Research Data Management — Activities and Challenges
Report on the Knowledge Exchange Workshop
Wissenschafterzentrum, Bonn, 14-15 November 2011

(Photographs by: DFG/Lichtenscheidt)

On 14 and 15 November 2011 an international workshop on Research Data Management was held at the Wissenschafterzentrum in Bonn. Organised by the Knowledge Exchange working group on Primary Research Data, the event gathered over 50 experts from data centres, libraries, computational centres, funding organisations, publishing organisations and other higher education and research institutions who are working to improve research data management standards and to increase current levels of data re-use and exchange.

The conference in Bonn aimed to provide an update on running projects and policies being currently implemented by all four KE-member organisations dealing with Research Data Management (RDM) at a national level, as well as discussing ‘challenges’ related to research data management. One of the workshop highlights was the presentation and later discussion of the report "A Surfboard for 'Riding the wave': Towards a four country action programme on research data". This report is a follow-up to the "Riding the wave: How Europe can gain from the rising tide of scientific data" report published in October 2010 by the EU High Level Expert Group on Scientific Data.

Besides the presentation of the ‘Surfboard report’ by Maurits van der Graaf, the KE Workshop included a keynote lecture by Todd Vision as well as a five minute madness session where 17 data management projects from various disciplines and countries were presented¹, poster sessions introducing additional projects and a discussion of consecutive activities and challenges with regard to the action programme in the ‘Surfboard report’. The outcomes were summarised in a panel discussion.

Closing remarks were that data formatting and the preparation of data for submission to data repositories are found to be a time-consuming process. Nevertheless the value of experiment replication and fighting fraud provide a solid enough basis to work on a social contract for data sharing and re-use. Possible approaches for the future are to work on awareness raising and getting into the data workflow.

Introduction: Knowledge Exchange and the KE Working Group on Primary Research Data

An introduction² to Knowledge Exchange (KE) was delivered by Keith Russell. Knowledge Exchange started in 2005 as a joint effort by the JISC in the UK, DFG in Germany, SURFfoundation in the Netherlands and DEFF in Denmark ‘aiming to make a layer of scholarly and scientific content openly available on the Internet’.

¹ See the annex for a summary of projects presented in the five minute madness session
² All the presentations are available at: http://www.knowledge-exchange.info/Default.aspx?id=476
In his presentation, Keith Russell mentioned some of the activities KE has carried out in the past, such as the 20,000-signature request to the EU for guaranteed public access to publicly-funded research results, the analysis of costs and benefits of Open Access in the UK, the Netherlands and Denmark as well as studies on persistent identifiers.

Franziska Regner, from the Deutsche Forschungsgemeinschaft (DFG) and chair of the Knowledge Exchange Working Group on Primary Research Data, described the working group members and activities since its foundation in 2008. The working group’s vision can be summarized as follows: ‘research data should be freely accessible, re-usable and professionally curated on a long-term basis’. Central elements achieving this goal are the analysis of data storage requirements, metadata standards, accessibility, long-term preservation as well as appropriate organisational frameworks and services. Challenges attached to these tasks are four-fold: organisational, technical, legal and ‘cultural’, with awareness-raising and meeting discipline-specific needs highlighted as two important considerations in the work to be undertaken.

The legal challenges were addressed in a report which was recently released by KE on the current legal situation with regard to data sharing in the four KE partner countries. Finally, Franziska Regner underlined that data management is a challenge to be faced internationally, not just by the countries represented in KE but well beyond, as proven by the international character of milestone reports such as the OECD principles and guidelines for access to research data (2007) or the EU ‘Riding the wave’ report (2010).

Keynote speech

In line with this international nature of data management initiatives, the workshop’s keynote speech, ‘How many solutions does it take to change the fate of research data?’, was delivered by Todd Vision, researcher in genome evolution at NESCent, University of North Carolina at Chapel Hill, Dryad Project director and Dryad UK member.

Dryad is a US-based international project³ for research data management working closely with publishers in order to promote journal data policies. These policies will result in the collection of life sciences research datasets that get published as supplementary material in a single portal. International, publisher-engaging initiatives such as Dryad or the forthcoming OpenAIRE+ project are a very important contribution to the development of data management procedures and serve as a perfect complement for smaller projects for building research data management infrastructure at subject-based or institutional level.

In his presentation Todd Vision provided examples of the multiple advantages dataset sharing and re-use has for researchers, and how their present attitudes towards this research activity are reflected in various surveys. Still quite widespread reluctance to share data was analysed and proven to be contrary to research codes of conduct, since data sharing will allow mistakes to be detected⁴. The need to articulate some kind of recognition for research data publication was also addressed in the presentation. It was underlined that data authors are not necessarily the same as article authors.

Dryad is actually providing storage and persistent identification features to the scientific community so any kind of dataset - even those that might otherwise not get published - may be filed and preserved in this international, cross-disciplinary data repository. Some recent statistics on Dryad contents and usage were supplied by Todd Vision: over 1,000 data files stored as of Oct 2011, and 42,000 data file downloads monthly show evidence for service uptake. Datasets filed in Dryad are released under a CC-0 waiver and the success of the initiative is very much a consequence of the Joint Data Archiving Policy (JDAP) being adopted by an increasingly large number of journals (88 of them as of October, with Dryad UK shortly adding

³ It is also being developed in the UK as Dryad UK.
a significant number of additional ones). Coupling data submission to the manuscript submission to the publisher is also seen as a driver for Dryad success, with most of the authors being reported to release their datasets upon article publication. This is the case for over 90% of the stored datasets, which are usually small and mainly tables or spreadsheets. Providing such datasets makes the data transfer process easier.

Efforts for estimating data re-use and citation rates - as recommended in the incentives section of the Surfboard report - are already under way, with Piwowar, Carlson and Vision currently analysing data re-use rates for datasets filed in GEO, Pangaea and TreeBase data repositories.

Costs associated with data management and preservation were also addressed during the keynote speech: a reference was made to the ‘taxonomy of data archiving benefits’ as featured in the JISC-funded ‘Keeping Research Data Safe’ (KRDS2) study. A mention was also made of the need to emphasize the benefits for funding agencies as well as for researchers. It would be worthwhile to analyse the costs and benefits of storing research data. Information was provided on data management cost estimation for the Dryad project as a complement to the use cases featured at the KRDS2 Costs Data Survey for UK and Europe based research data management initiatives. The main conclusion from analysis of Dryad costs is that ‘moderate economies of scale are required’, with deposit costs under $50 (depending on specific curation needs) for an estimated 10,000 dataset deposit rate per year. This is well below the costs associated with supplementary data management by scholarly publishers. Data management cost estimation is very much work in progress and further studies are required, but from a return on investment point of view, Dryad experience suggests data management costs to be under 2% of article publication fees and around 0.2% of grant costs per article. This figure does beg the question ‘Is the data worth 2% of research investment?’.

Todd Vision concluded that data archiving is an essential activity that needs to be carried out internationally. Journals and learned societies also have a role to play in its implementation, which is as important as institutions. Finally, funding organisations are called to drive the marketplace by setting up policies for requiring and supporting wide availability and re-usability of research data.

The five minute madness session: selected projects

After the keynote speech a session was held highlighting specific data management projects being developed in the four KE countries mostly under the umbrella of national agency funding programmes. Institutional projects, as well as subject-focused ones were presented. A number of issues were highlighted in the presentations, such as long-term dataset preservation, management of sensitive data in the healthcare sector, the role of data journals in promoting data re-use and sharing among researchers or exploration of new ways for research data storage. A brief summary of the presented projects in this five minute madness session is featured in an annex at the end of this report.

Frank Scholze, from the Karlsruhe Institut für Technologie (KIT) Library Services, provided a ‘summary of thoughts’ to the session — along with a mention of the DFG-funded Re3Data Project being carried out at KIT for developing a registry of data repositories. Some of those thoughts were the following:

- Transfer solutions are needed that work well both for articles and datasets as supplementary material, either through international subject-specific platforms such as Dryad or in data journals.
- Further studies are needed on the cost of research data management — we have a growing amount of information on curation and management costs, but not so much on dataset production and gathering. Besides that, data re-creation may sometimes be more cost-effective than storage, that should also be analysed.
- IT systems have a role as service providers for research data management. Besides that, instrument manufacturers could also be of great help: we should aim for instruments talking a common language which will ease the challenges of managing the data they produce.

---

5 "Is data being re-used? Beginning to track 1,000 datasets from public repositories into the published literature", ASIST 2011 Poster
8 The JISC UMF programme is a good example
• We need a cultural change regarding data re-use and exchange, with an emphasis on incentives, rewards and codes of conduct (the ‘social contract’ of science). Differences exist also between the STM disciplines, where technical issues are mostly solved, and the SSH disciplines, where technology to some extent is still being put in place.

Poster session
A lively discussing developed during the poster session. This offered a chance to exchange more information after the five minute madness sessions and hear more about several other projects and initiatives involved in research data management.

“A Surfboard for ‘Riding the wave’” report: a summary
The second day of the workshop was organised around the presentation and discussion of the Knowledge Exchange report “A Surfboard for ‘Riding the wave’: towards a four country action programme on research data” by one of the authors Maurits van der Graaf.

This report was a response to the ‘Riding the wave’ report which provides a vision for the year 2030 on the issue of how Europe can gain from the rising tide of scientific data. It describes how a collaborative data infrastructure should function not only as a valuable asset for science and technology, but also for the economy and the society as a whole. It proposes the development of a scientific e-infrastructure that supports seamless access, use, re-use and trust of data.

“A Surfboard for ‘Riding the wave’” proposes a short- and medium-term roadmap for realising this goal. In order to better provide this action programme, the Surfboard report is structured along three major sections:

• Four main drivers were identified for achieving the ‘Riding the wave’ goal of developing a collaborative data infrastructure that will enable researchers and other stakeholders from education, society and business to use, re-use and exploit research data to the maximum benefit of science and society. These drivers are: incentives, training, infrastructure and funding.

• A description of the state of the art of research data management in general and these four key drivers in particular in the four Knowledge Exchange countries, as the starting point for a subsequent action programme for those four countries and potential synergies among them,

• An action programme along the four drivers (incentives, training, infrastructure and funding) to be carried out in the KE countries. This action programme aims to realise the objectives described at the ‘Riding the wave’ report.

The Surfboard report presentation went into the analysis of the four strategic drivers for an implementation of a research data management strategy:

The report defines four classes of incentives as necessary to achieve the challenge of making sharing datasets an accepted and integrated part of the academic culture, and examines the degree of implementation for each of them both in the four KE countries and internationally:

• Re-use and recognition of work aimed to publish and share datasets by establishing peer-reviewed data journals and a system for dealing with dataset citation

• Journal data availability policies

• Rules and codes of conduct by international scientific organisations to share data as a desirable good scientific practice

5 Several of the posters can be viewed at: http://www.knowledge-exchange.info/Default.aspx?ID=476
• **Requirements by funding organisations**, including key elements as data management plans, promotion of data sharing or re-use, requirements for long-term maintenance and preservation or funding the costs of data management and sharing.

Having agreed that data sharing is mainly about culture, **training** becomes the most pressing challenge for the implementation of a research data management strategy. The Surfboard report summarises this challenge as ‘changing curricula in universities and librarian training programmes’ and defines two main areas where training should be supplied:

• **Data skills in research domains**: ‘data skills should be made a core academic competency —the earlier the better— and data handling should be embedded in the curriculum’

• **Data librarians**: despite best practices being identified in the US and the UK, the role of data librarians in the new setting of supporting researchers in data publishing and data sharing ‘has yet very much to be developed’.

By defining a two-level challenge on gaps and connectivity issues, the Surfboard report deals with the complex issue of data management **infrastructure** requirements —with large differences between big science and long-tail science in this regard— by focusing on the need for infrastructure interoperability, integration and harmonisation and defining two main levels for work:

• **Institutional data infrastructure**, composed of virtual research environments that enable researchers to work together and create, collect and process research data, and institutional data repositories to store and share datasets. Both parts of the institutional data infrastructure -states the Surfboard report- ‘are presently in their infancy’.

• **Disciplinary or (inter)national infrastructure**, where although internationally organised data centres are gradually being developed, disciplinary data centres are often developed by scientific communities according to their own needs.

The report calls for filling in the gaps in data infrastructure, both at institutional and disciplinary/international level, so that no resulting datasets are left ‘homeless’, and to ensure a proper connection between institutional data infrastructure and research workflow and between institutional and national data infrastructures.

Finally, two main challenges are defined regarding **funding** issues, first of which is establishing who funds what and why and what the evaluation criteria should be. The second is how to minimize the costs of the data infrastructure while maximising the benefits for the researchers. In the section devoted to data management funding issues, the report provides a comprehensive picture of funding infrastructure (universities, other important research performers and some major funding organisations being the featured funders) in the four KE countries and focuses on two main aspects:

• **Costs and benefits** - ESFRI estimation for research data infrastructure is 10-15% of the overall cost of research infrastructure, with costs distributed as follows: 55% on outreach, data acquisition and ingest, 30% on access and 15% on storage and preservation. Besides that, an institutional data repository is estimated to cost 2.5 to 4 times what an institutional repository for publications does due to higher staff and equipment requirements. Cost-benefit analysis is being currently analysed for both national and international initiatives

• **Funding stages** - three stages for data management infrastructure development are identified in the Surfboard report: prototypes, emerging infrastructures used by early adopters and long-term infrastructures. An analysis is done of how those stages may get funded, along with examples of how data infrastructures are being presently funded in the KE countries.

After presenting the status quo and the challenges for the future, the Surfboard report outlines an **action programme on research data for the four Knowledge Exchange partner countries**, with a focus on the four key drivers: incentives, training, infrastructure and funding. Long-term strategic goals are defined for each of the key drivers, along with stakeholders to be involved to achieve those. Finally, mid-term objectives (3 to 5 years time) are also supplied. A number of possible concrete actions (e.g. to increase researchers’ incentives for data sharing, to promote training of researchers and librarians for data management and to develop a sound data infrastructure) are suggested per mid-term objective.
Finally, specific actions are also proposed in the report’s ‘Quick wins’ section to be carried out at Knowledge Exchange level, including the regular organisation of workshops for the exchange of experience and knowledge, the identification of infrastructure gaps by carrying out a 4-country survey for finding out where ‘homeless’ datasets lie, influencing journal editorial boards to issue data availability policies and carrying out awareness raising campaigns for additional rules and codes of conduct to be issued by institutions and learned societies.

Some interesting issues were addressed in the Q&A session that followed the Surfboard report presentation. Discussing cost-related aspects, stress was put upon the fact that researchers and funding agencies are already paying for not managing research data by losing important information, and floor and speaker also agreed on the fact that funding for research data management should not be perceived as ‘stolen’ from research budget.

**Breakout discussions**

After the Surfboard report presentation, four parallel breakout sessions dealt with the four main research data management challenges outlined in the report: incentives, training, infrastructure and funding. Four discussion groups were established to discuss proposals featured in the Surfboard report and additional suggestions from the attending experts. The groups reported back on the discussions.

**Incentives**

Incentives in the form of **scientific credit** to researchers are seen as critical for research data management initiatives to succeed. There are several ways in which such credit may be awarded, one of them being a formal recognition of research data management work in formal assessment exercises, either directly or through the account of data publications in a growing number of data journals (a directory of which should be made available, starting with Copernicus GmbH **ESSD**, and Ubiquity Press Journal of Open Archaeology Data, **JOAD**). The moment we start considering data as a part of research assessment, **data validation** and **data quality** become an issue. Ways to ensure those - both through peer-review of datasets and through citation counts - should be identified and established by data journals. Means for implementing **persistent identifiers** and citations at institutional level are also required for standardisation purposes.

Two recommendations were finally added to the set of proposals contained in the Surfboard report: first, getting research data management tasks included in the research evaluation workflow, and second, developing some system of **prizes and awards** as a way of rewarding and highlighting best practices in RDM.
Training

In order to design an effective approach to training in research data management, potential stakeholders involved in training efforts should first be identified. Four main groups are scoped: employers, publishers, information science schools and internships. The first goal of training schedules should be to create the motivation for the management and publishing of research data, and that includes raising awareness on the value of research data for students/researchers and institutions.

As Reviewers of project proposals are not specifically in the target group of professionals to be trained in the assessment of data management plans, the experts suggested that Knowledge Exchange could take on the responsibility to ensure funding bodies take this aspect into account when reviewing proposals.

On the training activities themselves, it was stressed that it is important to share educational resources and toolkits within the community. Some assessment guidelines for training activities should also be available. The Library of Congress DPOE pyramid\(^\text{10}\) (Digital Preservation, Outreach and Education initiative) is mentioned as an interesting framework for training strategies to be embedded into.

Finally, internships were thoroughly discussed as a means of achieving training. It is important that these create value to both parties (i.e. the trainee and the project).

Infrastructure

Four introductory questions were considered by the group in order to focus the discussion:

1. Can best practice disciplinary fields be identified, e.g. robotics, for transfer to other fields?
2. Which research areas would benefit most from setting up an ERIC\(^\text{11}\) (European Research Infrastructure Consortium)
3. Can a technical priority agenda be drafted?
4. How to ensure more timely availability and accessibility to data? Are there best practices?

Current data infrastructures were claimed to be ‘complicated, monolithic, isolated and not trustworthy’. In order to clarify the infrastructure requirements, a distinction should be made between data challenges (data modeling, management and services) and system challenges (VREs, interoperability and mediation software).

Research data infrastructure
Wilhelm Hasselbring, University of Kiel

- Integrate / couple
  - Institutional data infrastructures and
  - Disciplinary (inter)national data centers (WDC)
  - For instance, via workflow support
- Establish the research data infrastructure at the intersection of
  - Libraries and
  - Data / computing centers
- Provide domain-specific solutions
  - Custom-made solutions for specific disciplines (domains)

**Figure 2:** Slide presenting suggestions from the infrastructure breakout group

In the discussion, a survey on infrastructure components was suggested. The importance of shared services among institutions was also highlighted. Harmonisation of legal issues should also be a step prior

\(^{10}\) [http://www.digitalpreservation.gov/education/educationneeds.html](http://www.digitalpreservation.gov/education/educationneeds.html)

to facing interoperability and licensing barriers. The group proposed to publish preliminary guidelines for the building of data repositories.

Among the recommendations were the following:

- Doing usability studies as part of constructing infrastructures and defining requirements for discipline specific solutions,
- Identifying common parts of the workflow between different research groups,
- Funders need to act as a catalyst and encourage their communities to create a menu of RDM infrastructure and create awareness around this menu,
- Funding a project to produce example workflows in various disciplines
- Providing a list of already existing infrastructure

Funding

The break out group on funding came up with the following results. There is a need for business models and cost benefit analysis that will provide a basis for funding strategies. Best practices and success stories are also required, with emphasis on subject-specific approaches: a model is suggested that splits research data management into three disciplinary fields: physical sciences, biomedical sciences and social sciences & humanities.

A quantification of budget waste through data loss would be useful as a starting point, as well as an assessment of research data management practices by funders and an audit of effectiveness of RDM policies. The scale of funding was also discussed: international organisations, funders and institutions: who should fund what. Examples for funding strategies should be collected — such as a flat rate for publishing. There should be an agreement about the distribution of responsibilities among different stakeholders, and again success stories may play a critical role.

Panel discussion

A panel session featuring Todd Vision, Frank Scholze, David Shotton and Leo Waaijers closed the event. Differences in adoption and success of research data management by discipline were examined. Some of these contain highly heterogeneous data types. There is ‘not necessarily a lack of willingness’ to explain the differences between the disciplines. Data formatting and the preparation of data for submission to data repositories was found to be a time-consuming process, although experiment replication and fighting fraud provide a solid enough basis for a social contract for data sharing and re-use. The current situation was summarised as follows ‘we should be standing on two legs, awareness raising and getting into the data workflow.’ Suggestions for future steps were: ‘The road ahead should start at a low level, then undertaking a gap analysis. After this data centres could take care of orphan datasets from other KE countries’.
Annex: Five minute madness session

Selected projects in Research Data Management from KE countries

The presentations are ordered by the presenters’ surnames. An abstract of each presentation is provided.

A Research Collaboration and Archiving/Submission Interface: Minimal Intrusion Grid
(MiGrid, Jonas Bardino, The eScience Center, University of Copenhagen),
http://www.migrid.org/

A Virtual Research Environment is already being provided with multiple collaboration tools inside our Minimum intrusion Grid (MiG) software, but we would like to tightly integrate archiving and submission. Our proposed solution is to introduce a simple means of read-only archiving files directly in our MiG file manager. Submit portals can then be created using the existing web page publishing and file sharing features in combination with such an archiving solution.

Interfacing biology at the molecular level to personal healthcare sector information
(Soren Brunak, DTU Center for Biological Sequence Analysis, Copenhagen),
http://www.inbiomedvision.eu/

Going from genomics to (disease) phenotypes and back is the aim of this initiative. It is carried out within the INBIOMEDvision European FP7 Coordination and Support Action. This has the objective of promoting biomedical informatics by monitoring the scientific state-of-the-art and identifying common grounds between bioinformatics in biomedical research and medical informatics in healthcare and clinical research.

A paper published last August by Roque et al. in PLoS Computational Biology, ‘Using electronic patient records to discover disease correlations and stratify patient cohorts’, demonstrates how text mining extracts clinically relevant terms from patient electronic medical records (EMRs) and how mapping these to disease codes in the International Classification of Disease ontology (ICD10) helped construct phenotypic profiles for each patient. These profiles were used both to identify correlations between diseases (co-morbidity) and to cluster patients into groups with highly similar clinical features.

These kind of studies suggest that a new predictive, preventive, and personalised medicine is emerging, while at the same time raising the issue of sensitive healthcare data sharing on which projects such as ELIXIR, are currently working.

Digital Knowledge Store at Berlin-Brandenburgische Akademie der Wissenschaften
(BBAW-Telota, Alexander Czmiel, Berlin-Brandenburg Academy of Sciences and Humanities)
http://www.bбав.de/en/telota/resources/dkb

The Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) aims to explore and exploit opportunities advances in information technology have provided for new research in all disciplines. For this purpose, the initiative The Electronic Life Of The Academy (TELOTA) was started in 2001. Since then, extensive collections of scholarly resources, above all in the humanities, have been made available online to the research community and the broader public including circa 30 projects.

The presentations can be viewed at: http://www.knowledge-exchange.info/Default.aspx?ID=476
Archaeological and classical research data

(Reinhardt Förtsch, Universität zu Köln)
http://www.dainst.org/es/node/30724?ft=all

The Berlin based German Archaeological Institute (Deutsches Archäologisches Institut, DAI) supports research projects worldwide, in which a large amount of classical antiquity information in electronic version is collected and processed. The cited report (in German) summarizes the current status and challenges in dealing with these digital research data at the DAI. Besides focusing on the process of generating research data and the related standards that need to be considered, these issues are discussed in the context of interoperability, reusability and long-term archiving. Using various technical systems, both routine operations and first results are presented, and prospects for future development are identified.

The Smart Research Framework Programme

(Jeremy Frey, University of Southampton)
http://www.mylabnotebook.ac.uk/index.shtml

Providing tools to facilitate data management in research projects is now essential if researchers are to meet the needs of both funders and the expectations of the wider public for transparency, responsibility, collaboration and maximising the return on the investment in research by enabling reliable re-use of research outputs. The LabTrove service has demonstrated these features in a range of projects from Chemistry, Physics and Engineering and is being investigated in environmental and medical contexts. The blog3 demo service has proven that where a deeper semantic base is available in the researcher community a further level of integration is possible.

In the laboratory context a key feature of LabTrove and blog3 is the integration with experimental equipment and computational services, this is achieved using the LabBroker middleware service; this service will also be brought to a wider community.

Morph-D-Base

(Peter Grobe, Zoologisches Forschungsmuseum Alexander König)
https://www.morphdbase.de/

Morph-D-Base aims to provide a way to archive, share and disseminate data, provide detailed documentation and efficiently collaborate with research colleagues in biological morphology.

The need for standardisation of data and metadata has been recognised throughout different disciplines in biology – except for morphology. The goal of Morph-D-Base project and approach is to improve the quality of morphological data in different ways.

PubFlow - Excellence Cluster ‘Future Ocean’

(Wilhelm Hasselbring, Universität zu Kiel)
http://www.pubflow.uni-kiel.de/de
http://www.ozean-der-zukunft.de/english/research-areas/data-management/

The aim of the PubFlow Project is to develop a concept and a pilot application to work with research data in scientific workflows to increase productivity in the actual scientific work. In order to do so, various processes such as data collection, processing, archiving and publication are considered.
It is not the purpose of PubFlow to provide original contributions in the areas of long-term archiving or data citation, as is done for example by the PANGAEA data library. The contribution of this project is the workflow-based linkage and integration of such services, here from the perspective of scientific institutions. The first phase of PubFlow aims to develop research data publication processes for the disciplines of marine and earth sciences as discipline-specific forms of organisation and as preliminary pilot applications for workflow-based information infrastructure.

Beginners’ experiences in research data management: The Danish Perspective
(Alfred Heller, Danish Technical University Copenhagen)
http://forskningsdata.deff.wikispaces.net/Afrapportering

For research data management to be handled properly, the role of data professionals must be recognised, this includes researchers, data specialists, IT-people and librarians. Therefore carrier ways for these professionals are the key to success.

Under the Danish Digital Library (DDF) four main research libraries have collected first experiences regarding working with research data. The project was collecting international ‘best practices’, the development of a prototype for a digital repository based on Fedora-Commons, testing of existing software, especially DataVerse, a first trial with collecting data from researchers and evaluating standards for description of research datasets. The report, available at the Danish Research Data wiki, presents a large number of recommendations that have led to a follow-up project to be carried out in 2012 by all research universities in Denmark. In this project the readiness towards the handling research data will be examined by interviews at all levels of the university hierarchy.

Enhanced Journals... Made Easy
(Driek Heesakkers, Universiteit van Amsterdam)
http://www.surffoundation.nl/en/projecten/Pages/EJME.aspx
http://wiki.surffoundation.nl/display/vp/EJME+english

The Internet makes it possible to present publications in combination with related research data, as Enhanced Publications. The aim of the EJME project is to design a practical work process for publishing scientific/scholarly journals with enhanced publications. In order to do so, an agreement was established with two Open Journal System (OJS)-based journals from two different small academic presses for supplying them an EJME plugin for OJS that will allow authors to submit research data along with the article. In this way the data becomes integrated into the publishing workflow and gets published with the article. At the end of the pilot project, the work package and software produced by EJME will also become available for use by other authors, editors, and publishers.

Cost Model for Digital Preservation
(CMDP, Ulla B. Kejser, Royal Library of Denmark),
http://www.costmodelfordigitalpreservation.dk/

The Danish National Archives and The Royal Library are developing an economic model for digital preservation based on the OAIS Reference Model and the International Standard Cost Model Manual. The purpose of the project is to create a tool for estimating the long-term costs and benefits of digital preservation, from the moment when digital objects are ingested into the repository until they are made accessible to the users. The tool will enable institutions to compare different preservation alternatives and thus support planning and decision-making leading to the most efficient use of the available resources.
Controlled Access to Research Data - Securely Stored (CARDS) Project
(Ana van Meegen, Vrije Universiteit Amsterdam Library)
http://www.surffoundation.nl/en/projecten/Pages/CARDS.aspx

Researchers frequently run up against problems with managing their valuable research data. Many of those problems have to do with the safe storage of data and the controlled sharing of data. Without imposing any demands or getting in the way, the CARDS project aims to help researchers manage their data in what is for them the most efficient and effective way. This process will also be the basis for developing a digital information package. Amongst other things, this will include a template for a data management plan provided by DANS.

Open Data and Publications: Librarians at Tilburg University liaising with researchers
(Thomas Place, Universiteit van Tilburg)

The project was meant to help researchers link their datasets to their publications, thus creating ‘enhanced publications’. The experience gained by creating forty such publications will enable the library at Tilburg University to offer this as a standard service with which researchers can make datasets available and link them to publications. The library collaborated with the University’s Schools of Economics & Management and Social & Behavioural Sciences; there are no serious obstacles to their making their datasets publicly accessible.

DataFlow and beyond
(David Shotton, University of Oxford)
http://www.dataflow.ox.ac.uk/

DataFlow is creating a SaaS-based two-stage data management infrastructure that makes it easy for researchers and research groups to manage their research data. The system is locally managed using the researcher's own instance of DataStage, while allowing the institution to deploy DataBank easily to preserve and publish most valuable datasets. Published datasets are assigned DOIs to make them citable and to gain academic credit for the researcher.

Besides explaining the DataFlow Project, discussions were held on why researchers aren't sharing their data (with assessment focusing on articles and it being a time-consuming activity as main explanations) and on best practices to sharing data. Available datasets were mentioned as a criterium for awarding 'the five stars of online journal articles' and as a pressing issue in the Force11 Manifesto (the Future of Research Communication and e-Scholarship).

---

13 An example can be found at: https://bzv.uvt.nl/search?displayType=single&query=ir-uvt-nl:oai:wo.uvt.nl:167602
CellFinder Project
(Harald Stachelscheid, Charité Berlin)
http://www.cell-finder.org/

Large amounts of research data on stem cells are already available, but these are scattered, derived by diverse technologies, not standardised and are not available at the necessary integration level for e.g. cellome modelling. CellFinder is a stem cell data repository under development that will provide a high-quality virtual environment for stem cell research under the Open Source model. The platform shall facilitate the representation of various kinds of stem cells and describe their biological characteristics and possible medical applications.

BRISKit Project
(Jonathan Tedds, University of Leicester)
http://www2.le.ac.uk/offices/itservices/resources/cs/pso/project-websites/brisskit/brisskit

BRISKit (short for Biomedical Research Infrastructure Software Service kit) aims to design, deliver and begin to exploit a national data hosting service for researchers in the field of Biomedicine and Bioinformatics. Hosted at the University of Leicester in close partnership with the Leicester cardiovascular biomedical research unit and funded by the JISC UMF Shared Services and the Cloud Programme, BRISKit operates two data warehouses, one inside a hospital infrastructure, and the other outside. The data warehouse in the hospital can have sensitive patient data stored in it, which can range from questionnaires to clinical data. The external data warehouse will have anonymised versions of the patient data on it (pushed from the hospital side), allowing researchers to analyse clinical data while maintaining patient confidentiality.

VIDaaS Project
(James Wilson, University of Oxford)
http://vidaas.oucs.ox.ac.uk/

The ViDaas Project developed at Oxford within the JISC UMF Shared Services and the Cloud Programme, is designed to provide a Database as a Service to a wide range of Arts and Humanities researchers. ViDaas (Virtual Infrastructure with Database as a Service) is a project of two halves: the 'DaaS' part will develop software that enables researchers to build, edit, search, and share databases online, while the 'VI' part involves the development of an infrastructure enabling the DaaS to function within a cloud computing environment.

TLA@MPI-PL, The Language Archive at Max Planck Institute for Psycholinguistics
(Sebastian Drude, MPI-PL)
http://www.mpi.nl/research/research-projects/the-language-archive

The Language Archive is a new unit created last September at Max Planck Institute for Psycholinguistics in Nijmegen. The mission of this new unit covers the digitising and archiving of valuable language resources deposited by any researcher worldwide and the providing of access to archived language resources while respecting legal and ethical restrictions. Currently, the archive gathers about 60 Terabyte of well-described resources, 20.000 hours of digitised audio/video recordings, 73.000 metadata described sessions, 4.5 million annotated segments and about 47 lexica, including resources from about 200 languages. The Unit will continue to accept to digitise and archive other valuable data. The Language Archive is part of CLARIN, a European initiative to create a Common Language Resources Infrastructure, and will be using the CLARIN Component Metadata Infrastructure (CMDI) standard for the description of language resources.