficiary/Institution
Leonardo Candela, Gianpaolo Coro, Pasquale Pagano, Giancarlo Panichi	National Research Council of Italy (CNR)
Malcolm Atkinson, Rosa Filgueira	University of Edinburgh
Daniele Bailo	Istituto Nazionale di Geofisica e Vulcanologia (INGV)
Carl-Friedrik Enell	EISCAT Scientific Association
Markus Fiebig	Norsk Institutt for Luftforskning (NILU)
Florian Haslinger	Eidgenoessische Technische Hochschule Zuerich (ETHZ)
Maggie Hellström, Alex Vermeulen, Harry Lankreijer	Lund University
Robert Huber	University of Bremen
Sylvie Joussaume, Francesca Guglielmo	Centre National de la Recherche Scientifique, Institut Pierre Simon Laplace
Victor Mendez	Universitat Autònoma de Barcelona

Accepted by: Zhiming Zhao (Data for Science Theme Leader)

Deliverable type: REPORT

Dissemination level: PUBLIC

Deliverable due date: 31.12.2016/M20

Actual Date of Submission: 27.03.2017/M23

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182

ABSTRACT

Data processing is a very wide area or domain because of a series of characteristics including the contexts resulting from diverse application scenarios, the great variety of processes to be enabled, the large set of enabling technologies and solutions. One of the consequences of this large variety is that each software solution for data processing only manages to address parts, i.e. it is difficult to imagine a single solution that is equally suitable for any (or even most) application scenarios and contexts. This deliverable illustrates that scope and diversity by reporting detailed practices and requirements from seven of the ENVRI RIs. It then describes the design of a *single* data processing solution that will help meet a substantial range of requirements in a representative range of contexts. That approach is conceived to be (a) suitable for ***serving the needs of scientists involved in ENVRI RIs***, (b) ***open and extensible*** both with respect to the algorithms and methods it enables and the computing platforms it relies on to execute those algorithms and methods, (c) ***open-science-friendly***, i.e. it is capable of incorporating every algorithm and method integrated into the data processing framework as well as any computation resulting from the exploitation of integrated algorithms into a "research object" catering for citation, reproducibility, repeatability and provenance.

PROJECT INTERNAL REVIEWER(S)

Project internal reviewer(s):	Beneficiary/Institution
Keith Jeffery	NERC
Malcolm Atkinson	University of Edinburgh
Markus Stocker	University of Bremen

Document history:

Date	Version
15.11.2016	Draft for comments
29.11.2016	Integrating V. Mendez contribution on DIRAC
02.12.2016	Integrating S. Kindermann contribution on IS-ENES
12.12.2016	Integrating M. Fiebig contribution on ACTRIS
12.12.2016	Integrating M. Hellström contribution on ICOS
12.12.2016	Integrating C.-F. Enell contribution on EISCAT_3D
15.12.2016	Integrating D. Bailo contribution on EPOS
16.12.2016	Integrating R. Huber contribution on FixO3 and EMSO
13.02.2017	Release of first complete version
28.02.2017	Collection of Reviewer feedback and requests for changes
27.03.2017	Release of the revised version (responding to reviewers' comments and integrating comments and contributions)

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the editor (Leonardo Candela leonardo.candela@isti.cnr.it).

TERMINOLOGY

A complete project glossary is provided online here: envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between RIs, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environmental understanding and decision-making for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonize policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance trans-disciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the inter-RI (European and Global) level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

TABLE OF CONTENTS

TABLE OF FIGURES

1 Introduction

Data Processing or Analytics is an extensive domain including any activity or process that performs a series of actions on dataset(s) to distil information [Bordawekar et al. 2014]. It is particularly important in scientific domains especially with the advent of the 4th Paradigm and the availability of “big data” [Hey et al. 2009]. It may be applicable at any stage in the data life cycle, from QA to seismic event recognition, close to data acquisition to transformations and visualisations tailored for decision makers as results are presented. Data analytics methods draw on multiple disciplines including statistics, quantitative analysis, data mining, and machine learning. Very often these methods require compute-intensive infrastructures to produce their results in a suitable time, because of the data to be processed (e.g., huge in volume or heterogeneity) and/or because of the complexity of the algorithm/model to be elaborated/projected. Moreover, being devised to analyse dataset(s) and produce other “data”/information (than can be considered a dataset) these methods are strongly characterised by the “typologies” of their inputs and outputs. In some data-intensive cases, the data handling (access, transport, IO and preparation) can be a critical factor in achieving results within acceptable costs.

ENVRIplus WP7 ‘Data processing and analysis’ is called to design and develop a technical solution for data analytics that is suitable for the needs and contexts arising in environmental Research Infrastructures. In particular, Task 7.1. ‘Interoperable Data Processing, Monitoring and Diagnosis’ is devised to design and develop a solution for data processing aiming at making it significantly easier for scientists to conduct a range of experiments and analyses upon a great variety of data. Expanding the common data processing workflow modelled in the ENVRI project, this task focuses on the engineering and technological aspects of managing the entire lifecycle of computing tasks and application workflows for the efficient utilisation of services and facilities offered by existing e-Infrastructures. Distinguishing features of the service include enabling scientists to enrich the data processing environment by easily injecting new algorithms and methods to be also reused by others. Algorithms and methods can be produced by using programming languages (e.g. Java) or scripting languages scientists are familiar with (e.g. R scripts). The objective of this task is to provide common and cost-effective data processing services for environmental RIs with consideration of existing technologies in e-Infrastructures, data infrastructures and other relevant RIs, building on recent advances in data-intensive computation.

This deliverable describes the process and settings leading to the definition of the data analytics platform resulting from Task 7.1. In fact, the deliverable contains first a long and detailed discussion aiming at documenting what are (a) the existing technologies and solutions for data processing (including workflow management systems and data processing frameworks and platforms); (b) the envisaged data-processing-related patterns captured by the ENVRIplus reference model; and (c) the existing solutions seven environmental Research Infrastructures have currently in place for satisfying their data processing needs. After this discussion, the deliverable concludes by presenting a platform for data processing that has been specifically conceived to complement the offering of existing solutions and meet some needs arising in environmental Research Infrastructure. This platform is characterised by the following features: (a) it is suitable for ***serving the needs of scientists involved in ENVRI RIs***, (b) it is ***open and extensible*** both with respect to the algorithms and methods it enables and the computing platforms it relies on to execute the algorithms and methods, (c) it is ***open-science-friendly***, i.e. it is capable to transform every algorithm and method integrated into the data processing framework as well as any computation resulting from the exploitation of integrated algorithms into a “research object” catering for citation, reproducibility, repeatability, provenance, etc.

The remainder of this deliverable is organised as follows. Section 2 describes the settings characterising data processing in ENVRIplus. It describes (i) existing technologies and solutions for data processing, (ii) expectations and approaches currently characterising ENVRIplus Research Infrastructures, and (iii) how this challenging and wide concept is captured by the ENVRIplus Reference Model and how this platform is expected to contribute to the rest of solutions developed by ENVRIplus work packages (namely, WP5 'Reference model guided RI design', WP6 'Reference model guided RI design', T7.2 'Performance optimisation for big data sciences', WP8 'Data curation and cataloguing') to provide environmental Research Infrastructures with new solutions to basic needs. Section 3 describes the envisaged data analytics platform by presenting the architecture, the main functionalities it offers, and the possible exploitation models. Finally, Section 4 concludes the report.

2 ENVRIplus Data Processing in Context

Data Processing is certainly a wide concept embracing tasks ranging from (systematic) data collection, collation and validation to data analytics [Bordawekar et al. 2014] aiming at distilling and extracting new “knowledge” out of existing data by applying diverse methods and algorithms. The vast quantities of available data are further enlarging the need for effective data solutions that are capable to deal with the “big data” phenomenon where data are potentially many in quantity, diverse in typologies, and produced rapidly. Khalifa et al. [Khalifa et al. 2016] recently surveyed existing solutions for data analytics and observed that (a) “existing solutions cover bits-and-pieces of the analytics process” and (b) when devising an analytics solution there are six pillars representing issues to deal with (cf. Figure 1).

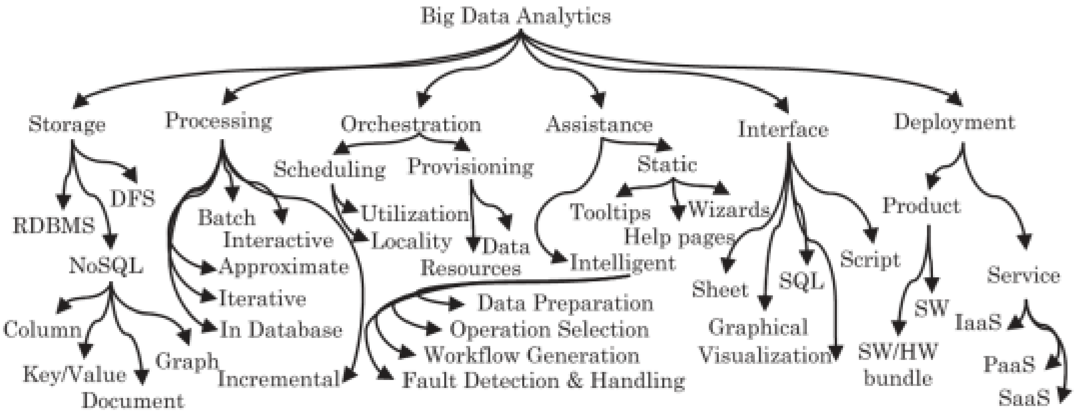


FIGURE 1. THE BIG DATA ANALYTICS ECOSYSTEM TAXONOMY BY KHALIFA ET AL. (2016)

The six pillars identified by Khalifa et al. are:

- *Storage*, i.e., how data to be analysed are going to be handled;
- *Processing*, i.e., how the pure processing task is going to happen;
- *Orchestration*, i.e., how computing resources are going to be managed to reduce processing time and cost;
- *Assistance*, i.e., how scientists and practitioners are provided with facilities helping them to perform their task;
- *User Interface*, i.e., how scientists and practitioners are provided with the data analytics system to run their analytics, monitor the execution and get back the results;
- *Deployment Method*, i.e., how the analytics system is offered to the end users.

This taxonomy is relevant to the Environmental Research Infrastructures ENVRIplus is targeting. However, as the remaining paragraphs of this section (namely Sec. 2.3) make clear, these Research Infrastructures are characterised by having great variety and are currently in different phases of their development. This makes the task of devising a shared solution for data processing and analytics quite challenging.

The rest of the section is aiming at describing an overall picture of the data processing and analytics that characterise the ENVRIplus application context. Section 2.1 gives and overview of existing solutions for data processing. Section 2.2 describes how processing tasks are captured by the ENVRI Reference Model [Hardisty et al. 2017], i.e. a model capturing basic concepts and relationships among them to build a picture characterising the ‘archetypical’ Environmental Research Infrastructure and to promote interoperability. Section 2.3 discusses the requirements characterising environmental Research Infrastructure and presents some solutions currently in use or planned by seven environmental RIs. Section 2.4 summarises the overall picture emerging from this overview and highlights the key findings.

2.1 Data Processing Technologies Overview

An extensive and detailed analysis of the needs and approaches characterising data processing for the sake of ENVRI Research Infrastructure has been presented in [Atkinson et al. 2015]. The technical report surveyed generic technologies. Here, we report an excerpt of the analysis and its major findings.

A lot of technologies and approaches have been developed to support data processing and analytics tasks including:

- ***High Performance Computing solutions***, i.e., aggregated computing resources to realise a "high performance computer" (including processors, memory, disk and operating system);
- ***Distributed Computing Infrastructures***, i.e., distributed systems characterised by heterogeneous networked computers that offer data processing facilities. This includes high-throughput computing and cloud computing;
- ***Scientific workflow management systems (SWMS)***, i.e., systems enacting the definition and execution of *scientific workflows* consisting of [Liew et al. 2016]: a list of tasks and operations, the dependencies between the interconnected tasks, control-flow structures and the data resources to be processed;
- ***Data analytics frameworks and platforms***, i.e., platforms and workbenches enabling scientists to execute analytic tasks. Such platforms tend to provide their users with implementations of algorithms and (statistical) methods for the analytics tasks.

These classes of solutions and approaches are not isolated, rather they are expected to rely on each other to provide end users with easy to use, efficient and effective data processing facilities, e.g., SWMS rely on distributed computing infrastructures to actually execute their constituent tasks.

In Europe, PRACE[1] represents the major initiative for High Performance Computing. Similarly, EGI[2] is a point of reference for distributed computing. These infrastructures represent a valuable asset that must be taken into account when devising any long-term solution for data processing for environmental Research Infrastructure since (a) they are conceived to offer a computing capacity that no other infrastructure can build on its own easily, timely and in a cheaper way (economy of scale) and (b) they are strategically designed to be an enabling backend for every Research Infrastructure.

The resources of many RIs and the resources of their partner institutions, and those hosting projects using RI services and data must also be recognised as a significant computational context. These are used to run continuous services that also contribute to RIs, to handle emergency response computations, and to support the work of their practitioners and researchers. In consequence, substantial portions of the computational requirements are met using these resources to avoid disrupting established methods developed and refined over years, to meet financial and organisational constraints, and to use contexts familiar to practitioners and innovators. These ICT resources, including their management, development and support are often subject to the priorities of their funders, such as national governments.

[1] PRACE Research Infrastructure http://www.prace-ri.eu/

[2] EGI Federated Infrastructure https://www.egi.eu/

2.1.1 Scientific Workflow (Management Systems)

Scientific workflows represent a key concept characterising data processing [Liew et al. 2016]. Over the last two decades, many large-scale scientific experiments take advantage of scientific workflows to model data operations such as loading input data, data processing, data analysis, and aggregating output data. The term workflow refers to the automation of a process, during which data is processed by different logical data processing activities according to a set of rules, along with the attendant tasks of, for example, moving data between workflow processing stages. Workflow management systems (WMS) [Bux and Leser 2013] aid in the automation of these processes, freeing the scientist from the details of the process, since WMS manage the execution of the application on a computational infrastructure. WMSs usually offer diverse solutions for (a) ***reusability***, i.e., how (and if) they make it possible to incorporate and reuse existing workflows (or part of them) when designing a new workflow, (b) ***performance***, i.e., what are the policies and practices they have in place to optimise the execution time of a workflow; (c) ***design***, i.e., what they offer for facilitating the development of a new workflow; and (d) **collaboration**, i.e., what mechanisms they support for the collaborative development of workflows.

Several scientific workflow management systems (SWMS) [Liu et al. 2015] [Liew et al. 2016] have been developed to offer a user-friendly way for designing and implementing computational scientific procedures under the workflow paradigm, providing GUIs and tools for easing the task of handling large and complex computational processes in science. Examples of ***task-oriented*** workflow systems, i.e., systems (a) having a predominant model with stages that correspond to tasks, and (b) organizing their enactment on a wide range of distributed computing infrastructures (DCI), normally arranging data transfer between stages using files:

- **Pegasus**[3] [Deelman et al. 2015] supports execution of workflows in distributed environments such as campus clusters, grids and clouds. Pegasus Workflow Management Service maps an application onto available resources pertaining to the cluster while keeping the internal and external dependencies of the workflow in order. Pegasus workflow has been used to power LIGO gravitational wave detection analysis[4].
- **Triana**[5] [Churches et al. 2006] is an open source graphical problem-solving environment that supports the assembly and execution of a workflow through a graphical user interface while minimizing the burden of programming.
- **Taverna**[6] [Wolstencroft et al. 2013] provides an easy to use environment to build, execute and share workflows of web services. It was initially developed for the enactment of bioinformatics workflows and is now more widely used. It emphasises usability, providing a graphical user interface for workflow modelling and monitoring as well as a comprehensive collection of predefined services.
- **Galaxy**[7] [Blankenberg et al. 2011] is a web-based system that aims to bring computational data analysis capabilities to non-expert users in the biological sciences domain. The main goals of the Galaxy framework are accessibility to biological computational capabilities and reproducibility of the analysis result by tracking the information related to every step of the process. The Galaxy workflow model does not follow the directed acyclic graphs

[3] Pegasus website pegasus.isi.edu

[4] https://pegasus.isi.edu/2016/02/11/pegasus-powers-ligo-gravitational-waves-detection-analysis/

[5] Triana website www.trianacode.org

[6] Taverna website www.taverna.org.uk

[7] Galaxy website galaxyproject.org

(DAG) paradigm, as it allows to define loops, being a directed cyclic graphs (DCGs) approach.

- **KNIME**[8] [Beisken et al. 2013] shares many characteristics with Taverna, with a stronger focus on user interaction and visualisation of results, yet with a smaller emphasis on web service invocation. Furthermore, KNIME focuses on workflows from the fields of data mining, machine learning, finance and chemistry, while Taverna is more concerned with integration of distributed and possibly heterogeneous data. A graphical user interface facilitates design and execution monitoring of workflows.
- **Kepler**[9] [Ludäscher et al. 2006] is a frequently used graphical SWMS. Similar to Taverna and KNIME, it provides an assortment of built-in components with a major focus on statistical analysis. Kepler workflows are written in MoML (a markup language in XML) or KAR files, which are an aggregation of files into a single JAR file. Kepler is built on top of the Ptolemy II Java library, from which it inherits the concepts of Directors and Actors. The directors control the execution of the workflow, while the actors execute actions when specified by directors.
- **Apache Airavata**[10] [Marru et al. 2011] is an open source, open community SWMS to compose, manage, execute, and monitor distributed applications and workflows on computational resources ranging from local resources to computational grids and clouds. Airavata builds on general concepts of service-oriented computing, distributed messaging, and workflow composition and orchestration.

Alternative approaches to task-oriented workflows are the ***stream-based workflows***. This mirrors the shared-nothing composition of operators in database queries and in distributed query processing that has been developed and refined in the database context. Data streaming was latent in the auto-iteration of Taverna, it has been developed as an option for Kepler, and it is the model used by Meandre [Acs et al. 2010], and by Swift (which supports the data-object-based operation using its own data structure). Data streaming pervaded the design of Dispel [Atkinson 2013]. Dispel was proposed as a means of enabling the specification of scientific methods assuming a stream-based conceptual model that allows users to define abstract, machine-agnostic, fine-grained and data-intensive workflows. The Python library dispel4py [Filgueira et al. 2016] implements many of the original Dispel concepts. It describes abstract workflows for data-intensive applications, which are later translated and enacted in distributed platforms (e.g., Apache Storm, MPI clusters, etc.).

Bobolang [Falt et al. 2014], a relatively new workflow system based on data streaming, has linguistic forms based on C++ and focuses on automatic parallelisation. It also supports multiple inputs and outputs, meaning that a single node can have as many inputs or outputs, as a user requires. Currently, it does not support automatic mapping to different Distributed Computing Infrastructures (DCIs).

Many RIs already use workflows to handle their data according to established working practices. Individual researchers, projects and RIs can gain substantial productivity gains from the automation workflows offer, particularly with respect to data management, data identification, metadata collection, provenance and curation procedures. Such automation of data-chores often reduces error rates, though human judgement always remains a crucial component of QA. Their access to such benefits depends on support, which saves them the effort of setting up the workflow management systems, which provides tools that help in the development, sharing, validation and optimisation of such encoded methods, and from ready access to previously developed workflow fragments that match their domain's requirements

[8] KNIME website www.knime.org

[9] Kepler website kepler-project.org

[10] Apache Airavata website http://airavata.apache.org

or trans-disciplinary requirements. The diversity of scientific workflow systems makes it difficult to meet this need, though EUDAT is attempting to develop and support a common enactment mechanism for such workflows. Within ENVRIplus we will seek to exploit such initiatives while meeting the RIs' diverse scientific-workflow requirements, e.g., by establishing good solutions as workflow fragments meeting the commonly required data-lifecycle actions.

2.1.2 Data Processing Platforms and Framework

In parallel with scientific workflows, a series of platforms and frameworks have been developed to simplify the execution of (scientific) distributed computations. This need is not new; it is actually rooted in high-throughput computing which is a well-consolidated approach to provide large amounts of computational resources over long periods of time. The advent of Big Data and Google MapReduce in the first half of 2000 brings new interests and solutions. Besides taking care of the smart execution of user-defined and steered processes, platforms and environments start offering ready to use implementations of algorithms and processes that benefit from a distributed computing infrastructure.

There exist many *data analytics frameworks and platforms*, including:

- ***Apache Mahout***[11] is a platform offering a set of machine-learning algorithms (including collaborative filtering, classification, clustering) designed to be scalable and robust. Some of these algorithms rely on Apache Hadoop, others are relying on Apache Spark.
- ***Apache Hadoop***[12] is a basic platform for distributed processing of large datasets across clusters of computers by using a MapReduce strategy. In reality this is probably the most famous open-source implementation of ***MapReduce***, a simplified data processing approach to execute data computing on a computer cluster [Li et al. 2014]. Worth to highlight that one of the major issues with MapReduce – resulting from the *"flexibility"* key feature, i.e., "users" are called to implement the code of map and reduce functions – is the amount of programming effort. In fact, other frameworks and platforms are building on it to provide users with data analytics facilities (e.g., Apache Mahout).
- ***Apache Spark***[13] [Zaharia et al. 2016] is an open-source, general-purpose cluster-computing engine which is very fast and reliable. It provides high-level APIs in Java, Scala, Python and R, and an optimised engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL[14] for SQL and structured data processing, MLlib[15] for machine learning, GraphX[16] for graph processing, and Spark Streaming[17].
- ***gCube Data Analytics***[18] [Candela 2013], [Coro 2014] is an open-source solution conceived to offer an open set of algorithms with the as-a-Service paradigm. The platform relies on a set of DCIs for executing the computing tasks including D4Science and EGI. This platform is equipped with more than 100 ready-to-use algorithm implementations which include features of significant value, including clustering, climate simulations, niche modelling, model performance evaluation, time series analysis, analysis of marine species and handling geo-referenced data. New algorithms can be integrated easily. The platform

[11] Apache Mahout website http://mahout.apache.org/

[12] Apache Hadoop website http://hadoop.apache.org/

[13] Apache Spark website http://spark.apache.org

[14] Spark SQL, DataFrames and Datasets Guide http://spark.apache.org/docs/latest/sql-programming-guide.html

[15] MLlib guide for the RDD-based API http://spark.apache.org/docs/latest/mllib-guide.html

[16] GraphX Programming Guide http://spark.apache.org/docs/latest/graphx-programming-guide.html

[17] Spark Streaming Programming Guide http://spark.apache.org/docs/latest/streaming-programming-guide.html

[18] gCube website www.gcube-system.org

comes with a development framework dedicated to this (Java algorithms as well as R scripts are well supported). Once integrated, each algorithm is automatically exposed via a REST-based protocol (OGC WPS) as well as via a web-based GUI that is a complete dashboard for executing computations by guaranteeing Open Science practices (e.g., every computation leads to a "research object" recording and making available every "piece" of the task).

- ***iPython/Jupyter*** [Perez & Granger 2007] is a notebook-oriented interactive computing platform which enacts to create and share "notebooks", i.e., documents combining code, rich text, equations and graphs. Notebooks support a large array of programming languages (including R) and communicate with computational kernels by using a JSON-based computing protocol. Similar solutions include: knitr [Xie 2015] which works with the R programming language and Dexy[19], a notebook-like program that focuses on helping users to generate papers and presentations that incorporate prose, code, figures and other media.

In addition to them, the DIRAC platform is worth discussing.

2.1.2.1Data Processing by DIRAC

By V. Mendez (DIRAC)

The DIRAC platform eases scientific computing by providing distributed computing resources in a manner transparent to the end-user. The platform has been demonstrated to be a robust and efficient solution for large communities, with a track record starting in 2002 at the LHCb experiment, integrating distributed computing clusters before the first data grid infrastructure was available. Since then, the platform has become a multiple community framework able to integrate Grid, Cloud and other distributed resources, providing computing and storage, in an interoperable manner from a single user interface. Thus, the lemma is: DIRAC, *the interware*. During last years, several international communities have chosen DIRAC to implement their compute and data management systems. This background ensures that new adopters attain proven scalability to meet their requirements using a well-supported solution that will be sustained and supported in the following years.

DIRAC is part of the EGI offering and the goal of this contribution is to:

- Retain communities and attract new communities into the dirac.egi.eu service;
- Provide simplified access to EGI resources and accelerate the ability of researches to undertake excellent science with DIRAC technical innovations;
- Transfer DCI skills and know-how to other medium and big communities and resource providers in the context of EGI.

Previous experience integrating scientific portals with DIRAC have demonstrated an important improvement in job efficiency. For example, WeNMR e-Infrastructure project reported a performance improvement between 70% to 99% over previous systems with DIRAC job submission[20]. This higher efficiency is made possible by two main assets:

- The pilot job [Casajus et al. 2010] approach proposed by DIRAC project in 2003 and broadly adopted now by other distributed computing solutions is an evolution from previous job push model to an improved pull model. In a push model, jobs are submitted to batch queues or grid WMS to be allocated in computing resources. On the other hand, a pull model submits just a pre-allocation container, which performs basic sanity checks of the environment, integrates heterogeneous resources (Grid, Cloud, Volunteer) and

[19] Dexy website http://www.dexy.it

[20] The HADDOCK WeNMR portal: From gLite to DIRAC submission in three hours. https://indico.egi.eu/indico/contributionDisplay.py?sessionId=19&contribId=62&confId=1994

when ready, pulls the job from a central queue. In this case, proactive actions can be taken in case of problems and redundant transfers are avoided.

- From the DIRAC WMS service side, the late binding of jobs to resources allows further optimizations and maintaining a central queue simplifies job resubmission (e.g. to recover failing jobs), easing the job management and helping to increase the rate of successful final job states.

Furthermore, DIRAC provides a High Throughput Computing PaaS, which improves the general job throughput compared with native management of grid storage. This is possible at the DIRAC configuration level, connecting compute and storage resources for particular use requirements. At the same time, the DIRAC File Catalogue includes replica, metadata and provenance functionality simplifying the development of scientific application adaptations for distributed environments [Smith & Tsaregorodtsev 2008].

DIRAC data and job management systems ensure proven production scalability up to peaks of 80,000 concurrently running jobs for the LHCb experiment [Tsaregorodtsev et al. 2004]. This is, by far, large enough for the computing requirements of environmental science in a sensible temporal horizon.

The first challenge is to improve scientific application throughput using the DIRAC engine. This is a new programming model for the communities which require DIRAC client installation within their services, to adapt necessary components to the DIRAC API for high-level data management. At the same time, the job execution wrappers have to be developed for each use case, which are also using the DIRAC API provided by the runtime environment in the client side. DIRAC has a simple workflow engine, which needs improvements and adaptions, to cover the VRE complex pipelines requested. Additional developing integration is also required in the VRE service side by using the DIRAC API.

2.2 Data Processing in ENVRIplus Reference Model and Architecture

The ENVRI Reference Model [Hardisty et al. 2017] has been developed (and will be in continuous evolution) to provide environmental Research Infrastructures with a unifying view on the “systems” they are developing in terms of processes, “sub-systems”, connections among them, etc. It is envisaged to be an authoritative document facilitating communication and collaboration among major players by removing barriers resulting from the use of diverse terminologies and/or the use of different models to capture a common concept.

The current version of the Reference Model is driven by the Research Data Management lifecycle[21]: (a) Data Acquisition, (b) Data Curation, (c) Data Publishing, (d) Data Processing, and (e) Data Use. Although there is a specific phase named “data processing”, in the reality ‘data processing’ tasks are present in other phases too (e.g. data curation). However, the following minimal functions are included in the Data Processing phase[22]:

- ***Data Assimilation***: Functionality that combines observational data with output from a numerical model to produce an optimal estimate of the evolving state of the system.
- ***Data Analysis***: Functionality that inspects, cleans, and transforms data, providing data models which highlight useful information, suggest conclusions, and support decision making.
- ***Data Mining***: Functionality that supports the discovery of patterns in large datasets.

[21] ENVRI plus. (2016). ENVRI RM V2.1, Model Overview, The Research Data Lifecycle. https://wiki.envri.eu/display/EC/Model+Overview#ModelOverview-ENVRI_Data_LifecycleTheResearchDataLifecyclewithinEnvironmentalResearchInfrastructures

[22] ENVRI plus. (2016). ENVRI RM V2.1, Model Overview, Common functions within a common lifecycle https://wiki.envri.eu/display/EC/Model+Overview#ModelOverview-CommonFunctionswithinaCommonLifecycle

- ***Data Extraction***: Functionality that retrieves data out of (unstructured) data sources, including web pages, emails, documents, PDFs, scanned text, mainframe reports, and spool files.
- ***Scientific Modelling and Simulation***: Functionality that supports the generation of abstract, conceptual, graphical or mathematical models, and to run instances of those models.
- ***(Scientific) Workflow Enactment***: Functionality provided as a specialisation of Workflow Enactment supporting the composition and execution of computational or data manipulation steps in a scientific application. Important processing results should be recorded for provenance purposes.
- ***Data Processing Control***: Functionality that initiates calculations and manages the outputs to be returned to the client.

The implication of this on the planned data analytics platform (cf. Sec. 3) is quite immediate. Since it is almost impossible to realise a platform having on board all the possible methods and approaches for the envisaged functions it must be easy for scientists and practitioners to plug their specific methods into the platform. Moreover, the platform should play the role of facilitator and multiplier. This means that whenever a method is added to the platform, it should be easy for actors different from those producing the method to make use of it, to benefit from it and to minimizing "reinventing the wheel".

The following roles are associated with the Data Processing phase[23]:

- ***Data Provider***: Either an active or a passive role, which is an entity providing the data to be used.
- ***Service***: A passive role, which makes functionality for processing data available for general use.
- ***Service Consumer***: Either an active or a passive role, which is an entity using the services provided.
- ***Service Provider***: Either an active or a passive role, which is an entity providing the services to be used.
- ***Service Registry***: A passive role, which is an information system for registering services.
- ***Capacity Manager***: An active role, which is a person who manages and ensures that the IT capacity meets current and future business requirements in a cost-effective manner.
- ***Data Processing Subsystem***: In the Science Viewpoint, the data processing subsystem represents a passive role of the data processing community. It is the part of the research infrastructure providing services for data processing. These services could require authorisation at different levels for different roles.
- ***Processing Environment Planner***: An active agent that plans how to optimally manage and execute a data processing activity using RI services and the underlying e-infrastructure resources (handling sub-activities such as data staging, data analysis/mining and result retrieval).

These roles suggest that the model underlying data processing is essentially "service oriented" and that both the entire platform and its constituents (those offering specific processing facilities) should adhere to this model and make themselves discoverable via a registry. For what regards the processing planner, this is a facility that can be split across two diverse

[23] ENVRI plus. (2016). ENVRI RM V2.1, Science Viewpoint, Roles in the Data Processing Community https://wiki.envri.eu/display/EC/SV+Community+Roles#SVCommunityRoles-roles_serRolesintheDataProcessingCommunity

components, the data processing solution envisaged here (cf. Sec. 3) and the one on optimisation resulting from T7.2 [Martin et al. 2016].

When analysing the Computational Viewpoint related with data processing (cf. Figure 2)[24], two typologies of components are suggested: (a) a ***coordination service*** that takes care of delegating "all processing tasks sent to particular execution resources, coordinates multi-stage workflows and initiates execution" and (b) a set of ***process controllers*** each representing a computational functionality of registered execution resources.

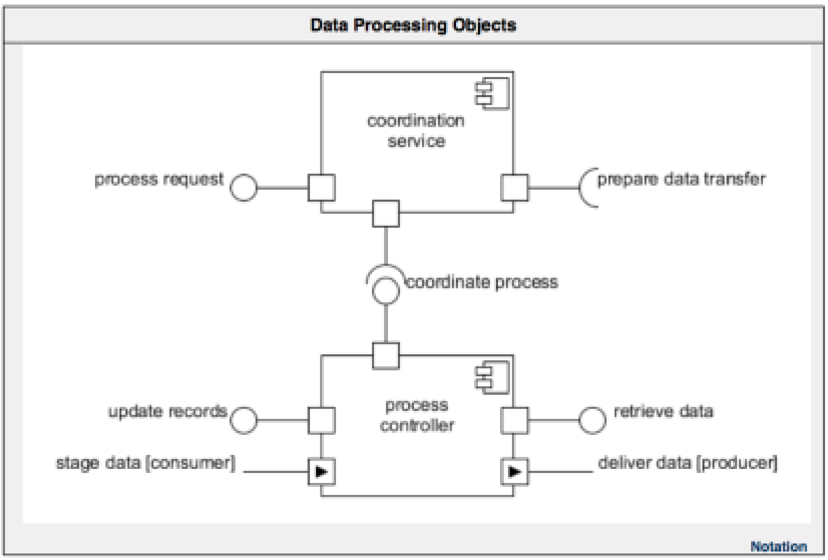


FIGURE 2. ENVRI REFERENCE MODEL: DATA PROCESSING COMPUTATIONAL VIEWPOINT

This model is underlying the envisaged data processing solution (cf. Sec. 3) where both the components are implemented by masters and workers.

2.3 Data Processing in ENVRI RIs

An extensive and detailed analysis of the needs and approaches characterising data processing in ENVRI Research Infrastructures has been presented in [Atkinson et al. 2015]. To capture such needs it was decided to focus on collecting major aspects that characterise each RI's data processing needs:

- **Input**, i.e., what are the characteristics of the dataset(s) to be processed? This includes dataset(s) typologies, volume, velocity, variety/heterogeneity, and access methods;
- **Analytics**, i.e., what are the characteristics of the processing tasks to be enacted? This includes computing needs quantification, implementation aspects including programming languages, standards and re-use potential;
- **Output**, i.e., what are the characteristics of the products resulting from the processing? This includes typologies, volume, variety, variety/heterogeneity, and availability practices.

An excerpt of the findings [Atkinson et al. 2015] is reported below.

Input As largely expected, RIs' needs with respect to datasets to be processed are quite diverse because of the diversity in the datasets that they deal with. Datasets and related practices are diverse both across RIs and within the same RI. For instance, in EPOS there are many communities each having its specific typologies of data and methodologies (e.g., FTP)

[24] ENVRI plus. (2016). ENVRI RM V2.1, Computational Viewpoint, Objects and Subsystems, Data Processing https://wiki.envri.eu/display/EC/CV+Data+Processing

and formats (e.g., NetCDF, text) for making them available. Time series and tabular data are two very commonly reported types of dataset to be processed yet they are quite abstract. Regarding “volume”, datasets vary from a few KBs to GBs, and sometimes TBs. In the large majority of cases datasets are made available as files while few infrastructures have plans to make or are making their data available through OGC services, e.g., ACTRIS. The need to homogenise and promote state-of-the-art practices for data description, discovery and access is of paramount importance to provide RIs with a data processing environment that makes it possible to easily analyse datasets across the boundaries of RI domains.

Analytics When moving to the pure processing part, it emerged that RIs are at diverse levels of development and that there is a large heterogeneity. For instance, the programming languages currently in use by the RIs range from Python, Matlab and R to C, C++, Java, and Fortran. The processing platforms range from Linux servers in the case of ACTRIS to HPC approaches exploited in EPOS. With respect to licences, software in use or produced tends to be open source and freely available (with some exceptions, e.g., Matlab). In the majority of cases there is almost no shared or organised approach to make available the data processing tools systematically both within the RI and outside the RI. One possibility suggested by some RIs is to rely on OGC/WPS for publishing data processing facilities. Some care needs to be taken balancing the benefits of common solutions with the need to support a wide range of working practices well. The platform should be “open” and “flexible” enough to allow (a) scientists to easily plug-in and experiment with their algorithms and methods without bothering with the computing platform, (b) service managers to configure the platform to exploit diverse computing infrastructures, (c) third-party service providers to programmatically invoke the analytics methods, and (d) to support scientists executing existing analytic tasks eventually customising/tuning some parameters without requiring them to install any technology or software.

Output In essence, we can observe that the same variety characterising the input exists for the output. In this case, however, it is less well understood that there is a need to make these data available in a systematic way (e.g. for reproducibility), including information on the entire process leading to the resulting data. In the case of EMBRC it was reported that the results of a processing task are to be made available via a paper (e.g. a scientific paper with data as additional material or a data paper [Candela et al. 2015]) while for EPOS it was reported that the datasets are to be published via a shared catalogue describing them by relying on the CERIF metadata format. In many cases, but by no means all, output resulting from a data processing task should be “published” to be compliant with Open Science practices. A data processing platform capable of satisfying the needs of scientists involved in RIs should offer an easy to use approach for having access to the datasets that result from a data processing task together. As far as possible it should automatically supply the entire set of metadata characterising the task, e.g., through a provenance framework. This would enable scientists to properly interpret the results and reduce the effort needed to prepare for curation. In cases where aspects of the information are sensitive, could jeopardise privacy, or have applications that require a period of confidentiality, the appropriate protection should be provided.

In the remainder of the section, we report a per-RI perspective on the data processing tasks with the goal of highlighting their specific understanding of this activity, their current development status and any plans for future activity.

The seven RIs covered in Sections 2.3.1 to 2.3.6 (two are discussed in 2.3.3) report their data processing in very helpful detail. Such detail is necessary for working with each RI and it underpins the detailed development of computation provision for those RIs. Readers may skip to the specific RI in which they are interested or return to these sections after reading the summaries and strategy in Section **Error! Reference source not found.**2.4 or 3 respectively.

2.3.1 Data Processing in ACTRIS

By M. Fiebing (NILU) on behalf of the ACTRIS community

ACTRIS measurement data are available through the ACTRIS Data Portal[25]. The data are handled by 3 highly specialised topic data repositories. At the start of ACTRIS-2, measurement data from about 60 sites and ~130 different atmospheric variables were included in the ACTRIS data centre. The data curation is closely linked to the networking activities, and to the calibration centres to facilitate and ensure standardized and comparable procedures throughout the infrastructure. Since August 2016, the ACTRIS data centre handles data from about 75 sites and ~130 different atmospheric variables, of these ca. 100 different trace gases, 10 different aerosol variables measured near the surface, 10 aerosol profile variables, and 8 cloud variables. The data are resulting from ca. 30 different methodologies, both near surface and remote observations, with time resolution ranging from seconds to 1 week.

The ACTRIS Data Portal is a metadata catalogue. Development, management and maintenance of the data flow to the ACTRIS Data Portal is a centralised task performed by NILU, and the portal is up and running close to 24/7. Figure 3 shows the main structure of the portal. The metadata catalogues are updated regularly, every night through various procedures, so potentially new data added to the topical data bases are available through the portal latest the following day. The structure is flexible, e.g. to add and change access to topic databases, implementation of various password and registrations procedures etc.

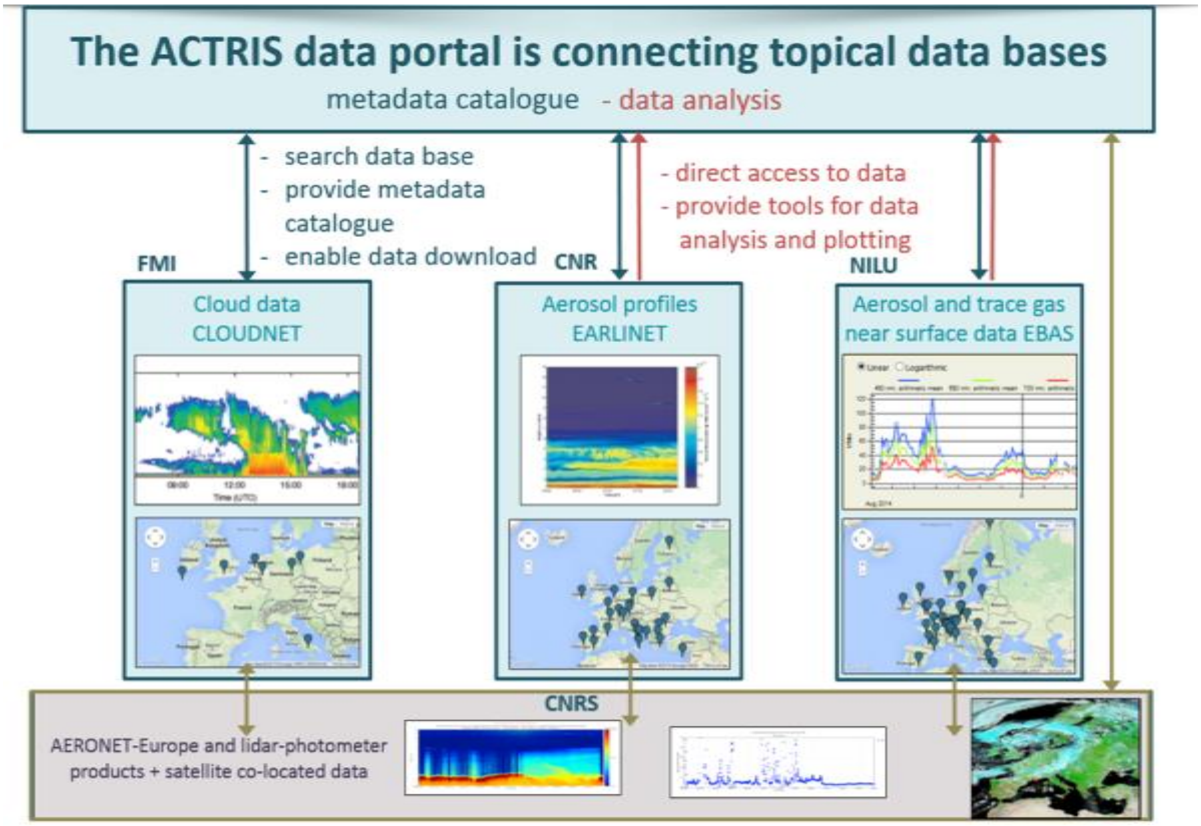

The ACTRIS data portal is connecting topical data bases
metadata catalogue - data analysis
- search data base
- provide metadata catalogue
- enable data download
- direct access to data
- provide tools for data analysis and plotting
FMI
CNR
NILU
Cloud data CLOUDNET
Aerosol profiles EARLINET
Aerosol and trace gas near surface data EBAS
CNRS
AERONET-Europe and lidar-photometer products + satellite co-located data

FIGURE 3. OVERVIEW OF THE CORE STRUCTURE OF THE ACTRIS DATA CENTER

[25] ACTRIS Data Centre http://actris.nilu.no

The data curation of the ACTRIS primary measurements data is organised in the 3 specialised data repositories:

- All aerosol and trace gas near surface data are archived in EBAS[26] under the responsibility of NILU;
- All aerosol profile data are archived in EARLINET database[27] under the responsibility of CNR;
- All cloud profile data are archived in Cloudnet DB[28] under the responsibility of FMI.

In addition, AERIS-ICARE is the forth topic database and offers satellite data support to facilitate products combining with ACTRIS ground data with Earth observation data.

All data repositories are linked in the ACTRIS Data Portal[29], and the ACTRIS measurements data are accessible also through the portal. Additionally, the portal provides access to secondary data. Secondary datasets are derived from primary measurement data by e.g. averaging, filtering of events, interpolation of data. Secondary datasets are usually the result of analysis for a targeted article, special studies or processed for model experiments. Primary datasets are regularly updated mainly due to extension of additional years. Secondary datasets are normally not updated over time.

The ACTRIS data management plan [Lund Myhre et al 2015] describes requirements and recommendations for ACTRIS datasets, the data flow, how the data is made available, and the data repositories. It presents a list with all ACTRIS atmospheric variables together with recommended methodology.

For data processing, every database discussed above has its own solutions and workflows.

Data Processing Workflow at the ACTRIS Data Centre Node CLOUDNET. Cloudnet comprises the cloud vertical profile component of ACTRIS. The original objective for Cloudnet was the routine-based automated evaluation of the representation of clouds in numerical weather prediction (NWP) models using observations derived from a combination of ground-based remote-sensing instruments, whereby the processing framework has been designed with this in mind. The Cloudnet processing suite covers the entire processing chain, from ingesting raw instrument data through post-processing to product generation. Due to the large data volume, there are two options available for data originators: the Cloudnet processing scheme can be installed on-site, with local generation of products and transfer of products to the central Cloudnet data server, or the raw instrument data can be transferred to the central Cloudnet server for processing. The Cloudnet processing scheme ensures that instrument data have been processed according to the procedures recommended by the ACTRIS Cloudnet community. This includes checking that the metadata for each raw data stream are present, and that measurement uncertainties are implemented for each individual measurement. These parameters are then propagated through the processing scheme. All Cloudnet products at each step are in NetCDF format. The standard Cloudnet station requires the following instruments: cloud radar, ceilometer or lidar, multi-channel microwave radiometer, and thermodynamic profiles from a radiosonde or NWP model. Cloudnet uses a level-based framework (Figure 4).

[26] EBAS http://ebas.nilu.no/

[27] AERLINET-ASOS http://access.earlinet.org/

[28] Cloudnet Data Archive http://www.cloud-net.org/data/

[29] ACTRIS Data Portal http://actris.nilu.no/

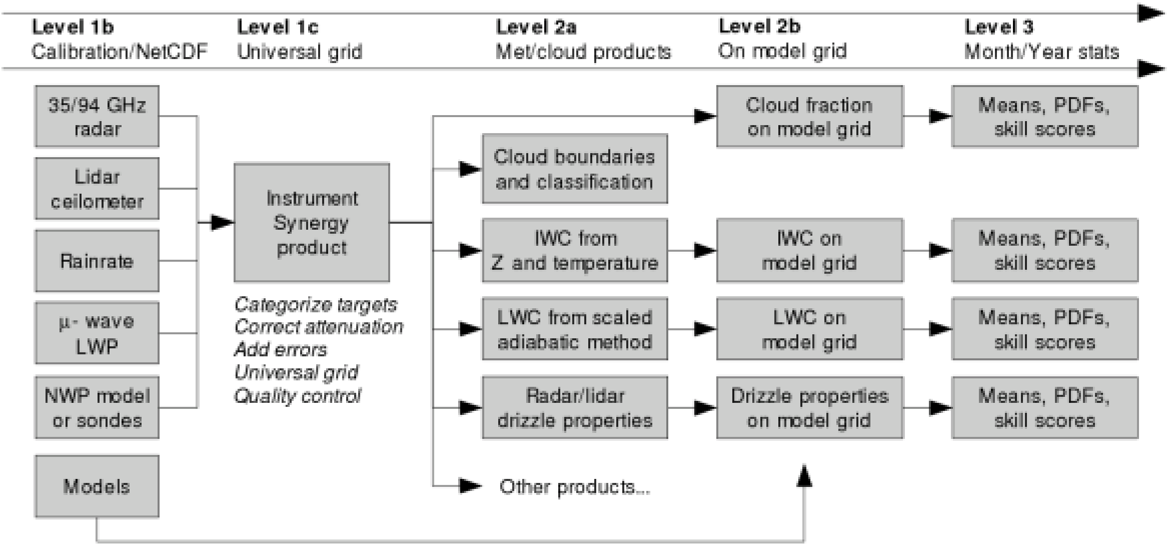


FIGURE 4. ACTRIS DATA PROCESSING WORKFLOW AT CLOUDNET

Level 1 deals with the processing of observations from different instruments and their subsequent combination to provide a single synergistic product (Level 1c) on a well-defined time-height grid. *Level 1c* is the basis from which all Cloudnet products are created. High resolution products (if possible at the native instrument resolution if possible) are created in *Level 2a*, and are used for all scientific studies. Specific products for model evaluation are created in *Level 2b*, where the high resolution products are averaged onto the grid of each individual model, and in *Level 3* and beyond, aggregated into monthly and yearly files containing a wide range of statistical measures. Measures include: means, distributions, and joint-pdfs for creating the contingency tables used for deriving the skill score of choice. From these files, a wide range of metrics are then routinely plotted and analysed.

Single-instrument processing. All individual observations are pre-processed, quality-checked, and Cloudnet-formatted by stage *Level 1b*. Addition of metadata and reformatting to NetCDF is performed if necessary, as is calibration and generation of quality flags. For most instruments, this pre-processing is now performed internally to the instrument. Instruments at certain sites are sometimes operated in modes that require temporal and spatial averaging, or identification and removal of non-vertically-pointing data. This level also includes NWP model output that has been transformed into Cloudnet format for ease of use in creating higher-level products.

Instrument combination and target categorisation. This product is the basis of the Cloudnet output. All instrument measurements and their associated uncertainties are present, together with the attenuation corrections applied to the cloud radar data. This product contains the metadata detailing the sources and processing history (e.g. processing software version). Additionally, the product contains the full target categorization (presence or absence of: cloud, cloud phase, precipitation, aerosol, insects etc.) and quality flags detailing the usability of the data (e.g. precipitation causing unknown attenuation which impacts the generation of the ice water content product).

Products. The standard procedure is to start from the target categorization dataset (**Level 1c**), as this contains the relevant information required for providing consistent uncertainty estimates for all products. This ensures that each algorithm has standard well-defined inputs, flags, and metadata, enabling reliable automated generation. The design of the processing framework also makes it simple to add new retrieval methods developed by the scientific community. These can then be applied across all sites and instruments. Again, the standard procedure is to start from the target categorization dataset (**Level 1c**). The new algorithm should respect the in-built quality control flags, and propagate the instrument uncertainties through to the retrieved parameters.

Higher-level products. Evaluating the representation of clouds in climate and (NWP) models is not straightforward. For NWP models, this task is compounded by the expectation of a good forecast, as well as the reliable representation of the specific cloud parameters themselves. Cloudnet has developed and implemented a comprehensive suite set of objective metrics for the evaluation of model cloud parameters, in continual joint collaboration with operational modellers. The set of evaluation metrics is designed to investigate both the climatological aspects required of a climate model, and the ability to forecast the correct cloud at the right time, a necessary validation for NWP. These products utilise the quality flags to understand the impact of conditional sampling arising from the discarding of profiles deemed unreliable (e.g. precipitation causing unknown attenuation).

The Cloudnet processing framework and workflow was historically optimised towards **near-real-time** (NRT) generation of products and associated Quicklook archives (i.e., sort of “landing pages” for accessing product snapshots). Automated processing for all stages, including higher-level products, is implemented. NRT generation requires the transfer of NWP model data to the Cloudnet server, and then to the station if processing is performed on-site, since most stations do not have access to, or are not located close to, operational radiosondes. The NRT stream of data and Quicklooks is fully automated. To enable version control and manual data curation, a second stream will be created. This enables instrument artefacts or malfunction spotted at a later date to be accounted for, and longer term recalibrations to be applied if necessary.

Data Processing Workflow at the ACTRIS Data Centre Node EBAS. The EBAS database is the ACTRIS Data Centre node for near-surface observations. EBAS has been in operation since the 1970ies, with basic funding secured through the EMEP protocol of the Convention on Long-Range Transport of Air Pollution (CLRTAP). It is also used by a number of other national and international monitoring programs including WMO Global Atmosphere Watch (World Data Centre for Aerosols (WDCA), World Data Centre for Reactive Gases (WDCRG)), Arctic Monitoring and Assessment Programme (AMAP), OSPAR, HELCOM, as well as by more than 50 different EU research projects. The database is tailored to host a wide range of atmospheric composition measurements and instrumental auxiliary data, mainly from fixed near-surface sites, but also mobile platforms. Data reporting standards have been established for more than 600 different chemical and physical atmospheric variables. EBAS hosts currently more than 50.000 datasets from 71 different countries, 1060 stations, 608 component types and 94 instrument types. Thus, it is probably the most comprehensive international database for atmospheric composition data from ground based instrumentation worldwide. EBAS users comprise the data provider communities themselves, e.g. for purposes of quality assurance, comparisons, and climatology, the atmospheric modelling communities, e.g. for model initialisation and verification, and authorities, e.g. for assessing air quality. EBAS consists of a relational database back-end that is designed for tracking versions of data, and includes a rich set of metadata comprising discovery metadata and rich set of application specific use metadata, e.g. on systematic and statistical uncertainties, operating procedures, quality assurance measures, sample treatment, etc. The backend is complemented by a web-interface (http://ebas.nilu.no) that allows search (criteria based), display, and download of the data holdings. Web-service interfaces to metadata and data are available as well.

The data processing workflow for EBAS can be separated in 3 branches:

- **Data from online instruments reported at regular (annual) schedule.** Data from online instruments are collected and temporarily stored at the station. In the responsibility of the instrument’s principal investigator (PI, also called data originator), the data are transcribed into Level 0 format containing all the original raw data in original time resolution, and annotated with discovery and use metadata. In a step of manual quality assurance applying flags for instrument malfunctions (flag invalidates data) and conditions such as local influence or calibration periods (flag is informative), Level 0a is reached. With the step of applying remaining calibrations and corrections, the data reaches Level 1a. The final, fully manually quality assured data product, Level 2, is reached by calculating hourly averages starting and ending at the turn of the hour, disregarding invalid data and copying any informative flags occurring in an hour. All 3 data levels (0a, 1a, and 2) are uploaded to the data centre, levels 0a and 1a for direct archiving, Level 2 inserted into the database and made publicly available.

- **Data from online instruments reported at near-real-time (NRT) schedule.** In the NRT branch of the data flow, the Level 0 data are assembled automatically by the data acquisition system at the stations in hourly intervals, and are pushed to the data centre at an hourly schedule. Automatic flagging results in data Level 0b. Applying remaining calibrations and corrections yields Level 1b, and hourly averaging Level 1.5, the final NRT data product. NRT data are targeted towards operational applications and open only for users able to assess data quality automatically in an operational setting.
- **Data from offline measurements.** Offline measurements by exposing a sample medium (e.g. a filter or a flask) to a sample air stream at a station and collecting a targeted part of the sample air stream. The sample medium is subsequently transported back to a laboratory, where it is analysed offline for the targeted constituent. The workflow consists of individual steps for pre-treatment of the sample medium, exposition of the filter medium at the field station, and sample analysis. Each step results in a protocol for establishing traceability of the measurement. At the institute of the PI, the information from these protocols is condensed into the final, manually quality assured data product containing not only the concentration of the targeted atmospheric constituent, but also discovery and use metadata, e.g. operating procedure, quality assurance measures, and special conditions occurring at any step in the process. The Level 2 data product is uploaded to the data centre.

Level 2 data submitted to the data centre are subject to an additional, semi-automatic and manual, quality assurance procedure at the data centre. Metadata and data are inspected for correctness of syntax, semantic consistency, and the data themselves inspected for outliers. Depending on the measured quantity, additional checks are applied, e.g. inter-annual consistency and consistency between measurements of different, collocated instruments (closure). These tests are refined in collaboration with international frameworks (EMEP, GAW), reflecting their needs and demands. The data submitter receives feedback of the outcome of the tests, either inter-actively for the initial checks, or subsequently for the additional checks. Upon removal of issues with the submission, the data are published.

Data citations services are scheduled to be implemented for all public data, i.e. Level 2 data. The data are to be tagged with primary Digital Object Identifiers (DOIs), i.e. at fixed granularity, i.e. one DOI for each annual dataset per instrument. Further, it is planned to take Persistent Identifiers (PIDs) into use to identify each pre-final data level item and each procedure or algorithm used for data processing in all versions.

Data Processing Workflow at the ACTRIS Data Centre Node EARLINET

EARLINET (European Aerosol research Lidar Network) is the aerosol vertical profile component of ACTRIS. EARLINET aerosol profile data are collected at each EARLINET station applying procedures compliant with the recommendations and procedures provided by the ACTRIS profile community, harmonised within EARLINET. Data are reported in NetCDF format to the EARLINET database in accordance with the reporting procedures. The EARLINET database contains at the moment only quality assured aerosol optical profile data, but a new structure of the database has been designed to meet the wide request from the users of more intuitive products from research communities and to face the even wider request related to the new initiatives of EU such as the Copernicus programme. The ACTRIS RI provides to the EARLINET data originator support for centralized data processing through the common standardized automatic analysis software developed within EARLINET, the Single Calculus Chain (SCC). The Single Calculus Chain (SCC) is an open source software for analysing aerosol lidar data to obtain aerosol optical property profiles from raw data. The SCC is an official EARLINET developed tool to accomplish the fundamental need of any coordinated lidar network to have an optimized and automatic tool providing high-quality aerosol property profiles. Main concepts at the base of the SCC are automation and full traceability of quality-assured aerosol optical products. The SCC has been installed on a single server hosted and managed by CNR-IMAA. The users can connect to the server and use or configure the SCC retrieval procedures for their data using a web interface. With this approach, it is possible to keep track of all the system configurations used to calculate each product and also to certify which configurations are quality assured. Moreover, in this way it is always guaranteed that the same and latest SCC version is used to produce optical products.

The EARLINET data processing workflow is schematically shown in Figure 3.

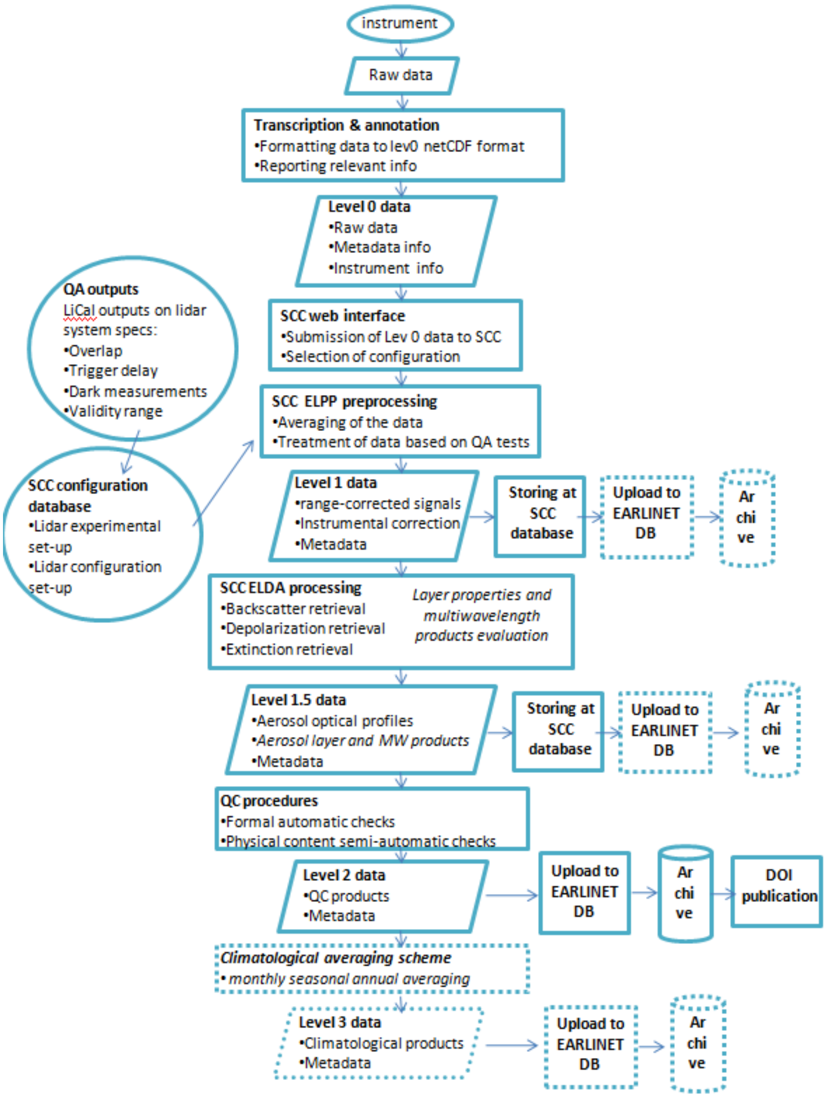


FIGURE 5. ACTRIS DATA PROCESSING WORKFLOW AT EARLINET

Level 1 contains pre-processed lidar data, i.e. a step in between the raw signal (Level 0 data stored at each station) and the optical properties, where all instrumental corrections are already implemented. The Level 1 data are the base for the retrieval of the optical properties contained in the Level 1.5 products. The Level 1.5 datasets are not quality checked, except for format aspects, and therefore released as soon as data originators submit them to the database. Afterwards, all Level 1.5 data pass through quality check procedures. Only the data that passed the quality checks go into Level 2 which is therefore the quality checked optical properties level. Finally, Level 3 data contain climatological datasets retrieved from Level 2 optical products.

Steps in between the different Levels are schematically described in Figure 5 referring to SCC, but data processing from raw data to aerosol optical properties can even be performed locally at each station using well documented quality-assured retrieval under the Data Originator's responsibility. These data undergo the QC procedures as well.

The processing steps can be summarized as follows:

- SCC web interface: This module represents the interface between the raw data originator and the SCC. The web interface provides a user-friendly way to interact with the SCC database by using any web browsers. The SCC database is a relational database which handles the large number of input parameters needed for the retrievals of aerosol optical products from lidar signal. Two different types of parameters are needed: experimental (which are mainly used to correct instrumental effects) and configurational (which define the way to apply a particular analysis procedure). Typically, this information is provided by the QA procedures applied through the Lidar Calibration centre (http://lical.inoe.ro). A web interface and other processing steps allow monitoring the processing status of each measurement.
- Pre-processor module (ELPP: EARLINET Lidar Pre-Processor): The ELPP module implements the corrections to be applied to the raw lidar signals before they can be used to derive aerosol optical properties. Following the EARLINET quality assurance program, some instrumental effects are corrected. The raw lidar signals have to be submitted in a NetCDF format with a well-defined structure. Outputs are stored in an internal SCC archive and then on the EARLINET database (currently available through agreement to external users).
- Optical processor module (ELDA: EARLINET Lidar Data Analyser): ELDA enables the retrieval of particle backscatter coefficients, the calculation of particle extinction coefficient and finally the computation of particle and volume linear depolarization profiles. New modules, currently under development, will allow for multi-wavelength and layer products. The final optical products are written in NetCDF files with a structure fully compliant with the EARLINET database, and then submitted to the EARLINET database.
- QC procedures: Level 1.5 data collected at the EARLINET database are subject to centralized quality check procedures. Technical checks are procedures for the control of the file content with respect to the file structure as defined in the EARLINET database. In addition, scientific content QCs are related to the content of the EARLINET files in terms of validity of the EARLINET measured parameters. The data submitter receives feedback of the outcome of the QC. Data compliant to all the QC requirements are made publicly available as Level 2 data. Level 2 data will be regularly published with Digital Object Identifiers (DOIs).
- Averaging for climatological products: Level 2 aerosol optical products will be aggregated into monthly, seasonal and annual datasets for both profiles and integrated quantities. Information about the number of collected samples, mean, median and standard deviation of the properties, as well as mean statistical error for each property are reported. Metrics of the comparison with reference datasets (as AERONET for AOD) will be reported whenever available, in order to provide information about data representativeness.

The implementation of data versioning and use of Persistent Identifiers (PIDs) is planned, taking advantage from the SCC design to support the full traceability concept.

Even though the three ACTRIS Data Centre nodes differ, in the sense that they reflect the different nature of the data they are handling and the different service demands and foci, the tasks in the data curation workflows are similar and demand similar solutions, offering synergy benefits. As Research Infrastructure, ACTRIS is now entering its Preparatory Phase. Overarching tasks for the Data Centre in this phase will include: (a) Joint transition from customised data policies to one common, widely used data license; (b) Introduction of persistent identifiers (PIDs) in all data centre nodes for version tracked identification of all data pre-products and all data processing tools, including record of provenance.

2.3.2 Data Processing in EISCAT_3D

By C.-F. Enell on behalf of the EISCAT_3D Community

The European Incoherent Scatter Scientific Association (EISCAT) is an international organisation conducting fundamental research into solar-terrestrial physics and atmospheric science. It has successfully operated incoherent scatter radars in northern Fenno-Scandinavia and on Svalbard for more than 35 years. Currently EISCAT is preparing the next generation incoherent scatter radar system to be constructed in northern Norway, Sweden and Finland, EISCAT_3D. In contrast to the present radars, which all use parabolic dish antennas looking in one direction at a time, EISCAT_3D will use multiple radar sites with array antennas, from which up to 100 simultaneous beams will be formed at high time resolution by means of digital signal processing. This will introduce significant challenges in the handling of data, which will be generated at rates and volumes comparable to such experiments as the Large Hadron Collider or the Square Kilometer Array.

Data from the legacy EISCAT radars are archived as power spectral domain (Level 2) data in Matlab mat version 4 compatible files. EISCAT offers online analysis and plotting of these files through a web based schedule system (Python CGI + Matlab software)[30].

Analysed data (Level 3, i.e. fitted physical parameters) are distributed through the Madrigal system[31]. Madrigal allows both web and API access. The Madrigal system offers plotting and calculation of derived parameters using the primary parameters and models. This system, however, is only suited for small volumes of data in time series format, and will not scale very well to the 3-D volumes of analysed data that will be produced by EISCAT_3D.

[30] http://www.eiscat.se/schedule/schedule.cgi

[31] Madrigal database at EISCAT www.eiscat.se/madrigal

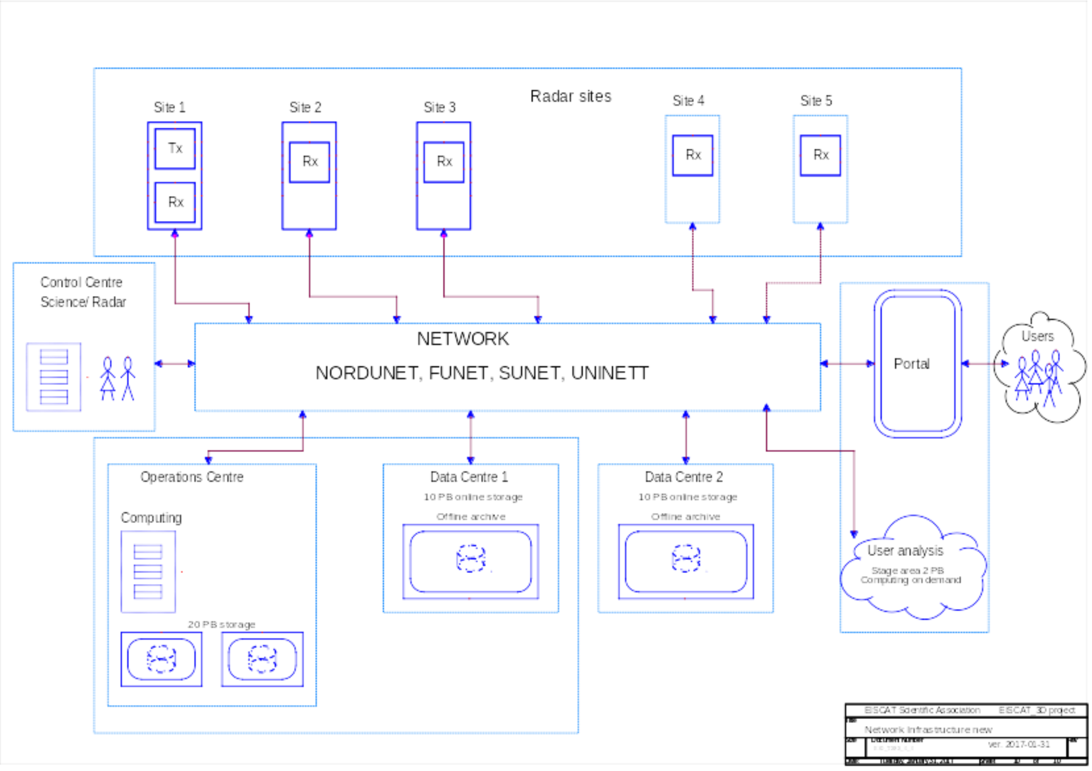


FIGURE 6. EISCAT_3D E-INFRASTRUCTURE: ACQUISITION, DATA TRANSFER NETWORK, DATA CENTRES FOR CURATION, CATALOGUING AND ARCHIVAL, USER ACCESS AND PROCESS COMPUTING

Figure 6 shows a top level block diagram of the planned e-infrastructure design of the EISCAT_3D system. The blocks correspond to some extent to the ENVRI reference model viewpoints: Data acquisition will take place at 3 to 5 radar sites. Beam-formed data will leave the sites through 100 Gb/s optical fibers provided by the Nordic NRENS. The sites will be monitored (from both engineering and scientific viewpoints) from a *control centre* (which does not have to be a physical location, but would rather be a gateway for system access). The *operations centre* represents the computation viewpoint, performing real-time analysis and visualisations, and it will also handle the first phase of data curation by adding necessary metadata. The data will then be archived, catalogued and curated at two redundant *data centres*, one of which should be closely connected to the operations centre.

The intention is that all data access and analysis should take place through a *user portal*, which will enable metadata discovery and analysis workflow management including reanalysis and data mining. The portal should have a web-based GUI as well as an API with web services and bindings for common programming languages.

As a prototype, a portal for existing EISCAT data has been implemented using DIRAC (cf. Sec. 2.1.2.1). This prototype supports:

- the plotting of data files;
- analysis tasks (by requesting a virtual machine running Matlab software on Octave).

A similar experimentation is ongoing by relying on DataMiner (cf. Sec. 3).

The archives and data processing systems envisaged in EISCAT_3D should:

- Use open standards; storage as HDF5 is planned;
- Store self-consistent metadata. Legacy EISCAT data files do not describe the format of the stored data, which is difficult to understand. At present, metadata must therefore be collected from several sources in order to analyse a data file. A metadata catalogue (e.g.

in an SQL database) is necessary for fast data discovery. In addition, EISCAT_3D data files should also contain self-sufficient metadata for setting up and running the analysis. This is addressed by the EISCAT_3D data model development in collaboration with the EGI-Engage Competence Centre for EISCAT_3D as well as the ENVRI+ reference model;

- Design analysis software in languages with open source implementations (e.g. C, Fortran, R, Python);
- Implement user authentication and authorization according to EISCAT embargo rules. Only authenticated users with access rights according to EISCAT statutes should have access to raw data and analysis resources whereas visual Level 3 data (visualizations of physical parameters) should be open;
- Use a process pipeline management system, to be accessed through the user portal;
- Allow (re)analysis of online data (fast archive) as well as reprocessing of archived data, whether on offline system (such as tapes) or slower access disk systems;
- Implement parallel real-time processing pipelines for simultaneous experiments with different purposes.

Real-time analysis and user (re)processing of data will likely initially run on the EISCAT_3D radar sites by means of switching between different VMs for real-time processing and (re)analysis when site computing is idle (e.g. during low duty cycle operation). Dedicated central computing hardware may be installed at the operations centre at a later stage.

2.3.3 Data Processing in EMSO and FixO3

By R. Huber on Behalf of the EMSO and FixO3 Communities

Currently, in EMSO and FixO3 there are no plans to design or develop common data processing (except e.g. some basic QC).

Currently, neither FixO3 nor EMSO has a common data processing facility in place, no common technical solution has been set up, and no common services exist for water column data. Seismic data will be channelled via IRIS/EPOS services.

Some processing tasks performed at EMSO for **seismic data** are:

- ***Application to micro-seismicity***: (a) Distinguish between P-wave (compressional wave) and S-wave (shear wave); (b) Short-Time Average (STA) window sensitive to seismic events; (c) Long-Time Average (LTA) window provides information about the amplitude of seismic noise.
- ***Application to seismic signal analysis***: Probability density function is performed with the algorithm proposed by [McNamara and Buland 2004] and [McNamara et al. 2009] for data quality check and for a statistically reliable estimate of background seismic noise. This noise follows a typical model [Peterson 1993] and probability density function is able to detect deviations from this model.
- ***Application to joint seismic and methane signal analysis***: STA/LTA trigger algorithm to detect high frequency seismic signals (Short Duration Events) linked to methane peaks of emission. The algorithm is based on the ratio between STA and LTA. STA is sensitive to the seismic events while LTA provides information about the temporal amplitude of the seismic noise at the site. A Short Duration seismic event is 'declared' when STA/LTA exceeds a pre-set value.

For **water column data**, the scientific demands towards both infrastructures are extremely heterogeneous and do not allow the provision of “fit for all” solutions. Typical data processing tasks which currently are performed by individual EMSO and FixO3 related working groups during their specific scientific data analysis are e.g. time-series analysis (variability, spectral analysis (cyclicity, seasonality), trends), in-situ verification of remote sensing trends, input and verification of data for climate modelling and multivariate statistics in general. Some of the most recent approaches performed within the relevant scientific community are:

- *Application to sea ice change*: (i) Analyse a sequence of satellite images and extract sea ice indicators which show climate change signal (area of drift ice + area of fast ice), (ii) construct time series of the indicators and compare with other climate data, (iii) interpret the results;
- *Application with proxies*: (i) Compare time series with local and larger scale influence, e.g. NAO, (ii) automatic signal matching, using dynamic time warping (DTW) in sediment core;
- *Application to foraminifera*: use of an age-depth model to correlate various depths in the sediment core with age (biostratigraphy, magnetostratigraphy, radiometric dating);
- *Application to long term salinity data*: (i) detect trends and attribution to climate change: check that the change is not due to internal variability only, and evaluate the relative contributions of multiple causal factors to a change or event with an assignment of statistical confidence. This requires some knowledge of internal variability and of the expected responses to external forcing (e.g. GHG-only), (ii) statistical analysis to account for climate modelling and observational uncertainties, (iii) construct gridded records from all available dataset and perform data processing: calculate 3-month seasonal median anomalies, apply 1-2-1 moving average filter, average seasonal anomalies, inversely weighted by error, exclude high error and/or low coverage seasons, interpolate gaps; (iv) Analyse features of the dataset (e.g. trend and variability);
- *Application to ocean biogeochemistry*: (i) comparison between control run and warming run (e.g. primary production) with particular focus on (a) give highest priority to data poor regions, regions sensitive to change, and variables with inadequate spatial and temporal resolution and (b) provide network designers with long-term climate requirements at the outset [Karl et al. 1995]; (ii) estimation of the number of years needed to detect climate change trend from natural variability varies according to the standard deviation of the noise, the estimated trend and the auto-correlation of the noise (Weatherhead et al., 1998);
- *Application to multisensory data – multivariate analysis*: Multivariate vs. univariate to detect outliers; Multivariate Statistical Process Control including the frequently applied method Principal Component Analysis; Redundancy in sensors is useful for detecting various types of changes;
- *Application to pH evolution*: Variables can sometimes be fitted to sinusoidal expressions, where the periodical term accounts for the high seasonal variability, while the inter-annual one marks the year-to-year trend; Compute correction for seasonal cycle considering the average value for a year; Compute the corresponding seasonal de-trended residual;
- *Application to greenhouse gases*: Use of Allan variance helps analysing noise and stability of the analyser; Low and high pass filtering to define inter-annual and short term variations (Transform data into the frequency domain using a Fast Fourier Transform (FFT), Apply low pass filter function to the frequency data, Transform the filtered data to the real domain using an inverse FFT); Wavelet analysis to produce de-trended seasonal cycles;
- *Application to atmospheric measurements*: Annual trends of greenhouse gases calculated by fitting observations by an empirical equation of Legendre polynomials and harmonic functions with linear, quadratic, and annual and semi-annual harmonic terms; use of

FLEXPART[32], a powerful tool to investigate transport and source regions; use of OsloCTM3[33] a global Chemical Transport model, developed over 25 years at UiO/Cicero;

- *Application to ocean carbon cycle modelling*: Decomposition of surface pCO2 trend to global ocean according to different drivers; Assessment of ecosystem parameters through data assimilation at various time-series; stations have proven to be a promising method to reduce model uncertainties. Data assimilation strategy (sequential ensemble method): (i) Repeat the 1-D optimization for each different sites (Stat. M, HOT, BATS, etc.), (ii) Determine sensitive parameters and constrainable with observations (NO3, PO4, Si, pCO2), (iii) Future predictions and comparison with non-data assimilated projections.
- Application to carbon cycle from station M in the Norwegian sea: study inter-annual changes: (i) comparing season by season gives a clearer picture, (ii) removing seasonality gives information on inter-annual trend, (iii) de-trending gives information on seasonality, (iv) eliminating processes (normalization of salinity, temperature, etc.), (v) Time period of study might influence the trend; Error bars that can reflect spatial variation (horizontal and vertical) and temporal variation (all seasons included);
- Use of statistical-physics to study extremes: Block maxima approach (annual temperature minima, daily wind speed maximum, etc.); Classical extreme value theory; Understanding the local transformation at each points is crucial to get the behaviour of extremes in climate-change-like problems; Statistics offers intuitive way to infer the shift in the mean and in the variability of weather extremes; Statistical Physics provides a methodological foundation for studying extreme events of complex fields and provides metrics for determining complexity and persistence of extremes;
- Scale-space techniques useful for a large number of different data sets: Kernel estimate; Spectral methodology for important periodicities in the series; Scale-space version for the spectral density to decide which peaks in a spectrum are real;
- Weather-health relationships: Functional form (spline function, linear threshold «hockey stick» model; Use of lag model (application for death related to heat wave); Poisson time series model; Long term trends in weather-health relationships (gradients and burdens) are influenced by multiple factors additional to meteorology – so difficult to interpret trends as marker of climate change;
- Non-stationarity in time series: (i) Non-stationarity, i.e. changes in the mean or variance of a process over long periods of time (or space); (ii) changes in relationships climate-ecosystem; (iii) Non-stationary relationships climate-ecosystems might be the rule rather than the exceptions.

2.3.4 Data Processing in EPOS

By D. Bailo (INGV) and F. Haslinger (ETHZ) on behalf of the EPOS Community

The EPOS overall RI has been designed to organize and manage the interactions among different EPOS actors and assets. To make it possible for the EPOS enterprise to work as a single, but distributed, sustainable research infrastructure, its structure takes into account technical, governance, legal and financial issues. Four complementary elements form the infrastructure: (i) the **National Research Infrastructures (NRIs)** contribute to EPOS while being owned and managed at a national level and represent the basic EPOS data providers. These require significant economic resources, both in terms of construction and yearly operational costs, which are typically covered by national investments that must continue during EPOS

[32] FLEXible PARTicle dispersion model https://www.flexpart.eu/

[33] Oslo CTM3 – A global chemical transport model http://folk.uio.no/asovde/osloctm3/

implementation, construction and operation; (ii) the **Thematic Core Services (TCS)** enable integration across specific scientific communities. They represent a governance framework where data and services are provided and where each community discusses its implementation and sustainability strategies as well as legal and ethical issues; (iii) the **Integrated Core Services (ICS)** represents the e-infrastructure consisting of services that will allow access to multidisciplinary resources provided by the NRIs and TCS. These will include data and data products as well as synthetic data from simulations, processing, and visualization tools. The ICS will be composed of the *ICS-Central Hub (ICS-C)* and *distributed computational resources including also processing and visualisation services (ICS-D)*. ICS is the place where integration occurs; (iv) the **Executive and Coordination Office (ECO)** is the EPOS Headquarters and the legal seat (ERIC) of the distributed infrastructure governing the construction and operation of the ICS and coordinating the implementation of the TCS.

The main concept is that the EPOS TCS data and services are provided to the ICS by means of a communication layer called the *interoperability layer*, as shown in the functional architecture (Figure 7). This layer contains all the technology to integrate data, data products, services and software (DDSS) from many scientific, thematic communities into the single integrated environment of the Integrated Core Services. The ICS represents the "core" of the whole e-infrastructure and those responsible for its implementation will provide the specification of the *interoperability layer*. The ICS is conceptually a single, centralized facility but in practice is likely to be replicated (for resilience and performance) and localized for particular natural language groupings or legal jurisdictions.

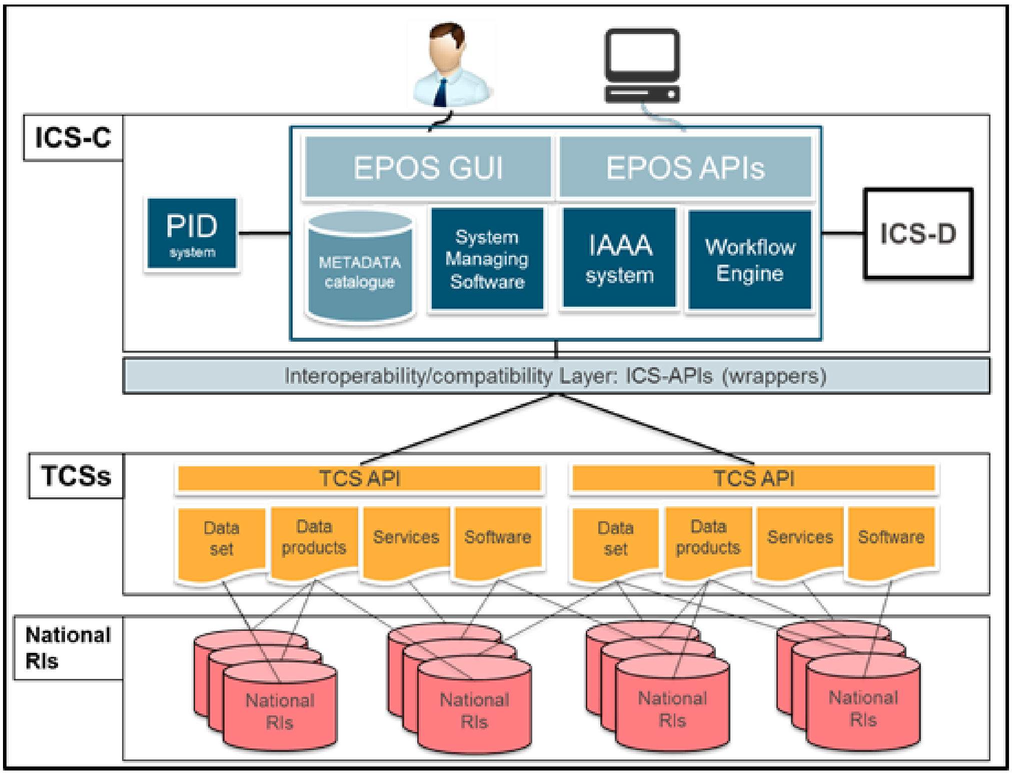


FIGURE 7. EPOS OVERALL ARCHITECTURE

A detailed description of the various components is beyond the scope of this deliverable. However, three components need to be better defined in order to understand how EPOS intends to deal with processing tasks: the *metadata catalogue*, the *workflow engine*, and the *ICS-D*.

ICS Metadata Catalogue. Metadata describing the TCS DDSS are stored using the CERIF data model. Metadata from the communities will be mapped to the metadata catalogue in order to create appropriate links between common concepts in different disciplines. This process involves the harmonization and interoperability of the various DDSS from the different TCSs through dedicated software modules. It requires TCS APIs for converting DDSS to the TCS specific metadata standard. It also requires ICS APIs (wrappers) to map and store metadata in the ICS metadata catalogue (i.e. CERIF). These TCS APIs and the corresponding ICS APIs collectively form the *interoperability layer*, which is the link between the TCSs and the ICS.

ICS Workflow engine(s) and provenance. A key aspect of the semi-automatic composition of software to meet a user request is the provision of a workflow to link together the software services as they access appropriate data. There exist many workflow engines and each of them fits different use cases and architectures (cf. Sec. 2.1). When devising the EPOS solution, the computational models (i.e. from the Computational Earth Science community) to be supported and the communities' requirements will be taken into account. Moreover, following the experience gained by the EPOS partners, in initiatives such as VERCE [Atkinson et al. 2015], particular attention will be dedicated to ***cross platform streaming libraries***. Moreover, by relying on CERIF, EPOS is planning to provide provenance information, since the linking entities associated with the role have, as attributes, both date/time start and date/time end. This handles versioning and – via the linking entity record – the relationship of one base entity instance (e.g. a dataset) to another. On the other hand, for a comprehensive traceability of the processes and agents that contributed to the generation of the research product, EPOS foresees the integration in CERIF with the W3C PROV-O standard [Lebo et al. 2013]. This will promote interoperability with other institutional data archives, fostering data preservation and curation across domains.

EPOS ICS-D. Integrated Cores Services – Distributed will include services from external computing facilities such as HPC (High Performance Computing) machines for modelling and simulation according to the requirements of the Computational Earth Science community, and HTC (High Throughput Computing) clusters for data intensive applications such as data mining. The data workflow will be managed by EPOS ICS-C in order to provide the end user with appropriate computational services, even though actual computations will be provided by ICS-D. Additional ICS-D services will provide visualization and processing capabilities. ICS-C will have to develop provisions for communicating with these external services in a seamless manner.

To better explain the typical needs arising in the EPOS domain, two user stories are exploited. The first one is simple and deals mostly with the discovery of the data (needed for further processing). The second one is more complex and deals with the actual processing of data.

Simple "friendly" user story. A seismologist likes to retrieve all earthquakes matching criteria such as:

- occurring in a given geographic area and time range, but excluding "false events" or anthropogenic induced events (such as explosions);
- occurring in a given area (defined by a bounding box and time range) filtered by magnitude (e.g. M>3);
- occurring in a given area (defined by a bounding box and time range) filtered by those which have been recorded by at least 5 stations within 150 km distance (all stations must be within that range);

so that she/he can carry out 'analytical tasks'[34] such as:

- visualize the catalogue in a map;

[34] In the reality, there is very little analytics here if we exclude some sort of "visual analytics".

- save the map view as a figure for publication, e.g., PNG, PDF, SVG, PS;
- download the complete earthquake catalogue (locations, phase readings, MT solutions) in QuakeML format;
- download waveforms starting 2 minutes before origin time (i.e., time of earthquake) and ending 10 minutes after that time

Complex "friendly" user story. A strong earthquake hit southern Italy near the Vesuvius volcano. A geoscientist likes to get different datasets, display and compare them. She/he wants to select a subset of the data that shows some specific trend and perform analysis on that subset. Once the dataset is identified, she/he likes to use the results in another context and prepare figures for publication or web presentation.

There may be relationships between large earthquakes affecting the local stress conditions and magma chamber underneath the volcano. Changes in stress can trigger volcanic activity.

The researcher's goal is to investigate possible relationships between different data types, analyse their statistical significance. Verify or reject the initial hypothesis and possibly propose new suggestions or conclusions. In order to perform this kind of investigation, the researcher is called to deal with in-situ stress measurements (time series), water level in surrounding boreholes (time series), real time GPS/GNSS (time series), amount of CO_2 production in boreholes near volcano (time series). From the data analytics perspective, the researcher is interested in the following facilities:

- General: time-aligned plots of time series data, personal workspaces for managing datasets of interest, save /export images produced, compare images. All of this could be tailored for various end uses via template designs;
- Earthquake analysis: plot waveforms and check automatic phase onsets (process online; data download, catalogue record download), do corrections of phase onsets (plot waveforms), relocate earthquake (using different velocity models – 1D, 3D), magnitude estimate, MT inversion (Compare MT solution with historical MTs of EQs in that area), Analyse static stress transfer, i.e. see if additional stress in magma chamber is significant;
- Co-seismic analysis: show InSAR images (map), show static displacement from GNSS after the earthquake (map), slip inversion (CES, modelling).

As described in the above use cases, the discovery is envisaged as a two-step process, where the user searches generically for the desired datasets and then refines his / her search. Then – as described in the complex user story – the user performs actual processing on the data. Such processing should, under the systemic point of view, connect directly to dedicated "EPOS computational" resource, or reuse existing computational nodes. This underlines the centrality of a workflow engine that must be able to interoperate with other workflow engines in 3rd party computational resources.

Processing for discovery. A first type of processing is done by the system at discovery time. In order to manage all the information needed to satisfy user requests, all metadata describing the Thematic Core Services (TCS) Data, Datasets, Software and Services (DDSS) will be stored into the EPOS ICS, internal catalogue. Such a catalogue, based on the CERIF model, differs from most metadata standards used by various scientific communities. The metadata to be obtained from the EPOS TCS will be mapped to the EPOS ICS catalogue. The expectation is that the various TCS nodes will have APIs or other mechanisms to expose the metadata describing the available DDSS in a TCS specific metadata standard that contains the elements outlined in the EPOS baseline document. It also requires ICS APIs (wrappers) or other to map and store this in the ICS metadata catalogue, CERIF. These TCS APIs and the corresponding ICS APIs collectively form the *interoperability layer* in EPOS, which is the link between the TCSs and the ICS.

According to the above general directives, when the user performs a discovery by means of the GUI (Graphical User Interface), the system will query the metadata catalogue in order to obtain the desired data, and then provide it to the user. Referred to as "internal processing for discovery", such operations can have different levels of complexity, according to the complexity of the query.

Processing actual data: ICS-D and CES. Once the user has all references to the data s/he wants, s/he can then perform actual processing. In systemic terms this entails complex actions triggered by the workflow engine. EPOS has no intention of building a supercomputer. Indeed, EPSO is willing to exploit existing resources, both in terms of "general processing facilities" (e.g. HPC centres, cloud) or "community specific computational facilities". In the first case, interoperability with the ICS can be obtained by assuming that appropriate computing interfaces are implemented on the computational facilities (e.g. OCCI). Computational facilities are seen, from EPOS ICS point of view, as distributed computational facilities, and hence referred to as ICS-D (ICS-Distributed). In the latter case, interoperability can be more complex to handle because of the community specific systems, established over many years together with long established scientific practices. In this case, it is expected to interoperate (i) via appropriate interfaces (aforementioned OCCI), (ii) by using interoperable workflow engine systems, and (iii) by the construction of a Computational Earth Science layer that encompass the several disciplines of the Solid Earth Science domain. The Computational Earth Science layer is still under construction. However, leveraging on previous efforts by the communities, first examples can be provided, as in the case of seismology.

Processing: the case of seismology. In seismology, 'data processing' in a general sense has historically always been undertaken at the researcher / group / institute level. Whoever had a need for specific processing also had to find her/his own resources for it (or enter into bilateral agreements with specific IT resource providers). Only over the last couple of years, the potential for 'shared' community resources for processing is being actively explored. The three examples of seismology where, currently, the provisioning of computational resources through the EPOS RI is being developed are:

- ***Seismological waveform storage and dissemination***: Seismic waveforms are densely and regularly sampled time-series of ground motion values at specific (often permanent) measurement sites, initially recorded as continuous streams. With a long history of open data sharing, these waveform streams are usually collected by the same institution that is also responsible for the measurement station, but then almost immediately made openly available. In Europe, a community data archive has been established at ORFEUS[35] since the mid-1980s. Responding to the increasing amount of available data, increase in data retrieval requests, and increase in computational capacity at national institutions, the European Integrated Data Archive EIDA[36] has been set up over the past years, replacing the previously single-sited archive with a federated approach. In EIDA, a limited number of competent (national) data centres are recognized as 'primary nodes' with the responsibility of holding and providing data from a subset of contributing networks and stations. Routing and mediating services are in place that allow transparent access to all data from any EIDA data portal. ORFEUS/EIDA is the core waveform service infrastructure for EPOS-Seismology. To further enhance the waveform services, and to address some specific challenges, a pilot project is currently under way in the framework of EUDAT2020 to utilize specific EUDAT services in the EIDA (and thus EPOS) context, among them: (i) Ensure data preservation and improve access services via B2SAFE; (ii) Improve discovery and support attribution via B2FIND; (iii) Enable access control and accounting via B2ACCESS; (iv) Enable massive data processing via B2STAGE. Various challenges on the

[35] ORFEUS Observatories & Research Facilities for European Seismology www.orfeus-eu.org

[36] ORFEUS European Integrated Data Archive (EIDA) www.orfeus-eu.org/eida

way have been identified, e.g. how to handle PID assignment to streams, segments, or collections, or how to ensure history / provenance tracking of dynamic data. From a data user point of view, a 'community processing' challenge that is currently not addressed is the ***on-the-fly computation of specific (user provided) metrics (on the data) that then would be used as selection filters***. This would require a direct connection of the data storage system with appropriate computational resources that allow a relatively free definition of a metric calculation algorithm to be applied to the data.

- ***Complex waveform modelling:*** The increased availability of complex 2D- and 3D- models of physical parameters (wave speed, etc.) in the Earth, together with advances in modelling codes and computational resources, today allow a relatively detailed synthetic calculation of full seismic waveforms (over a limited frequency range). However, these calculations remain computationally demanding, usually requiring HPC resources, and are embedded in rather complex workflows. Recently, the VERCE project[37] succeeded in building a first community platform that is now operating in pilot mode[38]. This waveform modelling platform is integrated as service in EPOS-Seismology. It provides not only an interface (workspace) for forward calculation of synthetic waveforms (using dispel4py [Filgueira et al. 2015] as a workflow engine), but also the capacity to calculate misfits between synthetic and actual waveforms, that are accessed from global data repositories [Atkinson et al. 2015]. Access to the HPC resources for the computationally intensive synthetic calculation is currently secured via bilateral agreements between the VERCE operator and specific HPC providers. The misfit analysis toolset is more data intensive and can utilize grid and cloud resources (e.g. EGI). Within EPOS, these computational resources shall in future be organized as ICS-D. Also in this context, identifier management and provenance handling, metadata interoperability, and data (and model) staging are challenges that need to be addressed in an overall interoperable and sustainable framework, where shared / integrated community resources will play a key role.
- ***Seismic Hazard Computation:*** Probabilistic Seismic Hazard Assessment (PSHA) concerns the calculation of expected ground shaking levels per region and time interval. It requires the combination of diverse input and background information (e.g., seismicity rates, geological information, geodetic strain, fault zones, attenuation models), capturing uncertainties by following a logic-tree approach and utilizing Monte-Carlo simulations at various steps along the way. Overall, this is a computationally intensive (both in terms of I/O and processing power) task, in particular for any realistic target (country level and larger). Today, PSHA are usually executed at national level by a responsible / designated institution, or by regional or global initiatives like SHARE/EFEHR[39] or GEM[40]. A service within EPOS-Seismology, namely EFEHR, provides an IT platform that hosts the current European seismic hazard model developed in the SHARE project[41], and allows user access to the model and the underlying data as well as the configuration files for the hazard computation by using the GEM-developed OpenQuake package[42]. Today, no community facilities for data processing by users are provided, even though these were envisioned from early on. While users may often not want to execute full-scale PSHA calculations, a processing platform that would allow selective modification of input parameters, logic-

[37] VERCE: Virtual Earthquake and seismology Research Community e-science environment in Europe www.verce.eu

[38] VERCE Science Gateway https://portal.verce.eu

[39] EFEHR The European Facilities for Earthquake Hazard & Risk www.efehr.org

[40] GEM Global Earthquake Model www.globalquakemodel.org

[41] SHARE Seismic Hazard Harmonization in Europe www.share-eu.org

[42] OpenQuake https://www.globalquakemodel.org/openquake

tree setups, or specific algorithms, on existing hazard models would constitute a significant scientific benefit.

Overall, an integrated approach to data processing is now beginning to emerge in seismology, which should in future allow the rather obvious interconnection of various processing tools and elements such as those described above. The coordinated IT system developments within EPOS (ICS-C and ICS-D), and the development of interoperable interfaces that provide uniform workflow, metadata, and provenance management as elements of the Computational Earth Science layer are key for a sustainable data processing environment for EPOS-Seismology.

2.3.5 Data Processing in ICOS

By M. Hellström, A. Vermeulen and H. Lankreijer on behalf of the ICOS Community

ICOS (Integrated Carbon Observation System[43]) is a pan-European research infrastructure for observing and understanding the greenhouse gas (GHG) balance of Europe and its adjacent regions. The major task of ICOS is to collect, process, curate and make available high-quality GHG and environmental data to a wide range of end users.

ICOS handles both observational data (collected by its own networks of measurement stations) and data associated with atmospheric and ecosystem modelling (mainly performed by external research groups). The requirements differ somewhat: the observations are mainly sensor time series data in tabular form, which have to be calibrated, quality controlled, and gap filled. The evaluation of greenhouse gas and energy fluxes involves quite complex calculations. Modelling, on the other hand, involves both, preparation of input data (combination of many different data sources), various calculation steps, and processing of the output data. Model runs typically require access to high-throughput and high-performance computing facilities.[44]

ICOS strives to optimize both our own internal data processing as well as the use of our data products as input to various types of research by end users [Hellström et al. 2016]. Figure 8 outlines the data flow, from observation station networks, via Thematic Centres and the Carbon Portal, to end user communities.

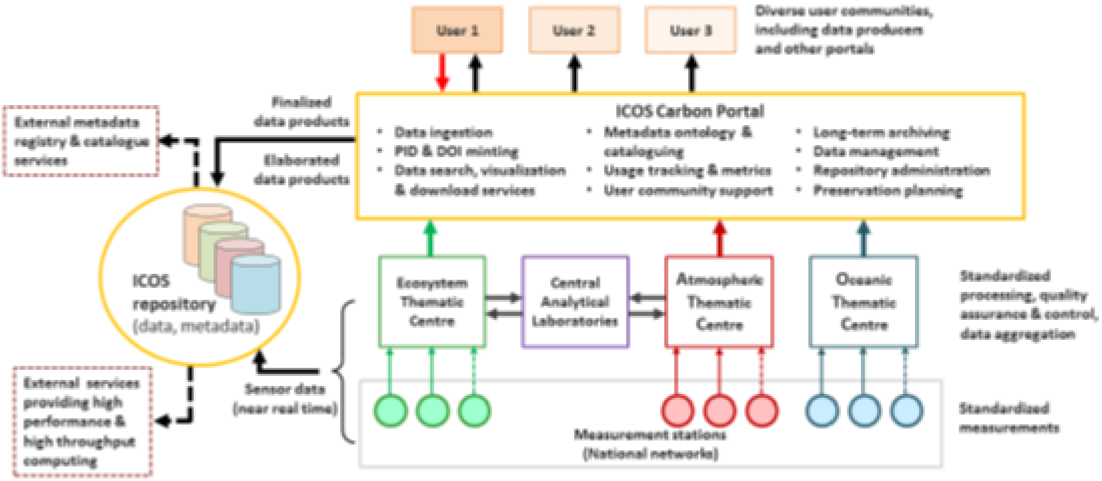

User 1
User 2
User 3
Diverse user communities, including data producers and other portals
ICOS Carbon Portal
Data ingestion
PID & DOI minting
Data search, visualization & download services
Metadata ontology & cataloguing
Usage tracking & metrics
User community support
Long-term archiving
Data management
Repository administration
Preservation planning
Finalized data products
Elaborated data products
External metadata registry & catalogue services
ICOS repository (data, metadata)
Ecosystem Thematic Centre
Central Analytical Laboratories
Atmospheric Thematic Centre
Oceanic Thematic Centre
Standardized processing, quality assurance & control, data aggregation
Sensor data (near real time)
External services providing high performance & high throughput computing
Measurement stations (National networks)
Standardized measurements

FIGURE 8. THE DATA FLOW IN ICOS

ICOS data products include [ICOS 2014, ICOS 2016]:

[43] ICOS Integrated Carbon Observation System https://icos-ri.eu/

[44] Processing in ICOS Wiki article https://wiki.envri.eu/display/EC/Processing+in+ICOS

- Level 0: data in physical units either directly provided by sensors or converted from engineer units (e.g. mV, mA) to physical units. They may have been filtered by a quality check (e.g. thresholds).
- Level 1: automatically quality checked and calibrated data, which are available for internal ICOS use (for station PI's quality checks, and input to further processing at Thematic Centres) as well as near-real-time (NRT) data products for external consumers (global monitoring and forecasting services).
- Level 2: the finalized quality checked and aggregated ICOS RI data products, produced by the Thematic Centres.
- Level 3: data products, requiring substantial contributions of ICOS observational data, that are produced by (external) researchers.

ICOS ingests data products into a storage repository as soon as they are available. Most data are also replicated at either the Thematic Centres or the Carbon Portal, allowing for fast retrieval in connection with ICOS-internal processing. Data can also be staged from the repository to computational centres external to ICOS, for processing by both internal and external users.

All ICOS data objects and their associated (basic) metadata are catalogued, allowing efficient search, discovery and access via the ICOS Carbon Portal. Here, human users are provided access to a number of visualisation tools, including comparative time series graphs and animations of tempo-spatial flux and concentration maps. The search & retrieval system is designed for both human and machine interaction, and thus supports calls from automated workflow engines.

As a special service to the atmospheric modelling community, ICOS annually prepares datasets formatted according to the ObsPack [Masarie et al. 2014] standard, containing greenhouse gas concentration data from ICOS stations. There are plans to extend this data preparation service to also cover greenhouse gas and energy flux data for the ecosystem and Earth system modelling communities.

To facilitate user interaction with, and analysis of, ICOS data products, the Carbon Portal aims to provide ICOS data users with access to them by Jupyter Notebook technology (cf. Sec. 2.1.2). In these, users can interact with ICOS data products and perform (simpler) analyses, including extraction of time series and statistics over different temporal and spatial selections.

ICOS is also interested in providing certain on-demand computational services. Through the involvement as a core community in EUDAT2020, ICOS is currently involved in a pilot project investigating the possibility to seamlessly connect EUDAT long-term storage services (using B2STAGE to stage data from and to B2SAFE) with computational services offered by the EGI FedCloud.

A general characteristic of ICOS is diversity, not only among data products but also concerning data producers and data users. Thus, interoperability between different processing platforms and data formats is crucial. Also discovery and easy access – by both human and machine users – to data and related (descriptive) metadata is very important.

However, each "branch" of ICOS – Atmosphere, Ecosystem and Ocean – has developed its own domain-specific data processing chains, and as a consequence the procedures for data quality control and assurance are likewise different. This sets clear demands for comprehensive and detailed descriptions not only of the provenance and uncertainty of finalized data values, but also of the meaning of the used quality flagging systems. ICOS has chosen to set up a RDF-based ontology describing all aspects of ICOS data collection, processing and data contents. This "database of assertions" will be able to serve all ICOS data discovery services, including those directly supporting data processing.

The time between data collection and availability of Quality Assured/Quality Controlled data products at different levels of maturity can also differ greatly, depending on a number of constraints – including the accessibility of the instrumentation platform, achievable data transfer times, the complexity of calibration and sample analysis procedures, and the need to batch process data from e.g. an entire season. These factors greatly influence the data processing inside of ICOS. In general, Level 2 data sets can be expected to be available within a few months of data collection, although for some stations or geographical regions there may be longer delays. For the special case of NRT data, atmospheric concentrations are typically available to external users within 36 hours, while ecosystem exchange fluxes may be released within 1-2 weeks of collection. Ocean data take much longer to process, e.g. because of intermittent data transfer capabilities, or because some data require chemical analysis of water samples.

Finally, as also mentioned above, support for data processing may also include assisting end user communities by streamlining the transfer of ICOS data to the relevant computational platforms. The various ICOS end user communities have quite different needs and requirements. While some researchers are focussed on analysing well-defined, easily downloadable data (sub)sets for their scientific questions, others require access to large amounts of data which they process at large-scale computational facilities.

However, there is now a trend to turn to cloud computing also for relatively CPU- and storage-intensive calculations. As a first test of how ICOS can support cloud computing-based applications, we are now (through our engagement as a EUDAT2020 core community) looking into ways to enhance the interoperability of EUDAT storage services and EGI-based federated cloud computing resources. As a first test, ICOS is implementing calculations of so-called station footprints with the STILT modelling framework[45], and the subsequent visualization of the outcomes (as animations of map time series). While this study is still ongoing (as of December 2016), the results look very promising [Karstens et al. 2016].

2.3.6 Data Processing in IS-ENES

By S. Joussaume, F. Guglielmo (CNRS-IPSL, France) on behalf of the IS-ENES Community with contributions from S. Kindermann, F.Toussaint, T.Weigel (DKRZ, Germany), S.Denvil (CNRS-IPSL, France), Christian Pagé (CERFACS –IS-ENES partner-, France), Maarten Plieger, Wim Som de Cerff (KNMI)

Data processing in the IS-ENES RI is important at various stages in the data life cycle: (a) to pre-process data to conform to community interoperability standards; (b) at data centres near processing for generating derived (most often aggregated and size reduced) datasets; (c) for climate data evaluation and analysis both at data centres and also remotely, e.g. in the frame of the climate4impact.eu portal (C4I) of IS-ENES.

In the following we will concentrate on the (b) and (c) related processing aspects as (a) is very community specific and also has to be performed at the HPC centre producing climate model data.

Supporting "data processing" within IS-ENES. Currently, there is a strong need in climate research to replace the default "download and process at home" with climate data evaluation and analysis. Within IS-ENES as well as within the global Earth System Grid Federation (ESGF), integrated with the IS-ENES data federation, it was decided to expose data processing functionalities (e.g. at data centres) as web services based on the OGC WPS standard [OGC WPS].

[45] http://www.bgc-jena.mpg.de/bgc-systems/projects/stilt/pmwiki/pmwiki.php?n=Main.HomePage

Processing functionalities developed within IS-ENES as part of the climate4impact.eu portal[46] are exposed in the form of OGC WPS services. Functionalities currently supported include climate indices calculations, subsetting, regridding, and tailoring data to specific end-user needs. Basic processing functionalities to be exposed by the IS-ENES data centres (as part of ESGF) are currently being standardized in the ESGF compute working team. A framework to support the development and packaging of OGC WPS was developed by IS-ENES partners. Plans to use it (within IS-ENES) to provide processing functionalities for the COPERNICUS CDS (climate data store) are currently being discussed. This framework is an open source Python framework called Birdhouse[47,48]. First prototype WPS installations are accessible at IS-ENES partners (as well as at partners in the US). First interoperability experiments are on the way.

On the technical side, current developments aim at:

- Enabling scalable processing back ends (e.g. by integration with batch systems at partners to exploit their clusters and HPC resources);
- Enabling scalability and failover aspects (roll-out of WPS instances on demand based on docker and automatic switching to another WPS provider in case of system failure at a site);
- Providing a generic security solution for rights delegation (based on tokens and OAUTH);
- Providing a generic deployment and software dependency management solution supporting the integration of analysis code into WPS frameworks (conda packaging[49], docker).

On the non-technical side, ongoing discussions concern:

- The definition with research groups of a standardized workflow to provide and package code to be deployed as part of WPS instances;
- Authorization and accounting (as the demand for I/O intensive data near processing cannot be satisfied by the available resources and strong regulations are necessary).

Process chaining. In a first step, processing services are called separately and service chaining etc. is done by the end user (or by a portal). Later, better support for automatic service chaining and support for workflow orchestration are required. One specific type of service in these chains/workflows is provided by “conversion/adaptation” components, to transform input data of a specific format/content to a format/content a specific processing component expects. A major outcome is that reproducible derived data products for climate science can be generated. The initial main focus is on data products derived from climate model output data (also observational data is involved in the context of climate model data evaluation).

Overall, the generic model is planned to be organised as follows:

- Large IS-ENES data centres (e.g. DKRZ, IPSL, BADC) hosting large climate data collections (multiple PB) as well as other IS-ENES resource centres, e.g. those supporting the climate impact community, expose OGC WPS compliant APIs for service description, discovery and invocation of data near processing services;
- Available services are registered in a joint catalogue (e.g. OGC CSW);
- End users can interact with these WPS endpoints via command line as well as with the help of graphical interfaces;
- Portals integrate these WPS endpoints to provide higher level services to end users; these portals again can expose functionality in the form of OGC WPS services.

[46] Climate4impact project portal https://climate4impact.eu

[47] Birdhouse Python http://birdhouse.readthedocs.io

[48] Birdhous GitHub repository https://github.com/bird-house

[49] A package based on the Anaconda platform https://www.continuum.io/anaconda-overview

Data provenance and relation to RDA. Increasing volumes of climate data (especially supporting the next round of climate model inter-comparison experiments -CMIP6-) will integrate handle based PIDs. Thus, data provenance support based on PIDs is on the roadmap to be integrated as part of WPS deployments. In the context of the RDA, IS-ENES has provided the generic processing workflow as an input for the RDA data fabric IG. Members of IS-ENES are also leading and contributing to PID related working groups (data collections, PID typing). For the storage of provenance information for the results of processing tools, first prototypes exist to produce W3C PROV [Lebo et al. 2013] based provenance descriptions. Also, IS-ENES partners involved in the EUDAT project are currently evaluating the use of graph databases as a backend for provenance storage. IS-ENES partners also develop dedicated climate big data analysis backend supporting efficient data operations on high volume and high-dimensional climate data sets, see e.g. the Ophidia framework[50].

2.4 Data Processing in ENVRIplus: Findings

The heterogeneity characterising existing systems (cf. Sec. 2.1) and Research Infrastructures (cf. Sec. 2.3) makes it evident that when discussing data processing “solutions” there are different angles, perspectives and goals to be taken into account. When analysing technologies from the scientist-perspective, the following trends should be considered:

- Technology should be “ease of (re-)use”, i.e., it should not distract effort from the pure processing task. Scientists should be exposed to technologies that are flexible enough to enable them to quickly specify their processing algorithm/pipeline. It should not require them to invest effort in learning new programming languages or in deploying, configuring or running complex systems for their analytics tasks. Methods and algorithms are expected to be reused as much as possible, thus data processing should enable them to be “published” and shared. This might call for:
- “as-a-Service” rather than “do-it-yourself”, i.e., scientists should be provided with an easy to use working environment where they can simply inject and execute their processing pipelines without spending effort in operating the enabling technology. This makes it possible to rely on economies of scale and keep the costs low.
- Solutions should be “hybrid”, i.e., it is neither suitable nor possible to implement one single solution that can take care of any scientific data processing need. Certain tasks must be executed on specific infrastructures while other tasks are conceived to crunch data that cannot be moved on other machines from where they are stored.

These trends actually suggest that scientists are looking for “workbenches” / “virtual research environments” / “virtual laboratories” [Candela et al. 2013b] providing them with easy to use tools for accessing and combining datasets processing workflows that behind the scene / transparently exploit a wealth of resources residing on multiple infrastructures and data providers (according to their policies). Such environments should not be pre-cooked / rigid, rather they should be flexible to enable scientists to enact their specific workflows. They should provide their users with appropriate and detailed information enacting to monitor the execution of such a workflow and be informed of any detail occurring during the execution. Finally, they should promote “open science” practices, e.g., they should record the entire execution chain leading to a given result, they should enact others to repeat/repurpose an existing process.

When analysing the overall solution to be developed and operated by RIs, the following aspects (going beyond the technology) are worth being considered:

[50] Ophidia http://ophidia.cmcc.it/

- Provide support for research developers who produce and refine the code and workflows that underpin many established practices, scientific methods and services. Without their efforts in understanding issues, in explaining software behaviour, and improving quality, scientists would struggle to continue to handle existing methods and explore new opportunities. They need tools that inform them about the operational use of their products and technologies that protect their invested effort as platforms evolve. They are in danger of being undervalued, overwhelmed by complexity and the pace of change, and of being attracted to the "big data" industry.
- Provide support for operations teams who need to keep the complex systems within and between RIs running efficiently as platforms change and communities' expectations rise while funders become more miserly. The tools and support they need are similar to those discussed in the previous bullet. They are not the same as the e-Infrastructure providers, they deploy and organise above those resources, but depend on them.
- Provide support for scientific innovators. They need to play with ideas, work on samples in their own favourite R&D environment, and then test their ideas at moderate and growing scale. The provided facilities should allow them to move easily between developing ideas and proto-deployment, and eventually, when their ideas work out, to production deployment.
- The majority of researchers do not want to innovate, they just want to get on with their daily job. As much care as possible must be invested in protecting their working practices from change. However, if tools become available, e.g. driven from provenance data, which help their work by removing chores, such as naming, saving, moving and archiving data, without them feeling they have lost responsibility for quality, then they will join in, and that eventually leads to fewer errors and better curated data [Myers et al. 2015].
- There are some computational processes that require expert oversight while they are running, that can save substantial waste or steer to better results.

All in all, data processing is strongly characterised by the "one size does not fit all" philosophy. There is no, and there arguably never will be, a single solution that is powerful and flexible enough to satisfy the needs arising in diverse contexts and scenarios.

The tremendous velocity characterising technology evolution calls for implementing sustainable data processing solutions that are not going to require radical revision by specialists whenever the supporting technologies evolve. Whenever a new platform capable of achieving better performance compared to existing ones becomes available, users are enticed to move to the new platform. However, such a move does not come without pain and costs.

Data analytics tasks tend to be complex pipelines that can require combining multiple processing platforms and solutions. Exposing users to the interoperability challenges resulting from the need to integrate and combine such heterogeneous systems strongly reduces their productivity.

There is a need to develop data processing technologies that address the problem by abstracting from (and virtualising) the platform(s) that take care of executing the processing pipeline. Such technologies should go in tandem with optimisation technologies [Martin et al. 2016] and should provide the data processing designer with fine-grained processing directives and facilitate detailed specification of processing algorithms.

3 ENVRIplus Data Processing Platform

According to the discussion and findings emerging from Section 2, it is evident how the large variety of needs and expectations arising in ENVRI environmental Research Infrastructures as well as the plethora of existing solutions makes the goal of T7.1 quite challenging. In fact, T7.1 is called to design and develop a solution for data processing aiming at making it significantly easier for scientists to conduct a range of experiments and analyses upon a great variety of data. In order to maximise the possible exploitation scenarios and promote its uptake by ENVRIplus environmental Research Infrastructure, the ENVRIplus data processing platform was designed to be driven by the following key principles:

- ***Extensibility***: the platform should be “open” with respect to (i) the analytics techniques it offers and support and (ii) the computing infrastructures and solutions it relies on to enact the processing tasks. It should be based on a ***plug-in architecture*** to support adding new algorithms / methods, new computing platforms;
- ***Distributed processing***: the platform should be conceived to execute processing tasks by relying on “local engines” / “workers” that can be deployed in multiple instances and execute tasks in parallel and seamlessly. The platform should be able to rely on computing resources offered by both well-known e-Infrastructures (e.g. EGI) as well as resources made available by the Research Infrastructure to deploy instances of the “local engines” / “workers”. This is key to make it possible to “move” the computation close to the data;
- ***Multiple interfaces***: the platform should offer its services via both a (web-based) graphical user interface and a (web-based) programmatic interface (aka API) thus to enlarge the possible application contexts. For instance, having a proper API facilitates the development of components capable to execute processing tasks from well-known applications (e.g. R, KNIME);
- ***Cater for scientific workflows***: the platform should be both exploitable by existing WFMS (e.g. a node of a workflow can be the execution of a task / method offered by the platform) and support the execution of a workflow specification (e.g. by relying on one or more instances of WFMSs);
- ***Easy to use***: the platform should be easy to use for both (a) algorithms / methods providers, i.e., scientists and practitioners called to realise processing methods of interest for the specific community, and (b) algorithms / methods users, i.e., scientists and practitioners called to exploit existing methods to analyse certain datasets;
- ***Open science friendly***: the platform should transparently inject open science practices in the processing tasks executed through it. This includes mechanisms for capturing and producing “provenance records” out of any computing task, mechanisms aiming at producing “research objects” so as to make it possible for others to repeat the task and reproduce the experiment (yet guaranteeing any policy regulating “access” and exploitation of the experiment and its results, i.e. the “owner” associates a well-defined policy and licence governing future exploitation and access).

The ENVRIplus data processing platform was not conceived to be developed from scratch. The rationale behind this is manifold including (i) there are so many solutions for data processing that developing a completely new one is neither feasible (because of the resources and time) nor reasonable (existing solutions are mainly conceived to be open and extensible), (ii) there will never be a single solution suitable for any application context or community, (iii) there are a plethora of solutions and e-Infrastructures to leverage on.

The ENVRIplus data processing platform stemmed from the data processing solution developed during the ENVRI project. The new data processing platform, named DataMiner, is an ***open-source computational system*** part of the ***gCube system*** [Assante et al. 2016].

From the end user perspective, it offers a **collaborative-oriented working environment** where users:

- can easily **execute and monitor data analytics tasks** by relying on a rich and open set of available methods either by using a dynamically generated **web-based user-friendly GUI** or by using a RESTful protocol based on the **OGC WPS Standard**;
- can easily **share & publish their analytics methods** (e.g. implemented in R, Java, Python, etc.) to the workbench and make them exploitable by an automatically generated web-based GUI and by the OGC WPS protocol;
- are provided with a **"research object"**[51] **describing every analytics task** executed by the workbench enabling for **repeatability, computational reproducibility, reuse, citation** and **provenance**. These research objects are a set of files organized in folders and containing every input & output, an executable reference to the method as well as rich metadata including a PROV-O provenance record;

The data analytics framework is integrated with a **shared workspace** where the research objects resulting from the analytics tasks are automatically stored together with rich metadata. Objects in the workspace can be shared with coworkers as well as published by a catalogue with a license governing their uses. Moreover, the framework is conceived to operate in the context of one or more **Virtual Research Environments**, i.e. it is actually made available by a dedicated working environment offering (besides the framework and the workspace) additional services including those for managing users, creating communities, and supporting communication and collaboration among VRE members.

The data analytics framework is conceived to give access to two typologies of resource:

- a **distributed, open & heterogeneous computing infrastructure** for the real execution of the analytics tasks. This distributed computing infrastructure is capable to exploit resources from the EGI infrastructure.
- the pool of **methods** integrated in the platform, i.e. each method integrated in the framework is made available as-a-Service to other users according to the specific policy.

3.1 ENVRIplus Data Processing Architecture

The high-level architecture of the Data Processing is Platform is depicted in Figure 9 and Figure 10.

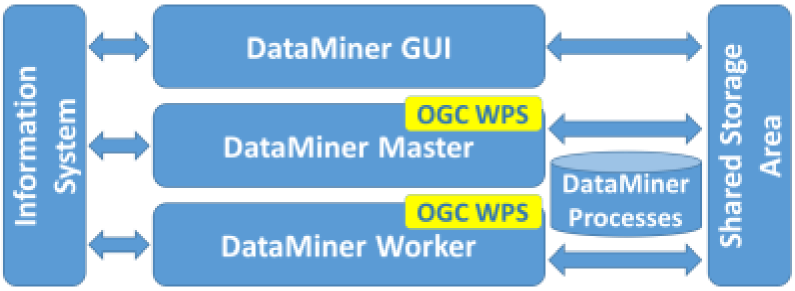


FIGURE 9. DATAMINER OVERALL ARCHITECTURE

The system consists of the following components:

- The ***DataMiner GUI***: a web-based user interface enacting users to select an existing process, execute it, monitor the execution and access to the results (cf. Sec. 3.2);

[51] These are packages of files worth being considered as a unit from the research activity perspective, e.g. they are expected to contain the entire set of data and metadata needed to capture a research activity and its results. From a conceptual point of view this is equal to the objects characterizing the http://www.researchobject.org/ initiative.

- The ***DataMiner Master***: this web service is in charge to accept requests for executing processes and executing them, either locally or by relying on the DataMiner Worker(s) depending from the specific process. The service is conceived to work in a cluster of replica services and is offered by a standard web-based protocol, i.e. OGC WPS;
- The ***DataMiner Worker***: this web service is in charge to execute the processes it is assigned to. The service is conceived to work in a cluster of replica services and is offered by a standard web-based protocol, i.e. OGC WPS;
- The ***DataMiner Processes***: this a repository of processes the platform is capable to execute. This repository if equipped with a set of off-the-shelf processes and it can be further populated with new processes either (a) developed from scratch in compliance with a specific API or (b) resulting from annotating existing processes (cf. Sec. 3.3).

These components are glued together thanks to (a) an information system that enable each service instance to dynamically discover other existing service instances and be informed on their capabilities and (b) a shared storage area the service instances use to exchange the data they are operating on, e.g. the datasets to be processed or the datasets resulting from a processing activity.

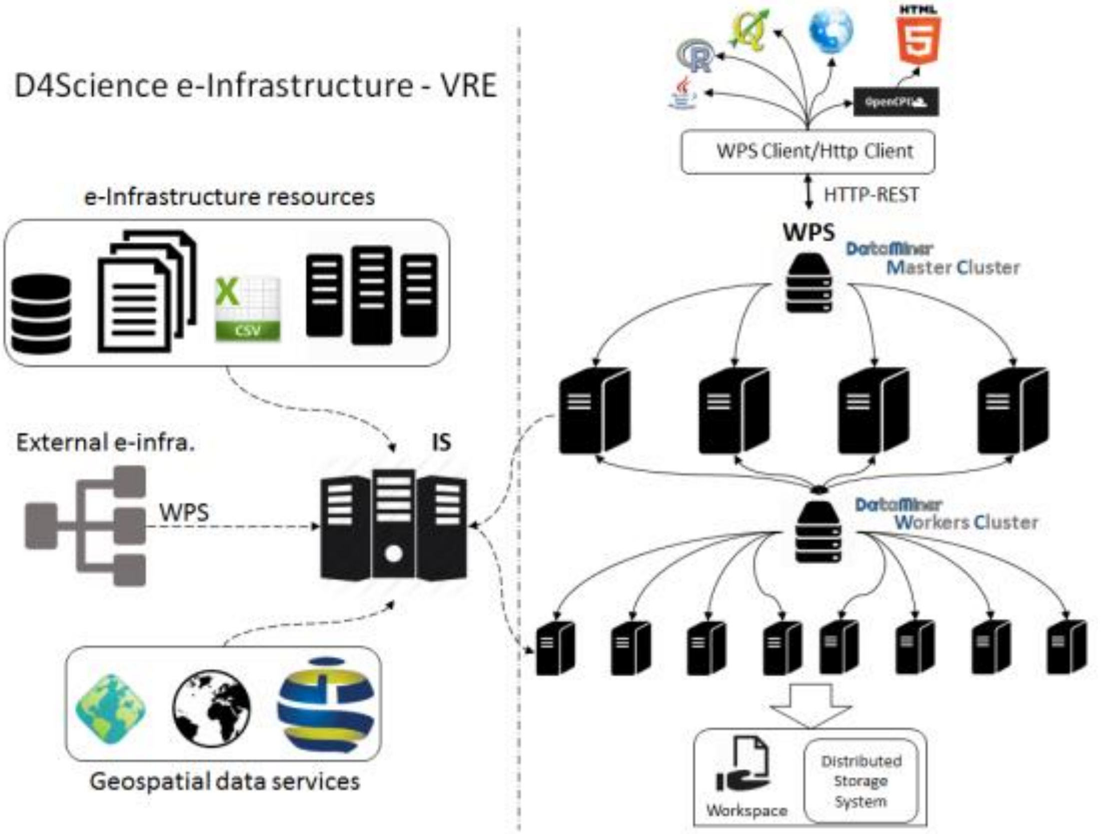


FIGURE 10. DATAMINER DATA PROCESSING SYSTEM

Figure 10 is oriented to describe how the system works by highlighting how the various instances interact. In a typical deployment scenario, the Master cluster is made up of a number of machines managed by a load balancer that distributes the requests uniformly to the machines hosting DataMiner Master instances. Each machine is endowed with a DM service that communicates with the Information System. The balancer is indexed on the IS and is the main access point to interact with the DMs. The machines of the Worker cluster have less local computational power and serve distributed computations. DataMiner is based on the 52North WPS implementation[52]. It is developed in Java and the Web service runs on an Apache Tomcat instance endowed with gCube system libraries.

When a WPS request comes to the Master cluster balancer, it is distributed to one of the DataMiner Master instances forming the cluster. Each DataMiner instance host the processes the service is capable to execute, these processes are likely to be provided by several developers and providers. In particular, two kinds of algorithms are hosted: “local” and “cloud” algorithms. Local algorithms are directly executed on the DataMiner Master instances and possibly use parallel processing on several cores and a large amount of memory. In contrast, cloud algorithms use distributed computing with a Map-Reduce approach and rely on the DataMiner Worker instances in the Worker cluster.

All the DataMiner instances (be them Masters or Workers) are exposed by the OGC WPS protocol. However, DataMiner WPS implementation adds a number of features to the standard 52North implementation it relies on. First, it returns a different list of processes according to the “application context” in which the service is invoked. In fact, the overall DataMiner framework is conceived to operate in multi-tenancy settings where the same instance can serve many “application contexts” (cf. Sec. 3.5).

Whenever an algorithm is added to the DataMiner Processes repository, it is conceptually added to all the instances of the DataMiner relying on the specific repository. However, a specific DataMiner instance will expose and make executable only the processes it is configured to make available in a specific “application context”. This approach is flexible and well suited with the Virtual Research Environment development model envisaged by gCube [Assante et al. 2016].

The exposition of DataMiner processes by the WPS standard allows a number of thin clients to use the processes and to transparently rely on a distributed computing infrastructure to execute them. Third party software (e.g. the well-known QGIS[53] and ArcMap[54] for geospatial data manipulation) can be able to retrieve the capabilities of a WPS service and run remote processes. Further, clients for R and Java have been developed[55] as well as the WPS service can manage HTTP-GET requests (thus, a process can also be invoked by using a common Web browser). Finally, by relying on OpenCPU[56] it is possible to transform WPS objects into Javascript objects and allows for fast building of HTML applications.

For authentication and authorization, DataMiner relies on a token-based mechanism[57] [Assante et al. 2016], i.e. the user is requested to acquire a “valid token” before being authorized to execute processes. This token is passed via basic HTTPS-access authentication, which is supported by most WPS and HTTP(S) clients. The token identifies both a user and an “application context” and this information is used by DM to query the IS about the capabilities to be offered in that context, i.e. the processes the user will be able to invoke with that authorization.

The DataMiner is conceived to rely on a shared workspace built on top of the storage area. In particular, it interfaces with the gCube-based workspace [Assante et al. 2016] for accessing input data. Inputs can also come from external repositories, because a file can be provided either as an HTTP link or embedded in a WPS execution request. The outputs of the

[52] 52North. The 52north wps service 2016. http://52north.org/communities/geoprocessing/wps/

[53] QGIS. A free and open source geographic information system 2016. http://qgis.org/en/site/

[54] ArcMap. Arcgis for desktop 2016. http://desktop.arcgis.com/en/arcmap/

[55] National Research Council of Italy. gCube wps thin clients 2016. https://wiki.gcube-system.org/gcube/How_to_Interact_with_the_DataMiner_by_client

[56] OpenCPU. Producing and reproducing results 2016. https://www.opencpu.org

[57] National Research Council of Italy. gCube token-based authorization system 2016. https://wiki.gcube-system.org/gcube/Authorization_Client_Library

computations are written onto the Distributed Storage System and are immediately returned to a client at the end of the computation. Afterwards, an independent thread also writes this information on the Workspace. Indeed, after a completed computation, a Workspace folder is created which contains the input, the output, the parameters of the computation, and a provenance document summarizing this information. This folder can be shared with other people and used to execute the process again. Thus, the complete information about the execution can be shared and reused. This is the main way by which DataMiner fosters collaborative experimentation. The DataMiner processes can access the resources available in an "application context" by querying the IS. For example, it is possible to discover geospatial services, maps, databases, and files. The DataMiner Java development framework simplifies the interaction with the IS. Since the IS interface is HTTP REST too, it could be managed by the processes directly. Further, the DataMiner development framework provides methods to transform heterogeneous GIS formats into a numeric matrix and thus simplifies the effort to process geospatial data.

DataMiner can also import processes from other WPS services. If a WPS service is indexed on the IS for a certain "application context", its processes descriptions are automatically harvested, imported, and published among the DataMiner capabilities for that "application context". During a computation, DataMiner acts as a bridge towards the external WPS systems. Nevertheless, DataMiner adds provenance management, authorization, and collaborative experimentation to the remote services, that being standard WPS do not support any of them.

3.2 The DataMiner web-based GUI

DataMiner offers a web-based GUI to its users (Figure 11).

On the left panel (Figure 11 a), the GUI presents the list of capabilities available in the specific "application context", which are semantically categorised (the category is indicated by the process provider). For each capability, the interface calls the WPS *DescribeProcess* operation to get the descriptions of the inputs and outputs. When a user selects a process, in the right panel the GUI on-the-fly generates a form with different fields corresponding to the inputs. Input data can be selected from the Workspace (the button associated to the input opens the Workspace selection interface). The "Start Computation" button sends the request to the DM Master cluster, which is managed as explained in the previous section. The usage and the complexity of the Cloud computations are completely hidden from the user, but the type of the computation is reported as a metadata in the provenance file.

A view of the results produced by the computations is given in the "Check the Computations" area (Figure 11 b), where a summary sheet of the provenance of the experiment can be obtained ("Show" button, Figure 11 c). From the same panel, the computation can be also re-submitted. In this case, the Web interface reads the XML file containing the PROV-O information associated to a computation and rebuilds a computation request with the same parameters. The computation folders may also include computations executed and shared by other users.

Finally, the "Access to the Data Space" button allows obtaining a list of the overall input and output datasets involved in the executed computations (Figure 11 d), with provenance information attached that refers to the computation.

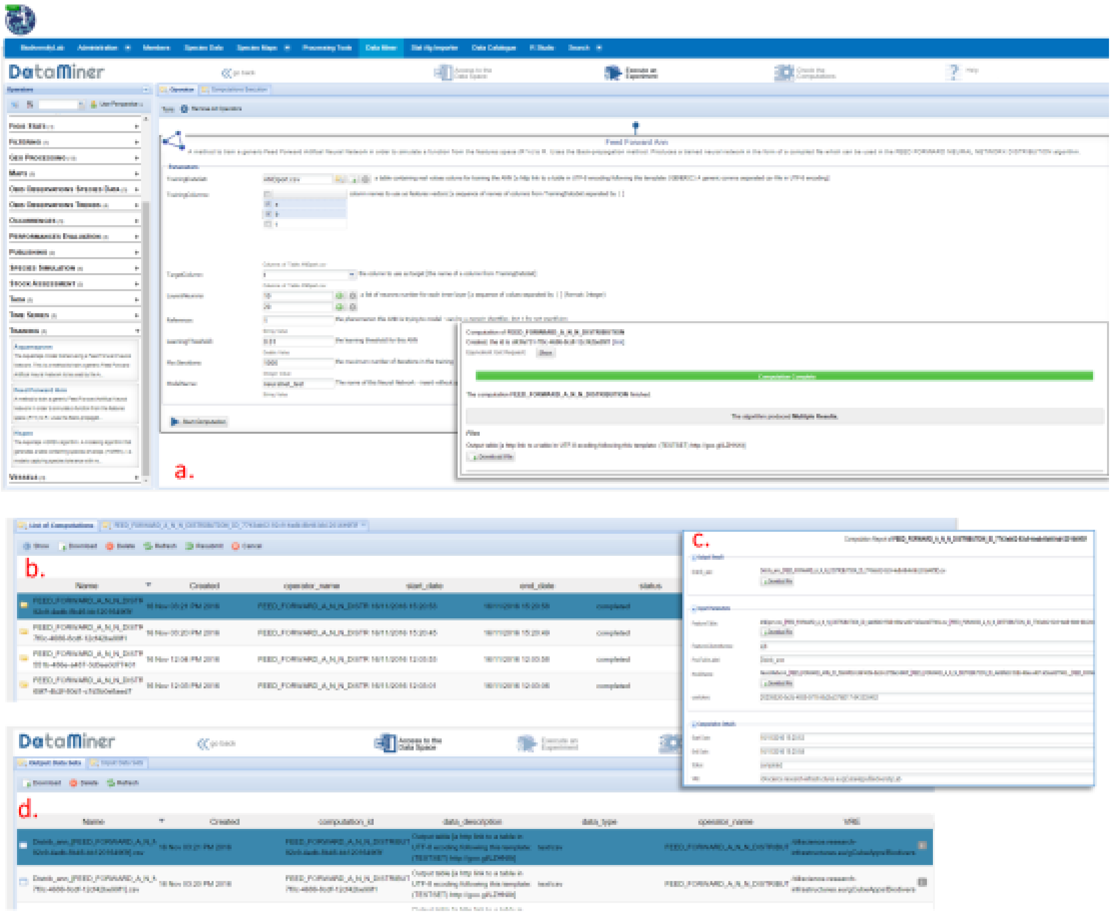
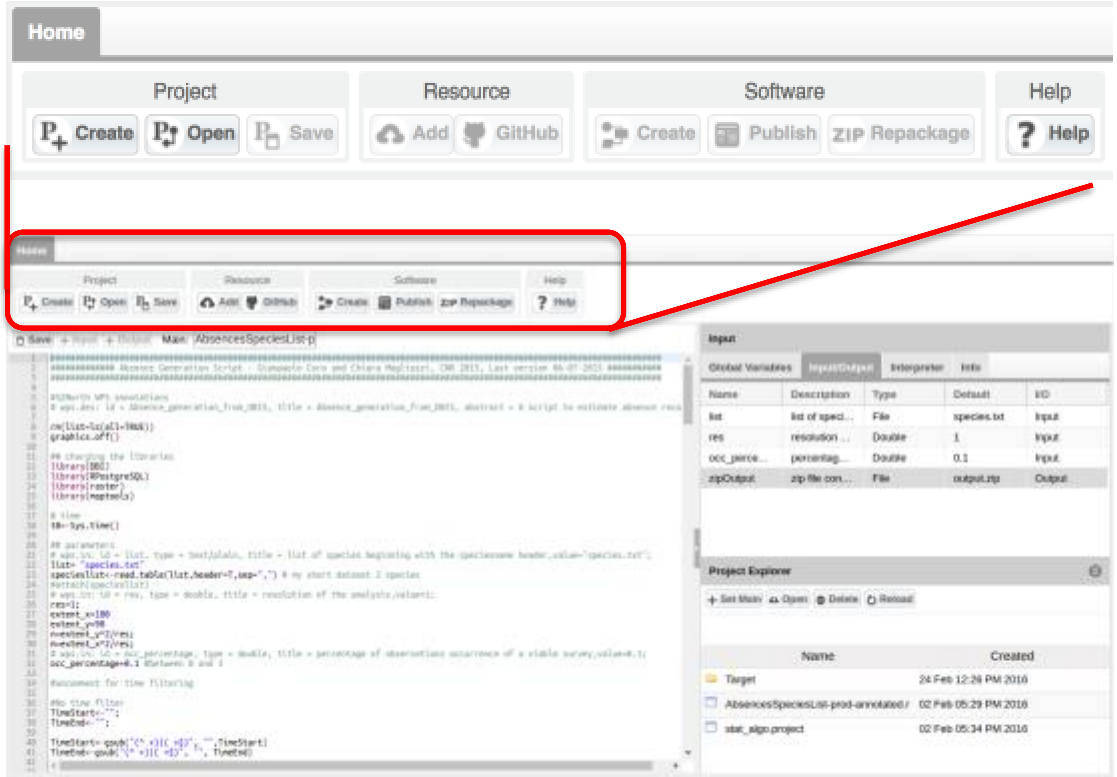


FIGURE 11. INTERFACE OF THE GCUBE DATAMINER SYSTEM

3.3 Importing new processes

Prototype scripting is the base of most models in environmental sciences. Scientists making prototype scripts (e.g. using R and Matlab) often need to share results and provide their models and methods for use by other scientists. They will encounter new data and may run in different contexts, which may require careful engineering to accommodate the wider scope. To help meet this aim, DataMiner lets them publish scripts as-a-Service, possibly under a recognized standard (e.g. WPS). The Statistical Algorithms Importer (SAI) is an interface that allows scientists to easily and quickly import R scripts onto DataMiner. DataMiner in turn publishes these scripts as-a-Service and manages multi-tenancy and concurrency. Additionally, it allows scientists to update their scripts without following long software re-deploying procedures each time. In summary, SAI produces processes that run on the DataMiner Cloud computing platform and are accessible via the WPS standard.

The SAI interface (Figure 12) resembles the R Studio environment, a popular IDE for R scripts, in order to make it friendly to script providers.



```
########################################################################
############ Absence Generation Script - Gianpaolo Coro and Chiara Magliozzi, CNR 2015, Last version 06-07-2015 ##########
########################################################################

#52North WPS annotations
# wps.des: id = Absence_generation_from_OBIS, title = Absence_generation_from_OBIS, abstract = A script to estimate absence reco

rm(list=ls(all=TRUE))
graphics.off()

## charging the libraries
library(DBI)
library(RPostgreSQL)
library(raster)
library(maptools)

# time
t0=Sys.time()

## parameters
# wps.in: id = list, type = text/plain, title = list of species beginning with the speciesname header,value="species.txt";
list= "species.txt"
specieslist<-read.table(list,header=T,sep=",") # my start dataset 3 species
#attach(specieslist)
# wps.in: id = res, type = double, title = resolution of the analysis,value=1;
res=1;
extent_x=180
extent_y=90
n=extent_y*2/res;
m=extent_x*2/res;
# wps.in: id = occ_percentage, type = double, title = percentage of observations occurrence of a viable survey,value=0.1;
occ_percentage=0.1 #between 0 and 1

#uncomment for time filtering

#No time filter
TimeStart<-"";
TimeEnd<-"";

TimeStart= gsub([illegible],TimeStart)
TimeEnd= gsub([illegible], TimeEnd)
```

Name	Description	Type	Default	I/O
list	list of speci...	File	species.txt	Input
res	resolution ...	Double	1	Input
occ_perce...	percentag...	Double	0.1	Input
zipOutput	zip file con...	File	output.zip	Output

Name	Created
Target	24 Feb 12:26 PM 2016
AbsencesSpeciesList-prod-annotated.r	02 Feb 05:29 PM 2016
stat_algo.project	02 Feb 05:34 PM 2016

FIGURE 12. INTERFACE TO IMPORT NEW METHODS IN DATAMINER

The *Project* button allows creating, opening and saving a working session. A user uploads a set of files and data on the workspace area (lower-right panel). Upload can be done by dragging and dropping local desktop files. As a next step, the user indicates the “main script”, i.e. the script that will be executed on DataMiner and that will use the other scripts and files. After selecting the main script, the left-side editor panel visualises it with R syntax highlighting and allows modifying it. Afterwards, the user indicates the input and output of the script by highlighting variable definitions in the script and pressing the *+Input* (or *+Output*) button: behind the scenes the application parses the script strings and guesses the name, description, default value and type of the variable. This information is visualised in the top-right side *Input/Output* panel, where the user can modify the guessed information. Alternatively, SAI can automatically compile the same information based on WPS4R[58] annotations in the script. Other tabs in this interface area allow setting global variables and adding metadata to the process. In particular, the *Interpreter* tab allows indicating the R interpreter version and the packages required by the script and the *Info* tab allows indicating the name of the algorithm and its description. In the *Info* tab, the user can also specify the VRE in which the algorithm should be available.

Once the metadata and the variables information have been compiled, the user can create a DataMiner as-a-Service version of the script by pressing the *Create* button in the Software panel. The term “software”, in this case indicates a Java program that implements an as-a-Service version of the user-provided scripts. The Java software contains instructions to automatically download the scripts and the other required resources on the server that will execute it, configure the environment, execute the main script and return the result to the user. The computations are orchestrated by the DataMiner computing platform that ensures the program has one instance for each request and user. The servers will manage concurrent

[58] 52North. WPS4R. https://wiki.52north.org/Geostatistics/WPS4R

requests from several users and execute code in a closed sandbox folder, to avoid damage caused by malicious code.

Based on the SAI Input/Output definitions written in the generated Java program, DataMiner automatically creates a Web GUI (cf. Section 3.2).

By pressing the *Publish* button, the application notifies DataMiner that a new process should be deployed. DataMiner will not own the source code, which is downloaded on-the-fly by the computing machines and deleted after the execution. This approach meets the policy requirements of those users who do not want to share their code.

The *Repackage* button re-creates the software so that the computational platform will be using the new version of the script. The repackaging function allows a user to modify the script and to immediately have the new code running on the computing system. This approach separates the script updating and deployment phases, making the script producer completely independent on e-Infrastructure deployment and maintenance issues. However, deployment is necessary again whenever Input/Output or algorithm's metadata are changed.

To summarise, the SAI Web application enables an R script with as-a-Service features. SAI reduces integration time with respect to direct Java code writing. Additionally, it adds (i) multi-tenancy and concurrent access, (ii) scope and access management through Virtual Research Environments, (iii) output storage on a distributed, high-availability file system, (iv) graphical user interface, (v) WPS interface, (vi) data sharing and publication of results, (vii) provenance management, and (viii) accounting facilities.

3.4 Supporting Workflow Management Systems

New Science paradigms require solving complex multi-disciplinary problems using heterogeneous expertise and large-scale computational systems. Many experiments are made up of a sequence of processes that are possibly hosted by different computing platforms. Conducting one experiment can be difficult due to the heterogeneity of computing platforms in terms of capabilities, input/output formats and data provisioning methods. Workflow Management Systems (WMSs) are able to combine processes provided by different services into workflows in a flexible way. WMSs can be valid solutions to manage the complex experiments fostered by modern scientific approaches. Indeed, WMSs allow users with basic programming experience to combine algorithms and perform complex analyses. The as-a-Service published processes are atomic steps of the workflows and can be executed by remote systems. Formal definitions of input and output types and of other metadata allow users to understand and reuse algorithms in other workflows.

WPS processes are often supported by these systems and are naturally suited to be used in workflows that combine several of them. Thus, complex processes can be built as workflows using WPS processes in cascade. For example, the generation of a species distribution map usually requires (i) a data preparation phase, (ii) a probability calculation phase, and (iii) a map generation process. In the first phase, tables are produced based of environmental information collected at locations where the species presence/absence assessment was performed. In the second phase, another model calculates a presence probability value for each location, according to its environmental characteristics. Finally, a map is produced by another process that transforms each assessed location into a geographic information system standard geometry representation. Each of the steps in this specific workflow is a process that can be hosted and parallelized by different computational systems.

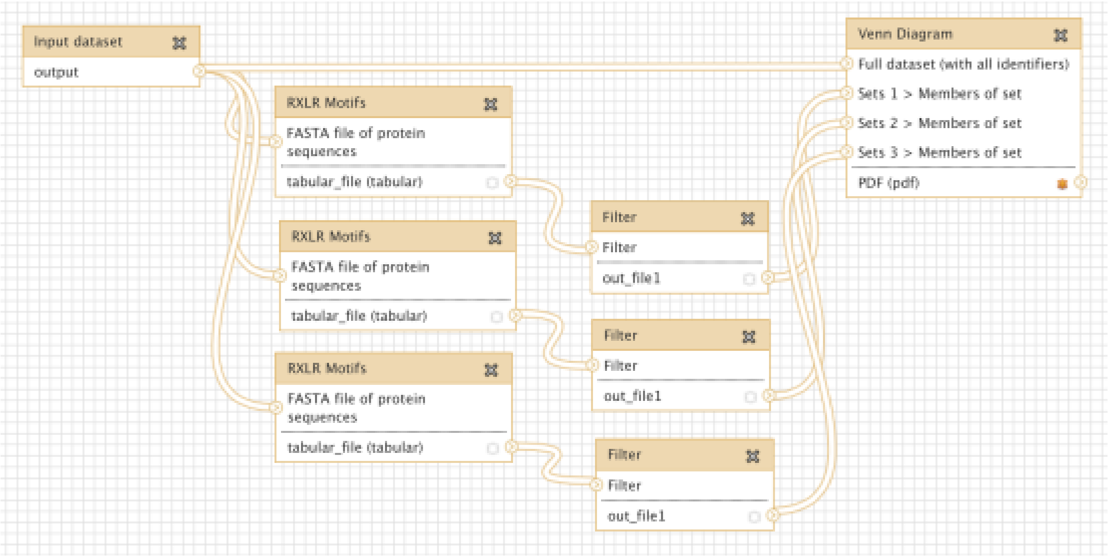


FIGURE 13. EXAMPLE OF WORKFLOW FROM THE GALAXY WORKFLOW MANAGEMENT SYSTEM

In the course of the ENVRIplus project, the DataMiner platform will be enhanced with the possibility to manage processes as Workflows. This will be managed in two ways: first, WMS supporting the WPS standard will be offered to the ENVRIplus community to design workflows based on the DataMiner-hosted processes. Should the WMSs not satisfy community needs, a new interface tailored for DataMiner will be developed. Second, the most common WMSs will be supported as far as possible. These may not be using the DataMiner processes but rather external services or even other WMS engines. For example, a KNIME Workflow will be processed by a proper KNIME engine, possibly invoking DataMiner processes and other external e-Infrastructure services.

Overall, in compliance with modern Science approaches, the evolution of WMSs will be closely followed in order to detect, as soon as possible, the birth of a standard representation of workflows, which is currently far to achieve [Liew et al. 2016].

3.5 Exploitation Scenarios

DataMiner is conceived to support the following scenarios / exploitation models:

- ***Full platform as-a-Service***: the entire DataMiner platform is operated by a service provider (e.g. D4Science) and the community / Research Infrastructure establishes a collaboration agreement to use it by well-defined service level agreements. In this scenario, the service provider can establish its own collaboration agreements with other research infrastructures to operate the service it is responsible for (e.g. this is the case of D4Science that has established a collaboration agreement with EGI to deploy some DataMiner instances on EGI sites);
- ***Full platform as-a-Software***: a community / Research Infrastructure can decide to exploit the DataMiner technology (open source) to set up its own instance of the technology. In this case, the community / Research Infrastructure faces hardware resource costs needed to operate the platform, as well as IT personnel costs needed to deploy and operate the technology. In this case, the community can set up an agreement with other infrastructures (e.g. EGI) thus to reduce the costs related with hardware resources. Still, the costs related to DataMiner technology deployment and operation remain;
- ***Platform as-a-Service with Community Contribution***: the DataMiner core components are operated by a service provider and the platform is complemented by some instances (namely, the workers) operated by the Research Infrastructure on its own resources. This

make it possible for RIs, to deploy these nodes close to where the data to be processed are actually stored.

3.6 Data Processing, Optimisation and ENVRIplus Data for Science Theme facilities

The DataMiner facility is contributing to form the ENVRIplus set of facilities falling under the Data for Science theme together with data identification and citation services [Hellström et al. 2017], performance optimization [Martin et al. 2017], data curation [Jeffery et al. 2017 b] and cataloguing [Loubrieu et al. 2017].

Data identification and citation services will cater for unique identifiers for the processes enacted by DataMiner that will be associated with any metadata related with the specific process and used for citation purposes.

Data curation and cataloguing will cater for a catalogue where the DataMiner processes are going to be described to make them discoverable and accessible. This activity is largely simplified by the offering of DataMiner processes by WPS. In fact, it is sufficient for catalogue providers to implement a lightweight harvester capable to rely on this standard to collect rich metadata on each available process.

Finally, performance optimisation is certainly the service that might have more impact on DataMiner. In fact, by relying on its facilities for infrastructure planning and provisioning it is possible to carefully identify resources suitable for hosting DataMiner components (mainly the Worker) thus to make the instance capacity compliant with the application requirements (including requirements arising in near-real time cases). The DataMiner feature of being able to transparently rely on components (both the Master and the Worker) that can be hosted and operated by diverse providers cater for transparent exploitation (from the end-user perspective) of resources offered by diverse infrastructures (e.g. DataMiner can be configured to have pieces hosted by EGI Fedcloud premises).

4 Concluding remarks

Data Processing is a wide concept embracing tasks ranging from (systematic) data collection, collation and validation to data analytics aiming at distilling and extracting new “knowledge” out of existing data by applying diverse methods and algorithms. When devising a solution suitable for ENVRIplus environmental Research Infrastructures it is immediate to realize that it is almost impossible to envisage a solution that is powerful and flexible enough to satisfy the needs arising in diverse contexts and scenarios.

This deliverable first discusses the settings characterising data processing in the context of ENVRIplus then propose a solution for data processing that focuses on specific needs and contexts. The larger part of this document is dedicated to provide the reader with a long and detailed discussion aiming at documenting what are (a) the existing technologies and solutions for data processing (including workflow management systems and data processing frameworks and platforms); (b) the envisaged data-processing-related patterns captured by the ENVRIplus reference model; and (c) the existing solutions seven environmental Research Infrastructures have currently in place for satisfying their data processing needs. This part concludes by reporting the following findings worth considering when devising data processing solutions for ENVRIplus: (i) technology should be “ease of (re-)use”, i.e., it should not distract effort from the pure processing task; (ii) the “as-a-Service” provision mode should be preferred to the “do-it-yourself”, i.e., scientists should be provided with an easy to use working environment where they can simply inject and execute their processing pipelines without spending effort in operating the enabling technology. This makes it possible to rely on economies of scale and keep the costs low; (iii) solutions should be “hybrid”, i.e., it is neither suitable nor possible to implement one single solution that can take care of any scientific data processing need; (iv) support for research developers who produce and refine the code and workflows that underpin many established practices, scientific methods and services should be provided; (v) support for operations teams who need to keep the complex systems within and between RIs running efficiently as platforms change and communities’ expectations rise while funders become more miserly should be provided; (vi) support for scientific innovators (playing with ideas, working on samples in their own favourite R&D environment, and then testing their ideas at moderate and growing scale) should be provided; (vii) as much care as possible must be invested in protecting researchers working practices by smoothly injecting novel approaches enacting them to perform their daily tasks.

The second part of the deliverable documents a solution for data processing and analytics. The deliverable captures the major challenges characterising the design of such a solution given the settings resulting from the specific application context (environmental Research Infrastructures) as well as from the great variety and heterogeneity characterising the “data processing domain. The resulting platform (DataMiner) is conceived to (i) *be extensible*, i.e., the platform is “open” with respect to the analytics techniques it offers / support and the computing infrastructures and solutions it relies on to enact the processing tasks; (ii) *promote distributed processing*, i.e. the platform executes processing tasks by relying on “local engines” / “workers” that can be deployed in multiple instances and execute tasks in parallel and seamlessly; (iii) *be offered by multiple interfaces*, i.e., the platform offers its facilities by both a (web-based) graphical user interface and a (web-based) programmatic interface (aka API) in OGC WPS; (iv) *cater for scientific workflows*, i.e., the platform is exploitable by existing WFMS as well as should support the execution of a processing task captured by a workflow specification; (v) *be easy to use*, i.e., the platform is easy to use for both algorithms / method providers and algorithms / method users; (vi) be *open science friendly, i.e.*, the platform transparently inject open science practices (provenance recording, repeatability) in the processing tasks executed through it.

References

[Ács et al. 2010] B. Ács, X. Llorà, L. Auvil, B. Capitanu, D. Tcheng, M. Haberman, L. Dong, T. Wentling, M. Welge (2010) A general approach to data-intensive computing using the Meandre component-based framework. In Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science (Wands '10). ACM, New York, NY, USA, Article 8, 12 pages. doi: 10.1145/1833398.1833406

[Amaral at al. 2014] R. Amaral; R. M. Badia; I. Blanquer; R. Braga-Neto; L. Candela; D. Castelli; C. Flann; R. De Giovanni; W.A. Gray; A. Jones; D. Lezzi; P. Pagano; V. Perez-Canhos; F. Quevedo; R. Rafanell; V. Rebello; M.S. Sousa-Baena; E. Torres (2014) Supporting biodiversity studies with the EUBrazilOpenBio Hybrid Data Infrastructure. Concurrency and Computation: Practice and Experience, Wiley, 2014, doi: 10.1002/cpe.3238

[Assante et al. 2016] M. Assante, L. Candela, D. Castelli, G. Coro, L. Lelii, P. Pagano (2016) Virtual research environments as-a-service by gCube. PeerJ Preprints 4:e2511v1 doi: 10.7287/peerj.preprints.2511v1

[Atkinson 2013] Atkinson, M. (2013) Data-Intensive Thinking with DISPEL. The DATA Bonanza: Improving Knowledge Discovery in Science, Engineering, and Business (eds M. Atkinson, R. Baxter, M. Galea, M. Parsons, P. Brezany, O. Corcho, J. van Hemert and D. Snelling), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9781118540343.ch4

[Atkinson et al. 2015] M. Atkinson, M. Carpené, E. Casarotti, S. Claus, R. Filgueira, A. Frank, M. Galea, T. Garth, A. Gemünd, H. Igel, I. Klampanos, A. Krause, L. Krischer, S. Hoon Leong, F. Magnoni, J. Matser, A. Michelini, A. Rietbrock, H. Schwichtenberg, A. Spinuso, J.-P. Vilotte (2015) VERCE Delivers a Productive E-science Environment for Seismology Research. 2015 IEEE 11th International Conference on e-Science, pp. 224-236 doi: 10.1109/eScience.2015.38

[Atkinson et al. 2016] M. Atkinson, A. Hardisty, R. Filgueira, C. Alexandru, A. Vermeulen, K. Jeffery, T. Loubrieu, L. Candela, B. Magagna, P. Martin, Y. Chen, M. Hellström (2016) A consistent characterisation of existing and planned RIs. ENVRIplus Project Deliverable D5.1

[Beisken et al 2013] S. Beisken, T. Meinl, B. Wiswedel, L. F. de Figueiredo, M. Berthold, C. Steinbeck (2013) KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinformatics, 14:257, 2013. doi: 10.1186/1471-2105-14-257

[Blankenberg et al. 2011] D. Blankenberg, G.V. Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, J. Taylor, J. (2001) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists, In Current Protocols in Molecular Biology, 2001. doi: 10.1002/0471142727.mb1910s89

[Bordawekar et al. 2014] R. Bordawekar, B. Blainey, C. Apte (2014) Analyzing analytics. *SIGMOD Rec.* 42, 4 (February 2014), 17-28. doi: 10.1145/2590989.2590993

[Bux and Leser 2013] M. Bux and U. Leser (2013) Parallelization in Scientific Workflow Management Systems, CoRR 2013

[Candela et al. 2013] L. Candela; D. Castelli; G. Coro; P. Pagano; F. Sinibaldi (2013) Species distribution modeling in the cloud. Concurrency and Computation: Practice and Experience, Wiley, pp. 289-301 doi: 10.1002/cpe.3030

[Candela et al 2013 b] L. Candela, D. Castelli, P. Pagano (2013) Virtual Research Environments: An Overview and a Research Agenda. Data Science Journal, Vol. 12, p. GRDI75-GRDI81 DOI: 10.2481/dsj.GRDI-013

[Candela et al. 2014] L. Candela; D. Castelli; G. Coro; L. Lelii; F. Mangiacrapa; V. Marioli; P. Pagano (2014) An Infrastructure-oriented Approach for supporting Biodiversity Research. Ecological Informatics, Elsevier, 2014, doi: 10.1016/j.ecoinf.2014.07.006

[Candela et al. 2015] Candela, L.; Castelli, D.; Manghi, P.; Tani, A. Data Journals: A Survey. Journal of the Association for Information Science and Technology, doi: 10.1002/asi.23358

[Casajus et al. 2010] A. Casajus; R. Graciani; S. Paterson; A. Tsaregorodtsev (2010) DIRAC Pilot Framework and the DIRAC Workload Management System. Journal of Physics: Conference Series, Vol. 219, doi: 10.1088/1742-6596/219/6/062049

[Churches et a. 2006] D. Churches, G. Gombas, A. Harrison, J. Maassen, C. Robinson, M. Shields, I. Taylor, I. Wang (2006) Programming scientific and distributed workflow with Triana services. Concurrency and Computation: Practice and Experience, vol. 18 issue 10, Wiley, doi: 10.1002/cpe.992

[Deelman et al. 2015] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, K. Wenger (2015) Pegasus, a workflow management system for science automation, Future Generation of Computing Systems, vol. 46, 2015. doi: 10.1016/j.future.2014.10.008

[Falt et al. 2014] Z. Falt, D. Bednárek, M. Kruliš, J. Yaghob, F. Zavoral (2014) Bobolang: a language for parallel streaming applications. In Proceedings of the 23rd international symposium on High-performance parallel and distributed computing (HPDC '14). ACM, New York, NY, USA, 311-314. doi: 10.1145/2600212.2600711

[Filgueira et al. 2015] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Spinuso and S. Sanchez-Exposito, *dispel4py: An Agile Framework for Data-Intensive eScience*, e-Science (e-Science), 2015 IEEE 11th International Conference on, Munich, 2015, pp. 454-464.

[Filgueira et al. 2016] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Moreno (2016) dispel4py: A Python framework for data-intensive scientific computing. The International Journal of High Performance Computing Applications. doi: 10.1177/1094342016649766

[Hardisty et al. 2017] A. Hardisty, A. Nieva de la Hidalga, D. Lear, B. Magagna, M. Atkinson, K. G. Jeffery, P. Martin, Z. Zhao (2017) A definition of the ENVRIplus Reference Model. ENVRIplus Project Deliverable D5.2

[Hellström et al. 2016] M. Hellström, A. Vermeulen, O. Mirzov, J. Tarniewicz, L. Hazan, L. Rivier, S. Sabbatini, D. Vitale, D. Papale, S.D. Jones, B. Pfeil and T. Johannessen, Near Real Time Data Processing in ICOS RI, Proc. IT4RIs-2016 workshop, November 29, 2017, Porto, Portugal. doi: 10.5281/zenodo.204817

[Hellström et al. 2017] M. Hellström, M. Lassi, A. Vermeulen, R. Huber, M. Stocker, F. Toussaint, M. Atkinson, M. Fiebig (2017) A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations. ENVRIplus Project Deliverable D6.1, January 2017

[Hey et al. 2009] T. Hey, S. Tansley, K. Tolle (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[ICOS 2014] W. Kutsch et al.: Technical and scientific description of ICOS research infrastructure (ICOS RI). Available on request via the ICOS head office (e-mail info@icos-ri.eu).

[ICOS 2016] W. Kutsch et al. The ICOS RI data lifecycle plan. Work in progress. (Contact alex.vermeulen@icos-ri.eu for more information.)

[Jeffery et al. 2017 a] K. G Jeffery, M. Atkinson, Z. Zhao, Y. Chen, A. Nieva de la Hidalga, A. Hardisty, Y. Legre, L. Candela, D. Bailo, T. Loubrieu, B. Magagna (2017) A Development Plan for Common Operations and Cross-Cutting Services based on a Network of Data Managers and Developers. ENVRIplus Project Deliverable D5.4, January 2017

[Jeffery et al. 2017 b] K. G Jeffery, Z. Zhao, B. Magagna, A. Nieva de la Hidalga, L. Candela, C.-F. Enell, M. Hellström, A. Hardisty, C. Paxton, F. Toussaint (2017) Data Curation in System Level Sciences: Initial Design. ENVRIplus Project Deliverable D7.1, January 2017

[Lebo et al. 2013] T. Lebo, S. Sahoo, D. McGuinness (2013) PROV-O: The PROV Ontology. W3C Recommendation. http://www.w3.org/TR/prov-o/

[Li et al. 2014] F. Li, B. C. Ooi, M. T. Özsu, S. Wu (2014) Distributed data management using MapReduce. ACM Computing Survey 46, 3, Article 31 (January 2014), 42 pages. doi: 10.1145/2503009

[Liu et a. 2015] J. Liu, E. Pacitti, P. Valduriez, M. Mattoso (2015) A Survey of Data-Intensive Scientific Workflow Management. Journal of Grid Computing, vol. 13, issue 4, 2015. doi: 10.1007/s10723-015-9329-8

[Loubrieu et al. 2017] T. Loubrieu, F. Merceur, A. Chanzy, C. Pichot, D. Boulanger, F. André, M. Hellström, B. Magnana, A. Nieva De La Hidalga, Z. Zhao, P. Martin (2017) Interoperable cataloging and harmonization for environmental RI projects: system design. ENVRIplus Project Deliverable D8.3, January 2017

[Ludäscher et al. 2006] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E.A. Lee, J. Tao, Y. Zhao (2006), Scientific workflow management and the Kepler system. Concurrency and Computation: Practices and Experiences, 18: 1039–1065. doi: 10.1002/cpe.994

[Lund Myhre et al. 2015] C. Lund Myhre, L. Mona, M. Fiebig, G. D'Amico, F. Amato, E. O'Connor, J. Descloitres, A. M. Fjaeraa, T. Hamburger, A. Hirsikko, P. Laj (2015) ACTRIS Data Management Plan. ACTRIS Deliverable D10.1. October 2015 http://www.actris.eu/Portals/46/Publications/DataCentre/ACTRIS_Data_Management_Plan.pdf

[Marru et al. 2011] S. Marru, L. Gunathilake, C. Herath, P. Tangchaisin, M. Pierce, C. Mattmann, R. Singh, T. Gunarathne, E. Chinthaka, R. Gardler, A. Slominski, A. Douma, S. Perera, S. Weerawarana (2011) Apache airavata: a framework for distributed applications and computational workflows. In Proceedings of the 2011 ACM workshop on Gateway computing environments (GCE '11). ACM, New York, NY, USA, 21-28. doi: 10.1145/2110486.2110490

[Karl et al. 1995] T.R. Karl, V.E. Derr, D.R. Easterling, C. K. Folland, D. J. Hofmann, S. Levitus, N. Nicholls, D. E. Parker, G. W. Withee (1995) Critical issues for long-term climate monitoring. (1995) Climatic Change, 31, 185–221 doi: 10.1007/BF01095146

[Karstens et al. 2016] U. Karstens, I. van der Laan-Luijkx, M. Selander, O. Mirzov, R. Groth, A. Vermeulen and M. Hellström. The ICOS contribution to the EUDAT/EGI Pilot Activity. Work in progress. (Contact ute.karstens@nateko.lu.se for more information.)

[Khalifa et al. 2016] A. Khalifa; Y. Elshater; K. Sundaravarathan; A. Bhat; P. Martin; F. Imam; D. Rope; M. McRoberts; C. Statchuk (2016) The Six Pillars for Building Big Data Analytics Ecosystems. *ACM Comput. Surv.* 49, 2, Article 33 (August 2016), 36 pages. doi: http://dx.doi.org/10.1145/2963143

[Liew et al. 2016] Chee Sun Liew; M. P. Atkinson; M. Galea; Tan Fong Ang; P. Martin; J. I. Van Hemert (2016) Scientific Workflows: Moving Across Paradigms. *ACM Comput. Surv.* 49, 4, Article 66 (December 2016), 39 pages. doi: 10.1145/3012429

[Martin et al. 2016] P. Martin, Z. Zhao, M. Stocker, R. Huber, J. Heikkinen, A. Kallio (2016) Performance optimization for environmental RI projects: System Design. ENVRIplus Project Deliverable D7.3

[Masarie et al. 2014] K.A. Masarie, W. Peters, A.R. Jacobson and P.P. Tans: ObsPack: a framework for the preparation, delivery, and attribution of atmospheric greenhouse gas measurements, Earth Syst. Sci. Data, 6, 375-384, doi: 10.5194/essd-6-375-2014.

[McNamara and Buland 2004] D. E. McNamara, R. P. Buland (2004) Ambient Noise Levels in the Continental United States. Bulletin of the Seismological Society of America, DOI: 10.1785/012003001

[Myers et al. 2015] J. Myers, M. Hedstrom, D. Akmon, S. Payette, B. A. Plale, I. Kouper, S. McCaulay, R. McDonald, I. Suriarachchi, A. Varadharaju, P. Kumar, M. Elag, J. Lee, R. Kooper, L. Marini (2015) Towards Sustainable Curation and Preservation: The SEAD Project's Data Services Approach. IEEE 11th International Conference on e-Science, Munich, 2015, pp. 485-494. doi: 10.1109/eScience.2015.56

[OGC WPS] Open Geospatial Consortium. OpenGIS Web Processing Service. http://www.opengeospatial.org/standards/wps

[Perez & Granger 2007] F. Perez and B. E. Granger (2007) IPython: A System for Interactive Scientific Computing. In Computing in Science & Engineering, vol. 9, no. 3, pp. 21-29, May-June 2007. doi: 10.1109/MCSE.2007.53

[Peterson 1993] J. Peterson (1993) Observations and modelling of seismic background noise. U.S. Department of Interior Geological Survey, Report 93-322

[Smith & Tsaregorodtsev 2008] A.C. Smith; A. Tsaregorodtsev (2008) DIRAC: Data Production Management. Journal of Physics: Conference Series, Vol. 119, doi: 10.1088/1742-6596/119/6/062046

[Thain et al. 2005] D. Thain, T. Tannenbaum, M. Livny (2005), Distributed computing in practice: the Condor experience. Concurrency and Computation: Practices and Experiences, 17: 323–356. doi: 10.1002/cpe.938

[Tsaregorodtsev et al. 2004] A. Tsaregorodtsev; V. Garonne; I. Stokes-Rees (2004) DIRAC: a scalable lightweight architecture for high throughput computing. Fifth IEEE/ACM International Workshop on Grid Computing. doi: 10.1109/GRID.2004.22

[Vanden Berghe et al. 2015] E. Vanden Berghe, G. Coro, N. Bailly, F. Fiorellato, C. Aldemita, A. Ellenbroek, P. Pagano (2015) Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. Ecological Informatics; 28:29–41 doi: 10.1016/j.ecoinf.2015.05.004

[Wolstencroft 2013] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M.P. Balcazar Vargas, S. Sufi, C. Goble C. (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Research, 41 (W1): W557-W561, doi: 10.1093/nar/gkt328

[Xie 2015] Y. Xie (2015) Dynamic Documents with R and knitr, Second Edition. Chapman & Hall/CRC The R Series

[Zaharia et al. 2016] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica (2016) Apache Spark: A Unified Engine for Big Data Processing. Communications of the ACM, Vol. 59 No. 11, Pages 56-65 doi: 10.1145/2934664