Open data from physical model tests: Lessons learned from related initiatives

James Sutherland[1], Chris Awre[2], Gonzalo Malvarez[3], Anna Van Gils[4] & Quillon Harpham[1]

[1] HR Wallingford, Wallingford, Oxfordshire OX10 8BA, UK
[2] The University of Hull, UK.
[3] Universidad "Pablo de Olavide", Seville, Spain.
[4] Deltares, Delft, The Netherlands.
(Email: j.sutherland@hrwallingford.com, q.harpham@hrwallingford.com, c.awre@hull.ac.uk, gcmalgar@upo.es, anna.vanGils@deltares.nl)

E-proceedings of the 37th IAHR World Congress, August 13 – 18, 2017, Kuala Lumpur, Malaysia

Abstract

The HYDRALAB network of European physical model laboratories (www.hydralab.eu) has a range of facilities that includes flumes, basins, ice facilities, rotating tanks and environmental facilities. Each institution had its own data collection system, there are many proprietorial data formats, a shortage of meta-data and no central effort to curate or preserve this data in a findable, accessible, interoperable and reusable (FAIR) way. HYDRALAB+ (2015-2019) is a European Commission Horizon 2020 project to support this network, which requires FAIR data management. HYDRALAB is reviewing the steps taken to make data openly accessible in related disciplines, so that lessons learned can be applied to HDRALAB+. The chosen communities were: (i) the University of Hull's digital repository, (ii) EMODnet Baltic Checkpoint, (iii) OpenEarth and (iv) the FP7 projects PEGASO and MEDINA and the EU MED project COASTGAP. It is clear that no one solution can deal with all situations: different data types and requirements can best be dealt with by different approaches. Standards for meta-data should be applied, but no existing standard covers the range of situations faced by HYDRALAB. All can be extended in a bespoke manner (which can potentially be included in an update of the standard) but it is highly likely that more than one standard (and none) will be used in such a diverse community. This is perfectly acceptable, so long as the standard is published. There is also a clear need for guidance on the development of repositories where large volumes of data are collected and an understanding of how much needs to be made available on-line. Although there can be conflicts of interest between institutions that are developing policies for data management and projects that want a uniform approach to data management across all partners, systems today can generally accommodate this.

Keywords

Open data, open access, data management, Horizon 2020

Introduction

We are moving towards an era of more open science, consisting of open source software, open software architecture, open access to data, open access publication and open collaboration (Sutherland and Evers, 2013). Open access publication and open access to data are increasingly becoming requirements of funding agencies, including the European Commission (EC) Horizon 2020 programme, Research Councils in the UK, the Bill and Melinda Gates foundation and many others. The EC is taking extensive steps promoting the development and uptake of Open Science. For example, the EU Competitiveness Council (May, 2016) stresses that “open science entails amongst others open access to scientific publications and optimal reuse of research data”. Moreover, the outcomes of the Open Science Conference (April 2016) were summarised in the ‘Amsterdam Call for Acton on Open Science’ which includes the goals for 2020 of making all scientific publications fully open access and making the sharing of data the standard position (using definitions, standards and infrastructures). Open access to publications is an obligation in Horizon 2020, while open access to data is encouraged (with a pilot being extended to all areas in the 2017 work programme).

The HYDRALAB network of physical model laboratories (www.hydralab.eu) has 24 partners and 8 associated partners across Europe with a range of laboratory facilities that includes flumes, basins, ice facilities, the Coriolis rotating tank and environmental facilities. The HYDRALAB+ project aimed at strengthening the coherence of experimental hydraulic and hydrodynamic research by improving infrastructure with a focus on adaptation to climate change issues. HYDRALAB+ has three key objectives:

1. to widen the use of, and access to, unique hydraulic and hydrodynamic research infrastructures in the EU through the Transnational Access (TA) programme, which offers researchers the opportunity to undertake experiments in rare facilities to which they would not normally have access;
2. to improve experimental methods to enhance hydraulic and hydrodynamic research and address the future challenges of climate change adaptation, through our programme of Joint Research Activities (JRAs). The JRAs are undertaking R&D to develop and disseminate tools and techniques that will keep European laboratories at the forefront of hydraulic experimentation; and
3. to network with the experimental hydraulic and hydrodynamic research community throughout Europe and share knowledge, best practice and data with the wider scientific community and other stakeholders, including industry and government agencies. Some training will also be provided to the next generation of researchers.

HYDRALAB+ is a voluntary member of the H2020 Open Data Pilot. This requires participants to make their publications open access and their research data Findable, Accessible, Interoperable and Reusable – FAIR for short (EC, 2016). However, each institution had its own data collection system, there are many proprietorial data formats, a shortage of meta-data and no central effort to curate or preserve this data in a FAIR way. General guidelines on FAIR data management can be found in EC (2016), the H2020 online manual section on Open Access and Data Management and the H2020 Annotated Model Grant Agreement.

HYDRALAB is reviewing the steps taken to make data openly accessible in related disciplines, so that lessons learned can be applied to HDRALAB+. The chosen communities were:

- Hydra, the University of Hull’s digital repository. Hydra has been developed “to hold, manage, preserve and provide access to the growing body of digital material generated through the research, teaching and administrative activities of the University” (https://hydra.hull.ac.uk/). This system must cope with results from many different fields of research.
- EMODnet Baltic Checkpoint, which is (i) examining the data collection, observation, surveying, sampling and data assembly programs in the Baltic Sea basin (ii) assessing the usefulness of the data in 11 challenge areas in terms of data uncertainty, availability, accessibility and adequacy; and (iii) delivering the findings to stakeholders through an internet portal with dynamic mapping features and a stakeholder workshop.
- OpenEarth has developed workflows to deal with data management. The raw data coming from the measurement devices is stored in subversion together with a description of the format and the scripts

used to process the data. Processing data to standard formats is encouraged, as this allows other people to access and use the data easily. Also processed data can be stored, such as the significant wave height, or a velocity field derived from a PIV experiment.

- Interoperability principles to integrate coastal mapping in science and management, which have been developed and applied in various projects including FP7 projects PEGASO and MEDINA (www.medinaproject.eu) which dealt with integrating advance mapping tools for marine ecosystem indicators (www.medinageoportal.eu) and in EU MED project COASTGAP which further developed integrating data and information from various Regional Governments responsible for coastal engineering in various locations in the Mediterranean.

This paper summarises the approaches of these communities.

1. Data collection in HYDRALAB

Traditionally, each institution has used its own data acquisition, storage and analysis systems. There has been a shortage of meta-data collected, little standardisation and little consideration of data exchanges. The HYDRALAB initiatives that have considered data exchange and data management are summarized below.

The HYDRALAB-III Data Management Tools report (Wells et al., 2009) made a number of recommendations, including that HYDRALAB build on existing technologies, develop a strategic view for data management, establish best practice for the documentation and management of data, adopt a standards-based approach to data management (including adopting the EC's CERIF data model for metadata) and identify a limited number of data formats for data exchange.

The HYDRALAB-III (2006-2010) Joint Research Activity on Composite Modelling (described as the balanced use of physical and numerical models) emphasised the need for protocols for data exchange (Sutherland et al., 2012, Gerritsen et al., 2011).

HYDRALAB-IV (2010-2014) shared meta-data about TA experiments using the UK Environmental Observation Framework (http:www.ukeof.org.uk) which is implementing data services based around the INSPIRE data standards for Environmental Monitoring Facilities. This involved mapping of the data in the HYDRALAB database to the UK-EOF schema, a bulk transfer of HYDRALAB data into the UK-EOF and accessing data from the UK-EOF catalogues through its Representational State Transfer application programme interface (RESTful api). As the HYDRALAB database was mapped to the UK-EOF model which in turn is being mapped to INSPIRE EMF (by UK EOF) UK-EOF was used as an intermediary data model (Figure 1).

The following concepts have been implemented and demonstrated:

- Separation of the website from data services, which enables different websites to make use of Hydralab data;
- Delegation of services. Hydralab has adopted the UK-EOF catalogue, thereby enabling it to concentrate on its areas of expertise and jurisdiction. A third party (UK-EOF) is responsible for data back-up, authentication, web services and demand management.
- Benefits of adopting standard data models and web services.

Transnational Access Data Management Plans and Data Storage Reports have common formats, with free-text entries into the sections. Information is required on instrumentation, data acquisition and measured parameters, but this can be descriptive and need not include any description of data formats or meta-data.

These documents provide much of the information that is useful for discovery metadata, and normally describe the location of instruments and the conditions run for each test.

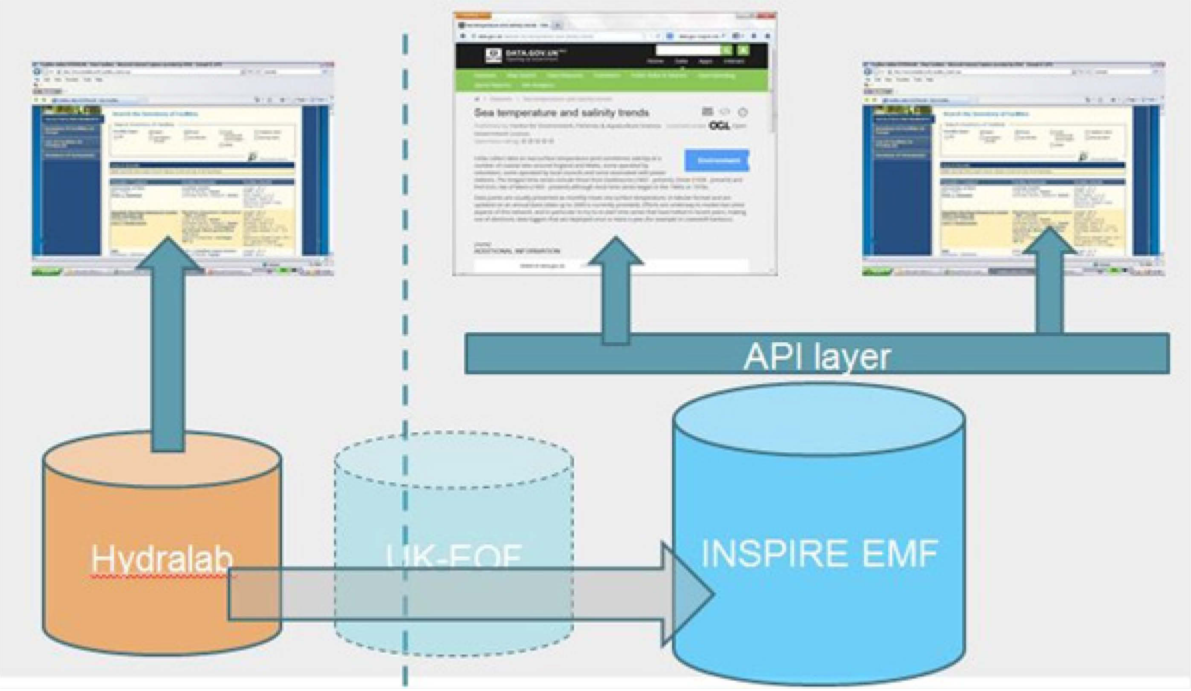

Figure 1: HYDRALAB IV used UK-EOF as an intermediary data model

2. Data management in related communities

2.1. Hydra digital repository at the University of Hull

Current research data management (RDM) initiatives at Hull are based on three main trends;

- The amount of data is growing;
- Data management is required across more disciplines; and
- there is an increasing perception of the value of data.

The Information Services team provide support for research data management throughout the data life-cycle. This includes providing guidance, training on data management http://libguides.hull.ac.uk/researchdata and access to templates for Data Management Plans, such as http://dmponline.dcc.ac.uk. The University of Hull participates in Hydra, a multi-partner, open source initiative that can be applied to all areas of university research. Hydra is based on 2 assumptions:

1. “No single system can provide the full range of repository-based solutions for a given institution’s needs, yet sustainable solutions require a common repository infrastructure.
2. No single institution can resource the development of a full range of solutions on its own, yet each needs the flexibility to tailor solutions to local demands and workflows.”

Therefore, the University of Hull has enhanced Fedora as a digital repository system, with a range of customised ‘Hydra heads’ to suit different research communities’ needs. Its original use at Hull was as a repository for theses, but it is most commonly used for open access journal articles. It’s use is not compulsory for data, but researchers at Hull must put a record in Hydra to record data generated. In this way, data can be managed through a variety of systems, then metadata shared for discovery through a single point. The repository has disk space to archive data, provided by the university as an investment in the future.

2.2. EMODnet Baltic Checkpoint

EMODNET (2009-2020) is the European Marine Observation and Data Network, a long term marine data initiative from DG-MARE underpinning its Marine Knowledge 2020 strategy. This has seven data lots: bathymetry, biology, chemistry, geology, human activities, physics and sea bed habitat. However, the value of data is only realized when it is used and for that, it has to be fit for purpose. The EMODnet Baltic Sea Checkpoint is one of a series of regional projects set up “to assess the quality and adequacy of the current observation monitoring data … by testing the data against specific end-user challenges.” The data adequacy assessments check data accessibility, completeness and coverage, resolution and precision. Limitations in several of the datasets have been identified, when used for particular challenges. The EMODnet Baltic Checkpoint showed that making data available is not enough, it must be understandable and sufficient to meet the user’s needs.

2.3. OpenEarth

The OpenEarth is an open source initiative to develop tools to handle data and models in earth science and engineering (https://publicwiki.deltares.nl/display/OET/OpenEarth). Five levels of data are defined:

- Raw data, collected by scientists;
- Standard data – either NetCDF with CF (Climate and Forecast) metadata conventions or a PostgreSQL implementation with PostGIS depending on the data type. The CF conventions provide descriptions of what the data in each variable represents.
- Tailored data – which is derived from one or more standard data sources to meet the needs of the professional user, e.g. significant wave height from a surface elevation time series, or a velocity field derived from a PIV experiment.
- Graphics of data, using OGC standards; and
- Catalogue of meta-data records.

The raw data coming from the measurement devices is stored in subversion together with a description of the format and the scripts used to process the data. Conversion to standard formats allows other people to easily access and use the data. For example, the dataset from the large-scale Dutch beach nourishment project the sand motor, is available on line at https://zandmotordata.nl/ using OpenEarth.

2.4. Inter-operable mapping in marine and coastal science

Data interoperability has been implemented in many EC projects, including PEGASO (www.pegasoproject.eu) MEDINA (www.medinaproject.eu) and EU MED project COASTGAP (http://coastgap.facecoast.eu/). The Medina Electronic Infrastructure and Pegaso Spatial Data Infrastructure both rely on OGC standards to cope with many, diverse sensors and users. They also rely on INSPIRE concepts and methods and had to be reliable. The Medina project liaised between mapping people and scientific laboratories. It used INSPIRE directive to standardize meta-data, although there was difficulty in convincing people to collect meta-data and share their data.

The Medina E-Infrastructure (MEI) was a mapping tool (or spatial data infrastructure) to disseminate Medina products and enhance GEOSS (Group on Earth Observations System of Systems) use in marine monitoring. It incorporated INSPIRE, GEOSS and OGC standards. The main interface of the MEI was a map viewer, which had a variety of tools (such as query, measure distance, time slider, synchronization, split screen and zooming) for exploiting the results, as illustrated in Figure 2.

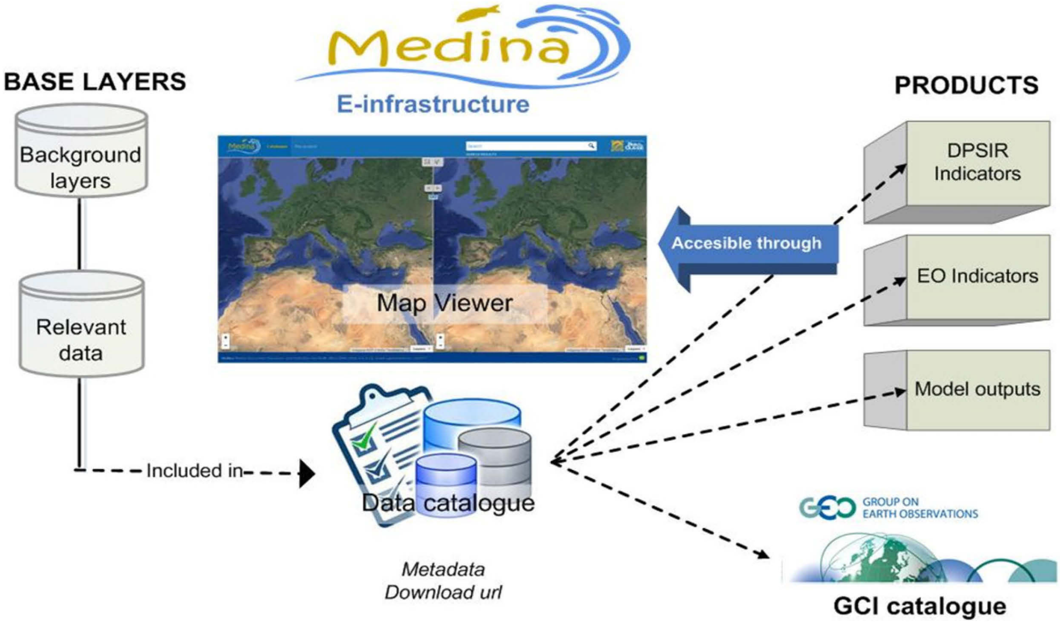

Figure 2: Schematic view of Medina E-infrastructure (www.medinaproject.eu)

Recommended technologies for future developments include:

- the ISO 19156:2011 standard on Observations and Measurements (O&M) which defines a conceptual schema for observations (such as point time series) and for features involved in sampling.
- Sensor Observation Service (SOS) is a web service interface specification for discovery and access of sensor observations.
- Internet of Things (European research cluster on the internet of things, 2015).
- WaterML 2.0 information model for the representation of water observations data.

The use of existing standards is recommended, as there is often no need to reinvent the wheel. Flexibility is required to accommodate the wide scope of data, which might be inconsistent. Adequate resources must be devoted to data management.

2.5. CEDEX physical model test standardisation

The CEDEX laboratory has developed a standardized way of recording physical model test data. Raw data from every experiment is catalogued, classified and stored, with the information organized into a general test record card, test configuration record card, raw data files, and reports. CEDEX have definitions and protocols for each stage of the process and have effectively created their own ontology. This is in Spanish (which improves the usability by technicians, scientists and engineers) but does not conform to international standards.

3. Discussion

Many different approaches to data management have been tried in the different hydraulic and environmental research communities considered. In many, there is a reluctance among scientists and engineers to provide meta-data or to adopt international standards. Data management often has a low priority for researchers and demands skills that are not commonly taught as part of an education in research. However, increasing volumes of data are being collected, which makes it more difficult to manage in an ad-hoc way, and the principles of FAIR data management (EC, 2016) being increasingly demanded by funding bodies.

Therefore, there is a pressing need to raise awareness of, and provide training in, data management within the hydraulic and environmental research communities. This should include training PhD students in data management as a matter of course, but also will require the development of data management skills within research teams. This may well require the employment of people with a background in Information Technology, who already have skills needed for data management. However, ways must also be found to reward those who devote time to data management and, thereby have less time for data analysis, synthesis and paper writing. Associating a data manager with the digital object identifier (DOI) of a dataset and collecting citation data when the data is used would be a practical way of doing this.

Journals are now encouraging authors to publish the data that went into a paper. This is already mandatory for some journals and is expected to become more common. The journal may specify a list of acceptable data repositories. However, this raises the question of what data should go into this repository? Should it be all the data collected in an experiment, or just the data used in the paper? Presumably the latter, but this means that in many cases, all data should go into one repository and some might have to be copied into another, which might have different rules. Some institutions, such as the University of Hull, are developing data repositories for all their projects, while some projects, such as PEGASO, want data from many institutions to go into a single repository. Although this could lead to conflict between different organisations and projects as to which repository to use, in practice, the requirements are generally flexible enough for data to be in one repository, with a corresponding a meta-data record in the other.

What level of processing is required for the data put in a repository? If a graph shows two non-dimensional variables plotted against one another, is it just the numbers used to form the non-dimensional variables that must be published, or the time series used to create those variables? Do we need to supply raw data and calibrated / filtered data, or can we just put the raw data in the repository and leave the reader to work it out for themselves? Ideally the scripts used for calibration, filtering and processing should be included in the repository, but this means making them open as well. The requirements are limited today: providing the numbers used to form the non-dimensional variables from the example above is likely to be sufficient. However, the trend is towards requiring more and this is likely to continue.

The publication of data-processing scripts alongside datasets makes the treatment of data and the production of derived parameters (that go into graphs and tables) transparent and open. Some data management systems, such as OpenEarth, already encourage the storage of data and processing scripts in version control systems, such as subversion, which allow the data-processing chain to be followed. This is likely to become more common in future.

Data in a repository must be understandable, which would be helped by the use of documented (preferably standard) formats and meta-data. However, data will continue to be collected in a variety of formats and flexibility is necessary to deal with the variety of data types. There is not one solution for all circumstances or all data types, although initiatives like OpenEarth can be used to convert an increasingly-wide range of data formats into standard formats.

Data that is not in a standard format should still be defined and a minimum level of meta-data is required for the data to be useful in the future (i.e. data must be understood to be interoperable). There is likely to be a move towards the standardisation of data, which is tending to occur at two levels: the structure of the data and its technical implementation (Sutherland and Evers, 2013). Definitions of data structure are independent of the file encoding. For example, ISO19115 outlines the data structure of spatial metadata with its XML encoding given in ISO19139. The supporting (use and discovery) metadata can be given in separate files to the values themselves. This is exhibited in formats such as CSML and XDMF, both of which offer a binary file type (such as HDF5) for high volumes.

Also, directives such as INSPIRE provide a legal and technical framework for data interoperability. It includes specifications for the data, discovery, use and download services and is aimed at making the finding, using and sharing of data easier across the EU. However, for any practitioner wishing to offer a dataset to the wider community, the set of standards on offer is incomplete, overlapping and highly esoteric.

Good solutions are based on the use of standards, which enable people to work together by establishing common rules and protocols. In recent years there has been a great increase in commonly used standards, approved by bodies such as the International Standards Organisation (www.iso.org), the Open Geospatial Consortium (www.opengeospatial.org), or the World-Wide-Web Consortium (www.w3.org) for web applications. The adoption of internationally accepted standards greatly improves interoperability and removes the need to reinvent the wheel. However, existing standards lack depth. For example, a standard might not specify sufficient details to meet the specialist needs of the HYDRALAB community, but it should be possible to expand it to meet those needs, with the hope that the extra HYDRALAB features could be considered for inclusion in the standard in due course.

No single institution can resource the development of a full range of solutions on its own. The adoption of standards allows each organisation to concentrate on its areas of expertise and jurisdiction. A third party is then responsible for maintenance and development of the software covered by the standard. In the example where HYDRALAB made data available through UKEOF, UKEOF became responsible for data back-up, authentication, web services and demand management.

4. Conclusions

There is a move towards FAIR data management (*Findable*, *Accessible*, *Interoperable* and *Reusable* – EC, 2016) that, combined with the greatly increasing volumes of data being collected, will drive the adoption of good practice in data management in the coming years. This will lead to the adoption and development of international standards throughout the data life-cycle.

Standards for meta-data should be applied, but no existing standard covers the range of situations faced by HYDRALAB. All can be extended in a bespoke manner (which can potentially be included in an update of the standard) but it is highly likely that more than one standard will be used in such a diverse community. This is perfectly acceptable, so long as the standard is published.

There is also a clear need for guidance on the development of repositories where large volumes of data are collected and an understanding of how much needs to be made available on-line. Although there can be conflicts of interest between institutions that are developing policies and local infrastructures for data management and projects that span different institutions and might want a uniform approach to data management across all partners, systems today tend to be sufficiently flexible to accommodate this.

5. Acknowledgements

The work described in this publication was supported by the European Community's Horizon 2020 Programme through the grant to the budget of the Integrated Infrastructure Initiative HYDRALAB+, Contract no. 654110. We would like to acknowledge the contributions made by our invited participants Chris Awre, Gonzalo Malvarez and Jens Murawski, who contributed their expertise from outside HYDRALAB community.

References

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

European Research Cluster on the Internet of Things (2015). IoT Semantic Interoperability: Research Challenges, Best Practices, Recommendations and Next Steps. European Commission, 48 pp.

Gerritsen, H., Sutherland, J., Deigaard, R., Mutlu Sumer, Fortes, J.E.M., Sierra, J.P. and Schmidtke U. (2011). Composite modelling of the interactions between beaches and structures. *Journal of Hydraulic Research*. 49: sup1, 2-14.

Sutherland, J. and Evers, K.-U. (2013). Foresight study on the physical modelling of wave and ice loads on marine structures." *Proceedings of the 35th IAHR World Congress*, Chengdu, China.

Sutherland. J., Gerritsen, H. and Taveira Pinto, F. (2012). Assessment of test cases on composite modelling. *Proceedings of 10th International Conference on Hydroinformatics*, HIC 2012, Hamburg, Germany.

Wells, S., Sutherland, J. and Millard, K. (2009). Data management tools for HYDRALAB – a review. *HYDRALAB report NA3-09-02.* Internet: available from http://www.hydralab.eu