hydralab+ Adaptation for Climate Change

WP10 JRA3
Facilitating the re-use and exchange of experimental data

Task 10.1 Critical review

D10.2 Critical Review of data flux between laboratory models, numerical models and field case studies

Status: Public document from 21 February 2018

Version: 3

Date: 21 February 2018

EC contract no 654110, HYDRALAB+

DOCUMENT INFORMATION

Title	Critical review of data flux between laboratory models, numerical models and field case studies
Lead Authors	Paul Cleverley, Lesley Mansfield, James Sutherland, Quillon Harpham (HR Wallingford)
Contributors	DELTARES, CNRS, LNEC, LUH, Samui, UHULL, UPORTO
Distribution	Public from 21 February 2018
Document Reference	DOI: 10.5281/zenodo.1182553

DOCUMENT HISTORY

Date	Revision	Prepared by	Organisation	Approved by	Status
24 July '17	1.0	Cleverley, Mansfield, Sutherland	HR Wallingford		Project only
30 Jan '18	2.0	Harpham, Sutherland	HR Wallingford		Project only
21 Feb '18	3.0	Harpham, Sutherland	HR Wallingford	Hamer – HYDRALAB+ Coordinator -	Public

ACKNOWLEDGEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654110, HYDRALAB+.

DISCLAIMER

LICENSE

CITATION

Suggested citation:

Cleverley, P., Mansfield, L., Sutherland, J. and Harpham, Q., 2018. Critical review of data flux between laboratory models, numerical models and field case studies. HYDRALAB+ deliverable D10.2. http://dx.doi.org/10.5281/zenodo.1182553

EXECUTIVE SUMMARY

The HYDRALAB+ project is aimed at strengthening the coherence of experimental hydraulic and hydrodynamic research undertaken across its partner organisations. This report is D10.2 of the HYDRALAB+ project, entitled "Critical Review". It is one of the outputs of Work Package 10 – JRA3: Facilitating the Re-use and Exchange of Experimental Data. It examines the state of the art in a number of areas relating to HYDRALAB+ with particular attention given to:

- data standards and protocols currently in use in the HYDRALAB+ community;
- the flow of data between the three communities (laboratory modelling, numerical modelling and field case study); and
- the effectiveness of the mechanisms in use for validation and verification of data.

A questionnaire was used to gather information about data management from project scientists. This revealed a lack of knowledge about data standards, protocols and other data management topics within the HYDRALAB+ community. This situation can be improved by education about the nature of data management and how important it is to the science itself but at the same time by raising the profile of data management as a profession in science. Differences in terms of data standards, formats, protocols and tools can be expected at cultural boundaries.

Many different data formats are used within the HYDRALAB community, so it would be impossible to recommend a set of formats to adopt. Rather, it is on the interfaces and translations between formats that data management should focus. To achieve greater openness in terms of data sharing researchers should concentrate on the structure and versioning of their data, avoiding any prescriptive licensing of third party software.

It's important that HYDRALAB recognizes the – likely exponential – increase in the storage, search and retrieval requirements for data.

For physical experiments the main effort reported for ensuring valid data was expended in calibrating the observation equipment.

Web crawlers, scrapers and indexers require standardized metadata to allow data to be discovered. Some communities have standardized some metadata, but a plethora of metadata standards also makes data difficult to find. There are many metadata standards and, as a result, a non-expert may find difficulty identifying which metadata standards to adopt for their specific data management requirements. Existing standards can be extended where needed to incorporate new ontologies if required. Detailed metadata standards like INSPIRE can serve as the base for this. However, the user experience of INSPIRE has recently been described as being abysmal.

Unless some clear and overwhelming evidence is forthcoming to indicate a pressing need for a new hydrological and hydraulic vocabulary and ontology we recommend the wider dissemination of existing structures (such as the IAHR list of sea state parameters). This is not to argue against vocabularies or ontologies or any standards. Rather, we suggest that researchers learn to use existing standards where possible, extend existing standards where necessary and only invent new standards when absolutely necessary. This principle of parsimony would make the use of data in hydrological and hydraulic research less intimidating for the non-expert in data management. The use of standards should be documented carefully in the Data Storage Report.

Much valuable information about experiments is in lab books, planning documents and analysis code. Access to this data would be beneficial to future researchers.

A number of recommendations for improving data management in the HYDRALAB community have been made:

- We propose that HYDRALAB+ researchers make use of suitable data repositories and ensure that each dataset is allocated a DOI. In a wish to provide a service without being prescriptive, we aim to provide an interface on the HYDRALAB+ website for researchers to the Zenodo Horizon 2020 data repository using Zenodo's Application Programming Interface.
- We should provide training on Data Management to HYDRALAB+ participants and provide links to existing online training resources.
- We recommend that the existing Data Storage Reports adopted by the project be reviewed to incorporate any appropriate aspects from the Data Management Plans which follow the H2020 template and also through the associated DMPOnline facility for the generation of data management plans. Since HYDRALAB+ has an established usage of Data Storage Reports, the DMPOnline facility is not planned to be used directly at this stage. We recommend that data papers (which can be based on the existing HYDRALAB Transnational Access Data Storage Reports) be written and published in data journals (with links to the associated data packages).
- We support the development of metrics that include citations to data and data papers as this would assist in changing the culture of science to recognise the importance of open data.

CONTENTS

Tables

Figures

1 INTRODUCTION

The purpose of this critical review is to examine the state of the art in the exchange of data between and among three distinct communities of scientific research. It has built on the previous work in this area conducted by the HYDRALAB group and has given particular attention to the *data standards* involved and the *effectiveness* of the *verification* and *validation* of data exchanged between communities. It then considers potential routes towards improvements in both data standards and data exchange.

1.1 WORK PACKAGE DESCRIPTION

A mix and combination of laboratory modelling, numerical modelling and field case study is required for the effective advancement of environmental hydraulics (van Os et al, 2004, Gerritsen et al., 2011).

Each of the domains is methodologically strong, with considerable advances in quantification possible in both the field and the laboratory and the significant growth over recent decades in power and effectiveness of numerical simulation.

However, the links between the three methodological approaches are weak. This is important as the links between laboratory, field and numerical model are critical for substantive advancement in our understanding of complex systems and thus interdisciplinary-based prediction.

For example, consider a fictional scientist (A) in Addis Ababa wishing to communicate with an equally fictional scientist (B) in Ulan Bator. The two wish to discuss the values of a (fictional) variable they both commonly refer to as "H_{m0}". The variable "H_{m0}" is the result of a (fictional) algorithm applied to a measurement – it involves other parameters and arithmetic and logic and so on. How do these scientists know they are talking about the same thing? Typically the answer is "It's just common knowledge."

Let's now assume that scientist A has invented a slightly modified algorithm for producing "H_{m0}" and failed to inform scientist B that this is what she used when producing the values for "H_{m0}" and one can see how confusion about such "common knowledge" can arise. Ontologies and their semantics change over time because they are, at heart, labels given to algorithms. If everyone understands and agrees the algorithm – all well and good. But algorithms develop.

In the real world, there can still be some confusion within the community between:

- H_s = significant wave height
- $H_{1/3}$ = average of the highest 1/3rd of wave heights
- H_{m0} = spectral significant wave height, defined as $4 \times \mathrm{sqrt}(m_0)$ with m_0 the zeroth order spectral moment.
- H_σ = estimate of significant wave height = $4.\sigma_\eta$ where σ_η is the standard deviation of the surface elevation.

The IAHR list of sea state parameters (1989, table 2) describes H_s as "significant wave height defined as the average of the highest one-third of the wave heights or it can be estimated as H_{m0} (recommended) or H_σ." In deep water, $H_{1/3}$ H_{m0} and H_σ will have very similar values, but in shallow water these values will diverge. It is important therefore to use $H_{1/3}$ H_{m0} or H_σ rather than Hs to avoid ambiguity.

This critical review builds on previous work by the HYDRALAB group on data standards and sharing and examines the state of the art in the flow of data between laboratory, field and numerical simulations. The review examines the effectiveness of validation and verification processes that drive comparisons and confidence in predictions. It highlights how these processes can be developed and improved. Protocols, standards and techniques will then be developed later in the project with the aim of improving the effectiveness of data flow in the future.

1.2 THE DOMAINS

The three domains with which HYDRALAB+ is concerned are:

- *field* - refers to research conducted in the real world in an uncontrolled environment
- *laboratory* - refers to research conducted in the real world in a controlled environment
- *virtual* - refers to research conducted in a virtual world (computer simulation, numerical model, etc.) in a controlled environment

1.3 THE ATTRIBUTES

HYDRALAB+ is concerned with the assessment of three major attributes relating to the flow of data between the communities in question:

- *validation* - "The assurance that a product, service, or system meets the needs of the [...] stakeholders. It often involves acceptance and suitability with external customers. Contrast with *verification*." *[Project Management Body of Knowledge]*
- *verification* - "The evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition. It is often an internal process. Contrast with *validation*." *[Project Management Body of Knowledge]*
- *effectiveness* - "In general, efficiency is a measurable concept, quantitatively determined by the ratio of useful output to total input . Effectiveness is the simpler concept of being able to achieve a desired result, which can be expressed quantitatively but doesn't usually require more complicated mathematics than addition." *[Wikipedia]*

1.4 THE DATA

The information about the data flow between communities with which HYDRALAB+ is concerned includes:

- *data standard* - refers to an agreed prescription for the *semantic* structure of the content of a data package; in other words the data standard refers to how each element of the data is structured and what each part of the structure actually means;
- *data format* - refers to an agreed prescription for the *physical* structure of the data package;

- *data protocol* – refers to an agreed prescription for the mechanism for exchanging data packages between systems;
- *data tool* - is a software application used for processing one or more specific types data transforming data from one format, protocol or standard to another.

2 EXISTING GUIDELINES

The efficient exchange of valid and verifiable data between research communities is key to the development of a mature and progressive scientific research environment (as humorously illustrated by NYU Health Sciences Library, 2012 animation[1]).

A set of guidelines for data archiving exists within the OpenAIRE [2] initiative. These provide instruction for data archive managers to expose their metadata in a way that is compatible with the OpenAIRE infrastructure.

Furthermore, the H2020 Programme Guidelines on FAIR Data Management in Horizon 2020[3] play a significant role in HYDRALAB in all its incarnations. This states that:

> *Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and re-usable (FAIR), a DMP should include information on:*
> *- the handling of research data during and after the end of the project*
> *- what data will be collected, processed and/or generated*
> *- which methodology and standards will be applied*
> *- whether data will be shared/made open access and*
> *- how data will be curated and preserved (including after the end of the project).*

A template for a H2020 Data Management Plan is provided, which has been converted into an on-line form by the Digital Curation Centre. In addition, their website interface for data management plans provides useful information on the suggested content of DMPs. The Digital Curation Centre eloquently states the lesson to be learned:

> *For research teams to enjoy the full benefit of the research data that is produced, institutions must put in place skilled digital curators and effective curation lifecycle management. This will help to ensure that important digital research data is adequately safeguarded for future use.*[4]

Validation of existing metadata can be achieved through the implementation of or use of the third party schema validations such as provided through the Schematron validation language.[5] This language allows a non-procedural specification of rules for validating data against a schema. It is particularly useful for validating XML data (or metadata) against a standard schema (such as GML,

[1] https://www.youtube.com/watch?v=dl6C_GrZrbE
[2] https://guidelines.openaire.eu/en/latest/data/index.html
[3] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[4] http://www.dcc.ac.uk/digital-curation
[5] https://en.wikipedia.org/wiki/Schematron

WaterML, etc.). Another example of the development of automatic validation systems is described in “GML Validation Based On Norwegian Standard” [6]

[6] http://hdl.handle.net/11250/144130

3 PREVIOUS HYDRALAB WORK

There have been a number of previous incarnations of the HYDRALAB network. HYDRALAB+ is the latest. This section presents some of the previous work relating to data storage, discovery and retrieval.

3.1 DATA MANAGEMENT TOOLS FOR HYDRALAB – A REVIEW

As part of HYDRALAB-III, Wells et al (2009) reviewed some of the many technologies, methodologies and standards that might be adopted to address the data management issues of the HYDRALAB consortium, or other physical modelling laboratories in the field of environmental hydraulics. The summary of this report is given below:

> *The members of the HYDRALAB consortium have significant investment in their existing project management structures and their science and engineering methodologies. Any HYDRALAB standards or recommendations should allow the organisations to continue to make use of their existing working practices, technologies and facilities without requiring internal conformance to a single rigid standard. This makes the importance of establishing common interface and data exchange criteria key to improving collaboration between HYDRALAB partners. It does not matter what tools or software a partner uses internally as long as other partners know that when data or metadata are exchanged they will be exchanged in a predefined format.*
>
> *Using standard formats and technologies would encourage and facilitate better communications with the wider community outside the HYDRALAB consortium and ensure that the HYDRALAB consortium contributes to the wider aims of the EU's programs such as INSPIRE and GMES. The move towards adoption of such standards benefits not only from the impetus of the HYDRALAB project itself, but also from other external drivers towards greater collaboration between organisations making more likely progress on such issues.*
>
> *The following suggestions are proposed:*
>
> 1. *The HYDRALAB participants should seek to make maximum use of existing information management technologies, methodologies and models where ever possible.*
> 2. *Develop a high level strategic view for the future of integration within HYDRALAB and identify significant, but small and achievable, steps to move the project forward towards those strategic goals. Knowing the goals, partners can align changes within their organisations with those goals as opportunities occur, alongside specific consortium wide initiatives to make progress on specific topics.*

3. *Establish a body of "best practice" with regard to the documentation and management of project data. If the requirements of Section 5.1 of the Specifications for Data Management report (v1.1, March 2007) can be met and suitable discovery metadata are available (via a common data model such as CERIF) this would represent a significant advance and leave partners free to adopt whatever internal structures best suit their practices and culture.*

4. *Adopt the EC recommended CERIF data model as the working basis for metadata and project data interchange within the HYDRALAB consortium. This will provide a well- documented basis for each organisation to exchange data with its partners in HYDRALAB and beyond. The EU INSPIRE and GMES directives will also provide impetus to data harmonisation.*

5. *Adopt a standards based approach to data and metadata management. There are many methodologies available to fulfil the data management needs of the HYDRALAB consortium. However standards, such as those from the OGC, rely on the input of experts from all over the world to develop a consensus and are well documented. Adopting standards for data and metadata has the additional benefit of opening up all the other standards based services that are available with little additional effort.*

6. *Make the integration and sharing of data and metadata as "cheap" as possible (in terms of time and effort) for those generating the data and metadata. Any implementation should strive to avoid placing additional administrative burdens on those generating and working with data and metadata.*

7. *Identify any "Master Data" sets within HYDRALAB and establish methods for the control and management of them. It is likely that this would mean the establishment of a central HYDRALAB system to store the "master data" that can be automatically populated or accessed by the partners' systems as required.*

8. *Investigate a common data format (or limited number of formats) for data exchange and which HYDRALAB partners might consider adopting for internal data storage over time. Partners can then use any existing internal formats if desired, knowing that data from HYDRALAB partners will always be available in a single (or few) defined formats.*

9. *Learn from the work of others – in particular study the working NEESGrid example from the US science community. This may offer the HYDRALAB consortium the ability to create a working collaborative system with significantly less effort than building such a system from scratch, in particular if access could be negotiated to the software developed within the NEES project. There is no need to adopt such a system wholesale, and*

indeed it is likely that if the NEES system builders themselves were to begin again they would not make all the same choices – HYDRALAB can take advantage of such hard won knowledge and experience. Collaboration with the other organisations such as NEES on infrastructure technologies may also help to foster closer scientific collaborations.

3.2 Composite Modelling

In HYDRALAB-III the Joint Research Activity 'Composite modelling of the interactions between beaches and structures' (CoMIBBS) and a subsequent IAHR Working Group on Composite Modelling looked at different ways of efficiently combining physical and numerical models (Gerritsen et al, 2009, 2011, Gerritsen and Sutherland, 2011, Sutherland and Barfuss 2011). Composite modelling techniques considered included

- Traditional model nesting where a physical model is a detailed representation of a system, which is modeled at a larger scale (and at a more general level) in a numerical model;
- Numerical modeling can assist in the design of physical models by helping to set the location and type of boundary conditions that are to be applied. Numerical pre-modeling also provides information about potential problems associated with the design, thereby reducing the number of physical modeling configurations necessary during the physical modeling portion of the study.
- Physical model representation of one element of a system, with the results being parameterized for use in a numerical model.
- Modelling the model can allow a numerical model to be calibrated or corrected using the physical model results. The calibrated or corrected numerical model is then available to undertake additional model runs that would be too time consuming in a physical model or were only considered after the physical model has been decommissioned.

Of interest here is the recognition that data exchanges need to be specified in detail for these techniques to work well. However, at this stage, no attempt was made to specify data formats, data standards or data protocols for this data exchange.

3.3 Remote Access to Data and Experiments

In HYDRALAB IV work was undertaken as part of Task 10.3 Remote Access to Data and Experiments under Task 10.1 'Organisation of a central data store of experiments'. This task developed the data structures and procedures for the sharing of information between research installations and research groups. The EC INSPIRE Methodology for Data Product Specification was used as the framework for the data model developed. As such, we depended upon the completion of the specifications and the development of an implementation of the specifications. In particular, we have worked with the United Kingdom Environmental Observation Framework (UK-EOF) community (http://www.ukeof.org.uk/), which has been implementing data services based around Inspire data standards for Environmental Monitoring Facilities (EMF).

The objective of this work was to widen access to data on HYDRALAB facilities and experiments, so as to ensure that data that exists about HYDRALAB facilities is not unique to the HYDRALAB

community, but is made more widely available and accessible to other communities who may have an interest. To achieve this we moved away from the situation where HYDRALAB data was only available for publication via the HYDRALAB webpage, to a situation where HYDRALAB data could be available to publish in a range of on-line services, one of which is the HYDRALAB website. The emphasis has been on Transnational Access (TA) projects, as data from these projects becomes freely available to outside parties two years after the experiments finish. However, in order for the re-use of TA data to be maximized, people outside HYDRALAB have to be able to find this data.

Deliverable D10.3 (Millard, Cleverley & Sutherland, 2013) reported on initial testing of the software that accesses data from the UK-EOF catalogue through its application programme interface (api). Deliverable 10.6 (Sutherland, Millard and Cleverley, 2014) subsequently undertook a full mapping of the data in the HYDRALAB database (that is used to generate content for the HYDRALAB website and for reporting to the EC) to the UK-EOF schema and then effected a bulk transfer of the HYDRALAB data into the UK-EOF database.

The development and testing of the UK-EOF HYDRALAB catalogue represented a significant step forward for the HYDRALAB community in the development of modern data models and the provision of modern data services. This benefited the Hydralab community by separating the website from data services, delegating services to experts in this field, and making information more accessible.

The success was only partial, largely because of delays in finalising the Inspire Environmental Monitoring Facilities (EMF) specification and the resulting UK-EOF Schema. As an early user of the draft schema, HYDRALAB contributed to the development of the published schema.

Greater benefits would be achieved through a full mapping of HYDRALAB activities, facilities, programmes and networks onto the corresponding UK-EOF elements, and examples of how this could be approached were provided.

3.4 WEB PORTAL

The existing HYDRALAB web portal[7] will be modified to incorporate changes relevant to the need to facilitate data exchange between domains. Alongside the HYDRALAB database, there are plans to implement a user interface to the Zenodo data repository[8].

3.5 HOW IS DATA EXCHANGED NOW?

Currently the HYDRALAB community has been using the HYDRALAB website for sharing data and metadata from previous incarnations. In addition, the UKEOF[9] Catalogue was used in the HYDRALAB IV incarnation for storing and retrieving metadata relating to HYDRALAB activities, facilities and programmes.

The search page at http://hydralab.eu/research--results/ta-projects/ will give information relating to HYDRALAB Transnational Access projects, including information on how to access the results. One

[7] http://hydralab.eu
[8] Zenodo is a research data repository created by OpenAIRE and CERN in 2013.
[9] UK Environmental Observation Framework - http://www.ukeof.org.uk/

example of the working exchange of data from HYDRALAB-III is the Barrier Dynamics Experiment II (BARDEX II) located here:

http://hydralab.eu//research--results/ta-projects/project/11/

as shown in Figure 1. The link provides a data management report and the instruction to contact the experiment provider to request access to the data.

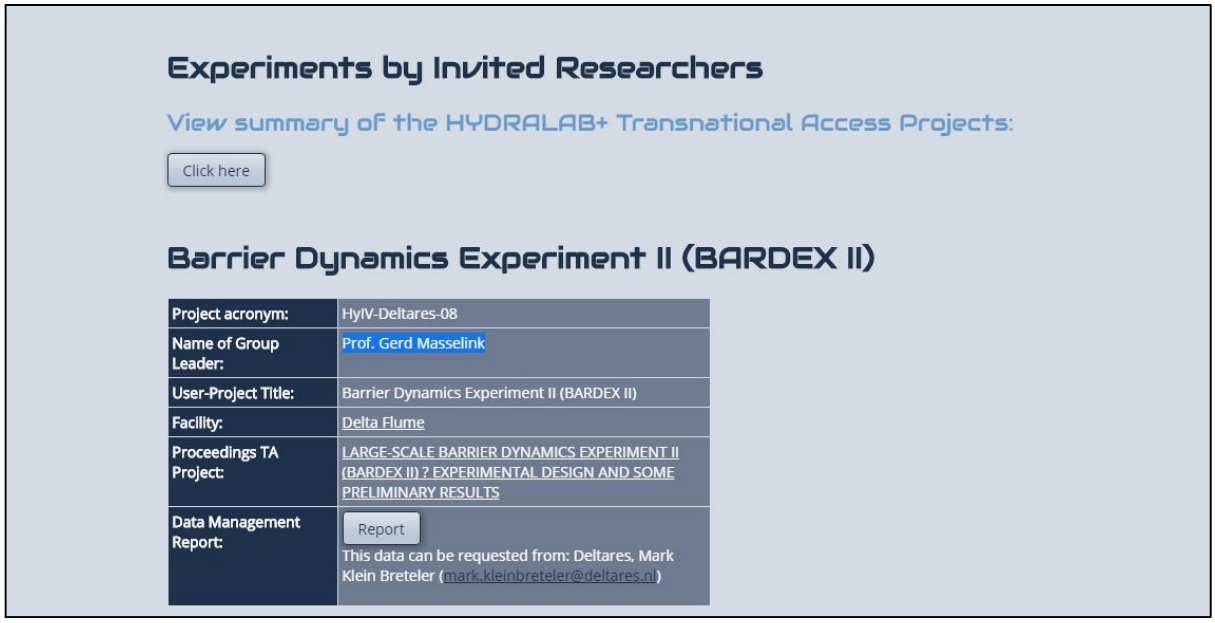

Figure 1 BARDEX II catalogue entry from HYDRALAB IV

This is relatively time consuming and requires a number of levels of redirection to get access to the data used in the experiment. While there are adequate metadata in the HYDRALAB database, the actual data itself are unavailable without the - potentially drawn out - process of contacting the provider, requesting the data, agreeing licence terms, determining the cost of translating the data if needs be from one form to another, negotiating an agreement and mechanism for transmitting, delivering and then ingesting the data.

4 THE STATE OF THE ART

This sections reviews the standards, protocols, formats and tools currently in use and available to the HYDRALAB community. It also looks at the flow of data between the domains (laboratory, field and numerical model) in the community and the effectiveness of the validation and verification of data when it is translated between domains. It goes on to highlight problems and issues which need addressing.

We conducted an online survey using Survey Monkey (questions and answers are listed in section 7 Questionnaire. The aim was to canvas the Hydralab+ community for information regarding issues with exchanging data between different domains.

4.1 DATA STANDARDS, PROTOCOLS AND TOOLS IN USE

This section examines the common standards, formats and protocols in use in the HYDRALAB+ community.

4.1.1 Standards

The list of data standards in common use in the HYDRALAB+ community includes but is not limited to those listed in Table 1.

Standard	Description
WaterML 2.0	WaterML 2.0 is a standard information model for the representation of water observations data
GML	GML serves as a modeling language for geographic systems as well as an open interchange format for geographic transactions on the Internet
Other OGC	e.g SensorML The primary focus of the Sensor Model Language (SensorML) is to provide a robust and semantically-tied means of defining processes and processing components associated with the measurement and post-measurement transformation of observations.
INSPIRE (EMF)	INSPIRE Environmental Monitoring Facilities

Table 1 List of standards in common use in HYDRALAB+ community

It is worth noting that only one of the survey respondents appeared to fully understand the question relating to data standards, i.e. "*To what data standards do the data packages adhere (what is it) ? e.g. OGC standards, GML, WaterML 2.0, etc*".

There is an implication that the question was poorly worded or the majority of HYDRALAB+ respondents are insufficiently trained in data standards and knowledge to help improve the FAIR compliance of HYDRALAB data or both.

The respondents themselves may not be the appropriate people to answer the question; in addition, the team conducting the experiments and activities may not have access to sufficient resources to answer the question completely.

Suffice to say that most of the people tasked with the job of answering a question about data standards in use in their experiments were unable to do so.

4.1.2 Formats

The list of data formats in common use in the HYDRLAB+ community includes but is not limited to those listed in Table 2.

Type	Format
Binary	netCDF[10], .MAT
Text	CSV, TSV, TXT
Image	JPG, PNG, TIFF
Video	Various unspecified formats
Audio	No audio formats were specified
Proprietary	XLS, XLSX (Excel), DOC, DOCX (Word), PPTX (Powerpoint), Vectrino, PIV

Table 2 List of data formats in common use in HYDRALAB+ community

There is a wealth of data formats available from specific proprietary formats with a complex internal structure to a very basic de facto standard ASCII text comma separated values (.CSV).

One significant aspect of any data format is the facility with which it can be read from and written to. Proprietary formats are often efficient because they are designed to couple closely with the proprietary code that reads and writes them. Common formats like .CSV are perhaps less efficient due to the fact they have to be able to be read from and written to in an open and less prescriptive and complex manner.

From the point of view of FREE Data, it would be impossible to recommend a single format – or even a set of formats – which the HYDRALAB community should adopt given the experimental nature of the community activities. It is on the interfaces and translations between formats that data management should focus.

4.1.3 Protocols

The list of data protocols in common use in the HYDRALAB+ community includes but is not limited to those listed in Table 3.

[10] While netCDF is also an OGC data standard it is also a binary data format.

Protocol	Description
HTTP	HyperText Transfer Protocol
FTP	File Transfer Protocol
WCS	Web Coverage Service[11]
WMS	Web Map Service
Proprietary	box.com, onedrive.com for file exchange (strictly speaking these are tools)
USB	Universal Serial Bus – hardware communication protocol

Table 3 List of data protocols in common use in HYDRALAB+ community

4.1.4 Tools

The list of software tools in common use in the HYDRALAB+ community is extensive and covers a wide range of open source and proprietary third party and in-house developed software. We do not intend to enumerate all the tools currently in use as these change and develop constantly. We can, however, classify these in terms of a modified model-view-controller design pattern[12].

This design assigns the responsibility for handling the human computer interface to a system known as the "view"; for handling the data management to a system known as the "model" and the overall management of the system to the "controller", with these roles being described in Table 4.

Function	Type examples
MODEL	Source code control systems such as SVN and GitHub, relational databases such as POSTGRES and ORACLE, RDF[13] triple stores, netCDF data files, etc. (NB this is not the same thing as a numerical model – "model" here means an abstract or concrete representation of a data structure)
VIEW	User interfaces to data – including web sites, visualization tools, proprietary data entry and visualization tools.
CONTROLLER	Software and hardware which connects the MODEL and the VIEW, for example, numerical models, real-time data capture and monitoring and other software which functionally transforms data from one MODEL to another

Table 4 Model – view – controller design pattern

Some software tools can fall into the VIEW and CONTROLLER categories. Some software - such as ArcGIS – can function across these multiple domains and act as all three.

[11] While WCS and WMS also represent data standards but here represent the protocol extensions such a standard requires
[12] https://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller
[13] https://www.w3.org/RDF/

It's worth noting also that some software is very tightly coupled across two or three of these functions. netCDF, for example, contains a data model and a controller (a library of tools for reading and writing the data); ArcGIS has a proprietary model, a view and a controller; MATLAB is also coupled across all these functions as is MS-Excel.

To achieve greater openness in terms of data sharing, options include: adopting a tightly coupled (and often proprietary) solution for all data processing and data exchange (such as MATLAB); or abandoning the cost and effort required to ensure consistent software versions across the domains and concentrate instead on the structure and versioning of the data, avoiding any prescriptive licensing of third party software.

However implemented and whatever labels may be used, the functionality of humans interacting with data requires tools and applications which perform the functions of a Model-View-Controller template.

4.2 The Flow Of Data Between Communities

Responses to the survey question *"Have you experienced any issues or problems associated with the transfer or exchange of this data package?"* include a number of negative responses. These indicate in some cases that no problems were encountered in transferring data and in others that research activities had not yet produced any data.

One respondent said that data sharing was not a problem between researchers in the same institution because reporting and metadata were sufficient. This highlights a cultural aspect to the sharing and flow of data between communities. Where a cultural boundary exists one may expect differences in terms of data standards, formats, protocols and tools. For the avoidance of doubt "cultural" does not mean solely "national" but also includes corporate and technical cultures. It is these interfaces where the focus of attention would pay dividends.

Two other respondents both reported difficulties with the volume of data in particular high resolution global satellite data and particle image velocity data. Further research into the capacity planning for sharing such research would pay dividends. It is unlikely that the quantity of data produced in scientific research will decrease in volume – quite the opposite. It's important that HYDRALAB recognizes the – likely exponential – increase in the storage, search and retrieval requirements in these areas.

4.3 The Effectiveness Of Validation and Verification

This section examines the validation and verification tools and techniques currently in use by the HYDRALAB+ community to validate data resulting from experiments and assesses their effectiveness.

The responses to the survey question on effectiveness included assessments of *"low"* with one respondent commenting that this would have been improved if the user had *"been involved from the beginning"*; observations that effectiveness is difficult to judge in experiments where non-standard measurement developments are being carried out as there is nothing to compare to. Others pointed out that no data had been collected yet so they could not assess its effectiveness.

An interesting observation was that the effectiveness can be determined by the number of publications which reference the data – hence it is too soon in the project to determine.

4.3.1 Validation

This section examines the validation tools and techniques currently in use by the HYDRALAB+ community to verify data resulting from experiments.

One experiment in the Laboratory domain compared its results data with data from Field observations in order to validate.

Two Laboratory experiments referred to the precise calibration of the observing equipment as being the basis of the validation of their data.

One requested feedback and comments from users of the data in order to validate.

One respondent validated the data by *"… organizing the data package in a simple and understandable way, identifying the most significant data, providing description of the data, and comparing the data with relevant information in the literature (if available)."*

In summary, the responses indicated that for physical experiments the main effort reported for ensuring valid data was expended in calibrating the observation equipment. Some responses were non-specific, such as *"Validation is performed through data post-processing*", but indicated that some validation processing did occur.

Validation is a way of answering the question "how do we know it works?" and justifying that process. To that end certain assumptions will always have to be made – e.g. the machine works; the power supply won't vary; signal drop outs won't occur, and so on. This at some point will require comparison of data with some expectation. The expectation requires a described rationale (the validation algorithm) for the validation itself to have validity.

For some validations (e.g. date and time validations) standard algorithms may exist. For brand new data objects and values this may require the development of validation algorithms. Where such algorithms exist, they should be referred to and where new validation algorithms are developed they should be documented and published ideally in the Data Storage Report unless there is a residual intellectual property that needs protecting.

4.3.2 Verification

This section examines the verification tools and techniques currently in use by the HYDRALAB+ community to verify data resulting from experiments.

If the description of "verification" is *"The evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition. It is often an internal process. Contrast with validation."* then the current suite of HYDRALAB+ activities have some way to go to identify or make clear their own regulations, requirements, specifications and imposed conditions.

Most of the answers from the survey were generic and often conflated with validation. Again, while this may be an issue with the wording of the question and a desire on the part of the respondents to answer the question, it does highlight an acknowledgement on the part of the researchers of the need for data management and its aspects such as verification. Moreover, it indicates some difficulty in describing and documenting the data management aspects.

4.4 IDENTIFIED PROBLEMS

According to Hsu et al, (2015) there are four main challenges, currently, in scientific data management particularly in the arena of Earth sciences.

1. Lack of metadata standards;
2. Insufficient workflow documentation and communication for experimental repeatability;
3. Inadequate data storage resources;
4. Lack of incentives and training.

These challenges highlight the need for data storage and sharing to meet the needs identified in the F.A.I.R. (findable, accessible, interoperable and re-usable) acronym:

- a lack of metadata standards can make data difficult to find;
- insufficient documentation and communication can make data inaccessible;
- inadequate data storage resources contribute to problems with interoperability of data as does poor or no implementation of standards; and
- poor incentives and training make data difficult to reuse.

4.4.1 Lack of metadata standards

It could be argued that a plethora of metadata standards also makes data difficult to find. There are many metadata standards and, as a result, a non-expert may find difficulty identifying which metadata standards to adopt for their specific data management requirements.

Web crawlers, scrapers and indexers require standardized metadata to allow data to be discovered. Some communities have standardized some metadata. Hsu et al (2015) build on Dublin Core, DataCite, HYDRALAB guidelines which all move towards standard structures and openness.

Some organisations utilize machine readable metadata, to proprietary standards. An example is given in Appendix I.

In conclusion, it is likely that the introduction of more, and more specific, metadata standards may obfuscate rather than clarify. Existing standards can be extended where needed to incorporate new ontologies if required. Detailed metadata standards like INSPIRE can serve as the base for this.

4.4.2 Insufficient workflow documentation and communication

While this is often cited as being a difficult area, it has not been raised as an issue in the survey. It's possible that this is a seemingly obvious and instinctive area of difficulty. It may attract more

attention than it really deserves and– in practice – it may not represent as big a problem as one thinks nor a particularly arduous task to address and rectify.

However, much valuable information about experiments is in lab books, planning documents and analysis code. Access to this data would be beneficial to future researchers.

4.4.3 Inadequate data storage resources

The FAIR approach extends the storage problem to include storage, discovery and retrieval and to that end this is clearly a difficulty if only because many human interactions are, currently, required in order to find, access, interoperate and reuse the data from any given experiment. As a result, there will always be some delays in determining if a necessary human resource is available to decide whether to and also to provide the data.

The BARDEX project in HYDRALAB IV produced a quantity of data of over 1 Terabyte. HYDRALAB IV had to create a specific web server (hydralab.info) to store and serve the data and allow people to download it.

This highlights the nature of physical sciences which appear to increasingly rely on large datasets and detailed statistical analysis to reach conclusions. The planning for the storage requirements should form part of the original experiment design and be documented in the Data Storage Report.

4.4.4 Lack of Incentives and training

Preparation of metadata, workflow documentation and preparation for transfer to a repository is time-consuming, detailed work and not, conventionally, in a scientists' skill set.

There does seem to be resistance to spending time ensuring data is FAIR, exemplified by one comment from a respondent to the survey question relating to validation that *"Reasonable - good. But my interest is using the data rather than checking this*."

This is likely to be because the FAIR aspects of the data following the research is perceived as less important than the experiment itself and actually using the data. It's moot whether using unvalidated data can ever be very useful. However, the implication that a common assumption exists that data management is generally of less significance than actually designing or conducting an experiment is quite a significant fact.

Researchers currently seem to lack much skill for, or training in, data management and sharing.

Furthermore, A significant and notable – if anecdotally evidenced - difficulty in terms of finding and accessing data resulting from experiments is the human factor. Resistance to sharing data has been referred to in the previous section.

This may be a reaction to the confusion non-experts may understandably have surrounding large scale metadata initiatives like INSPIRE. Navigating the metadata standard ocean for a scientist who perceives their main priority is the "science" and not the data sharing can be an extremely arduous task. Indeed, the user experience of INSPIRE has recently been described as being abysmal[14].

[14] INSPIRE workshop, Open Geospatial Consortium TC Meeting, TU Delft, 23rd March 2017.

This situation can be improved by education about the nature of data management and how important it is to the “science” but at the same time by raising the profile of data management as a profession in science.

5 POTENTIAL IMPROVEMENTS

5.1 DATA STANDARDS, FORMATS, PROTOCOLS AND TOOLS IN USE

This section examines the common standards, formats and protocols in use in the HYDRALAB+ community.

It is worth making a clear distinction here between the terms “standard”, “format” and “protocol”. Within the context of this document:

- a “standard” is an agreed prescription for the semantic structure of the content of a data package
- a “format” is an agreed prescription for the organizational structure of the data package
- a “protocol” is an agreed prescription for the mechanism for exchanging data packages
- a “tool” is piece of software used for reading, transforming and writing data using standards, formats and protocols.

5.1.1 Lack of metadata standards

While this hitherto may have been a problem, the recent development of internationally agreed metadata standards (such as OGC, INSPIRE, et al.) lead us to conclude that the absence of such standards is less a problem than the usability and dissemination of the standards and their adoption by and adaptation to the community.

It is important to clarify the distinction between adopting and adapting a standard and extending it.

A standard (such as XML) can be used to store any kind of data. In practice then, the content of an XML file is meaningless until an agreed semantic structure has been imposed on the standard, effectively eXtending the Markup Language.

The content *inside* the XML angle brackets <> (the names of the elements and their attributes) represents the semantic structure of a given standard. The GML example below is a case in point.

```
<gml:coordinates>45.67, 88.56</gml:coordinates>
```

<gml:coordinates> provides an agreed construct to hold a spatial coordinate pair. As such this represents the class of an object.

The content *between* the tags (in this case, the values 45.67, 88.56) represents an instance of a gml:coordinates object, just like the value of a field in a database table, or holes in a punched card.

Adopting GML as a standard would mean using the <gml:coordinate> tag to store and exchange values.

Extending the GML standard would mean inventing a new tag with a new name and possibly deriving some of its description and functionality from a previous tag (a 3D coordinate perhaps).

The current state of the art in data processing is unlikely to require a brand new OGC or other international standard for describing and exchanging data between field, laboratory and computer. Rather the existing standards are likely to suffice if used properly and documented carefully in the Data Storage Report.

The Work Package 4 JRA 3 training event in data flow between laboratory modelling, numerical modelling and field case study will endeavor to improve the existing data skills set of researchers and by making these sessions available online we will provide a resource for the future.

5.1.2 Insufficient workflow documentation and communication

In the USA, the Consortium of Universities for the Advancement of Hydrologic Science[15] has created the CUAHSI HIS (Hydrological Information System): a data catalogue and database system designed to help researchers share time series based water data. The adoption of documentation standards would help reduce the need for detailed and time consuming documentation for custom experimental data.

The OpenEarth system, developed by Deltares, is another data management system based on the principle that *"We believe that science and engineering have become so data-intensive that data management is beyond the capabilities of individual researchers"*[16]

Standard documentation and diagramming techniques could be adopted providing a mechanism for communicating appropriate levels of understanding between stakeholders.

5.1.3 Inadequate data storage resources

Large scale data repositories are increasing in number and the availability of storage for open source data will continue to increase.

DataCite's re3Data data repository catalogue at the time of writing contains links to over 1,500 research data repositories and has published version 3.0 of the "Metadata Schema for the Description of Research Data Repositories" (Rücknagel et al., 2015). This appears to suggest that a lack of adequate data storage resources may be a thing of the past.

We propose that HYDRALAB+ researchers make use of suitable data repositories and, in a wish to provide a service without being prescriptive, we aim to provide an interface on the HYDRALAB+ website for researchers to the Zenodo Horizon 2020 data repository[17].

5.1.4 Incentives and training

With regard to training, we aim to hold Early Career Researcher meetings on data – and provide this training to HYDRALAB+ participants. Examples of such data management training for researchers are:

- University of Minnesota, 'Managing your data', https://www.lib.umn.edu/datamanagement,
- JISC Managing Research Data, http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

[15] https://www.cuahsi.org/
[16] https://publicwiki.deltares.nl/display/OET/Data
[17] https://zenodo.org/

- MANTRA Research Data Management Training, http://datalib.edina.ac.uk/mantra/,
- ESIP, Data Management for Scientists, http://commons.esipfed.org/datamanagementshortcourse

Incentives for better data management will naturally emerge over time as research with better planned data management will begin to receive a greater share of funding as a result of their data management.

Moreover, the allocation of Digital Object Identifiers, or DOIs to datasets allows a dataset to receive a citation in a paper or report. DataCite (http://datacite.org/) for example, allocates DOIs that take you to a public web page with meta-data about the associated dataset and a direct link to the data itself. The allocation of DOIs to quality physical model datasets will support researchers by helping them to find, identify and cite these datasets with confidence. The allocation of a DOI to a dataset is built in to many repositories, such as Zenodo and we recommend that HYDRALAB participants only bank data in a data repository approved by OpenAire[18] with a DOI. In addition, the development of data journals (such as Data in Brief, Geoscience Data Journal, Earth System Science Data, Dataset Papers in Science and Journal of Visualised Experiments) which exist to share the details about experiments and provide links to the datasets, will allow researchers to gain publications and citations for papers that describe the data that they are sharing. We recommend that data journal papers (which can be based on the existing HYDRALAB Transnational Access Data Storage Reports) be written and published in data journals (with links to the associated data packages).

The inclusion of citations to data in the metrics produced by organisations such as the Web of Science, which are used to compare the performance of academics (in particular) would assist in changing the culture of science to recognise the importance of data and encourage its sharing.

5.2 The Flow Of Data Between Communities

Through the implementation of a two way relationship between the data and all associated documentation and conclusions using DOIs and by the discipline of writing effective Data Storage Reports we hope to achieve two goals:

- A consistent approach to data discovery and retrieval to the HYDRALAB community and the scientific community at large.
- A base from which HYDRALAB can develop over time allowing searching across current and historical metadata records.

Investigation of big data tools and data warehousing solutions (e.g. Apache Hadoop[19] and NoSQL[20]) which provide solutions over and above that offered by the likes of Zenodo will prove useful.

However, the specialism required for constructing and maintaining such systems is likely to lead to the conclusion that third party solutions will allow HYDRALAB partners to focus on the science rather than the details of data curation.

[18] https://www.openaire.eu/participate/deposit-publications-data
[19] http://hadoop.apache.org/
[20] https://en.wikipedia.org/wiki/NoSQL

5.2.1 Data flow between the field and the laboratory

Task 10.4 will develop and standardize methods and protocols for the exchange of data between field and laboratory, using case studies which will prove useful elsewhere in Hydralab+. *Table 5* indicates the areas that will be investigated and the related work packages.

Description	Links to other work packages
Effect of seagrass patchiness on suspended sediment concentration and wave attenuation	Relevant to Work Package 9.3 and 9.4
Impact of fauna on changes to suspended sediment concentration	Relevant to Work Package 8.4 and 8.5
Impact of suspended sediment concentration on eel grass health and behaviour and implications for sediment suspension hydraulics	Relevant to Work Package 8.4
Quanitfy differences between measurement techniques across field and flume studies	Relevant to Work Package 9.1
Examine how best to use data produced by models	

Table 5 Description of field experiments in HYDRALAB+ and links to other work packages

5.2.2 Data flow between the laboratory and numerical simulation

Task 10.5 will examine the relationship between the laboratory and numerical modelling simulations with particular regard to the matching of boundary conditions. Methods, protocols and standards relating to data validation and verification will be examined with a view to developing specific guidance for the flow of data between the physical laboratory and the virtual world of numerical modelling.

5.3 The Effectiveness Of Validation and Verification

Improvements in this area could be achieved in a number of ways but all essentially come down to developing the data skills necessary to apply data standard compliance validation against data from new experiments and the reprocessing of existing data. Such a task is not trivial – it can be observed, for example, in the music industry as the “digital remastering” of recorded music from the past.

The effectiveness of the validation and verification is, to a degree, subjective. It will after all be determined by the nature of the experiment and the data and, to some extent, the expected and foreseeable use of the data. There is little point, for example, ensuring that double precision numbers are used where only integer numbers are ever going to be required.

6 SUMMARY

It's certainly possible to develop specific ontologies, (e.g. OntoSoft[21]) for particular domains. However, ontologies do appear to multiply in a relatively haphazard and unpredictable way. The issue is one of common understanding within a given context and is a result of the fact that labels change their meaning over time and space and cultures (see, for example, the example about the definition of H_{m0} and H_s in Section 1.1).

The desire to invent new vocabularies and ontologies to overcome confusion between existing agreed and de facto standards, if indulged, is doomed to increase confusion, thereby having the opposite effect to that desired. Unless some clear and overwhelming evidence is forthcoming to indicate a pressing need for a new hydrological and hydraulic vocabulary and ontology we recommend the wider dissemination of existing structures (such as the IAHR list of sea state parameters).

This is not to argue against vocabularies or ontologies or any standards. Rather, we suggest that researchers learn to use existing standards where possible, extend existing standards where necessary and only invent new standards when absolutely necessary.

This principle of parsimony would make the use of data in hydrological and hydraulic research less intimidating for the non-expert in data management.

A number of recommendations for improving data management in the HYDRALAB community have been made:

- We propose that HYDRALAB+ researchers make use of suitable data repositories and ensure that each dataset is allocated a DOI . In a wish to provide a service without being prescriptive, we aim to provide an interface on the HYDRALAB+ website for researchers to the Zenodo Horizon 2020 data repository using Zenodo's Application Programming Interface.
- We should provide training on Data Management to HYDRALAB+ participants and provide links to existing online training resources.
- We recommend that the existing Data Storage Reports adopted by the project be reviewed to incorporate any appropriate aspects from the Data Management Plans which follow the H2020 template and also through the associated DMPOnline facility for the generation of data management plans. Since HYDRALAB+ has an established usage of Data Storage Reports, the DMPOnline facility is not planned to be used directly at this stage.
- We recommend that data papers (which can be based on the existing HYDRALAB Transnational Access Data Storage Reports) be written and published in data journals (with links to the associated data packages).
- We support the development of metrics that include citations to data and data papers as this would assist in changing the culture of science to recognise the importance of open data.

[21] https://doi.org/10.1016/j.envsoft.2017.01.024

7 QUESTIONNAIRE

A Survey Monkey questionnaire was built and distributed to the partners identified in Work Package 10.1, namely, HRW, Aalto, CNRS, HSVA, LBORO, UHULL.

7.1 QUESTION 1: CONTACT INFORMATION.

Results omitted for anonymity.

7.2 QUESTION 2 : WORK STREAM ID

Participants were asked to name their workstream ID (e.g. H+_HRW_JRA1_002 - this is the unique identifier for your experiment or research activity. If you don't know it please just write the name of the activity or experiment). Their responses are listed below.

Respondent No.	Response text
1	JRA1 task 8.2
3	Field experiment Rødsand, several WPs
4	Biostabilization by biofilms
5	RECIPE
6	COMPLEX and TA
7	EU-FAST
8	H+_UC_JRA2_ferrofluid
9	Experimental work in JRA1 and JRA2
11	Involved in JRA1/JRA2/JRA3
12	We have not conducted any tests in HYDRALB yet so this is now hypothetically answered
13	Innovative approaches for measuring organism stress and behavioural integrity in flume facilities (JRA1, Task 8.4)
14	HY+_HSVA-01_KVAERNER
15	Multiple: H+-HRW-02-Kleinhans & H+_HRWallingford_01_Troch
19	COMPLEX

7.3 QUESTION 3: HOW WOULD YOU CLASSIFY THIS EXPERIMENT/ACTIVITY?

Classification	Response Percentage	Response Count
Field (research conducted in the real world in an uncontrolled environment)	10.5%	2
Laboratory (research conducted in the real world in a controlled environment)	73.7%	14
Numerical (research conducted in a virtual world in a controlled environment e.g computer simulation or numerical model)	5.3%	1
Question skipped	10.5%	2

7.4 QUESTION 4: IN WHAT DATA FORMAT(S), OR PHYSICAL STRUCTURE, ARE THE DATA PACKAGES PROVIDED?

Suggested examples were e.g. csv, text, .png, proprietary, etc. Responses are given below.

Respondent No.	Question 4 Responses
1	mat, text, and netcdf - depending on type of measurement
4	text
5	csv, png, jpeg,
6	matlab binary files and text files
7	a mix due to different data sources, described in an xlsx: tif, csv, netcdf
8	proprietary Vectrino's data, text, images/videos, .mat
9	All possible formats, also specific to manufacturer specifications; depends als on what software is used
11	proprietary (e.g. Vectrino data / laser scan data / PIV data)
12	csv, ascii
13	Data packages will be provided in different formats depending on their contents. The formats that should be used are .mat (matlab data), .xls/.csv, and text. Please, note that we are still designing the lab experiments, therefore the listed formats might vary in future.
14	docx,xlsx,jpg,pptx
15	Mixture of proprietary, csv & text
17	.mat
18	Text
19	.txt

7.5 QUESTION 5: TO WHAT DATA STANDARDS DO THE DATA PACKAGES ADHERE (WHAT IS IT) ?

Example responses given were OGC standards, GML, WaterML 2.0, etc. Respondent responses are below.

Respondent No.	Question 5 Responses
1	?
7	spatial data and processed wave measurements are according to OGC standards, other data have no standards
8	not known
9	I do not understand this question.
11	n/a
12	Data is in ascii format
13	No data have been collected as of now, therefore this information is not available.
15	Unknown
17	none
18	none
19	n/a

7.6 QUESTION 6: WHAT, IF ANY, DATA PROTOCOLS ARE USED IN THE EXCHANGE OR TRANSFER OF THIS DATA PACKAGE?

Respondent No.	Question 6 Responses
1	mainly physical transfer (hard dirc), also ftp.
5	ftp
6	ftp
7	ftp, wcs, wms
8	no protocols
9	Not sure that I understand - USB to extract it from local data machines? Besides this, all other formats are possible...
11	smaller files exchanged via box.com or onedrive.com

12	http (most likely)
13	Not known at the moment as there are no data to share
15	Unknown
17	FTP
18	USB storage unit
19	n/a

7.7 QUESTION 7: HOW WOULD YOU ASSESS THE VALIDATION OF THIS DATA PACKAGE?

Respondents were provided with the following supplementary information:

> *'Validation can be described as "The assurance that a product, service, or system meets the needs of the [...] stakeholders. It often involves acceptance and suitability with external customers. Contrast with verification." [Project Management Body of Knowledge]'*

Their responses are given below.

Respondent No.	Question 7 Responses
1	measurement procedure is checked during tests data is used and analyzed - here possible errors can be detected
4	Comparison to field data
7	Reasonable. This has been a partly iterative process, as part of the first data were not useful and at the end this was better but we still did not really have what we wanted.
8	Validation is performed through data post-processing.
11	Vectrino and laser scan data are provided with a calibration. PIV data is calibrated for each experiment using a two-level target. The calibration process is carried out to achieve an accuracy of less than one-pixel for 3D measurement
12	By requesting feedback and comments.
13	By organizing the data package in a simple and understandable way, identifying the most significant data, providing description of the data, and comparing the data with relevant information in the literature (if available).
15	Same data collection techniques as we use for commercial work - calibration of instruments carried out as required.

7.8 QUESTION 8: HOW WOULD YOU ASSESS THE VERIFICATION OF THIS DATA PACKAGE?

Respondents were provided with the following supplementary information:

> *'Verification can be described as "The evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition. It is often an internal process. Contrast with validation." [Project Management Body of Knowledge]'*

Their responses are listed below.

Respondent No.	Question 8 Responses
1	the outcomes (reports) are checked in the Deltares ISO system
7	Reasonable - good. But my interest is using the data rather than checking this.
8	Verification of data quality is performed through data post-processing
11	For Vectrino data this is based on SNR and correlation thresholds. For laser scan data, 'stray' data are removed. For PIV data there are internal processing routines to identify and replace erroneous data. Standard approaches are used (such as local median verification)
12	We calibrate equipments carefully
13	By verifying that: (i) all instruments worked properly during data collection; (ii) experimental protocol is adequate and followed during the conduction of experiments; (iii) post-processing and analysis of data comply with established methodological standards.
15	Conditions requested from hardware (pump system, wave maker) are known. Post-processing determines if close to these values (+/- 5%) has been achieved.

7.9 QUESTION 9: HOW WOULD YOU DESCRIBE THE OVERALL EFFECTIVENESS OF THIS DATA PACKAGE?

Respondents were given the following supplementary information:

> *'Effectiveness can be described as "In general, efficiency is a measurable concept, quantitatively determined by the ratio of useful output to total input . Effectiveness is the simpler concept of being able to achieve a desired result, which can be expressed quantitatively but doesn't usually require more complicated mathematics than addition." [Wikipedia]'*

Their responses are given in the following table.

Respondent No.	Question 9 Responses
1	?
4	low effectiveness
7	Low. It would have been more efficient if I (the user) would have been involved from the beginning.
8	Overall effectiveness is difficult to judge, if non-standard measurement development are carried out. There is nothing to compare to.
11	Effectiveness will depend upon the output required which varies by experiment
12	Later we can see how many publications have come out from this
13	Not known, as no data have been collected as of now.
15	Unknown - data hasn't been collected yet.

7.10 QUESTION 10: HAVE YOU EXPERIENCED ANY ISSUES OR PROBLEMS ASSOCIATED WITH THE TRANSFER OR EXCHANGE OF THIS DATA PACKAGE?

Respondent No.	Question 10 Responses
1	no
4	no
7	Yes: high-resolution global satellite data is not easily transferable.
8	Data transfer within the different researchers in the institution is not difficult, since data are usually accompanied by enough reporting and metadata.
11	For PIV data we have significant issues with the volume of data.
12	We have not conducted any tests so we will see after the data delivery
13	Not applicable
15	Unknown - data hasn't been collected yet.

8 GLOSSARY

OGC	Open Geospatial Consortium "an international industry consortium of over 521 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards." *http://www.opengeospatial.org/ogc*
PIV	Particle Image Velocimetry – an optical method of flow visualization and measurement
VECTRINO	High-resolution acoustic Doppler velocimeter
DOI	Digital Object Identifier – a unique identifier identifying a specific dataset within the context of an agreed community.
UKEOF	United Kingdom Environment Observation Framework – an initiative "to improve coordination of the observational evidence needed to understand and manage the changing natural environment."[22]
CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science, Inc. – "a research organization representing more than 130 U.S. universities and international water science-related organizations"[23]
INSPIRE	Infrastructure for Spatial Information in Europe - *"aims to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment."*[24]
GML	Geography Markup Language - *"XML grammar defined by the Open Geospatial Consortium (OGC) to express geographical features."*[25]

[22] http://www.ukeof.org.uk/
[23] https://www.cuahsi.org/
[24] http://inspire.ec.europa.eu/about-inspire/563
[25] https://en.wikipedia.org/wiki/Geography_Markup_Language

9 REFERENCES

Gerritsen, H., Sutherland, J., Deigaard, R., Sumer, B.M., Fortes, J., Sierra, J.-P. and Prepernau, U., 2009. "Guidelines for composite modelling of the interactions between beaches and structures." HYDRALAB report JRA1.4, pp. 71.

Gerritsen, H. and Sutherland, J., 2011. "Composite Modelling". Chapter 6 (pp. 171 – 219) of L.E. Frostick, S.J. McLelland and T.G. Mercer (Eds), Users guide to physical modelling and experimentation: experience of the HYDRALAB Network. CRC Press/Balkema, Leiden, the Netherlands, ISBN: 978-0-415-60912-8 (Pbk), ISBN: 978-1-4398-7051-8 (e-Book).

Gerritsen, H., Sutherland, J., Deigaard, R., Mutlu Sumer, Fortes, J.E.M., Sierra, J.P. and U. Schmidtke, 2011. "Composite modelling of the interactions between beaches and structures." Journal of Hydraulic Research. 49:sup1, 2-14. http://dx.doi.org/10.1080/00221686.2011.589134

Hsu, L., Martin, R.L., McElroy, B., Litwin-Miller, K. and Kim, W., 2015. Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. Geomorphology, 255: 180-189.

IAHR (1989). List of sea state parameters. J Waterway, Port, Coastal and Ocean Eng. 115(6): 793-808. http://ascelibrary.org/doi/pdf/10.1061/(ASCE)0733-950X(1989)115:6(793)

Millard, K., Cleverley, P. and Sutherland, J., 2013. 'Data model report, implementation for information exchange.' HYDRALAB IV Deliverable D10.3

Min, Heng, 2010. GML Validation Based on Norwegian Standard. MSc Thesis, Department of Computer Science and Media Technology Gjøvik University College, http://hdl.handle.net/11250/144130

NYU Health Sciences Library, 2012, "Data Sharing and Management Snafu in 3 Short Acts" Internet: available from https://www.youtube.com/watch?v=dl6C_GrZrbE

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., Kirchhoff, A. , 2015. "Metadata Schema for the Description of Research Data Repositories"

Sutherland, J. and Barfuss, S., 2011. "Composite Modelling, combining physical and numerical models." Proc 34th IAHR World Congress, Brisbane, Australia, p 4505-4512.CD-ROM, ISBN 978-0-85825-868-6.

Sutherland, J., Millard, K., and Cleverley, P., 2014. 'System validation and evaluation, central data store for experiments.' HYDRALAB IV Deliverable D10.6

Van Os, A.G., Soulsby, R.L. and Kirkegaard, J. (2004). The future role of experimental methods in European hydraulic research: towards a balanced methodology. Journal of Hydraulic Research, Vol. 42(4): 341-356.

Wells, S., Sutherland, J. and Millard, K., 2009. "Data management tools for HYDRALAB – a review." HYDRALAB report NA3-09-02. Also HR Wallingford Report TR 176.

APPENDIX I

An example of a machine readable, proprietary metadata standard provided by Deltares. The structure is as follows:

- Sto: Some general information also about structure of the binary file that contains the data.
- General: General information about the project.
- Series: For each instrument a series section is available with the position and some other information

```
STO

DATATYPE,R4

ACCESS,DIRECT

FILEFORMAT,BINARY

RECL,    4

LINK

END:STO

GENERAL

PROJECT,1220204.PRJ

START,11:34:20

STOP,12:36:50

MEASUREMENT,NAME=122020,ID=00

DATE,13-10-2015

CREATION,DATE,2015:11:09,TIME,17:46:17

SEQUENCE

UNRELATED

NOVALUE,-.100000E+11

END:GENERAL

 ZERO-CROSSING,UPWARDS

SERIES,WHM01:WAVES

DIMENSION,m
```

```
X,  108.500,m
Y,    0.000,m
Z,    0.000,m
NUMBER,00001119
END:SERIES
SERIES,WHM03:WAVES
DIMENSION,m
X,  114.500,m
Y,    0.000,m
Z,    0.000,m
NUMBER,00001099
END:SERIES
SERIES,WHM04:WAVES
DIMENSION,m
X,  117.500,m
Y,    0.000,m
Z,    0.000,m
NUMBER,00001099
END:SERIES
```