



Adaptation For Climate Change

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

Task 10.2 Data Standards and Licenses

D10.3 Data Standards Report

Status: Public document from 22 February 2018

Version: 3

Date: 22 February 2018



EC contract no 654110, HYDRALAB+



DOCUMENT INFORMATION

Title	Data Standards Report
Lead Authors	Quillon Harpham, Paul Cleverley, James Sutherland, Lesley Mansfield (HR Wallingford)
Contributors	DELTARES, CNRS, LNEC, LUH, Samui, UHULL, UPORTO
Distribution	Public from 22 February 2018
Document Reference	DOI: 10.5281/zenodo.1182560

DOCUMENT HISTORY

Date	Revision	Prepared by	Organisation	Approved by	Status
08 Nov '17	1.0	Harpham, Cleverley, Mansfield, Sutherland	HR Wallingford		Restricted
30 Jan '18	2.0	Harpham, Sutherland	HR Wallingford		Restricted
22 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

This document reflects only the authors' views and not those of the European Community. This work may rely on data from sources external to the HYDRALAB project Consortium. Members of the Consortium do not accept liability for loss or damage suffered by any third party as a result of errors or inaccuracies in such data. The information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and neither the European Community nor any member of the HYDRALAB Consortium is liable for any use that may be made of the information.

LICENSE

This report is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

CITATION:

Suggested citation:

Harpham, Q., Cleverley, P., Sutherland, J. and Mansfield, L., 2018. Data Standards Report. HYDRALAB+ deliverable D10.3, <http://dx.doi.org/10.5281/zenodo.1182560>

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 D10.3 of the HYDRALAB+ project, entitled 'Data Standards Report'. It is one of the outputs of Work Package 10 – 'JRA3: Facilitating the Re-use and Exchange of Experimental Data' and is tasked with examining available data standards and licenses appropriate to the HYDRALAB+ domain and making recommendations for the uptake of proposed or selected standards, protocols and licenses. It has a wide scope and seeks to provide a comprehensive data management framework for the community. This includes guiding scientists into sensible choices for data formats, the provision of sufficient supporting information describing data outputs, community vocabularies in particular for parameter names and units, appropriate licenses and suitable embargo periods for experiment results.

Instead of insisting on a set of prescriptive data formats, this report gives a set of principles and guidelines for experimenters to follow, supported by recommendations where leading formats and structures exist. The intention is to respect the technologies and intentions at each of the partner organisations, whilst providing positive practices to ensure that the experimental data produced is interoperable and reusable. The recommendations respect the wide variety of data management options for formats and structures; metadata, vocabularies and ontologies; and licenses and embargo periods. Where appropriate, specific technologies have been offered. They do not seek to impose an unrealistic set of rules and regulations which must be followed, rather they offer a set of sensible, modern principles and resources to move the community forwards together and bring it in line with other similar communities currently iterating their own data management practices. They also dovetail with the project's usage of the Zenodo data repository for the storage of experiment results datasets.

A report output targeted at the scientists themselves articulates the recommendations as a set of simple statements with increasing levels of supporting information from this report. These statements can be summarized as follows:

- selecting a format for data which respects natural structure and size, and also will be accessible to other scientists now and in the future;
- including sufficient metadata to allow data to be interpreted, if possible with an established metadata standard;
- taking all parameter names and units from established vocabularies;
- including an open license, with as few restrictions as possible;
- enforcing an embargo period only if necessary and not more than two years;
- creating data;
- and storing data in the Zenodo repository.

The tendency towards larger and larger data sets will accelerate as the technology for observing and recording data about the real world (including that used in laboratory experiments) gets cheaper and improves in quality. However, the current assessment of the requirements of HYDRALAB+ activities suggest that the Zenodo data repository standard limit of 50GB per dataset will suffice for many datasets. Where large optical (level 0) raw datasets are collected, it may be more appropriate to store calibrated, processed (level 1) datasets to reduce the disk space required and make the data more usable.

CONTENTS

Document Information	2
Document History	2
Acknowledgement.....	2
Disclaimer	2
License	3
Citation:	3
Executive Summary	4
Contents.....	5
1 Introduction	7
1.1 Context and Objectives	7
1.2 Scope	8
2 Data Formats and Structures	9
2.1 Approach	9
2.2 Motivation	9
2.3 MIME Type.....	11
2.4 Structure.....	12
2.5 Conversion.....	14
3 Data volume	15
4 Metadata, Vocabularies and Ontologies.....	17
4.1 Approach	17
4.2 Metadata, Vocabulary and Ontology Standardisation.....	18
4.2.1 WaterML2	21
4.2.2 NetCDF.....	22
4.3 Data Repository Keywords	23
5 Licenses and Embargo Periods	25
5.1 Licenses	25
5.2 Open Access and Embargoes.....	27
5.2.1 OA to publications	27
5.2.2 OA to data	28
6 Recommendations.....	29
6.1 Zenodo.....	29
6.2 Data and Metadata Format	30

6.3	License and Embargo Period	31
7	Acronyms and Abbreviations	32
8	References.....	34

1 INTRODUCTION

1.1 CONTEXT AND OBJECTIVES

This report is HYDRALAB+ project deliverable D10.3 'Data Standards Report'. Part of HYDRALAB+ Work Package 10, it is tasked with examining available data standards and licenses appropriate to the HYDRALAB+ domain and making recommendations for the uptake of proposed or selected standards, protocols and licenses. The work package is intended to ensure the "interoperability" part of the FAIR (Findable, Accessible, Interoperable, Reusable) mnemonic (EC, 2016, Wilkinson et al., 2017) facilitating the exchange of data between different domains within the HYDRALAB+ community. Overall, the activities of this task can be summarized as:

- Providing easy-to-use tools and guidance to make the data produced by experiments easy to find, easy to understand and easy to re-use; and
- Convincing researchers to put effort into data storage and accessibility.

The other activities of the work package, to date, are contributing towards an overall context for the implementation of improved data management practices which will aid this interoperability.

The HYDRALAB+ project deliverable D10.2 'Critical Review' (HR Wallingford, 2017b), under Task 10.1 'Critical review of data flux between laboratory models, numerical models and field case studies', performed an examination of three areas in particular:

- 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.

This analysis revealed that typical scientists within the HYDRALAB+ community had a lack of knowledge about data management practices and principles such as standards and protocols. Many different local and community standards, formats protocols and tools were in use with inevitable issues identified at cultural boundaries. Attempts had been made amongst some communities to standardize metadata, but this has been hampered by the variety of such standards and associated vocabularies on offer. The HYDRALAB+ community has a well-established practice of producing Data Storage Reports incorporating some description of the data itself. The Critical Review (D10.2) examined previous recommendations from D10.1 'Data Management Plan' for HYDRALAB+ community scientists to use Data Management Plans (DMPs) (HR Wallingford, 2017a). The review concluded that it is appropriate to continue using Data Storage Reports, but incorporate any appropriate aspects from standardized Data Management Plans and, in particular, the DMP Online implementation. These actions implied the creation of certain metadata to support these plans including common identifiers and Multipurpose Internet Mail Extension (MIME) types for discrete datasets.

Alongside this, D10.4 – ‘Data Repository Rules (including DOIs)’ – supports use of the Zenodo repository¹ for storing the experimental data managed by the associated Data Storage Reports. This includes the provision of Digital Object Identifiers (DOIs) and other metadata elements representing the data stored. Zenodo has been funded by OpenAire (<https://www.openaire.eu/>), which is an EU funded project and has therefore been tailored with Horizon2020 projects in mind. The platform is free to use and provides an Application Programming Interface (API) to allow integration with the HYDRALAB+ website.

1.2 SCOPE

This deliverable has a wide scope and seeks to provide a comprehensive data management framework for scientists involved in HYDRALAB+ research. Attention is given, not only to the needs of those creating the data, but specifically to the subsequent usage of it (both for its original intended purpose and for exploitation of results beyond the original purpose for which the data were collected). There is a need to ensure that those who have created the data can themselves understand it in the future and new users are able to easily access and interpret archived data. This challenge to the data creators puts them in the position of a new user who is attempting to find, access and process the data that they create.

This includes guiding scientists toward sensible choices for file formats which will address the various competing requirements such as the overall size of the data package, the usability of the data package and the efficiencies of storage and speed of data transfer/download. Certain common data formats are considered and offered where appropriate. It is also important that scientists provide sufficient supporting information (e.g. metadata) describing their data and the circumstances in which it was created. This situation is complicated by the different approaches taken by different domains, in particular overlap between data storage and metadata provision in native formats and through on-line resources. Sensible choices must be made in this regard to enfranchise all users, minimizing duplication and maximizing system (and data) usability.

Vocabularies for categories, general terminology and phenomena / parameter names and units are commonly implemented to support discovery and aid understanding and will also be considered. These offer increased usability and longevity at the expense of initial local effort. This includes file type descriptions such as MIME type. It is also necessary to address the question of appropriate licenses and suitable embargo periods for experiment results.

To encourage and facilitate the sharing of data between domains in the HYDRALAB+ community, the overall thrust will be to avoid exchanging data in proprietary formats except where such formats are common and do not require significant investment in the purchase of software licenses to be able to read or write such formats. The recommendations must offer a clear and practical way forwards, respecting the current and varied practices throughout the HYDRALAB+ community as it seeks to reach the levels necessary for comfortable interoperability and re-use. Overall, the standards, processes and procedures followed must be compatible with the usage of the Data Storage Reports and a data repository, such as Zenodo, to store and provide reference to the data package itself.

¹ <https://zenodo.org/>

2 DATA FORMATS AND STRUCTURES

2.1 APPROACH

The programming language pioneer Niklaus Wirth famously defined computer programs as algorithms plus data structures. A research activity or scientific experiment is – at one level – no different. An algorithm or experiment is not much use without data. More accurately, neither algorithm nor experiment is much use without appropriately structured data.

The organisation of data into appropriate and useful structures is key to efficient and effective information processing. The operative word here is “useful”. Utility is determined by context and, in different contexts, different structures may be more or less useful. Deliverable D10.2 Critical Review (HR Wallingford, 2017b) shows a list of data formats in common use in the HYDRALAB+ community. This list includes some open and some proprietary formats. Included in the summary results is the following statement:

“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.”

Hydralab Deliverable D10.2 Critical Review, Executive Summary (HR Wallingford, 2017b)

Moreover, by prescribing specific data structures and formats, HYDRALAB+ would have to create and maintain a reference list of prescribed structures and formats. It would also need to determine how to deal with the situation where the prescription is not complied with – would there be sanctions? If so, what would they be and would they be enforceable? Any benefits of implementing such prescription would not be attained without this sanction. In effect, with no sanction, researchers could store data in non-compliant formats and the only choice then would be to extend the reference list to include these formats. This would make little practical sense.

As such, this work will not attempt to define a set of mandatory data structures and formats to be adhered to, rather it will seek to lead researchers into sensible choices in this regard, building on the use of particular formats where appropriate and established. Instead of insisting on a set of prescriptive data formats, it gives a set of principles and guidelines for experimenters to follow, supported by recommendations where leading formats and structures exist. These guidelines respect the technologies and intentions at each of the partner organisations, whilst providing positive practices to ensure that the data produced is interoperable and reusable.

2.2 MOTIVATION

The choice of which data structure or format to use in any given context can have many drivers including (but not limited to) personal knowledge (“*I know this structure so I’ll use this even though there may be something more suitable*”), expedience (“*I don’t have time to use a more complex normalized structure*”) and interoperability (“*I want others to be able to reproduce and confirm the*

results of my experiment"). Criteria to be considered for the adoption of a data format includes that it be as open and as well-understood as possible. In addition, it should be affordable and structurally appropriate to the data being exchanged. One other useful metric is the number of systems which can natively read and write the format in question. Good advice is given by UCL Research Data Management Best Practice guide to selecting file formats:²

"Four key questions in choosing formats:

- *How do you plan to use the data that you produce? How will you store, share and analyse your data?*
- *Do you have any funding for new software, if it is required?*
- *Do your peers expect your data to appear in certain formats? Do you have access to expertise in particular software or to best practice information for your discipline or research area?*
- *Does your funder have expectations regarding how you present your data?"*

These points also highlight some other motivations for selecting data formats and structures.

It is also possible to distinguish between different levels of data³ which may be produced by laboratory experiments:

- Level 0: raw data (e.g. PIV images) produced by the instruments used in the experiment.
- Level 1: standard (or processed) data, such as velocity vectors obtained from the PIV images. Standard data is typically produced by calibration and filtering processes applied to the raw data.
- Level 2: tailored data (or experiment results) which are used as the basis for the conclusions of the research.
- Level 3: graphics of data used in the presentation of results.
- Level 4: catalogue of data.

Those producing data packages for the use of others must decide which level of data that they wish to offer.

- Level 0 data can be very large and / or in a proprietary format and / or dependent on specific software packages to interpret. It can also be time-consuming to process.
- Level 1 data should be easier to view and use than level 0 data but may also lose important information held by level 0 data. It may also be more likely to be valuable intellectual property and often requires less storage space (for example if it is in the form of a grid of x,y velocity vectors, rather than the level 0 PIV image)
- Level 2 data would typically be published in an associated report or Journal paper.
- Level 3 data would be published in reports and papers.

² <http://www.ucl.ac.uk/library/research-support/research-data/best-practices/guides/formats>

³ <https://publicwiki.deltares.nl/display/OET/Data>

- Level 4 data should be held in a central, searchable catalogue.

HYDRALAB+ scientists are expected to put results into papers or reports and store these in a separate repository with a separate DOI. As such, this discussion concerns level 0 and level 1 data, which will be stored separately and is expected to be linked to the report or paper.

The UK Data Archive⁴ publishes a suggested list of formats for long term preservation of data. This is limited however, and omits some later structures and formats such as netCDF. In the HYDRALAB+ context, we are discussing the depositing of sets of data, resulting from experiments and research activities, into data repositories with the aims of reference and reproducibility. So in essence, the decisions should be based less on what is convenient for the creator of the data but on what is convenient for the consumer. If it is difficult to find and access the data, it is less likely that consumer will expend the effort required to use it.

The question remains how to get experimenters and researchers to bear this in mind when selecting appropriate data structures and formats for long term storage and retrieval. Perhaps the most significant single improvement in the exchange of data – between systems generally as well as between HYDRALAB+ scientific domains – could be achieved by the researcher/experimenter asking themselves the question:

“How will I access this data a year from now?”

Considering themselves as the most likely future user of the data emphasises the need to structure it helpfully and use an appropriate format.

One such data format assessment measure can be the FAIR data management principles of ‘Findable’, ‘Accessible’, ‘Interoperable’ and ‘Reusable’ (EC, 2016). For example, exchange of data in ASCII Comma Separated Values format (CSV) meets these principals strongly because it is a *de facto*, well understood format which is open, free to use and suitable for the exchange of small to medium sized flat structured data sets (so small to medium sized time series could be exchanged using the CSV format). Furthermore, there are many systems which can natively read and write CSV files providing a widespread base for selection. CSV ticks the FAIR acronym boxes of Accessible, Interoperable and Reusable. Where, however, the nature of the data acquisition requires proprietary equipment (in the form of, say, field observation equipment or third party computer software) the format in which the data is stored may well be proprietary and hence opaque to other systems. In addition, it is more likely that licensed software will be required to read, write and transform the data.

2.3 MIME TYPE

The Multipurpose Internet Mail Extension (MIME) standard⁵ is maintained by the Internet Assigned Numbers Authority (IANA). Originally designed to provide metadata for email attachments, the standard is used by a number of protocols to maintain a registry of well-understood media types while allowing for extensions and innovation. This allows the envelope of a given data set to be

⁴ <http://www.data-archive.ac.uk/create-manage/format/formats-table/>

⁵ <https://www.iana.org/assignments/media-types/media-types.xhtml>

described simply, in order for computer systems (and humans) understand how to deal with the dataset in question.

By ensuring that an appropriate MIME type is selected for each dataset stored in a repository (even if the MIME type is an experimental one) at least a *de facto* standard definition would have been used to describe a method of transmitting the data to other systems.

2.4 STRUCTURE

The decision about which data structures or formats to use to store datasets is influenced by the nature of the data and the internal natural relationships between the data items⁶. Moreover, this discussion is complicated by the overlap between file formats for storing data and associated supporting information such as metadata. Flat file structures or multivariate complex relationships and anything in between (including complete relational databases or object stores or RDF triple stores⁷) can represent a single data set.

There are no established rules or best practices governing the selection of different data structures and any associated guidelines may differ between different research establishments and funding bodies. Indeed, such guidelines may change significantly over time as better information technology becomes available. Therefore, the choice of how the dataset is represented should remain with the researcher / experimenter since they must be able to justify the choice as part of their experimental or research methodology.

For example, a self-describing, highly structured, highly compressed netCDF file structure may be selected for storing experimental data where the storage capacity and retrieval systems are limited – hence efficient storage wins out over ease of access; it is harder (more complex, time-consuming) for later experimenters to retrieve and read the data but it is at least efficiently stored.

⁶ [https://en.wikipedia.org/wiki/Cardinality_\(data_modeling\)](https://en.wikipedia.org/wiki/Cardinality_(data_modeling))

⁷ <https://www.w3.org/2001/sw/wiki/RDF>

Table 1 show a top ten list of data formats across a range of data repositories (including Zenodo). This ranking includes netCDF, a common format for scientific data storage and exchange. It should also be noted that the popularity of ZIP compressed files serves to hide the internal structure of the dataset compressed into the zip file itself.

Table 1 Top 10 Formats associated with published dataset.(format is MIME type – file extension) in Scientific Data Repositories - Assante et al.(2016)

Format	3TU.Datacentrum	CSIRO	Dryad	Figshare	Zenodo
#1	app./x-netcdf (3070 – 80%)	app./fits – sf (143,376 – 26%)	text/plain – txt (4926 – 16%)	n/a – xls (267,222 – 83.3%)	n/a – sav (1798 – 15.2%)
#2	app./zip (559 – 14.6%)	image/png – png (95,072 – 17.3%)	Excel 2007 – xlsx (3793 – 12.3%)	n/a – pdf (16,996 – 5.3%)	n/a – txt (1243 – 10.5%)
#3	text/plain (57 – 1.5%)	app./fits – rf (94,028 – 17.1%)	text/csv – csv (3099 – 10%)	n/a (12,968 – 4%)	n/a – png (1059 – 9%)
#4	app./octet-stream (27 – 0.7%)	app./fits – FTp (92,888 – 16.9%)	app./zip – zip (2834 – 9.2%)	n/a – docx (4868 – 1.5%)	n/a – fits (1043 – 8.8%)
#5	app./x-hdf5 (22 – 0.6%)	app./fits – cf (82,204 – 14.9%)	Excel – xls (2074 – 6.7%)	n/a – doc (4511 – 1.4%)	n/a – zip (878 – 7.4%)
#6	app./x-gzip (19 – 0.5%)	n/a – adf (9833 – 1.8%)	n/a – n/a (2007 – 6.5%)	n/a – xlsx (4012 – 1.2%)	n/a – gz (616 – 5.2%)
#7	video/x-msvideo (10 – 0.3%)	n/a – dat (4920 – 0.9%)	text/plain – nex (1191 – 3.9%)	app./zip – zip (1988 – 0.6%)	n/a – csv (532 – 4.5%)
#8	video/mpeg (9 – 0.2%)	n/a – nit (4911 – 0.9%)	app./pdf – pdf (1097 – 3.6%)	n/a – csv (1422 – 0.4%)	n/a – csv (269 – 2.3%)
#9	app./x-gzip (8 – 0.2%)	image/tiff – tif (3926 – 0.7%)	app./x-gzip – gz (734 – 2.4%)	n/a – jpg (1395 – 0.4%)	n/a – itp (260 – 2.2%)
#10	app./zip (4 – 0.1%)	n/a – 001 (1637 – 0.3%)	app./x-fastq – fastq (728 – 2.4%)	n/a – cif (1108 – 0.3%)	n/a – ods (205 – 1.7%)
Distinct	53	1876	868	524	961

2.5 CONVERSION

Software to transform data from one format or structure to another is becoming increasingly prevalent. Safe Software's Feature Manipulation Engine (FME)⁸ is a case in point. Other third party tools and libraries proliferate and online services are available.

A key consideration when selecting (or developing) any conversion tool is precision: for example, a dataset with numerical values stored as double precision may be converted using single precision arithmetic resulting in lower resolution results or errors that cannot be recovered. Another consideration is text handling. Narrative text may be truncated. It may be converted using incorrect language character sets. It may even require translation into different languages with all the concomitant pitfalls any language translation brings.

⁸ <https://www.safe.com/>

3 DATA VOLUME

It is possible that the production of data from HYDRALAB+ activities will produce datasets of a size which begins to exceed the capacity of some commonly used mechanisms of storage and exchange. Indeed, the tendency towards larger and larger data sets will accelerate as the technology for observing and recording data about the real world (including that used in laboratory experiments) gets cheaper and improves in quality. The concept of 'big data' has become increasingly important for the EC, governments, businesses and public bodies as data is now a key asset for our economy and for society. According to the EC⁹:

Generating value at the different stages of the data value chain will be at the centre of the future knowledge economy.

The principles of FAIR data management (Wilkinson *et al*, 2016) and the requirement for open access to publications and data (EC, 2016, 2017) are intended to make our data more usable in the future, so that additional value can be generated from publically-funded data. In particular (Wilkinson *et al*, 2016) state that:

the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.

So, if we are to get the most out of data (by including data in meta-analyses of multiple datasets, for example) it needs to be managed in such a way that a computer (and not just an expert user) can find, access and manipulate the data.

The sheer quantity of data generated may require the use of software tools such as HDFS¹⁰, HADOOP¹¹ and NoSQL¹² to serve the requirements of the HYDRALAB+ community. Leveraging existing installations as services (such as Amazon Web Services) is likely to be the preferred approach in the near to medium future as the cost/benefit of large scale investment in data processing hardware is hard to justify, given the transient nature of such technology. Furthermore, the development of new techniques for analysing big data sets and being able to apply rigorous analysis to large and relatively unstructured data, such as topological data analysis, will encourage the drive to capture more data even in a domain hitherto considered non-sociological.¹³ To support this, large scale publicly available data repositories are emerging (Amazon Web Services¹⁴ is one example, while JASMIN¹⁵ in the UK is another) so in the future it will be feasible to host ever larger data sets on public service data repositories.

In practice, this may prove to be the most challenging of all the aspects of HYDRALAB+ data management as experiments begin to produce larger and larger data sets. Our current assessment

⁹ <https://ec.europa.eu/digital-single-market/en/policies/big-data>

¹⁰ <https://wiki.apache.org/hadoop/HDFS/>

¹¹ <https://wiki.apache.org/hadoop/FrontPage>

¹² <https://en.wikipedia.org/wiki/NoSQL>

¹³ <https://www.wired.com/2013/10/topology-data-sets/>

¹⁴ <https://aws.amazon.com/public-datasets/>

¹⁵ <http://jasmin.ac.uk/>

of the requirements of HYDRALAB+ activities suggest that the Zenodo data repository standard limit of 50GB per dataset will suffice for the time being, especially since each activity is usually able to spread its data production over multiple datasets.

4 METADATA, VOCABULARIES AND ONTOLOGIES

4.1 APPROACH

In an article on the format registry problem, McGrath (2013) highlights the complexities and difficulties of attempting to formalize a fast developing technology.

“File format identification is an important issue in digital preservation. Several noteworthy attempts, including PRONOM, GDFR, and UDFR, have been made at creating a comprehensive repository of format information. The sheer amount of information to cover and the constant introduction of new formats and format versions has limited their success. Alternative approaches, such as Linked Data and offering limited per-format information with identifiers that can be used elsewhere, may lead to greater success.”¹⁶

Notably, the amount of information and the constant introduction of new formats and format versions has limited the success of such formalization. This is notable because the problem also applies to the desire to produce domain relevant data standards, ontologies and vocabularies – the rapid rate of change in any developing area of science or technology precludes attempts to formalize much of the underlying language and terminology. As new concepts emerge some words will change their meaning to adjust, organically, to an emerging consensus. Accordingly, any standardization of vocabularies and ontologies must respect this and be applied in areas of high stability and where they can add intrinsic value to understanding and clarity.

In January 2011 António Cardoso Neto published a web article entitled *“A Brief Description of Some Standards for Hydrological Information”*¹⁷. This article examined the state of the art in 2011 and documents many standards for metadata, time-series and other data (therein lies the problem - by attempting to solve a standard language interchange issue the number of possible standards is now enormous and – significantly – they overlap).

The World Meteorological Organization published an article from its 14th session in 2012¹⁸ looking at the issue of data, access and exchange. This was following a concern expressed in its 2008 edition of the *“WMO Guide to Hydrological Practices”* to effect that *“There are currently no standards for data exchange formats for hydrological data”*. The article goes on to discuss, among other topics, OGC WaterML 2.0 and its extension of the Observations and Measurements (O&M) standard¹⁹. A good summary²⁰ of WaterML 2.0 is included but the following highlights the commonality of the issues at hand:

“Conformance to a common model such as O&M supports increased interoperability through standardisation of certain terms for observational metadata, including: feature-of-interest, observed-property, procedure, temporal metadata and result. This allows data from

¹⁶ McGrath, G (2013). The format registry problem: <http://journal.code4lib.org/articles/8029>

¹⁷ <http://www.whycos.org/wordpress/?p=353>

¹⁸ <http://www.wmo.int/pages/prog/hwrrp/chy/data-access-exchange.php>

¹⁹ ISO 19156:2011 Geographic information -- Observations and measurements. 2011.

²⁰ http://www.wmo.int/pages/prog/hwrrp/documents/DataOperationsandManagement_v1-0.pdf

disparate sources to be mutually understood and reused, is applicable to many exchange scenarios and makes it easier to share data across subject domains.”

So where the registry problem is indeed a problem it does not perforce mean ongoing attempts to provide versioned standards are pointless. We therefore consider some useful attempts to formalize the exchange of hydrological and related data.

4.2 METADATA, VOCABULARY AND ONTOLOGY STANDARDISATION

Reaching a common understanding of any terms is rarely easy – if, indeed, it is ever entirely possible. To meet this challenge, structured attempts to formalize the exchange and understanding of environmental datasets (including hydrological and related data) vary in approach and scope. Some focus on the structure and content of supporting metadata, others combine metadata elements with the data itself. Some focus on providing lists and categories whilst others concentrate on defining and structuring phenomena names and units. Initiatives also vary in domain coverage, each arising from a community dedicated to experimentation, earth observation or numerical modelling: in HYDRALAB+ terms: laboratory, field and numerical simulation.

The Dublin Core Metadata Initiative (DCMI²¹) proposes four levels of interoperability:

- Level 1 (Shared term definitions) - interoperability among metadata-using applications is based on shared natural-language definitions;
- Level 2 (Formal semantic interoperability) - interoperability among metadata-using applications is based on the model provided by RDF (Resource Description Framework) which is used to support Linked Data²²;
- Level 3 (Description Set syntactic interoperability) - applications are compatible with the Linked Data model and, in addition, share an abstract syntax for validatable metadata records; and
- Level 4 (Description Set Profile interoperability) - the records exchanged among metadata-using applications follow, in addition, a common set of constraints, use the same vocabularies, and reflect a shared model of the world.

The Dublin Core is perhaps one of the most successful and widely used ontologies partly because it tries not to encompass too much in its scope.

The UK Ordnance Survey developed a specific hydrological ontology *Ordnance Survey Hydrology Ontology V2.0* the purpose of which is to:

“describe in an unambiguous manner the inland hydrology feature classes surveyed by Ordnance Survey with the intention of improving the use of the surveyed data by our customers and enabling semi-automatic processing of these data”²³

and its stated scope is:

²¹ <http://dublincore.org/>

²² https://en.wikipedia.org/wiki/Linked_data

²³ <http://webarchive.nationalarchives.gov.uk/20090216171416/http://www.ordnancesurvey.co.uk/oswebsite/ontology/>

“Permanent topographic features involved in the containment and transport of surface inland water of a size of 1 metre or greater including tidal water within rivers. Functional, topological and meteorological relationships between these features are included. Physical characteristics are described to a level sufficient to discriminate between the concepts but no further.”

Where some parts of the ontology may be useful in other domains (for example, coastal, offshore or deep ocean water) the scope is decidedly limited and consequently the use of this ontology would need to be communicated clearly in any metadata describing a particular dataset so that the researcher does not waste time examining data which was outside the scope of their enquiry. One is – effectively – accepting the definition of the scope of Hydrology as determined by the Ordnance Survey at the time of drafting the ontology. Simply put, others may have a different idea of the scope.

The Open Geospatial Consortium has a working group specifically for the Hydrological domain: the Hydrology Domain Working Group or HydroDWG. It aims to bring together interested parties

“...to develop and promote the technology for greatly improving the way in which water information is described and shared.”²⁴

The group is working towards four parts for WaterML2:

- WaterML2: Part 1 - Timeseries
- WaterML2: Part 2 - Ratings, Gaugings and Sections
- WaterML2: Part 3 - Hydrologic Features
- WaterML2: Part 4 - GroundWaterML 2 (GWML2)

As of writing parts 1, 2 and 4 are published.

With any data format there is an inevitable compromise between efficiency of storage, speed of transmission or exchange and accuracy of understanding. Metadata itself is data that is relatively less efficiently stored and transmitted data used to define the content of efficiently stored and transmitted data. However, the problems facing the relative domains in HYDRALAB+ (field, laboratory and computer-generated data) are less concerned with the efficiency of exchange (as evidenced in deliverable D10.2 Critical Review) than with the commonality of understanding of data exchanged between domains.

Table 2 gives three established metadata standards, used to describe geospatial datasets in general.

²⁴ http://external.opengis.org/twiki_public/HydrologyDWG/

Table 2 Geographical and hydrological related metadata standards

Metadata Standard	Description
ISO 19139 (and 19115)	An XML (“eXtensible Markup Language”) standard, ratified by ISO. The objective was to help standardize the exchange of geographical data. It can be extended to fit specific needs and requirements.
CSDGM	Content Standard for Digital Geospatial Metadata – developed by USA Federal Geographic Data Committee.
Dublin Core	A vocabulary of fifteen properties for use in resource description – initially biographical in nature numerous extensions have been developed.

A variety of common vocabularies exist and are being continually developed to structure and describe environmental phenomena and units. These include:

- SeaDataNet²⁵, a “pan-European infrastructure for Ocean and Marine Data Management”. In addition to a set of aggregated data products; metadata catalogues of marine organisations, datasets, projects, observing systems, research cruises and data description (CDI); SeaDataNet gives a vocabulary library including the SeaDataNet Parameter Discovery Vocabulary and Agreed Parameter Groups – extensively categorized vocabularies for terms covering a broad spectrum of disciplines of relevance to the oceanographic and wider community, in particular to describe and categorize marine data phenomena.
- CF Standard Names²⁶, a list of climate and forecasting parameter names expressed in a standard form and accompanied by a description and canonical unit. It is intended for use within atmosphere, surface and ocean disciplines with model generated data and comparable observational datasets. Also provided is a related set of basic discovery metadata.
- CSDMS Standard Names²⁷, a list of surface dynamics parameter names expressed in a standard form and motivated by the need to pass standard parameters between numerical model components. CSDMS Standard Names uses a similar approach to CF Standard Names with the intention of creating unambiguous and easily understood standard variable names or preferred labels according to a set of rules.
- ITTC ‘Symbols and Terminology List’²⁸ defines many standard names for the testing of marine structures, including terms for waves and fluid flows. It comes from the International Towing Tank Conference (ITTC): an international association of organisations involved in ship and marine structure testing.

²⁵ <https://www.seadatanet.org/>

²⁶ <http://cfconventions.org/standard-names.html>

²⁷ http://csdms.colorado.edu/wiki/CSDMS_Standard_Names

²⁸ <https://www.ittc.info/downloads/quality-systems-manual/>

Structured lists of parameter names, such as provided by these initiatives, provide the simplest form of vocabulary control. Parameter (or phenomena) names can be included within data or metadata structures by a simple reference and mappings between different vocabularies can be made. The domain coverage provided by these vocabularies offers a large number of parameter name and unit combinations for use by experimenters within HYDRALAB+.

The actual structure of the data to which the metadata may refer has many options, two of which – WaterML2 and NetCDF are discussed in more detail below.

4.2.1 WaterML2

Notwithstanding the hidden file formats within zip files, notably absent from the list of popular data structures in

Table 1 is WaterML2²⁹. The description of this standard reads:

“WaterML2 is a new data exchange standard in Hydrology which can basically be used to exchange many kinds of hydro-meteorological observations and measurements. WaterML2 has been initiated and designed over a period of several years by a group of major national and international organizations from public and private sector, such as CSIRO, CUAHSI, USGS, BOM, NOAA, Kisters and others. WaterML2 has been developed within the OGC Hydrology Domain Working group which has a mandate by the WMO, too. ”

Interestingly, the list of objectives of WaterML2, while including *“providing a common exchange format for hydrological time-series”* and *“provide the option to fully store information including information regarding quality, validity/interpolation, and remarks”* feels the need to exclude a specific objective,

“currently it is NOT an objective to provide a comprehensive format with a minimum of characters”

So WaterML2 is an accessible, interoperable and reusable exchange format for time-series data and as such would be a good candidate for uptake by the distinct HYDRALAB+ domains. However, there is an awareness of the compromise between efficiency and comprehension.

The more general TimeseriesML³⁰ standard was developed from work originally undertaken within the OGC WaterML2.0: Part 1 – Timeseries activity. It defines an XML encoding that implements the OGC Timeseries Profile of Observations and Measurements [OGC 15-043r3], with the intent of allowing the exchange of such data sets across information systems. It is intended to be re-used to address a range of data exchange requirements and so aims at meeting the HYDRALAB+ WP10 objective of producing interoperable data. It is also more generic than WaterML2.0: Part 1 – Timeseries and so may be more applicable to the scientists working within the wide ranging HYDRALAB+ framework. It is also clear that TimeseriesML, like WaterML2, has not been developed with the objective of minimising characters.

The issue with comprehensive, XML encoded standards aimed at a global audience, such as WaterML2: Timeseries and TimeseriesML, is that they can be impenetrably difficult to understand. Asking scientists who are not information management professionals to produce data from their experiments in such formats requires a long learning curve. Adopting them, therefore, requires commitment from the organisations to provide an environment such that this activity is accessible to its practitioners in a reasonable timescale. However, once locally produced examples are prevalent and local terminology adopted, the learning curve can be reasonable.

4.2.2 NetCDF

The Network Common Data Form, or netCDF, is described as follows:

²⁹ <http://www.waterml2.org/>

³⁰ <http://www.opengeospatial.org/standards/tsml>

“NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.”³¹

“NetCDF (network Common Data Form) is a set of interfaces for array-oriented data access and a freely distributed collection of data access libraries for C, Fortran, C++, Java, and other languages. The netCDF libraries support a machine-independent format for representing scientific data. Together, the interfaces, libraries, and format support the creation, access, and sharing of scientific data.”³²

The position of netCDF in table 1 demonstrates the popularity and prevalence of it amongst the scientific community.

Targeted at array-based data, netCDF is an example of a data format which includes a strong element of metadata as part of a single package. As such, it overlaps requirements for data formats, and metadata structures and content. It stores the data itself in a binary format making it a practical choice for large – but not ‘Big Data’ – datasets. It supports access by computers which store integers, characters and floating point numbers in different ways. Data can be appended to a netCDF file and efficiency exists for accessing a small subset of the whole dataset. The netCDF package also includes tools for converting netCDF files into ASCII or text files.

NetCDF also has associated metadata conventions, an implementation of the afore mentioned Climate and Forecasting standard name table³³ (CF Standard Names). The “NetCDF CF Metadata Conventions” are designed to promote the processing and sharing of netCDF files, created with the NetCDF API:

“The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities.”

Overall, netCDF is a practical choice for storing large volume, array-based scientific data.

4.3 DATA REPOSITORY KEYWORDS

The deliverable “D10.4 Data Repository rules” (including Digital Object Identifiers (DOIs)) is being developed on the HYDRALAB+ website as an interface to the Zenodo data repository. As part of that interface, a keyword list is being developed to allow HYDRALAB+ users tag their data deposits with relevant contextual keywords for later indexing and searching.

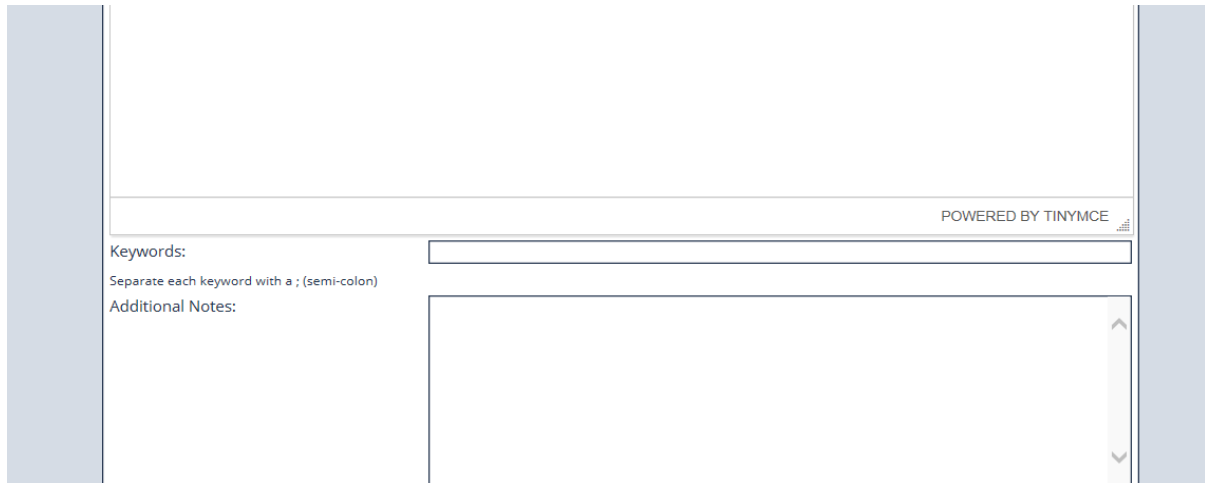
Such a vocabulary of keywords will be allowed to grow organically and is expected to feature HYDRALAB+ context specific terminology. Indeed, a facility to allow HYDRALAB+ partners to add and edit their own keywords will provide a mechanism for capturing the language terms in frequent use by the HYDRALAB+ community. It is intended that subsequent and continuing analysis and

³¹ <https://www.unidata.ucar.edu/software/netcdf/>

³² <https://www.unidata.ucar.edu/software/netcdf/docs/faq.html#whatisit>

³³ <http://cfconventions.org/index.html>

maintenance of this list will provide a valuable asset to increasing our understanding of the domain interfaces (i.e. field <-> lab <-> computer <-> field).



POWERED BY TINYMCE

Keywords:

Separate each keyword with a ; (semi-colon)

Additional Notes:

Figure 1. Transnational Access Projects, Add New Dataset Form, Keywords entry

Figure 1 shows the keywords entry in the form used when uploading a new dataset to the Hydralab+ Transnational Access Projects section on the Zenodo portal. At the time of writing, the implementation does not include keywords from a managed list. All keywords are selected and entered by the users on creating the dataset record in Zenodo.

5 LICENSES AND EMBARGO PERIODS

HYDRALAB+ deliverable D10.1 Data Management Plan stipulates the use of open licenses and suitable embargo periods for use with data generated or used or otherwise published as a result of HYDRALAB+ research and activities.

The description of work (Task 10.2 Data Standards and licenses) states:

“An approved license (or licenses) and a common embargo period, where the originators have exclusive access to their data before it is made open, will be chosen early on in the project.”

The Open Definition organisation provides a reference for licenses which conform to the principles laid out in the Open Definition³⁴. The conformant licenses³⁵ provide a commonly accepted mechanism for equably sharing data in the public domain.

An embargo period may be attached to data placed in a repository to protect Intellectual Property Rights (IPR) for an appropriate duration – this is referred to as a “Green” route to “Open Access” by the EC (2017) and HEFCE³⁶

An article from 2012 on the Scholarly Kitchen site³⁷ began a discussion on Open Data and embargo periods. The point is made that apparent disinterest in access can often conceal poor ease of access (in particular sufficiently useful cataloguing and indexing). By providing good – useful – metadata, cataloguing, indexing and searching facilities, the HYDRALAB+ branded interface to the Zenodo data repository will ensure that the selected embargo period for data sets becomes more relevant to the IPR of the data itself.

5.1 LICENSES

The Open Data Institute guides to Open Data Licensing states:

“Data that doesn’t explicitly have an open license is not open data.”³⁸

All HYDRALAB+ published data will become open data^{39,40} and licenses will conform to the open definition⁴¹ that is suitable for data. A list of suitable licenses can be found at <http://opendefinition.org/licenses/>.

Guides on licensing research data have also been provided by the H2020 Online manual⁴², DMP Online⁴³, the Open University⁴⁴, the University of Bath⁴⁵ and others. If a Creative Commons license is

³⁴ <http://opendefinition.org/od/2.1/en/>

³⁵ <http://opendefinition.org/licenses/>

³⁶ <http://www.hefce.ac.uk/rsrch/oa/whatis/>

³⁷ <https://scholarlykitchen.sspnet.org/2012/09/18/open-access-embargoes-how-long-is-long-enough/>

³⁸ <https://theodi.org/guides/publishers-guide-open-data-licensing>

³⁹ <http://opendefinition.org/guide/data/>

⁴⁰ <https://okfn.org/opendata/>

⁴¹ <http://opendefinition.org/od/2.1/en/>

used then it should be at least version 4, as earlier versions did not cater well for data. Licenses often allow licensors to impose a number of restrictions on the use of their data, which have advantages and disadvantages. These include:

- Attribution: the user of data must give due credit to the providers of the data whenever it is used, displayed or published. This is nice for the creator of the data but becomes a problem when a lot of datasets are used and the list of contributors become unwieldy. Some pieces of open source software have had hundreds of contributors, for example, while comparative studies of datasets in medicine could also potentially involve hundreds of contributors. Attributions can be put on a website, with a link provided by the user, which reduces the impact on a paper, but still leaves a potentially unwieldy procedure.
- No derivatives: the data may be redistributed, whole and unchanged, to anyone for any purpose. Licenses do not distinguish between using datasets to derive combined datasets, derived data or graphs. A no-derivatives clause could be interpreted as meaning that a user cannot derive graphs or parameterisations from datasets. It therefore potentially restricts the re-use of data, which goes against the principles of FAIR data management, so is discouraged.
- Share-alike / Copyleft: others can use the licensor's data to create new datasets / products that must be licensed under the same terms and conditions. This causes a problem when two or more data sources are used with different share-alike / copyleft licenses. Each license demands that the derived product is distributed through their license and their license only. This is impossible, so either some of the contributing datasets must be omitted or the licensor acts in breach of one or more license terms. This condition reduces the interoperability of the data, which goes against the principles of FAIR data management, so is also discouraged.
- Non-commercial: this allows others to use the licensor's data and build upon it for non-commercial purposes. According to creative commons, non-commercial use is 'primarily intended for or directed toward commercial advantage or monetary compensation'. A non-commercial clause is often used with a dual license (with free non-commercial use and potentially paid-for commercial use). Open data can be "freely used, modified, and shared by anyone for any purpose"⁴⁶. A non-commercial clause contravenes the definition of open data by restricting use, so should not be used for any dataset that is to be made open.

Overall, it is desired that licenses used by HYDRALAB+ experimenters be as simple and practical as possible, whilst fulfilling all necessary core requirements.

⁴² http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

⁴³ <http://www.dcc.ac.uk/resources/how-guides/license-research-data>

⁴⁴ <http://www.open.ac.uk/library-research-support/research-data-management/licensing-research-data>

⁴⁵ <http://www.bath.ac.uk/research/data/sharing-data/licensing/>

⁴⁶ <http://opendefinition.org/>

5.2 OPEN ACCESS AND EMBARGOES

It is well worth reading the Participant Portal H2020 Online Manual section on Open Access (OA)⁴⁷ which covers publications and data. Each is discussed in turn below.

5.2.1 OA to publications

Under article 29.2 of our grant agreement “each beneficiary must ensure open access to all peer-reviewed scientific publications”. The OA mandate comprises 2 steps:

1. “Beneficiaries must deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications. This must be done as soon as possible and at the latest upon publication.”
2. “Beneficiaries must provide open access to them, normally by offering self-archiving (Green Open Access) or through Open Access Publishing (Gold Open Access).”

Participant Portal H2020 Online Manual section on Open Access states explains the two main routes to OA as:

- “Self-archiving / 'green' open access – the author, or a representative, archives (deposits) the published article or the final peer-reviewed manuscript in an online repository before, at the same time as, or after publication. Some publishers request that open access be granted only after an embargo period has elapsed.
- Open access publishing / 'gold' open access - an article is immediately published in open access mode. The most common business model is based on one-off payments by authors. These costs, often referred to as Article Processing Charges (APCs) are usually borne by the researcher's university or research institute or the agency funding the research.”

The EC requires an embargo period of no more than 6 months in our field (similar to some funders⁴⁸) while many publishers require an embargo period of 12 to 24 months (e.g. ⁴⁹) despite knowing the views of the EC and other funders. To combat this the EC provides a model amendment⁵⁰ to the publisher's copyright agreement, which authors have to sign before publication, to allow self-archiving within 6 months. If publishing by Green OA, the model amendment should be used if your publisher has an embargo period of more than 6 months.

Clearly, there are stakeholders in the embargo domain including (but not limited to) publishers, funders and researchers. All of these have different reasons for wishing for different durations of embargo for different sets of data or publications. The tension between the stakeholders' different interests and requirements leads to a relatively contentious debate. However, Task 10.2.3 states:

“Selection of an open data license(s) and embargo period... An approved license (or licenses) and a common embargo period, where the originators have exclusive access to their data before it is made open, will be chosen early on in the project.”

⁴⁷ http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

⁴⁸ <http://www.rcuk.ac.uk/documents/documents/rcukopenaccesspolicy-pdf/>

⁴⁹ https://www.elsevier.com/__data/promis_misc/external-embargo-list.pdf

⁵⁰ http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-oa-guide-model-for-publishing-a_en.pdf

5.2.2 OA to data

The requirement for OA to research data is set out in section 29.3 of our grant agreement:

“the beneficiaries must:

- *deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:*
 - *the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;*
 - *other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan';*
- *provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).”*

As noted in Section 1.1, deliverable D10.4 – ‘Data Repository Rules (including DOIs)’ – supports use of the Zenodo repository for storing experimental data (although other suitable repositories are available and can be used – see <https://www.openaire.eu/search/data-providers>). Zenodo is set up to meet the requirements of H2020. The last point in making data available encourages researchers to make available ‘tools and instruments’ needed to validate results. In most cases this will be computer code, needed to read, filter and process data. In other words, you are encouraged to deposit

- raw data;
- the code needed to calibrate and filter raw data; and
- the code needed to process the calibrated data into the results presented in a paper.

A practical approach should be taken, particularly with very large raw datasets, such as those from Particle Image Velocimetry (PIV) and other optical techniques. A standard dataset on Zenodo can be up to 50GB, while a PIV experiment can capture over 1TB of data. Moreover, this volume of data takes a long time to process (often using proprietary software). In these consideration should be given to depositing the calibrated result files in a research data repository and noting that the raw data could be obtained from its owner, if required.

6 RECOMMENDATIONS

The stated purpose of Task 10.2 'Data standards and licenses' is to ensure that the data produced by HYDRALAB+ experiments is interoperable and so easily exchanged between researchers. The intrinsic benefits of achieving these improved data management practices within HYDRALAB+ will clearly have a further impact on all experiments undertaken at the partner organisations.

The needs of scientists who wish to use data provided by other scientists can be characterised as follows:

- Can I find and obtain data which may be useful to me?
- Can I open the dataset?
- Do I know what is in the dataset?
- Can I evaluate whether this data is useful?
- May I use this data?

The recommendations given here are targeted at scientists preparing data for others to use, so that the answers to the above questions are positive for those who may use the data. They concern data standards and licenses, outline a set of sensible data management principles, in particular with reference to the benefits of choosing open data structures and formats and communicating that choice to others via metadata. They respect the wide variety of data management options for formats and structures; metadata, vocabularies and ontologies; and licenses and embargo periods, many of which have been outlined as leading examples in the discussions above. Where appropriate, specific technologies have been offered, but sound data management principles designed to educate researchers and improve their management of data have also been included.

These recommendations also respect the wide variety of technologies embedded within the partner organisations. They do not seek to impose an unrealistic set of rules and regulations which must be followed, rather they offer a set of sensible, modern principles and resources to move the community forwards together and bring it in line with other similar communities currently iterating their own data management practices. They also dovetail with the project's usage of the Zenodo data repository for the storage of experiment results datasets.

The recommendations given in the subsections which follow are distilled down to a set of simple statements, articulated as data management recommendations targeted at the scientists themselves. The supporting information in this report is also available.

6.1 ZENODO

"Can I find and obtain data which may be useful to me?"

Those performing experiments as part of the HYDRALAB+ project are asked to store the dataset outputs of their experiments in the Zenodo repository. This process will automatically give the dataset a DOI to allow it to be uniquely referenced.

Recommendation 1: Store your results datasets in the Zenodo repository including the accompanying metadata. Do this through the Hydralab+ website form.

- This will give your dataset a unique DOI which you can use to reference it.
- For Transnational Access datasets or publications use the form provided on the HYDRALAB+ website (<http://hydralab.eu/>). Click on the 'Add Dataset' option under 'TAKING PART', 'Transnational Access Projects'.
- For Joint Research Activities or Networking, log into the hydralab website and go to the participant area at <http://hydralab.eu/participant-area/> then select "DOI datasets".
- The form itself contains some additional metadata required by the repository.

6.2 DATA AND METADATA FORMAT

"Can I open the dataset?"

It is important to select a good file format for storing data: one which is appropriate for the data structure and size; one which is usable and sustainable.

Recommendation 2: Select a format for your data which respects its structure and size.

- Does the data format match the natural data structure of your data (i.e. flat, hierarchical, multidimensional)?
- Does the data format allow you to comfortably store the entire final dataset?

Recommendation 3: Select a format for your data which will be accessible to other scientists, now and in the future.

- Is the data format broadly understood within your community and acceptable to funders?
- Is the data format supported by other communities and likely to be compatible with future common operating systems and applications?
- Is there a broad range of software that can read / write the data format?
- Are the terms and conditions of the license for the read / write software favourable? Is it free? Is it proprietary?
- Is the conversion process from the data format to / from other formats cheap and easy?

Leading formats for long time-series data include TimeseriesML or WaterML2: Part 1 - Timeseries. A leading format for larger, multi-dimensional array-based data is netCDF.

"Do I know what is in the dataset?"

All data should be accompanied by adequate metadata (supporting information) so that future users can understand and apply the contents. Sometimes adequate metadata is stored within the data structure format, sometimes an additional file is required.

Recommendation 4: Include sufficient metadata to allow your data to be interpreted by other users. If possible use an established metadata standard.

- The minimum information provided should be the fifteen elements given in Dublin Core⁵¹: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type.
- Use the MIME type vocabulary to describe the Format.
- Remember, the next person to use the dataset is likely to be you! How will you understand this data a year from now?
- Include a README.txt file to describe the files in your data package.

Leading metadata standards are Dublin Core (ISO 15836:2009) and ISO19115/19139.

Recommendation 5: Take all parameter names and units from established vocabularies.

- Avoid meaningless field names and remember to include the units.
- When you use a vocabulary to describe parameters, include a reference to its on-line record.

Leading vocabularies include SeaDataNet, CF Standard Names, CSDMS Standard Names and ITTC Symbols and Terminology List.

“Can I evaluate whether this data is useful?”

Recommendation 6: Include information which helps others evaluate whether it is useful to them.

- Include information such as a brief overview, the objectives and context of the work, brief conclusions and outstanding questions. This will help potential users quickly understand whether your data would be useful for them to investigate further.
- Include links to more comprehensive reports and papers which reference the data package. Link from the papers back to the data package.

6.3 LICENSE AND EMBARGO PERIOD

“May I use this data?”

An appropriate license and embargo period must be included and articulated alongside each dataset stored as part of HYDRALAB+.

Recommendation 7: Include an open license, with as few restrictions as possible, to allow others to use your data.

- A suitable list of licenses is given here: <http://opendefinition.org/licenses/>. The default license for HYDRALAB+ is the Creative Commons Attribution 4.0 International (CC BY 4.0) license⁵².

Recommendation 8: If an embargo period is required, for HYDRALAB+ experiment and research activity publication, you should select an embargo period of not more than two years.

- The embargo period should be included in the data storage report.

⁵¹ <http://www.dublincore.org/documents/dces/>

⁵² <https://creativecommons.org/licenses/by/4.0/>

7 ACRONYMS AND ABBREVIATIONS

API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
BOM	Bureau of Meteorology
CF	Climate and Forecasting
CSDGM	Content Standard for Digital Geospatial Metadata
CSDMS	Community Surface Dynamics Modeling System
CSIRO	Commonwealth Scientific and Industrial Research Organisation (Australia)
CSV	Comma Separated Values
CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science, Inc.
DCMI	Dublin Core Metadata Initiative
DMP	Data Management Plan
DOI	Digital Object Identifier – a unique identifier identifying a specific dataset within the context of an agreed community.
FAIR	Findable, Accessible, Interoperable and Reusable
FME	Feature Manipulation Engine
HEFCE	Higher Education Funding Council for England
IANA	Internet Assigned Numbers Authority
IPR	Intellectual Property Rights
ITTC	International Towing Tank Convention
JRA	Joint Research Activities
MIME	Multipurpose Internet Mail Extension
netCDF	network Common Data Form
NOAA	National Oceanic and Atmospheric Administration
O & M	Observations and Measurements

OGC	Open Geospatial Consortium
USGS	United States Geological Survey
WMO	World Meteorological Organisation
XML	eXtensible Markup Language

8 REFERENCES

Assante, M. et al., (2016). Are Scientific Data Repositories Coping with Research Data Publishing?. Data Science Journal. 15, p.6. DOI: <http://doi.org/10.5334/dsj-2016-006>

EC (2016) H2020 Programme, Guidelines on FAIR data Management in Horizon 2020. Version 3.0, July 2016.

EC (2017) H2020 Programme, Guidelines to the rules on open access to scientific publications and open access to research data in Horizon 2020. Version 3.2, March 2017.

HR Wallingford (2017a). Data Management Plan. HYDRALAB+ deliverable 10.1.

HR Wallingford (2017b). Critical Review of data flux between laboratory models, numerical models and field case studies. HYDRALAB+ deliverable D10.2.

McGrath, G (2013). The format registry problem: <http://journal.code4lib.org/articles/8029>

Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Nature <https://www.nature.com/articles/sdata201618.pdf> DOI: 10.1038/sdata.2016.18