Interoperability in physical model testing

James Sutherland[1], Quillon Harpham[2]

[1] HR Wallingford, Howbery Park, Wallingford, Oxfordshire OX10 8BA, UK (Email: j.sutherland@hrwallingford.com)
[2] HR Wallingford, Howbery Park, Wallingford, Oxfordshire OX10 8BA, UK (Email: q.harpham@hrwallingford.com)

Proceedings of the 7th International Conference on the Application of Physical Modelling in Coastal and Port Engineering and Science (Coastlab18), Santander, Spain, May 22-26, 2018

Abstract

The funders of research programmes, such as Horizon 2020 are increasingly requiring that the resulting publications and data are made openly available. The EC, for example, requires FAIR (Findable, Accessible, Interoperable, Reusable) data management. This is promoted through a set of principles and guidelines for experimenters to follow. These respect the technologies and intentions at each organisation, 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 use of data repositories for the storage of data and papers.

Keywords

Data management, open access, interoperable, metadata, license

1. Introduction

The funders of research programmes, such as Horizon 2020 (EC, 2016, 2017) are increasingly requiring that the resulting publications and data are made openly available. The EC requires FAIR (Findable, Accessible, Interoperable, Reusable) data management (EC, 2016, Wilkinson et al., 2017) and this paper examines how the principle of interoperability can be applied to laboratory hydraulic modelling, within the context of the HYDRALAB+ project (www.hydralab.eu).

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.

However, a critical review of data flux between laboratory models, numerical models and field case studies (Cleverley et al., 2018) has 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.

In order to address this lack of knowledge, this paper provides a data management framework (Harpham et al., 2018) for scientists involved in hydraulic laboratory research (Figure 1). Attention is given, not only to the needs of those creating the data, but specifically to the subsequent usage of it. 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 is to put themselves in the position of a new user who is attempting to find, access and process the data that they create.

This is a need to guide 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.

Figure 1: Experiment in the Fast Flow Facility at HR Wallingford

Vocabularies for categories, general terminology and phenomena / parameter names and units are commonly implemented to support discovery and aid understanding and have been 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, data should not be exchanged 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 use of a data repository, such as Zenodo, to store and provide reference to the data package itself.

A number of issues are considered in more detail in Sections 2 to 6, before a discussion leads into a set of recommendations.

2. Data formats and structures

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.

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. 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.

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.

One 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.

On the other hand, 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 and comes with associated metadata.

Moreover, where the nature of the data acquisition requires proprietary equipment 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.

The Multipurpose Internet Mail Extension (MIME) standard 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 described simply, in order for computer systems (and humans) understand how to deal with the dataset in question. The use of the MIME type vocabulary to describe file formats is recommended.

3. Data volume

Modern measurement techniques, such as terrestrial and underwater laser scanners, acoustic velocity meters and optical techniques, such as 3D-3C Particle Image Velocimetry (PIV) are producing datasets of a size which begins to exceed the capacity of some commonly used mechanisms for 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 EC (https://ec.europa.eu/digital-single-market/en/policies/big-data): "Generating value at the different stages of the data value chain will be at the centre of the future knowledge economy."

However, these large datasets provide practical problems in managing the data and making it open access (downloadable). A practical approach should be taken. A standard dataset on the Zenodo repository can be up to 50GB, while a PIV experiment can capture over 1TB of data relatively quickly. Moreover, this volume of data takes a long time to process (often using proprietary software). In these cases consideration should be given to depositing the just the calibrated result files in a research data repository and noting that the raw data could be obtained from its owner, if required.

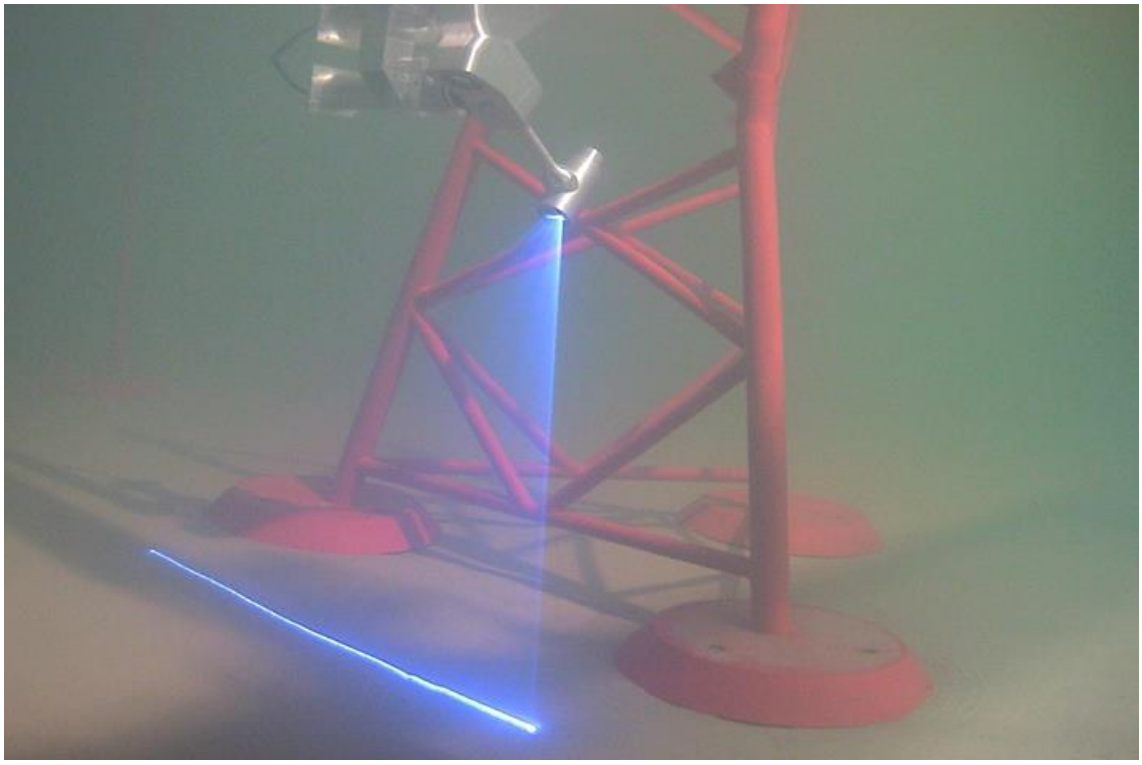

Figure 2: Modern data acquisition systems, such as this Underwater Laser Scanner, collect large volumes of data

In time, the sheer quantity of data generated may require the use of software tools such as HDFS , HADOOP and NoSQL to serve the requirements of hydraulic laboratories. To support this, large scale publicly available data repositories are emerging (Amazon Web Services is one example, while JASMIN in the UK is another) so in the future it will be feasible to host ever larger data sets on public service data repositories.

4. Metadata, vocabularies and ontologies

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.

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, which is 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 (Cleverley *et al*, 2018) than with the commonality of understanding of data exchanged between domains.

Many metadata standards exist. Dublin Core (http://dublincore.org/) is perhaps one of the most successful and widely used standards partly because it tries not to encompass too much in its scope. It consists of a

vocabulary of fifteen properties for use in resource description – initially biographical in nature numerous extensions have been developed.

Describing the variables measured using standard vocabularies reduces confusion over what has been measured and increases the interoperability of a dataset. A variety of common vocabularies exist and are being continually developed to structure and describe environmental phenomena and units. These include:

- SeaDataNet - https://www.seadatanet.org/ - is 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 Names - http://cfconventions.org/standard-names.html - 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 Names - http://csdms.colorado.edu/wiki/CSDMS_Standard_Names - 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' - https://www.ittc.info/downloads/quality-systems-manual/ - 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.

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.

Users and owners of vocabulary sets are often tempted to create a set of categories to accompany the parameter / phenomena names. This activity might be seen as inevitable and useful for understanding trends and performing quantitative assessments of activity within such categories. However, placement of a data entity in a category can often be arbitrary and misleading. Moreover, users often do not understand the categorisation – which may be written in esoteric terms – or allocate a category haphazardly to save time or for purposes of expediency. Therefore it is recommended that keywords be attributed to data entities rather than forcing them into a categorised arrangement. These keywords can then easily be picked up by search engines or, if necessary, engines creating reported statistics.

5. Licenses and embargo periods

The Open Data Institute guides to Open Data Licensing[1] states that "data that doesn't explicitly have an open license is not open data." Without a license, the recipient does not know what (s)he can do with the date received. Without an open license, the recipient is likely to face restrictions on the use of a dataset. 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 manual[2], DMP Online[3], the Open University[4], the University of Bath[5] and others. If a Creative Commons (CC) license is 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 (BY): 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. Despite this, the CC-BY 4.0 license is used by HYDRALAB+.
- No derivatives (ND): 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 (SA): 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 though their license and their license only. This is impossible, so share alike licenses are also discouraged.
- Non-commercial (NC): this allows others to use the licensor's data and build upon it for non-commercial purposes. Open data can be "freely used, modified, and shared by anyone for any purpose" . 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 experimenters be as simple and practical as possible, whist fulfilling all necessary core requirements.

Researchers may want to impose an embargo period on their data, so that they have the first opportunity to produce results and papers from their experiments. The length of these should be as short as possible to allow data to become widely available as soon as possible. It can be argued that an embargo period contravenes the spirit of open access, but they remain popular as a means of giving those who worked on an experiment the first opportunity to publish and generate impact from their work. An embargo date can be set

[1] https://theodi.org/guides/publishers-guide-open-data-licensing

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

[3] http://www.dcc.ac.uk/resources/how-guides/license-research-data

[4] http://www.open.ac.uk/library-research-support/research-data-management/licensing-research-data

[5] http://www.bath.ac.uk/research/data/sharing-data/licensing/

in a repository like Zenodo, so that data (or a paper) will automatically become available on that date. Metadata about the paper or data will be available in the meantime.

6. Open access through a repository

Data and papers can be made open access by uploading them onto a suitable repository, such as those listed at see https://www.openaire.eu/search/data-providers. One such repository is Zenodo (https://zenodo.org/) which is funded by OpenAire (https://www.openaire.eu/) 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 the platform to be integrated with a project website. This has been done with the HYDRALAB+ website and ensures that publications and data submitted to Zenodo through the HYDRALAB+ website are included in the HYDRALAB+ community on Zenodo.

The Horizon 2020 model grant agreements require open access to all peer-reviewed publications, which can be achieved in two steps:

1. deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript in a repository; and
2. provide open access by offering self-archiving (Green Open Access) or through Open Access Publishing (Gold Open Access).

The requirement for Open Access to research data is set out in section 29.3 of the H2020 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)."

7. Discussion

Data that is interoperable is easily exchanged between researchers, but this requires data management practices that are not familiar to many laboratory researchers. 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 below 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. 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 HYDRALAB+ 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.

8. Conclusions

In order to make data easier to exchange, the following set of eight data management recommendations should be followed. Supporting information can be found in Harpham *et al.* (2018).

Recommendation 1: Store your results datasets in a recognized research data repository, such as Zenodo. Include the accompanying metadata. This will give your dataset a unique DOI which you can use to reference it and will ensure long-term preservation of your dataset.

Recommendation 2: Select a format for your data which respects its structure and size. Chose a data format that matches the natural data structure of your data (e.g. flat, hierarchical, multidimensional). Consider if the data format allows 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. This involves consideration of the following questions:

- 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?

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 Core: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type – see http://www.dublincore.org/documents/dces/ for details.
- Use the MIME type vocabulary to describe the Format.
- Get someone who was not involved in the experiment to look at your dataset. If they cannot understand it, it is not clear enough.

- Include a README.txt file to describe the files in your data package.

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

- Avoid meaningless field names and remember to include the units.
- Where possible, avoid allocating your data package to a category, instead describe it with a comprehensive set of well-constructed keywords.
- 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.

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.

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) license .

Recommendation 8: If an embargo period is required, it should be as short as possible and certainly no more than two years. The embargo period should be included in a data storage report, any data paper and in the research repository metadata (so your data is automatically made available when the embargo ends).

Acknowledgement

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

References

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

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.

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

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.