DRIHM2US

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

D5.1: Overview, assessment, plan and prioritisation of development needs

Abstract: This document identifies the development needs for science, technology and training required for the further development and deployment of the hydro-meteorological infrastructure.

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	5.1
Deliverable Name	Overview, assessment, plan and prioritisation of development needs
Work Package Number	5
Work Package Name	Sustainable International Research Infrastructure
Deadline	28/02/2015
Version	2.0
Dissemination Level	PU
Nature	R
Lead Beneficiary	DELTARES

Document History

Date	Version	Description
4th Feb 2014	0.5	Initial draft version focusing on primary needs
10th Apr 2014	1.0	Updated based on reviews
11th Jan 2015	1.5	Updated version expanding secondary needs
28th Feb 2015	2.0	Final version including reviewer's comments

Table of Contents

List of Figures

List of Tables

1 Executive Summary

This work package 5 focuses on sustainability aspects of developing an international research infrastructure (e-platform) that enables persistent and effective sharing of data and models (including the execution thereof) across earth science disciplines, institutions, and national boundaries. We have worked along two parallel tracks: work package 2 worked towards a "common architecture model" [3] and opportunity and gap analysis [6], and meanwhile the outlines of [5] and plan for [7] an organization needed to support and develop such a platform were developed in the other tasks of the current work package. This development action plan brings these two together and focuses on the development needs related to science (scientific operations), technology (ICT operations), and training (education centre) – the three core elements of both the platform and the associated organization.

This report has been delivered in two stages. In the first stage (as represented by the first chapters of this report) we collected, based on the generic reference framework, the scientific innovation cycle, and a categorization of the audience of the educational centre, a list of 92 needs in total. The second part of this report condenses these needs into a prioritized list of 26 activity cluster each aimed at one of the four themes "data management", "simulation management", "data visualization", and "general user requirements". From this we make the final step towards the action plan listing short-, medium-, and long-term priorities with an overview of resources required and available.

2 Introduction

Building on the DRIHM project[1] in which a prototype infrastructure[2] for hydro-meteorological research (HMR) for Europe is being developed, DRIHM2US set out to coordinate research activities in this domain and the wider earth sciences across Europe and the USA. Such an international research infrastructure should ensure persistent availability and effective sharing of data and models across scientific disciplines, institutions, and national boundaries. To accomplish this task we have developed, together with the US partners of the SCIHM project (Standards-based CyberInfrastructure for HydroMeteorology), a vision on the architecture of an international infrastructure for earth sciences research.

Based on joint prototype developments, and expert networking meetings [4], we have aligned our developments and sharpened our vision. In work package 2 (Architecture Harmonization Analysis and Planning) we started by defining the concept of a "generic reference framework", and applied it in an overview of current approaches for infrastructures related to HMR [2]. From this we worked towards a "common architecture model" [3] and opportunity and gap analysis [6], which guide the technical developments.

This work package (WP5) focuses on the sustainability of the international research infrastructure. Most of the deliverables address the organizational aspects of an organization to support and develop such an infrastructure, but this report focuses on the development needs related to science, technology and training aspects of the e-platform [1]. We assess the interdependencies, and determine secondary needs, resources (required and available) and required actions.

In Chapters 3–5 match those of the first stage version of this document as delivered at month 16 with minor textual corrections. In these chapters we address, in line with the proposed organizational substructure as outlined in [5], the development needs for scientific operations

---

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

[2] Also referred to as an e-platform, e-science environment, or cyberinfrastructure.

[3] See for example https://www.youtube.com/watch?v=s84mwHDKyF4 for a scan of an urban

(science), ICT operations (technology), and education centre (training) respectively. For each group we define a number of categories, and within these categories the needs are numbered; this numbering is not based on prioritization. Hence, each need is characterized by a triplet: group (SC for science, IT for technology, and TR for training), a two-character indication of the category, and unique number of the need within the category. From the three perspectives we arrive at a total of 92 expressed needs.

In this second and final stage version of this report, we continue with a more thematic look at the expressed needs. In Chapter 6 we analyse those 92 needs by consecutively focusing on the four themes "data management", "simulation management", "data visualization", and "general user requirements". Per focus area we describe the interdependencies and condense the relevant needs into a prioritized list of 26 activity clusters in total.

Chapter 7 collects the activity clusters into short-, medium-, and long-term priorities of a development action plan, which includes the identification of resources. The report ends with a brief summary and conclusions in Chapter 8.

3 Needs of scientific operations

We start by specifying the needs of the science side since that is the prime purpose for which the infrastructure is developed. As described in [5], the "Scientific Operations" cover the functional aspects of the 'management of the scientific applications on the ICT infrastructure' and technical support associated with the implementation of scientific applications on the infrastructure.

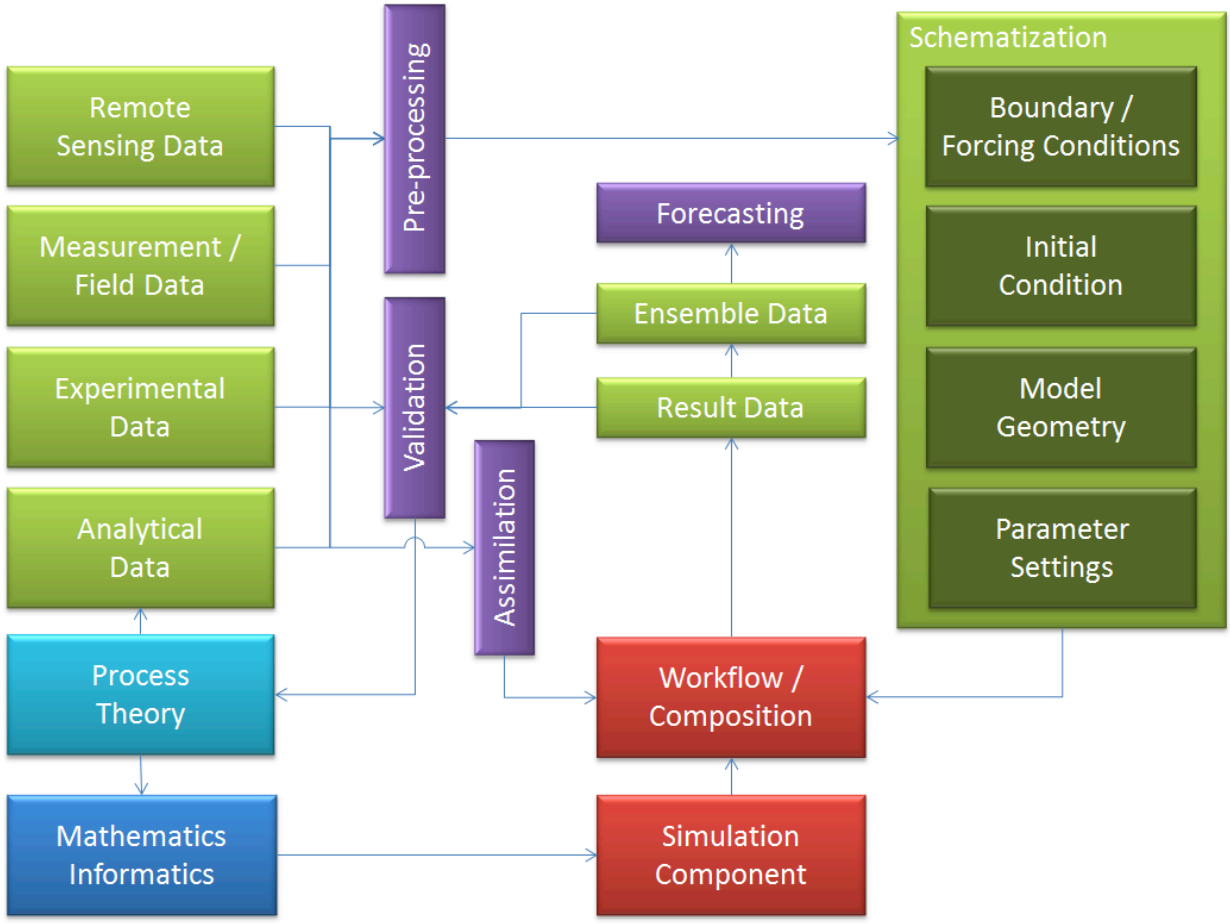

Figure 1: Scientific innovation cycle

Figure 1 shows the scientific innovation cycle in which — based on a theory of the processes involved— a simulation component is developed, possibly included in a workflow or model composition, filled with data to generate a model instance, and executed to give results, which

are compared with observational and analytical data to validate it and verify the theory. Once the software component or model instance has been validated, it may be used to forecast the behaviour of the system under investigation and estimate the impact of events or modifications to the system. Blocks representing data are indicated in green, while theory is indicated in blue, and software components and activities using (other) software components are indicated in red and purple respectively.

It's clear that data plays an important role in the scientific cycle. The data used in the environmental sciences comes from various sources such as theory (mostly for verifying correctness of the simulation components), laboratory or scale model experiments, field measurements, and on the largest scales increasingly from satellites. The data amounts are rapidly increasing due to cheaper electronics to build measurement devices, improved ICT technologies to collect and process data, and by new data collection methods such as laser scanning and new satellite sensors. For example, laser scanning can be used to rapidly (100 thousands of points per second) capture local 3D geometry in high resolution[3], while Google Earth Engine[4] has recently started to provide access to an unprecedented amount of 40 years of Landsat (and other) satellite data. The earth sciences have to deal with big and heterogeneous data. Besides the data preparation steps carried out before the start of the simulation (to generate the geometry and initial conditions), data may be provided in real-time during the simulation; examples include boundary or forcing conditions, and state updates via data assimilation.

From the diagram in Figure 1 we have derived a set of needs for the scientific use of the platform. These needs are listed in Table 1 where they have been subdivided into six categories: data management and processing, simulation management, data-model integration, data visualization, real-time operational use, and user support. Not all services requested may be needed for all use cases in hydro-meteorology or the wider earth sciences.

[3] See for example https://www.youtube.com/watch?v=s84mwHDKvF4 for a scan of an urban environment, and https://www.youtube.com/watch?v=XPg6tkmA8dM for a scan of a natural environment.

[4] https://earthengine.google.org

Table 1: Scientific operations needs

Id	Topic
SC.DP	Data Management and Processing
SC.DP.01	Search and Locate Data
SC.DP.02	Upload and Download Data
SC.DP.03	Set Access Permissions for My Data
SC.DP.04	Store Data: Short – Long Term
SC.DP.05	Remotely Use Data Available on 3rd Party Servers
SC.DP.06	Serve Data for Remote Access
SC.DP.07	Track and Reference Data
SC.DP.08	Filter and Combine Data
SC.DP.09	Convert Data
SC.DP.10	Manage Ensemble Data
SC.DP.11	Version Control
SC.SI	Simulation Management
SC.SI.01	Search and Locate Simulation Components
SC.SI.02	Add and Document Simulation Components
SC.SI.03	Track and Reference Simulation Components
SC.SI.04	Chain Components to Workflows
SC.SI.05	Couple Components into Compositions
SC.SI.06	Track and Reference Workflows and Compositions
SC.SI.07	Prepare Input for Components, Workflows, Compositions
SC.SI.08	Store Configured Components, Workflows, Components
SC.SI.09	Set Access Permissions for My Models
SC.SI.10	Run Components, Workflows, Compositions
SC.SI.11	Configure and Run (Multi-Model) Ensembles
SC.SI.12	View Diagnostic Information of (Failed) Simulations
SC.DM	Data-Model Interaction
SC.DM.01	(Auto) Calibrate Parameter Settings
SC.DM.02	Assimilate Data into Simulations
SC.DM.03	Perform Uncertainty Analysis

SC.VS	Data Visualization
SC.VS.01	Visualize Data in Graph Plots
SC.VS.02	Visualize Data in Map Plots
SC.VS.03	Visualize Data in Slice Plots
SC.VS.04	Visualize Data in 3D Plots
SC.VS.05	Visualize Ensemble Data
SC.VS.06	Add Data to Plots
SC.VS.07	Animate Data in Plots
SC.VS.08	Configure, Store, and Recreate Plots
SC.VS.09	Trace-back Data from Plots
SC.VS.10	Export Plots for Reporting
SC.RT	Real-Time Operational Use
SC.RT.01	Get Guaranteed Availability
SC.RT.02	Get Guaranteed Run Performance
SC.RT.03	Get Guaranteed Data Up/Download Performance
SC.RT.04	Report on Recent User Activity
SC.US	User Support
SC.US.01	Simple to Access (Single Sign On)
SC.US.02	Simple to Use (for Starters)
SC.US.03	Flexible to Use (for Advanced Use)
SC.US.04	Stability, Longevity, and Support

4 Needs of ICT operations

In [5] we have outlined the scope of the "ICT Operations" component of the support organization as the part that deals with the following high-level activities: 'management of the ICT services, management of the scientific applications on the ICT infrastructure', 'provision of technical support to users', and 'ICT incident management'. The 'management of the scientific applications on the ICT infrastructure' should be interpreted here as purely the responsibility for ensuring the technical integrity and performance of the scientific applications from ICT

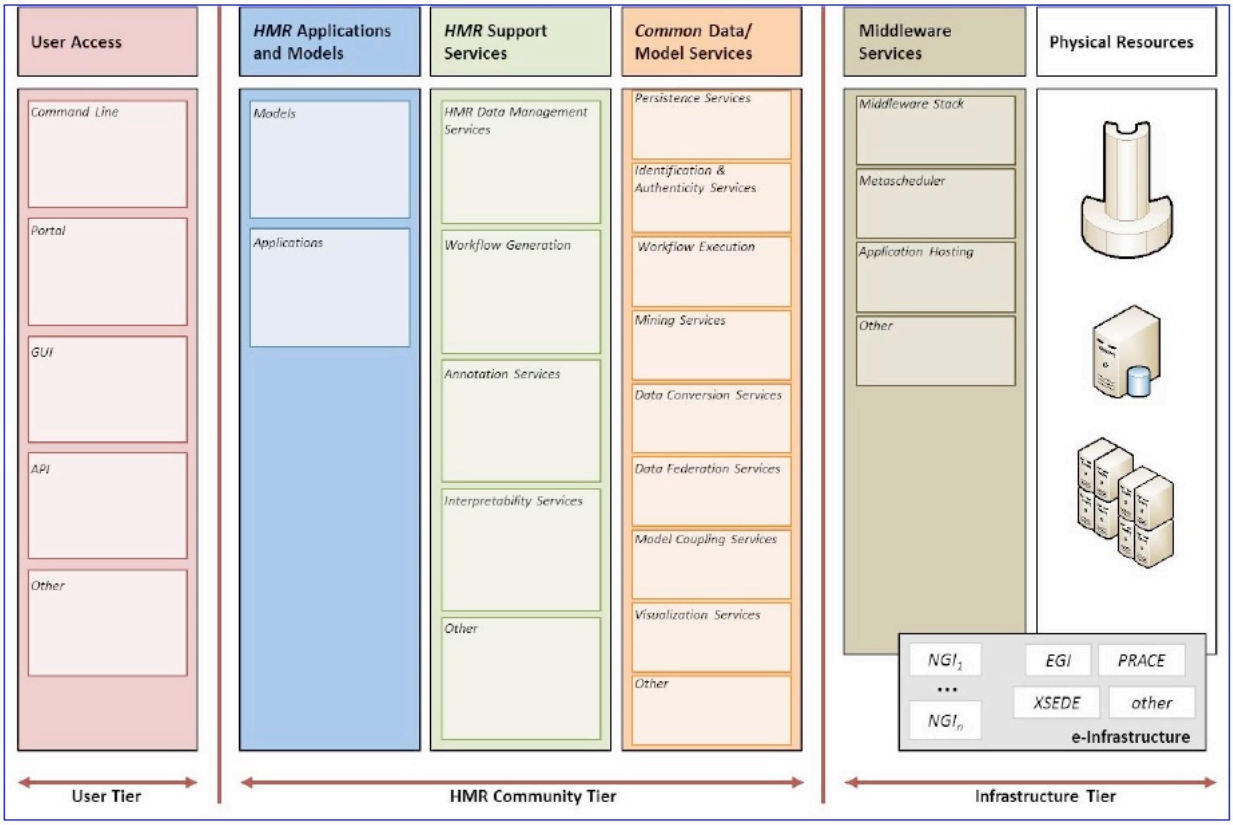

Figure 2: Generic Reference Framework (Fig. 2 of [2])

perspective. As such the main aim of the ICT operations is to hide as much as possible of the complexities of the underlying heterogeneous, distributed computational and storage resources

from the scientist user, while providing the user with maximum stability *and* flexibility. The base architecture to configure simulations, run them across systems, track and store results, and provide access to results and the configurations that created them, is all part of this task.

This chapter addresses the needs of this group by looking at the infrastructure from ICT perspective. For this we will build on the results of the Architecture Harmonization Analysis and Planning work package (WP2). There we are using the "generic reference framework" as shown in Figure 2 to compare infrastructures and classify infrastructure components; this conceptual framework was introduced in [2]. From an ICT perspective, the framework is composed of different service layers that contain components such as networks, grids, data centers, collaborative environments, service registries, single-sign on mechanisms, certificate authorities, training and help-desk services. Here, we will use it to develop the list of primary ICT needs. The components of the framework represent almost one-to-one the needs for the ICT operations. They are listed in Table 2; the categories used correspond to the layers of the reference framework: physical resources, middleware services, common data/model services, HMR support services, HMR applications and models, and user access and support.

Table 2: ICT operations needs

Id	Topic
IT.PR	Physical Resources
IT.PR.01	Grid (Computing)
IT.PR.02	HPC (Computing)
IT.PR.03	Cloud (Computing)
IT.PR.04	Storage
IT.PR.05	Physical Resources for Other Services
IT.MI	Middleware Services
IT.MI.01	Middleware Stack
IT.MI.02	Metascheduler
IT.MI.03	Application Hosting
IT.MI.04	Other Middleware Services
IT.DM	Common Data/Model Services
IT.DM.01	Persistence Services

IT.DM.02	Identification & Authentication Services
IT.DM.03	Workflow Execution
IT.DM.04	Mining Services
IT.DM.05	Data Conversion Services
IT.DM.06	Data Federation Services
IT.DM.07	Model Coupling Services
IT.DM.08	Visualization Services
IT.DM.09	Other Common Data/Model Services
IT.SS	HMR Support Services
IT.SS.01	HMR Data Management Services
IT.SS.02	Workflow Generation
IT.SS.03	Annotation Services
IT.SS.04	Interpretability Services
IT.SS.05	Other HMR Support Services
IT.AM	HMR Application and Models
IT.AM.01	Applications
IT.AM.02	Models
IT.UA	User Access & Support
IT.UA.01	Command Line Interface (CLI)
IT.UA.02	Portal
IT.UA.03	Graphical User Interface (GUI)
IT.UA.04	Application Program Interface (API)
IT.UA.05	Ticket Tracking System
IT.UA.06	Other User Access & Support

5 Needs of education centre

Following [5] the 'Education Centre' covers the 'provision of training', 'provision of scientific workshops and conferences (including journals and publications)', and 'devising and implementing student curricula'. This Educational Centre is directly associated with the ICT infrastructure and covers only those educational needs that relate to its use and development.

For the sustainability of the infrastructure, it's important to have sufficient documentation and training material available for educating ICT developers about the technical aspects of the infrastructure in the proper scientific context. This first audience category is best served by technical information presented in a hierarchical context from general overview down to technical details and code. Experienced scientists and consultants (including the staff of operational forecasting centres) form the second category; the platform should make their daily work easier or give them more opportunities – the education centre needs to be able quickly and adequately answer their questions. However, by getting them on board, we can build curriculum components and attract an increasing number of students. Science students and general citizen scientists form the third and final audience category for the education centre. Students can effectively be reached via their curricula, hands-on training sessions aligned with popular conferences, and webinars and online training videos. Science blogs and lively demonstration videos can help to attract citizens, which may also become actively involved via the online training videos. These needs are summarized in Table 3, which include a final general category representing the need for an overall training website with user forum.

Table 3: Education centre needs

Id	Topic
TR.IT	ICT Developer
TR.IT.01	ICT Architecture Overview
TR.IT.02	Technical Reference Documentation per Component
TR.IT.03	Formal Specification of Standards Used
TR.IT.04	Documentation for Every API Used
TR.IT.05	Documentation of ICT Operations Procedures

TR.EX	Expert Scientist and Consultant
TR.EX.01	Overview of Science Features of the Platform
TR.EX.02	Quick Guide on Platform Use
TR.EX.03	Reference Guide for Standards Applied
TR.EX.04	Training on Adapting Software Components
TR.EX.05	Training on Improving Component Performance
TR.EX.06	Science Focused User Meetings
TR.ST	Science student or Citizen Scientist
TR.ST.01	Science Highlights of Platform Users
TR.ST.02	Webinars / Online Training Videos
TR.ST.03	Introduction Training Sessions at Conferences
TR.ST.04	Series of Demonstrators with Increasing Complexity
TR.GN	General
TR.GN.01	Support Website to Host Information
TR.GN.02	Discussion Forum

6 Interdependencies and Need Prioritization

In this chapter we'll focus on the interdependencies of the needs listed in the previous three chapters. Because it's impractical to discuss all 92 needs here at once, we'll discuss them in smaller groups. Per group we will indicate interdependencies, cluster needs and prioritize clusters to distinguish between primary and secondary needs.

6.1 Data Management and Processing

Figure 3 shows the interdependencies amongst needs of the scientific data management category and their interactions with the needs of other categories. The core element here is to store data (which may be observational data, simulation results, or model configurations) for short to long periods (SC.DP.04) which needs to be implemented as data management services (IT.SS.01) on some physical storage resources (IT.PR.04). These services should include besides local read and write access, and basic backup and restore facilities, also services for remote access to the data at a low level (such as, basic upload and download services; SC.DP.02) and be extendable with the possibility to serve available data via standardized protocols such as OPeNDAP, HIS, WCS (SC.DP.06). To enable smooth interaction with other systems, it should also not be necessary to upload data which is online available on 3rd party servers via such protocols (SC.DP.05).

Although open data would in general be preferred, it should be possible to set data access permissions (SC.DP.03) which links to identification and authentication services (IT.DM.02) which also enables data federation services (IT.DM.06). Basic conversion services (SC.DP.09, IT.DM.05) and various data processing and mining services (SC.DP.08, IT.DM04) can be applied to create derived data sets from the original data stored.

Running a Persistence Services (IT.DM.01) will enable the system to reference particular data sets and track their usage (SC.DP.07) in the system (provenance); this should be integrated with a version control system (SC.DP.11) to track changes in data sets. Although this may not be a prime need, the demand is growing (especially with climate researchers, and operational centres) and it would be good to include this aspect early on in the infrastructure design. This

connects to a data cataloguing and annotation and interpretability services (IT.SS.03, IT.SS.04) which are used to efficiently search and locate data (SC.DP.01) and to annotate data sets with special markers such as ensemble member identifiers (SC.DP.10).

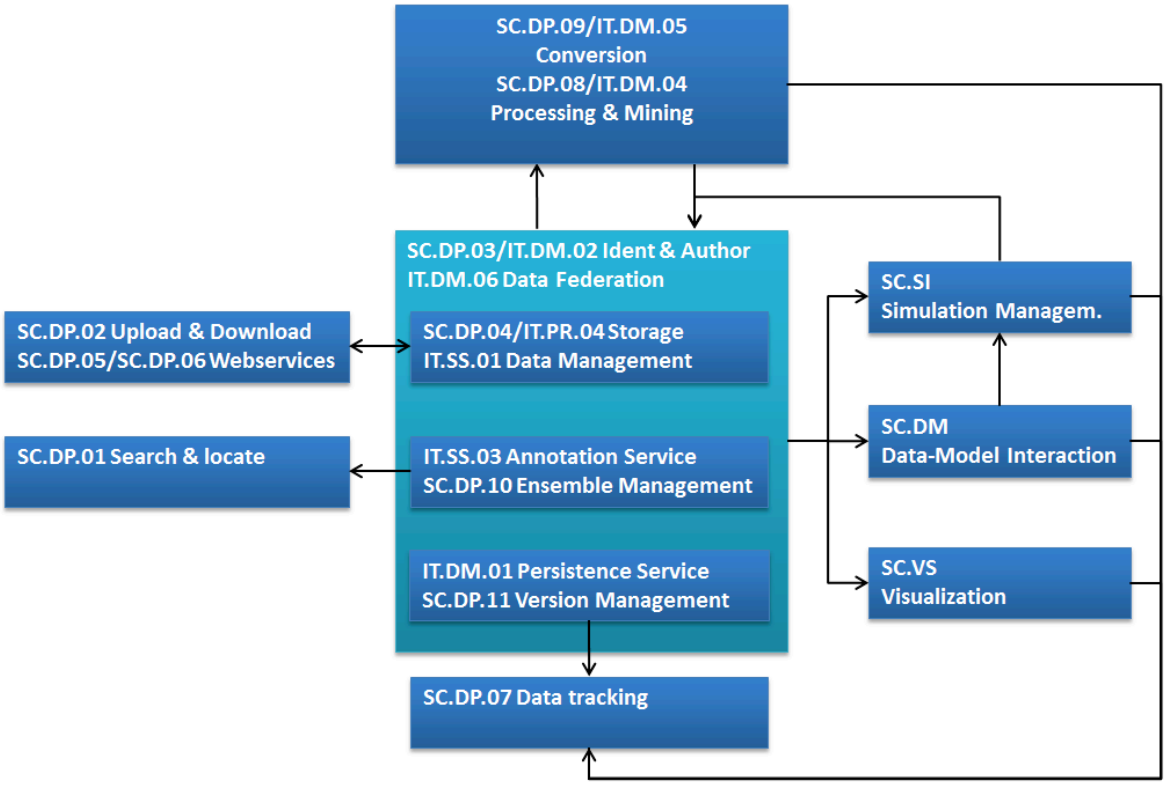

Figure 3: Interdependencies Data Management

Last but not least, it should obviously be possible to use the data for and store data obtained from running simulation models (SC.SI) possibly including data-model interaction (SC.DM), and to optionally visualize the data (SC.VS) in the portal – see also the discussion in Section 6.3. All these components should connect back to the data tracking facility such that for every graph it can be determined on which data it has been based. This results in a prioritized list of 7 activity clusters shown in Table 4.

Table 4: Prioritization for Data Management

Prio	Id	Topic
1	**DM1**	**Data Storage Services**

	SC.DP.04	Store Data: Short – Long Term
	IT.SS.01	HMR Data Management Services
	IT.PR.04	Storage
2	**DM2**	**Basic Data Access Services**
	SC.DP.02	Upload and Download Data
	SC.DP.03	Set Access Permissions for My Data
	IT.DM.02	Identification & Authentication Services
	IT.DM.06	Data Federation Services
3	**DM3**	**Simulation Support Services**
	SC.SI	Include in Simulation Management
	SC.DM	Include in Data-Model Interaction
4	**DM4**	**Data Documentation and Search Services**
	SC.DP.01	Search and Locate Data
	IT.SS.03	Annotation Services
	IT.SS.04	Interpretability Services
5	**DM5**	**Advanced Data Access Services**
	SC.DP.05	Remotely Use Data Available on 3rd Party Servers
	SC.DP.06	Serve Data for Remote Access
6	**DM6**	**Data Provenance Services**
	SC.DP.07	Track and Reference Data
	IT.DM.01	Persistence Services
	SC.DP.11	Version Control
7	**DM7**	**Data Processing Services**
	IT.DM.04	Mining Services
	SC.DP.08	Filter and Combine Data
	IT.DM.05	Data Conversion Services
	SC.DP.09	Convert Data
	SC.DP.10	Manage Ensemble Data

6.2 Simulation Management

Figure 4 shows the interdependencies amongst needs of the scientific simulation management category and their interactions with the needs of other categories. The core element here is the management of generic simulation components (applications) and instances thereof (models): simulation components must be stored (IT.AM) in a version controlled environment, such that new components can be added (SC.SI.02) and existing components can be updated. The software can be stored in a dedicated repository, or as part of the general data management solution (SC.DP); in order to store model configuration (instances of select simulation components with associated data) we'll need a connection between these two systems anyway. An annotation service (IT.SS.03) must be available to add labels and documentation (SC.SI.02), such that users can search and locate components (SC.SI.01).

Although open source and freeware models are in general preferred, it should be possible to set software component specific access permissions (SC.DP.03) which links to identification and authentication services (IT.DM.02). A Persistence Service (IT.DM.01) is needed, similarly as for the data, for identifying and referencing (versions of) software components and tracking their usage (SC.DP.07) in the system (provenance). This feature is even more important for software than for the data because in a *research* environment updates to numerical models tend to occur more frequently than new data. This connects back to the software cataloguing and annotation services (IT.SS.03) mentioned before.

Increasingly simulation components are not run as standalone components, but as part of a bigger integrated model composition using tight coupling (SC.SI.05, and IT.SS.07) while in other model chaining of components into a one-way loosely coupled workflow is sufficient (SC.SI.04, and IT.SS.02). The HMR workflows composed of meteorological, hydrological, and hydraulic models across heterogeneous hardware within the DRIHM project are examples of the latter, while the tighter coupling is needed for the two-way interaction of the 1D hydraulic and flood and 2D flood spreading configurations.

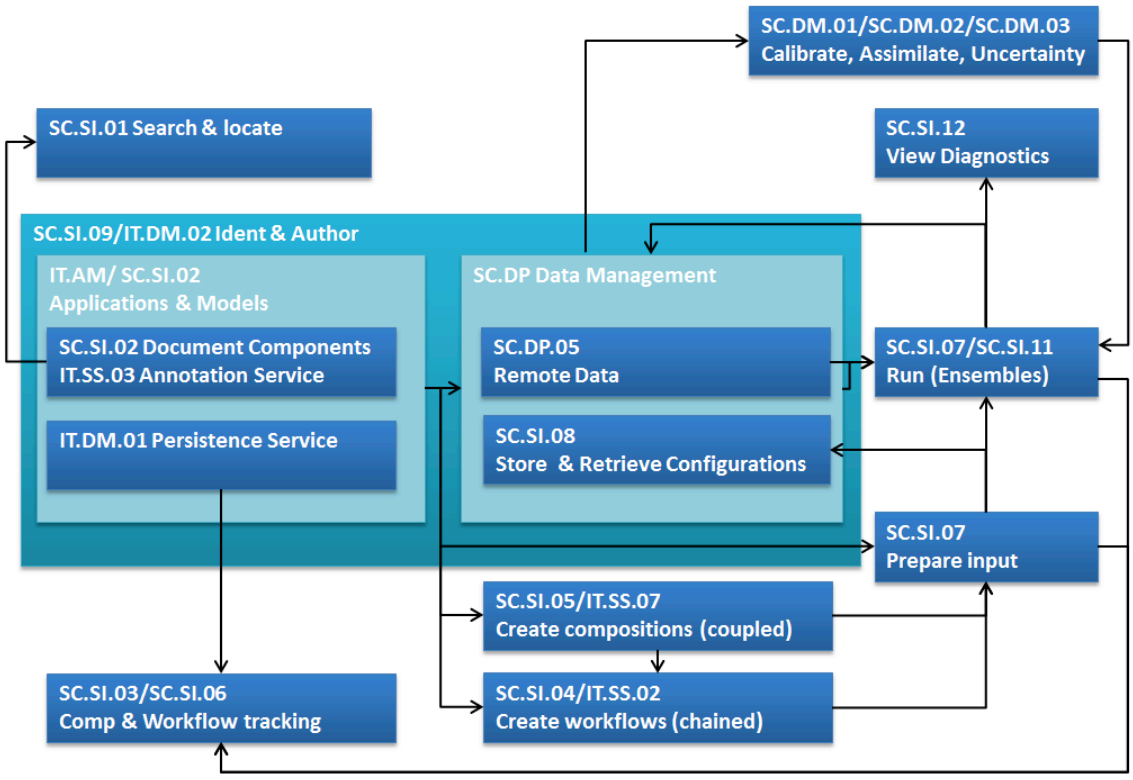

Figure 4: Interdependencies Simulation Management

For both file based and in memory coupling, we'll need standards for the data exchange protocols (file standard or API) as well as naming conventions for the quantities (standard name list or ontology). A large number of other coupling technologies exist, many of which have been listed in [8]; a for grid infrastructures valuable addition was developed by the MAPPER project[5]. There is no generic interoperability yet between the various coupling frameworks, although activities in that direction are on-going. Workflow interoperability has been addressed by the SHIWA project[6].

For single components, and chained or coupled versions thereof, we should be able to specify the input (SC.SI.07) which may point to local data or data on a remote server (SC.DP.05) and

[5] http://www.mapper-project.eu

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

store and retrieve such model instances (configurations; SC.SI.08), run the workflow (IT.DM.03) as a single instance (SC.SI.10) or ensemble (SC.SI.11), and verify execution status and diagnostics (SC.SI.12) while the model results are stored again on the platform for further processing. For running the models we'll obviously need access to computing resources ranging from grid (IT.PR.01) to HPC (IT.PR.02) and cloud (IT.PR.03) and a range of middleware services (IT.MI) which are not indicated in Figure 4. It should be noted that there is a considerable demand for entry level HPC resources with a limited number of parallel cores, i.e. between embarrassingly parallel and 1000+ core parallelization.

Configuring models and running models connects back to the tracking facility such that it's know which models are used and which output are generated with it. Building on the feature to run ensembles a quick-win would be to enable perform uncertainty analysis (SC.DM.03), with a bit more added intelligence the platform should enable (auto) calibration (SC.DM.01) of models and data assimilation (SC.DM.02). This last feature is again very important for future operational users of the platform. The foregoing leads to a prioritized list of 10 activity clusters shown in Table 5.

Table 5: Prioritization for Simulation Management

Prio	Id	Topic
1	**SM1**	**Model Storage Services**
	IT.AM	HMR Application and Models
	SC.DP	Data Management and Processing
2	**SM2**	**Single Model Component Services**
	SC.SI.02	Add and Document Simulation Components
	SC.SI.07	Prepare Input for Components, Workflows, Compositions
	SC.SI.08	Store Configured Components, Workflows, Components
	SC.SI.10	Run Components, Workflows, Compositions
	SC.SI.12	View Diagnostic Information of (Failed) Simulations
3	**SM3**	**Workflow Services**
	SC.SI.04	Chain Components to Workflows
	IT.SS.02	Workflow Generation
	IT.DM.03	Workflow Execution

	IT.SS.04	Interpretability Services
4	**SM4**	**Coupled Component Services**
	SC.SI.05	Couple Components into Compositions
	IT.DM.07	Model Coupling Services
	SC.SI.07	Prepare Input for Components, Workflows, Compositions
	SC.SI.08	Store Configured Components, Workflows, Components
	SC.SI.10	Run Components, Workflows, Compositions
5	**SM5**	**Basic Ensemble Services**
	SC.SI.11	Configure and Run (Multi-Model) Ensembles
	SC.DM.03	Perform Uncertainty Analysis
6	**SM6**	**Software Documentation and Search Services**
	SC.SI.01	Search and Locate Simulation Components
	IT.SS.03	Annotation Services
7	**SM7**	**Model Provenance Services**
	SC.SI.03	Track and Reference Simulation Components
	SC.SI.06	Track and Reference Workflows and Compositions
	IT.DM.01	Persistence Services
8	**SM8**	**Software Access Services**
	SC.SI.09	Set Access Permissions for My Models
	IT.DM.02	Identification & Authentication Services
9	**SM9**	**Data Model Integration Services**
	SC.DM.02	Assimilate Data into Simulations
	SC.DM.01	(Auto) Calibrate Parameter Settings
10	**SM10**	**Remote Coupling Services**
	SC.DP.05	Remotely Use Data Available on 3rd Party Servers

6.3 Data Visualization

Although data visualization features may be attractive to include for demonstration and quick data inspection, it can easily become a can of worms. Figure 5 indicates the interdependencies amongst needs of the data visualization management category and their interactions with the needs of other categories. The core component seems to be combination of visualization

services (IT.DM.08) providing — depending on the data — the ability to create line graphs (SC.VS.01), maps or horizontal slices (SC.VS.02), vertical slices (SC.VS.03), and 3D plots (SC.VS.04).

However, end users have usually already a dozen tools locally to create exactly the plot that they want, so what features are specific to the visualization services provided in the portal of the infrastructure compared to the services already available to the user locally? One argument could be the speed: instead of having to download the data, it could be quicker to visualize it online. However, tools such as OPeNDAP include features to query and select individual quantities (or subsets thereof) via web protocols, making this argument largely invalid. Another argument could be the support for the file formats used, but for an infrastructure based on common standards this is unlikely: there should be plenty of post-processing tools that also support those common standards.

The convincing argument could be that the visualization is linked directly to the core infrastructure and can provide more information than just the visualization itself: it should assist in the visualization based on the contextual information in the storage e.g. add ensemble context to a plot of a single member (SC.VS.05), or provide links back from the plot to the underlying data and software (SC.VS.09).

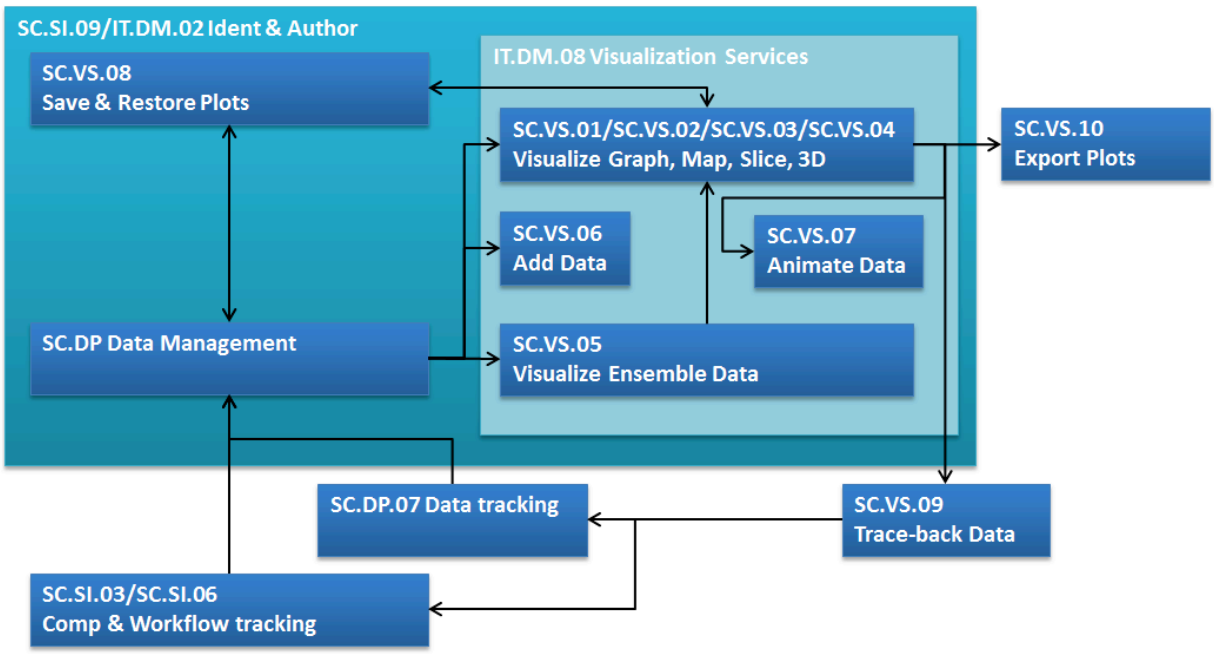

Figure 5: Interdependencies Data Visualization

This would, however, mean that the infrastructure should implement a full provenance framework to track data (SC.DP.07) and simulation components (SC.SI.03, SC.SI.06). CSIRO has implemented such a provenance framework[7] amongst others in the context of WIRADA (see [2]). This makes the trace-back functionality the core feature, and the visualization just one of the components to support it.

If you go down this route and offer a visualization service there will be a continuous stream of feature requests that depend on the earth sciences application considered. This will include requests to have the ability to add data to plot containing results of another quantity or other simulation run (SC.VS.06), to animate data (SC.VS.07), to export plots (SC.VS.10), and to adjust and configure plots (and subsequently save and restore settings; SC.VS.08). This results in a prioritized list of 5 activity clusters shown in Table 6.

Table 6: Prioritization for Data Visualization

[7] https://wiki.csiro.au/display/PROMS/The+Provenance+Management+System

Prio	Id	Topic
1	**DV1**	**Full Provenance Services**
	SC.VS.09	Trace-back Data from Plots
	SC.DP.07	Track and Reference Data
	SC.SI.03	Track and Reference Simulation Components
	SC.SI.06	Track and Reference Workflows and Compositions
2	**DV2**	**Basic Visualization Services**
	SC.VS.01	Visualize Data in Graph Plots
	SC.VS.02	Visualize Data in Map Plots
	SC.VS.03	Visualize Data in Slice Plots
	SC.VS.04	Visualize Data in 3D Plots
	IT.DM.08	Visualization Services
3	**DV3**	**Extended Visualization Services**
	SC.VS.07	Animate Data in Plots
	SC.VS.10	Export Plots for Reporting
4	**DV4**	**Advanced Visualization Services**
	SC.VS.06	Add Data to Plots
	SC.VS.05	Visualize Ensemble Data
5	**DV5**	**Visualization Management Services**
	SC.VS.08	Configure, Store, and Recreate Plots

6.4 General User Requirements

The general user requirements are mapped in Figure 6 to ICT and Training services. Simple access (SC.US.01) is required by all types of users; this is non-trivial to arrange across such a heterogeneous infrastructure with varying policies: a proper identification and authorization service is needed (IT.DM.02).

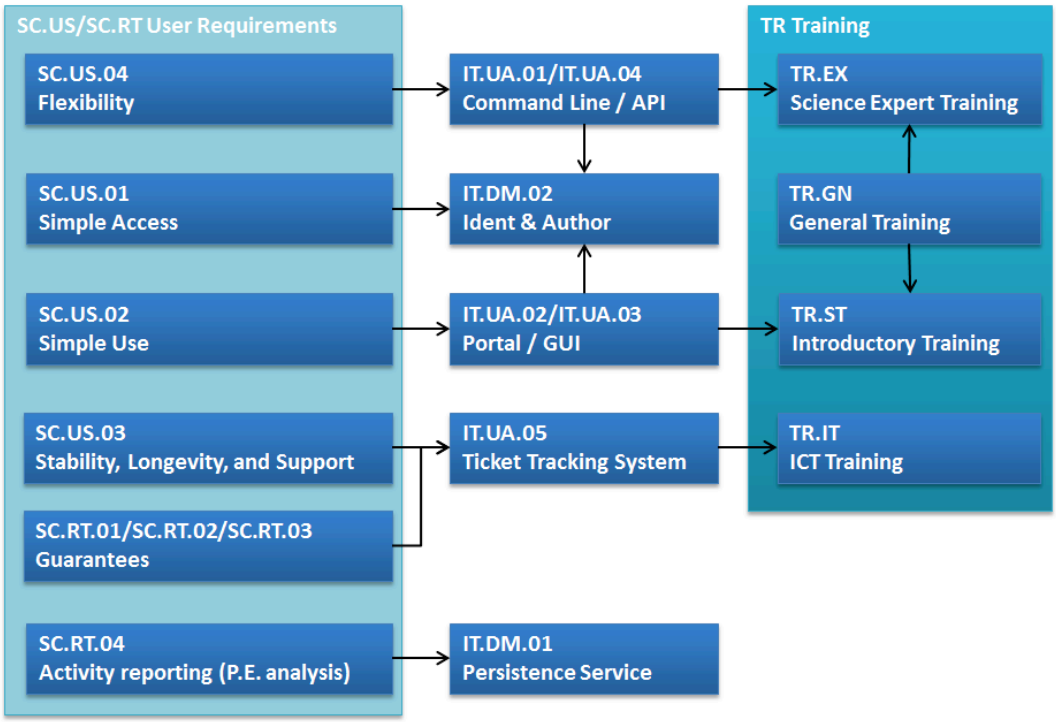

Figure 6: Mapping of User Requirements to ICT and Training Services

The portal (IT.UA.02) and embedded GUIs (IT.UA.03) should provide simplicity in use (SC.US.02) for beginning users, while command line (IT.UA.01) and API (IT.UA.04) should be available for experienced researchers that need flexibility (SC.US.04). Since these two groups use the infrastructure in a different way, they will need different training services (TR.ST and TR.EX, respectively).

The stability, longevity, and support requirements of the average researcher (SC.US.03) become more formal needs for guarantees (SC.RT.01—SC.RT.03) when operational forecasting centres depend on the service for their operations. These needs can only be met by a proper support organization, here represented by only the issue tracking system (IT.UA.05) and the training for the ICT staff maintaining the services. After an extreme event occurs, an operational centre frequently performs a post-event analysis of all the activities performed leading up to and during the event; the final user requirement is the desire for the infrastructure to assist in this type of analysis by means of a service that provides an overview of all "recent" activities. The foregoing leads to a prioritized list of 4 activity clusters shown in

Table 7.

Table 7: Prioritization for General User Requirements

Prio	Id	Topic
1	**GU1**	**Basic User Support Services**
	SC.US.01	Simple to Access (Single Sign On)
	IT.DM.02	Identification & Authentication Services
	SC.US.02	Simple to Use (for Starters)
	IT.UA.02	Portal
	IT.UA.03	Graphical User Interface (GUI)
	TR.ST	Training Science student or Citizen Scientist
	SC.US.04	Stability, Longevity, and Support
	IT.UA.05	Ticket Tracking System
	TR.IT	Training ICT Developer
2	**GU2**	**Advanced User Support Services**
	SC.US.03	Flexible to Use (for Advanced Use)
	IT.UA.01	Command Line Interface (CLI)
	IT.UA.04	Application Program Interface (API)
	TR.EX	Training Expert Scientist and Consultant

3	**GU3**	**Basic Operational Use Services**
	SC.RT.01	Get Guaranteed Availability
	SC.RT.02	Get Guaranteed Run Performance
	SC.RT.03	Get Guaranteed Data Up/Download Performance
4	**GU4**	**Advanced Operational Use Services**
	SC.RT.04	Report on Recent User Activity

6.5 Summary

In the previous four sections we have analysed the long list of 92 needs by consecutively looking at them from the perspective of one of the four themes "data management", "simulation management", "data visualization", and "generic user requirements". In the process we have derived from the initial unsorted list of needs, a total 26 activity clusters prioritized per theme. The overall prioritization grouped together in 7 services categories SC1-SC7 is given in Table 8 below.

Table 8: Overall Prioritization

Prio	Id	Topic
	SC1	**Basic Platform Services**
1	DM1	Data Storage Services
2	DM2	Basic Data Access Services
3	DM3	Simulation Support Services
4	SM1	Model Storage Services
5	SM2	Single Model Component Services
6	GU1	Basic User Support Services
	SC2	**Extended Platform Services**
7	SM3	Workflow Services
8	SM4	Coupled Component Services
9	SM5	Basic Ensemble Services

	SC3	**Documentation and Search Services**
10	DM4	Data Documentation and Search Services
11	SM6	Software Documentation and Search Services
12	DM5	Advanced Data Access Services
	SC4	**Provenance Services**
13	DM6	Data Provenance Services
14	SM7	Model Provenance Services
15	DV1	Full Provenance Services
16	SM8	Software Access Services
	SC5	**Visualization Services**
17	DV2	Basic Visualization Services
18	DV3	Extended Visualization Services
	SC6	**Advanced Platform Services**
19	SM9	Data Model Integration Services
20	DV4	Advanced Visualization Services
21	GU2	Advanced User Support Services
22	DM7	Data Processing Services
23	DV5	Visualization Management Services
24	SM10	Remote Coupling Services
	SC7	**Operational Use Services**
25	GU3	Basic Operational Use Services
26	GU4	Advanced Operational Use Services

7 Towards a Development Action Plan

In the previous chapters we have analysed the needs of the three pillars of the platform and associated support organization "science", "technology", and "training" and condensed them into a prioritized list of 26 activity clusters. The aim of this chapter is to map this list onto a development action plan with realistic and sustainable goals. This plan aligns with the Organization Implementation Plan [7] and the Future Integration Plan [10].

As the starting point we take the prototype environment as developed and provided by the DRIHM project. In the development of this infrastructure, significant effort was spent on chaining workflows with or without feedback loops across heterogeneous computing resources. This environment already addresses a large number of the activity clusters listed in Table 8 including but not limited to Basic User Support Services (GU1), Workflow Services (SM3), and Basic Visualization Services (DV2) and was shown to be interoperable with XSEDE resources in the US [9].

7.1 Short-Term Plan

The core objectives for the short-term (1-2 years) are to further strengthen the basis of the infrastructure by focussing on the Basic and Extended Platform Services and Documentation and Search Services; this includes work on topics such as single sign-on authentication, policy-based authorization of accessing restricted data, standardized data exchange and metadata catalogue. These services are important for building a larger active user community although still focused on the initial hydro-meteorological research community studying floods due to excessive amounts of precipitation. The initial support organization described in Chapter 5 of [7] is aimed at this short-term. The main short-term actions including access to resources are listed in Table 9.

Table 9: Short-Term Actions

Item	Action

1	Finalize negotiations with EGI on computing resources for initial community and demonstration purposes.
2	Migrate data storage from temporary solution to EGI resources.
3	Reorganize website from project focus towards infrastructure focus.
4	Create training material based on the current status of the infrastructure for growing the community.
5	Work with EGI, PRACE, and XSEDE to improve accessibility.

7.2 Mid-Term Plan

In the mid-term (2-3 years) our objectives shift to Provenance Services, Extended Visualization Services, and Advanced Platform Services; this includes work on topics such as expanding the infrastructure with more advanced features such as data assimilation, support for large data amounts, and additional models. During this phase we also aim to significantly broaden the scope of the infrastructure by including improved support for two-way coupling and more simulation components from the wider earth sciences. An initial list of actions focused at the mid-term objectives is given in Table 10.

Table 10: Short-Term Actions for Mid-Term Objectives

Item	Action
1	Extend platform with more features focused on high-weather events – ensembles, data assimilation. DRHIWE proposal submitted to Horizon2020 call expanding use cases to longer time scales, coastal storm risk, and droughts.
2	Improve the use of HPC resources by the broader earth sciences in collaboration with centres of excellence (DRIHM2US partner involved in EnCompAS proposal submitted to Horizon2020 call).
3	Address the needs of two-way coupling in the broader earth sciences. Collaborate with CSDMS, NSF, and European Commission to identify appropriate call.

7.3 Long-Term Plan

The longer term (4+ years) objective is to be able to serve besides the research community also operational forecasting centres by deploying Operational Use Services. Other on-going long-term actions are the collaboration with various communities on improving the already implemented services based on new technologies and new standards. A list of short-term actions aimed at enabling the long-term objectives is given in Table 11.

Table 11: Short-Term Actions for Long-Term Objectives

Item	Action
1	Collaboration with standards communities such as RDA, OGC, and EGCF to develop new and improved standards for data and workflow management on heterogeneous resources
2	Collaborate with international science communities such as EarthCube and CSDMS to provide unified international services.
3	Working with operational centres to draft a more specific list of service level requirements for the operational use of the infrastructure.
4	Investigate the possibilities of establishing the extended platform as a research infrastructure to be included in the ESFRI plans

8 Conclusion

Building on the work done in work package 2 and 5, this report addresses the infrastructure needs from the perspectives of science (scientific operations), technology (ICT operations), and training (education centre) – the three core elements of both the platform and the associated organization that we aims for. Based on the generic reference framework, the scientific innovation cycle, and a categorization of the audience of the educational centre we have collected a list of in total 92 needs: 44 for the science, 31 for the technology, and 17 for the training.

That concluded the first stage version of this report. In this extended second stage version, we continued by assessing the interdependencies and priorities amongst those 92 needs. To simplify the analysis we subdivided the needs into four themes "data management", "simulation management", "data visualization", and "general user requirements". Per theme, we briefly summarized the functionality and interdependencies, and condensed the long list of needs into a prioritized set of 7 service categories composed of 26 activity clusters. The service categories are listed in Table 12 with the associated prioritization focus in the development action plan formulated in Chapter 7.

Table 12: Service Categories

Id	Service Category	Priority for
SC1	Basic Platform Services	Short-Term
SC2	Extended Platform Services	Short-Term
SC3	Documentation and Search Services	Short-Term
SC4	Provenance Services	Mid-Term
SC5	Visualization Services	Mid-Term
SC6	Advanced Platform Services	Mid-Term
SC7	Operational Use Services	Long-Term

9 Acronyms and References

9.1 Acronyms and Abbreviations

Acronym / Abbreviation	Definition
API	Application Programming Interface
CSDMS	Community Surface Dynamics Modeling System
CSIRO	The Commonwealth Scientific and Industrial Research Organisation
CUAHSI	Consortium for the Advancement of the Hydrological Sciences, Inc.
DRIHM	Distributed Research Infrastructure for Hydro-Meteorology
DRIHM2US	Distributed Research Infrastructure for Hydro-Meteorology to United State of America
EGCF	European Globus Community Forum
EGI	European Grid Infrastructure
ESFRI	European Strategy Forum on Research Infrastructures
GEOSS	Global Earth Observation System of Systems
GUI	Graphical User Interface
HIS	CUAHSI Hydrologic Information System
HMR	Hydro-Meteorological Research
HPC	High Performance Computing
ICT	Information and Communications Technology
NGI	National Grid Initiative
NSF	US National Science Foundation
OGC	Open Geospatial Consortium
OPeNDAP	Open-source Project for a Network Data Access Protocol
PRACE	Partnership for Advanced Computing in Europe
RDA	Research Data Alliance
SCIHM	Standards-based Cyberinfrastructure for HydroMeteorology

XSEDE	Extreme Science and Engineering Discovery Environment
WCS	OGC Web Coverage Service

9.2 References

[1] Parodi, A. et al: DRIHM2US Description of Work (DoW), 2012.

[2] Schiffers, M. et al: DRIHM2US Report on an Assessment of Current e-Infrastructure Approaches for Hydro-Meteo Research in Europe and the US (D2.1), 2013.

[3] Schiffers, M. et al: DRIHM2US Report on a Common Architecture Model (D2.2), 2013.

[4] Harpham, Q. et al: DRIHM2US Domain Expert Networking Report (D3.3), 2013.

[5] Harpham, Q. et al: Outline of organisational structure (D5.3), 2014.

[6] Clematis, A. et al: Opportunity and gap analysis (D2.3), 2014.

[7] Jagers, B. et al: Detailed organizational implementation plan (D5.4), 2015.

[8] Jagers, H.R.A.: 'Linking Data, Models and Tools: An Overview' in the Proceedings of the iEMSs 2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada, David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)

[9] Harpham, Q. et al: DRIHM2US Interoperability Experiment Report (D3.2), 2015.

[10] Schiffers, M. et al: DRIHM2US Future integration report (D2.4), 2015.