DRIHM2US

DISTRIBUTED RESEARCH INFRASTRUCTURE FOR HYDRO-METEOROLOGY TO UNITED STATES OF AMERICA

D2.3: Opportunity and gap analysis

Abstract: This documents presents the analysis of the gaps identified using the results of tasks 2.1 and 2.2 and discusses the opportunity in terms of interoperability aspects.

DRIHM2US (G.A. n° 313122) is co-Funded by the EC under 7th Framework Programme

Document Information Page

Contract Number	313122
Project Name	Distributed Research Infrastructure for Hydro-Meteorology to United States of America
Project Acronym	DRIHM2US
Deliverable Number	2.3
Deliverable Name	Report on Opportunity and gap analysis
Work Package Number	2
Work Package Name	Architecture Harmonization Analysis and Planning
Deadline	01/04/2014
Version	1.0
Dissemination Level	PU
Nature	R
Lead Beneficiary	IMATI

Document History

Date	Version	Description
7th Jul 2014	0.1	Initial write up for internal review by IMATI.
1st August 2014	0.2	Draft version issued for review by project team.
31st August 2014	1.0	Final version

Table of Contents

1 Executive Summary

The objective of Work Package 2, Architecture Harmonization Analysis and Planning Sustainable International Research Infrastructure, is to form an assessment of the architectures supporting the technologies of present e-Infrastructure Approaches for Hydro-Meteorological Research (HMR) in Europe and the US with respect to the future requirements for interoperation.

This deliverable is based on the results of task 2.1, which assesses the most important projects and initiatives in this field, and of tasks 2.2, which sketches a generic architectural model.

The outcome of this task identifies opportunities and gaps with respect to the following items:

1. The Scientists
2. The Infrastructures
3. The HMR Models and Workflows
4. The HM Data

The results feed the definition of the scope and targets of the interoperability experiments of work package 3.

2 Introduction

The simulation of environmental systems requires the coupling of different models representing the natural processes that characterize the system. These models are often designed using different and in some cases incompatible approaches. Integrated environmental modelling aims at achieving systems modelling capabilities by assembling collections of linked components (i.e. models, services and data) supporting different aspects of the combined system.

Deliverable 2.1 [1] presents an evaluation of the most important e-Infrastructures for Hydro-Meteorological Research (HMR). The general finding of that deliverable is that there are several building blocks available, but first of all the possibility to successfully combine them has to be evaluated for future advanced HMR services and applications. At the heart of this challenge lies the ability to have easy access to model and data components, assembling them according to recognized standards and facilitating collaboration between the appropriate scientific and ICT communities. This is of particular importance because ICT methods are often aimed at satisfying general needs resulting in a gap between specific modelling chain requirements and available tools.

The identification of the possible gaps of e-Infrastructure approaches for HMR is a cumbersome task, and a first analysis was conducted within the Distributed Research Infrastructure for Hydro-Meteorology Study[1] (DRIHMS) in 2011 [2]. The analysis was based on the results of two questionnaires, one for the HMR community and one for the ICT community, augmented by additional expert interviews. Globally, about 300 respondents returned the questionnaire.

To summarize the results, HMR is increasingly challenged by the ability to exploit computational resources for online operations and by the ability to retrieve and access data from different sources/countries, stored in different formats, and to be used in combination with various hydrological/meteorological models. Therefore, the four ICT key aspects are

[1] http://www.drihms.eu

represented by a) data management services, b) the availability of High Performance Computing (HPC) resources, together with c) dedicated services for Workflow management and in general d) dedicated portals and user interfaces.

The critical aspect is the usability of such services. Nearly all respondents claim for a significantly more user-driven approach ("Most of the developers seem to develop what they want to develop, but not what is really needed. No one really seems to put the user into focus."). Technically, this is often expressed as a severe difficulty to add new application components (e.g., models) to existing workflows or as a lack of adequate data management.

The final outcome of DRIHMS project in 2011 was the production of the DRIHM White Paper [3]. Since then we achieved progresses in different directions. The results of the DRIHM project have demonstrated the possibility of coupling models in a feasible way, and how to provide seamless access to computational resources available across Europe. The investigations in this project extended our knowledge from European to US based systems (Deliverable 2.1) and provided architectural hints and model (Deliverable 2.2). In the following Sections we summarize the finding of Deliverable 2.2 (Section 3) and then each Section addresses items of the following list:

1. The Scientists
2. The Infrastructures
3. The HMR Models and Workflows
4. The HM Data

3 The Common Architecture Model for HMR Infrastructure

In this section we briefly summarize the results of Deliverable 2.2 [4], which represents the baseline for the analysis presented here. The deliverable does not provide a complete architectural model but it represents an important contribution as a discussion and implementation guideline not only for HMR e-Science infrastructures, but more in general for data- and HPC-intensive e-Science infrastructures.

Following the Model Driven Architecture adopted in the abovementioned deliverable, Figure 1 represents the Domain Model (also called Computation Independent Model – CIM in the Model Driven Architecture methodology), providing a view of a system from the computation independent viewpoint.

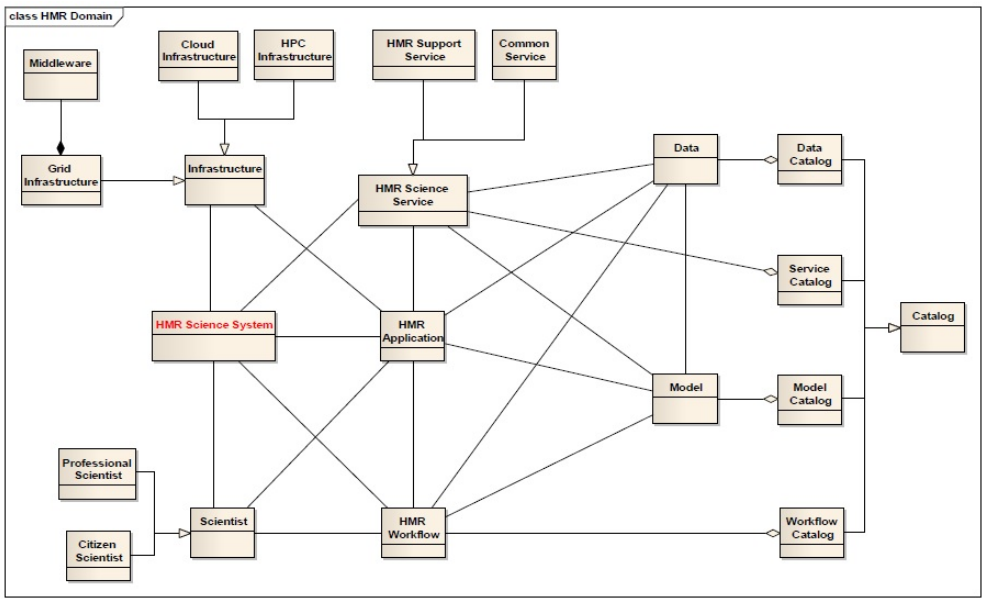

Figure 1 HMR Domain Model

The Domain model is closely related to the generic reference framework presented in Deliverable 2.1, shown here in Figure 2, and the architecture proposed in the DRIHMS White Paper [3].

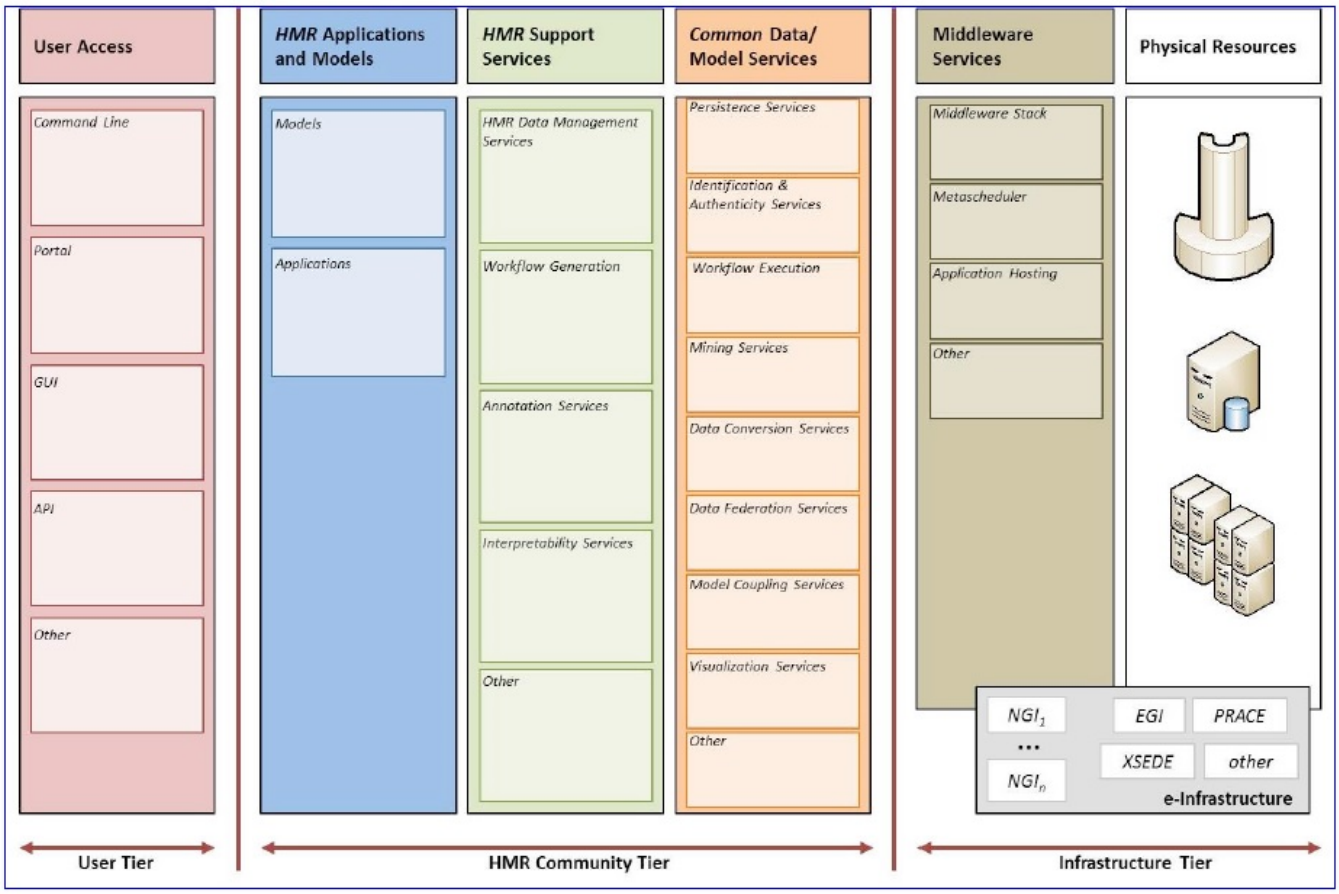

Figure 2 Generic reference framework

For the gap analysis the main components of Figure 3 (derived by Figure 1 with the addition of numbered boxes to better define them) have been analyzed together with the framework of Figure 2, because they represent two complementary views of HMR infrastructures.

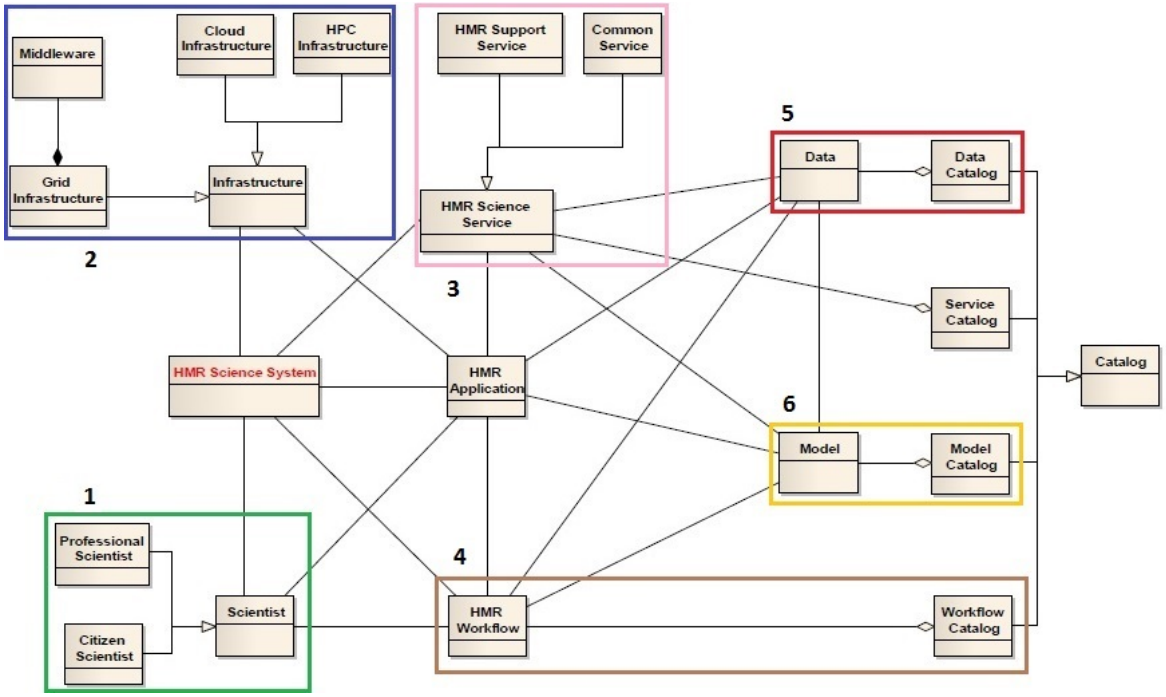

Figure 3 Main components of the HMR Domain Model

In particular the key component of Figure 3 is represented by the generic HMR Science System, which makes one or more HMR applications available:

- The **scientists** (component 1 of Figure 3) interact with the system through one or more user interfaces, corresponding to the User tier of Figure 2. These interfaces allow running the HMR application(s) provided by the system.
- The system is based on an **infrastructure** (component 2 of Figure 3), whose physical resources can be provided through middleware services, corresponding to the Infrastructure tier of Figure 2.
- An HMR application can correspond to a single **model**, a composition of fully coupled models (both represented by component 6 of Figure 3) or a **workflow** of models (component 4 of Figure 3). Models and workflows can exploit specific HMR or general purpose Services (component 3 of Figure 3). All these components correspond to the HMR Community tier of Figure 2.

- An HMR application processes and produces **data**, and possibly exploits services (specific or general purpose as above) to share, mine, store and transfer them. Also these services belong to the HMR Community tier of Figure 2.

Each of these four items (i.e. the scientists, the infrastructures, the HMR models and Workflows and the HM data) is analyzed in details in one of the following sections.

4 The Scientists

The first aspect in our analysis is represented by the support provided by the current systems to the scientists that use them to perform their experiments. But which scientists actually use these systems? In the DRIHM project three classes were identified [5]:

- Citizen scientists, i.e. enthusiast citizens that a) can act as data or computational resources providers, b) can be actively enrolled to process and analyze data, c) can improve and pass on knowledge and experience to others, making a contribution to the knowledge society beyond their immediate circle and life-spans;
- Scientists, which can setup and execute large-scale simulations;
- Expert scientists, which can provide new models or improve the existing ones, provide data for the other scientists (i.e. static or calibration data for hydrological or hydraulic models) and perform simulations using an extended set of services or computational resources. Public servants belong to this class.

As regards the first class, a solid consensus grew in the scientific community that the so-called citizen science could contribute to advancing our current knowledge in many disciplines. In particular HM researchers immediately pointed out the countless advantages of cooperating with citizen scientists [6]. An article presenting the DRIHM vision of near-future citizen science, as well as concrete examples of how it is already considerably contributing to hydro-meteorological research has been published on the project website[2]. However in this report we focus on the actual HM scientists.

What does a scientist expect from research infrastructures for integrated environmental modeling? It depends on the kind of scientists but, in general, the focus is on the interfaces they provide.

[2]http://www.drihm.eu/images/documents/CitizenScientist/Citizen-scientists%20weather%20observations%20-%20DRIHM%20vision.pdf

The models composing a HM workflow are usually configured in different, specific ways - from graphical user interfaces to a list of parameters created with a common text editor - requiring an extensive knowledge of each model.

HMR user interfaces have to be enhanced in such a way that the configuration of a complete workflow reduces the risk of configuration mistakes, i.e. non-overlapping time intervals between consecutive models, unrelated spatial domains and so on. This is especially relevant for newcomers but it is generally applicable for HM scientists, who are not expected to be familiar with all the available models.

Many typical use cases can be addressed by use of graphical user interfaces (GUIs). The most interesting solutions among the systems considered in Deliverable 2.1 are represented by DRIHM, CESM and partially WIRADA, whose GUI-based interfaces exactly focus on these aspects.

But do all users benefit from GUI-based user interfaces since they can restrict active, iterative development of the core model code base? Expert users often prefer to have the full control. Certain activities such as formulating input datasets, creating boundary conditions or calibrating models do not lend themselves to use of GUIs (which would be more typically used in mature models). For this aspect the most interesting solution is represented by ESMF, that provides mainly command line interfaces.

The last point to consider is represented by interoperability aspects between the services of different e-Infrastructures, whose common solution is represented by the provisioning of a specific set of APIs. The most interesting cases are represented by the CUAHSI HIS, which is able to serve its data via GUI for the scientists and via Web Services APIs for ICT systems, and by ESMF, that offers a Web service interface for model coupling via OpenMI. The solution provided by MAPPER addresses these aspects at a different level, because it leverages product specific access to models, services and resources.

In general, none of the considered e-Infrastructures provide all three kinds of interfaces, and the solutions implemented depend on the aim of the system. However, with respect to the situation analyzed in 2011, user-friendly interfaces for HM scientists have been developed [7].

5 The Infrastructures

From the HMR point of view, a HMR Science System has to be a place where scientists can do "science" without being hampered by "computer programming" issues.

Therefore great attention has to be paid to design solutions that provide easy, secure and consistent access to the underlying ICT infrastructure, which has to be tailored to the expected user communities. This is even more compelling in the case of non-ICT experts, as most of the HM scientists, who have to focus on the model results and not on technicalities such as job scheduling, data movements and so on.

Grid portals and Science gateways represent a feasible solution to hide all these complexities.

Grid portals emerged as a solution for the need of a simple and straightforward environment for Grid exploitation. In particular a portal is a Web-based user interface able to unify and compact the information search and utilization of hardware resources and software, and to hide the level of complexity of the Grid, enabling inexperienced users to access and use it. "Science gateways" can be considered the technological evolution of Grid portals.

The term has been coined by the TeraGrid project (now XSEDE) in the US[3], then has been picked up and is currently widely used in various e-infrastructure and e-science collaborations. The XSEDE definition is "A Science Gateway is a community-developed set of tools, applications, and data that are integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community. Gateways enable entire communities of users associated with a common discipline to use national resources through a common interface that is configured for optimal use. Researchers can focus on their scientific goals and less on assembling the cyberinfrastructure they require. Gateways can also foster collaborations and the exchange of ideas among researchers".

[3] https://www.xsede.org/gateways-overview

If we compare Grid portals and Science Gateways, we can see that in both cases users interact with dedicated and customized interfaces to run a specific software, disregarding how and where the software is executed. However a science gateway provides more specific features:

- it allows to consider a "larger" DCI also composed by non-Grid resources;
- it puts a major emphasis on tools and data with respect to the resources used to run the analyses tasks, because it is based on a community-designed interface that provides user-friendly access to several community-specific services.

In conclusion a Science Gateway represents the tier between the user and the computational resources, and it integrates the set of services (models, data and tools) specific for the community.

Several different toolkits exists, each with its pros and cons [8]. Disregarding CUAHSI HIS and WIRADA, which do not rely on computational resources to run HMR models, the other considered projects present the following architectural approaches.

CESM and ESMF are based on the XSEDE infrastructure. In this case the solution is straightforward because, as said before, the concept of science gateway is natively supported by this infrastructure and a detailed how-to about the science gateway creation is provided[4].

MAPPER tools are independent of any middleware technologies, although they favor the QosCosGrid middleware, now called the QCG middleware[5]. This is an integrated system offering advanced job and resource management capabilities to deliver to end-users supercomputer-like performance and structure. In particular the middleware provides remote access to computational resources in a single cluster or many clusters in Europe based on Unicore, gLite, Globus Toolkit, or queuing systems such as Sun Grid Engine (SGE), Platform LSF, Torque/Maui, PBS Pro, Condor, Apple XGrid and LoadLeveler.

[4] https://www.xsede.org/web/guest/for-developers

[5] http://www.qoscosgrid.org/trac/qcg

The DRIHM project offers a science gateway developed using gUSE (grid and cloud User Support Environment), an open source science gateway toolkit based on the Liferay portlet container.

Figure 4 THe multi-tier architecture of gUSE

Three different tiers compose the gUSE architecture. The Architectural tier is formed by the job submission service, the DCI-Bridge, which enables the access of a large number of infrastructures, including clusters (PBS, LSF), Grids (ARC, gLite, GT2, GT4, GT5, UNICORE), supercomputers (e.g. via GT5 or UNICORE), desktop grids (BOINC) and clouds (via CloudBroker Platform, Amazon EC2 or OCCI interfaces). Thus, the DCI-Bridge performs an abstraction of the different DCIs providing specific "plugs/connectors" to each cited infrastructure. The Middle tier contains the gUSE services that enable the creation, management, execution and monitoring of workflows, i.e. the execution of several computational or data manipulation. As regards the analysis performed in the DRIHMS project, this tier is exactly what respondents mean for general-purpose and non-dedicated services. This tier also provides the so-called Application Specific Module (ASM) API that enables a direct access to the provided services and allows to create dedicated HMR-friendly interfaces with the Presentation tier, where the graphical web interface of the customized Science Gateway [9] can be implemented.

One of the last versions of gUSE added the support for XSEDE resources[6]. Therefore this is the most advanced open source solution in terms of supported infrastructures [10].

Another solution is worth mentioning, i.e. the RADICAL Cybertools[7]. It is an abstractions-based suite of well-defined capabilities that are architected for scalable, interoperable and sustainable approaches to support science on a range of high-performance and distributed computing systems. The system currently consists of two components. The first one, named RADICAL-Pilot, is a scalable and flexible Pilot-Job system that provides flexible application-level resource management capabilities. The second one, named RADICAL-SAGA, is a lightweight interface that provides a standards-based interoperability across a range of computing systems. However, for the moment, the solution has been mainly used in the Bioinformatics field.

In conclusion considerable results were achieved in the user-friendly access of heterogeneous ICT resources with respect to the situation depicted in 2011.

However the request to have access to adequate computational resources has to deal with an important issue: the different basic research infrastructures – EGI and PRACE in Europe and XSEDE in the US - are based on different policies.

In details the GRID-based solutions, i.e. those supported by the EGI, grant the access to their computational sites to the members of a Virtual Organization (VO), therefore their use is straightforward for the scientists. Moreover most VOs allow the use of robot certificates. The chief difference between an individual and a robot certificate is that the robot certificate is essentially a single certificate shared by many (human) users that allow the submission of pre-defined and trusted software[8]. Also in XSEDE many gateways provide access to XSEDE resources through a community account rather than setting up unique XSEDE accounts for

[6] http://guse.hu/documentation/guse-3-6-7

[7] http://radical-cybertools.github.io

[8] https://wiki.egi.eu/wiki/Robot_certificates

each gateway user. The gateway user running under the community account typically has privileges to run only a limited set of applications, exactly as for the robot certificate[9]. But the most powerful resources of XSEDE, and all the PRACE resources, are available only to a limited set of scientists who get an access on them.

Therefore, if scientists are able to acquire a grant on HPC resources, the abovementioned systems are able to allow a user-friendly exploitation. Otherwise they can join a VO and use the (often less powerful) resources provided by the Grid infrastructures.

[9] https://www.xsede.org/for-developers

6 The HMR Models and Workflows

A HMR application is implemented as a workflow of possibly heterogeneous HM models, therefore the extensibility is a key feature for most HMR system. An expert scientist who wants to provide a new model will ask: "How do I adapt my model or forecasting workflow to use it on a research infrastructures for integrated environmental modeling?"

The wrong answer is "it depends on the particular infrastructure". The DRIHM model MAP process instead represents a suitable answer.

The main challenges faced in bringing together heterogeneous models on the DRIHM e-Infrastructure have been a) gaining a common understanding of how the different modeling tools work, b) translating the HMR requirements into effective ICT services, and c) resolving different terminologies used by the research groups. This was supported by the DRIHM iterative learning-by-doing approach, where the HM scientists carry out HMR activities while simultaneously working with ICT researchers to develop or adapt the necessary tools and practices. Learning-by-doing has been a well-chosen and positive methodology as each group begins to appreciate the issues faced by the others.

In particular HM scientists used to hard-wire a reduced set of HMR models, mainly due to not standardized interfaces and to the difficult access to computational resources other than their own. This means that, for example, only meteorological model i can be coupled with hydrological model j. Adding a different data format, replacing model j by model j_2, exploring sensitivities, porting the model on a different cluster, etc..., can involve considerable re-engineering and analysis and thus hampers progress in this scientific field.

The lessons learned are: a) when working on an heterogeneous infrastructure, a careful preparation of the model installation bundle is needed in order to minimize requirements on the underlying architecture; b) the ability to chain models strongly depends on the adoption of standardized interfaces; c) a meaningful simulation requires a coherent configuration of all the models involved in the simulation chain. While the third aspect can be addressed by the science gateway approach, a "MAP" process for each model can address the other two aspects. In details the MAP process is based on:

- M – Metadata, Documentation and Licence: each model must be supplied with metadata according to a given standard, appropriate documentation and a licence to use it.
- A – Adaptors (or Bridges) must be provided, which translate the model inputs and outputs from and to common standards.
- P – Portability: each model must be made portable, that is, not strongly tied to local infrastructure.

More details are provided in [11]. The most important point for the present analysis is that DRIHM uses interface standards for controlling the input and output to and from models. This allows models to more easily pass data to and from one another as part of integrated modelling compositions. Two file-based standards and one memory based standard have been adopted:

- NetCDF-CF 1.6[10] – for file-based inputs and outputs, which are structured as Grid or Grid Series. Grid is a single time snapshot of a gridded field, and data is attributed as a single time value for each point, while grid Series are time-series of gridded parameter fields, where data is structured around a grid with results produced at discrete time-steps.
- WaterML 2[11] – for file-based inputs and outputs, which are structured as Point Series, i.e. a time-series of single datum observations at a fixed location. Data varies in time, but not in space, being associated with a single fixed point.
- OpenMI 2.0[12] – for memory-based inputs and outputs.

This last point, i.e. the memory-based exchange of data, is of particular importance for allowing two or more HMR models to interact with each other during the execution of the

[10] http://cfconventions.org

[11] http://www.opengeospatial.org/standards/waterml

[12] http://www.openmi.org

simulation. To this extent the Open Modelling Interface Standard Version 2 (OpenMI) has been approved as a standard by the Open Geospatial Consortium (OGC)[13].

The original driver for the OpenMI was the European Water Framework Directive[14] and the requirement for an integrated approach to water management. The implementation of the Directive was expected to be very challenging and environmental managers would have in all probability needed support, in the form of decision support systems (DSS). As Earth systems are complex and interrelated, these DSS would need to bring together many (heterogeneous) models in order to better understand and predict the environmental impacts of events and policies. The European Commission therefore co-funded the research and development of a generic model interface, the outcome of which is the OpenMI.

As regards the other projects, ESMF organizes models into collections of components with standardized interfaces[15], arranged in hierarchical trees. Modelers can utilize ESMF toolkits for common tasks, and can draw from a pool of modeling components already present in the community. The result is obviously an increased code-sharing and collaboration across organizations, faster knowledge transfer and technical adaptation, and cost savings through code reuse.

CESM is a fully coupled, global climate model (mainly consisting of dynamic geophysical models simulating the atmosphere, ocean, land surface and sear-ice, and one central coupler component) that provides state-of-the-art computer simulations of the Earth's past, present, and future climate states. It uses NetCDF-CF as the standard data format for all output data, including the available post-processing tools.

---

[13] http://www.drihm.eu/index.php/forum/7-technology/89-ogc-extensions-to-web-coverage-service-standard#192

[14] http://ec.europa.eu/environment/water/water-framework/index_en.html

[15] http://www.earthsystemmodeling.org/documents/dev_guide.pdf

The main access point of the WIRADA system is the Hydrologist Workbench, which provides interfaces to hydrological models to integrate and orchestrate modeling tasks. It supports the WaterML 2.0 standard within the defined Water Data Transfer Format (WDTF)[16].

The focus of the hydraulic experiment suite in MAPPER is on the development of a complete mutiscale model for the entire irrigation canal network "La Bourne". This means the focus was more on the coupling of the involved sub-models in order to provide a flexible and powerful multiscale, multiphysics simulation of a system of irrigation canals, than on any other aspects[17].

The porting of a new model is however only the first step for executing a forecasting chain. The management of the whole workflow represents the following aspect to be analyzed. Many scientific workflow systems are available, and some of them are tightly coupled with science gateway toolkits as WS-PGRADE with gUSE, but in principle they are not interoperable. This means that porting a complete workflow in a different HMR system would require to learn a new workflow system and to re-create the workflow, with possibly significant efforts.

The most suitable solution is represented by the design of interoperability solutions allowing the workflow sharing. In this case the most important initiative is represented by the European FP7 "Sharing Interoperable Workflow for Large-Scale Scientific Simulation on Available DCIs" (SHIWA) project[18], that enables the workflow sharing among different workflow systems and e-Infrastructures. In particular the SHIWA Repository[19] stores the formal description of workflows and workflow engines plus the executables and data needed to execute them. These workflows can be exported, and subsequently imported, using the following workflow

[16] http://www.bom.gov.au/water/standards/wdtf/

[17] http://www.mapper-project.eu/web/guest/hydraulics

[18] http://www.shiwa-workflow.eu

[19] https://shiwa-repo.cpc.wmin.ac.uk/shiwa-repo/

management systems: ASKALON, Galaxy, GWES, Kepler, LONI Pipeline, MOTEUR, Pegasus, P-GRADE, ProActive, Triana, Taverna and WS-PGRADE workflows.

The European FP7 "Building a European Research Community through Interoperable Workflows and Data" (ER-flow) project has the aim to disseminate the achievements of the SHIWA project and to build workflow user communities across Europe. ER-flow provides application supports to research communities to develop and share workflows.

In conclusion in this case the execution of heterogeneous models in workflows using also heterogeneous e-Infrastructure is shown to be possible today, but this issues have to be faced with a division of responsibilities along the entire supply chain. Model developers have to prepare their software, for example by "MAPping" it, and ICT scientists have to deploy proper interoperable solutions in terms of workflow management systems.

7 The HM Data

Data represent probably the key issue for HMR systems. Data in fact

- can have a huge size, both as input and/or output;
- can be freely available or can have a controlled access;
- should be efficiently retrieved, therefore meaningful and complete metadata should be defined, including data provenance attributes for the results;
- normally need to be moved between repositories and computing infrastructure, thus requiring powerful interconnection bandwidth and dedicated ICT services;
- have to adhere to standards, in particular when data flow in forecasting chains composed by heterogeneous models.

The last point was discussed in the previous section, and is reported here for the sake of completeness. As regards the other points, the most interesting solutions among the considered projects are presented by CESM, whose Data Management and Data Distribution Plan is described in details in [12].

The two broad categories of CESM output data are development simulations and production simulations. Development simulations are primarily short-term (days of model time to a few years), executed to test and evaluate new or updated parameterization schemes, software engineering developments, improved input datasets, modernized model components, and other changes. There may be hundreds of these short simulations, and therefore the provenance information is critical. Production simulations are typically much longer in duration, from decades to millennia of simulated time. These production simulations are the most extensively used by the CESM user community, so aspects as data retrieval and data transfer apply.

CESM data is developed to follow the Climate and Forecast (CF) metadata conventions (the abovementioned NetCDF-CF[10]) as closely as possible. This system is adopted also by the ESMF project.

Provenance is indirectly addressed in that all simulations are strongly encouraged to follow good practice in recording the model configuration, computing platforms, input datasets, and other information. CESM uses a run database that lists all information necessary to reproduce a simulation. Additionally, records of all the steps involved in creating post-processed data are stored, both by attributes within each netCDF-CF file as well as log files of the post-processing analysis.

Due to the large volume of data generated, in the order of Terabytes, no centre involved in the project can support all CESM data. It is therefore necessary to coordinate the data storage and discovery services plus suitable access policies among the various sites where CESM data are archived. The Integrated Rule-Oriented Data System (iRODS)[20] for creating data grids, digital libraries, persistent archives, and real-time data systems is used for the data storage. As regards data transfer methods, CESM and ESMF rely on XSEDE, which recommend using one among Globus, the Globus Command Line Interface, globus-url-copy, uberftp, scp and sftp. Pros and cons are analysed on a dedicated page[21]. In general the use of Globus is suggested because it fast (the typical transfer rate ranges from 100 to 200 MBps[22]) and reliable. Moreover researchers with no ICT background can easily move large quantities of files, or move files of large size, using the Web GUI and developers who want to automate workflows can use the command line interface.

In Europe such aspects are tackled by the EUDAT initiative[23], whose objective is to build a collaborative data infrastructure as a pan-European solution to the challenge of data proliferation in Europe's research communities. In particular B2STAGE[24] is a reliable, efficient,

[20] http://irods.org

[21] https://www.xsede.org/data-transfers

[22] https://www2.cisl.ucar.edu/docs/transfer

[23] http://www.eudat.eu

[24] http://www.eudat.eu/#slide4

light-weight and easy-to-use service to transfer research data sets between EUDAT storage resources and the computational resources of EGI and PRACE.

As regards the data sharing and mining, the CESM "ecosystem" includes the Earth System Documentation (ES-DOC) project[25], which provides services supporting the earth system documentation creation, result analysis and dissemination. Also this system is adopted by the ESMF project.

Also the Purdue Environmental Data Portal[26] provides services to access the datasets managed by the Purdue multidisciplinary data framework. It allows researchers to perform map based data search, data browsing, metadata display, metadata query, data download and data visualization. In particular, besides the simulation result produced using the Purdue portal, the main datasets provided are a) the data produced by the national network of Next Generation Radar (NEXRAD), that comprises 159 Weather Surveillance Radar-1988 Doppler (WSR-88D) sites across the United States and in selected overseas locations; b) the Indiana Precipitation dataset, containing hourly precipitation data from 144 COOP rain gauges located in the states; c) the St. Joseph Watershed data, containing the stream flow data dynamically accessed via CUAHSI HIS web services.

CUAHSI HIS provides a family of web services, called WaterOneFlow[27], as a standard mechanism for the transfer of hydrologic data between hydrologic data servers (databases) and users computers. Web services streamline the often time consuming tasks of extracting data from a data source, transforming it into a usable format and loading it in to an analysis environment. Web services in fact format the data as WaterML. CUAHSI HIS provides also

[25] http://es-doc.org

[26]https://gridsphere.rcac.purdue.edu:10443/gridsphere/gridsphere;jsessionid=44CBF8BC7A6F2D498995EB9FDD298CFE?gs_action=doNewUser&cid=login

[27] http://his.cuahsi.org/wofws.html#wof

other services[28] as HydroDesktop, a free, open source application for finding, getting, analysing and using hydrologic data from the CUAHSI-HIS system. It works with HydroCatalog, which indexes the data to find out what data exists and where they are, then gets the data from HydroServers where they are stored and published, communicating using WaterOneFlow to retrieve a local copy of the data.

The abovementioned Water Data Transfer Format (WDTF)[16] was instead developed in the WIRADA project, to take a national approach to water information. The format helps water providers to efficiently deliver four million files of water observation data to the Australian Bureau of Meteorology each year. The Bureau uses this data to provide reports and updates on the state of Australian water resources. The most important aspect is hat WDTF supports also WaterML 2.0, making this solution interoperable with other systems.

DRIHM and MAPPER projects do not offer any particular service regarding data management, but they use dedicated and straightforward data repositories.

In conclusion data management aspects still represent an open issue. Several technologies and standards have been defined and developed, but there is no proper solution that addressed all the aspects listed at the beginning of this section.

[28] http://his.cuahsi.org/components.html

8 Conclusion

In this deliverable we presented an opportunity and gap analysis considering the HMR related e-Infrastructure assessed in Deliverable 2.1 and the common architecture model presented in Deliverable 2.2.

With respect to the situation analysed in 2011 within the DRIHMS project we can see that several new technologies and standards are available, but also that several issues remain open, in particular regarding the data management. The good news is that the issues are common to systems both in Europe and the US, therefore the major opportunity and challenge is the possibility to design common solutions, which moreover will promote interoperability among them.

The current gaps in research infrastructures for integrated environmental modelling in fact have to be faced with a division of responsibilities along the entire supply chain – from writing core model engines, to creating instances, to integrating with other models, to running and using results. The achievement of this goal requires a tight interaction between the HMR community and the ICT community. The learning by doing practical approach, adopted within the DRIHM project, represents a methodological example of how this cooperation can be fruitful.

9 Acronyms and References

Acronyms and Abbreviations

Acronym / Abbreviation	Definition
DRIHM	Distributed Research Infrastructure for Hydro-Meteorology Study
DRIHM	Distributed Research Infrastructure for Hydro-Meteorology
DRIHM2US	Distributed Research Infrastructure for Hydro-Meteorology to United State of America
EGI	European Grid Infrastructure
HM	Hydro-Meteorological
HMR	Hydro-Meteorological Research
HPC	High Performance Computing
ICT	Information and Communications Technology
MAPPER	Multiscale Applications on European e-Infrastructures
NetCDF	Network Common Data Form
NetCDF CF	NetCDF Climate and Forecast
NGI	National Grid Initiative
OGC	Open Geospatial Consortium
OpenMI	Open Modelling Interface
PRACE	Partnership for Advanced Computing in Europe
WaterML	Water Markup Language
XSEDE	Extreme Science and Engineering Discovery Environment

References

[1] DRIHM2US Consortium, Report on an Assessment of Current e-Infrastructure Approaches for Hydro-Meteo Research in Europe and the US, Deliverable 2.1 of the DRIHM2US Project, 2013

[2] Kranzlmüller D., Schiffers M., Clematis A., D'Agostino D., Galizia A., Quarati A. Parodi A., Morando M., Rebora N., Trasforini E., Molini L., Siccardi F., Craig G., Tafferner A., Towards a Grid Infrastructure for Hydro-Meteorological Research. Computer Science, Vol. 12. pp. 45-62, 2011. ISSN 1508-280

[3] Parodi A., Morando M., Rebora N., Trasforini E., Molini L., Siccardi F., Craig G., Tafferner A., Kranzlmüller D., Schiffers M., Clematis A., D'Agostino D., Galizia A., Quarati A., The DRIHMS White Paper, Aprile 2011, ISBN 978-88-906068-0-9.

[4] DRIHM2US Consortium, Report on a Common Architecture Model, Report D2.2 of the DRIHM2US Project, 2013

[5] DRIHM Consortium, Report on e-Science environment management, Deliverable 4.1 of the DRIHM Project, 2015.

[6] Daniele D'Agostino, Andrea Clematis, Alfonso Quarati, Tatiana Bedrina, Quillon Harpham, Arnold Tafferner, Caroline Foster, Antonio Parodi and Edoardo Mazza, Citizen-Scientists in Hydro-Meteorological Research Activities: The DRIHM Experience, Submitted to the 11th IEEE International Conference on eScience (eScience 2015).

[7] E. Danovaro, L. Roverelli, G. Zereik, A. Galizia, D. D'Agostino, A. Quarati, A. Clematis, F. Delogu, E. Fiori, A. Parodi, C. Straube, N. Felde, Q. Harpham, B. Jagers, L. Garrote, L. Dekic, M. Ivkovic O. Caumont, E. Richard, Setup an hydro-meteo experiment in minutes: the DRIHM e-infrastructure for hydro-meteorology research, IEEE 10th International Conference on e-Science (e-Science) Proceedings, pp. 47-54, IEEE Computer Society, 2014.

[8] D'Agostino, D., An overview of Grid portal technologies for the development of HMR science gateways. EGU General Assembly 2012, p.7308, 2012.

[9] Tamás Kiss, Péter Kacsuk, Róbert Lovas, Ákos Balaskó, Alessandro Spinuso, Malcolm Atkinson, Daniele D'Agostino, Emanuele Danovaro and Michael Schiffers, WS-PGRADE/gUSE in European Projects, Chapter 17 in Science Gateways for Distributed Computing Infrastructures, pp 235-254, Springer International Publishing, 2014.

[10] D. D'Agostino, A. Clematis, A. Galizia, A. Quarati, E. Danovaro, L. Roverelli, G. Zereik, D. Kranzlmuller, M. Schiffers, N. gentschen Felde, C. Straube, A. Parodi, E. Fiori, F. Delogu, O. Caumont, E. Richard, L. Garrote, Q. Harpham, B. Jagers, V. Dimitrijevic, L. Dekic, The DRIHM Project: a Flexible Approach to Integrate HPC, Grid and Cloud Resources for Hydro-Meteorological Research, International Conference for High Performance Computing, Networking, Storage, and Analysis 2014 – Supercomputing 2014 (SC14) Proceedings, pp. 536-546, IEEE Computer Society, 2014.

[11] Quillon Harpham and Emanuele Danovaro, Towards standard metadata to support models and interfaces in a hydro-meteorological model chain, Journal of Hydroinformatics Vol 17 No 2 pp 260–274.

[12] Gary Strand, Community Earth System Model Data Management: Policies and Challenges, Procedia Computer Science, Volume 4, 2011, Pages 558-566.