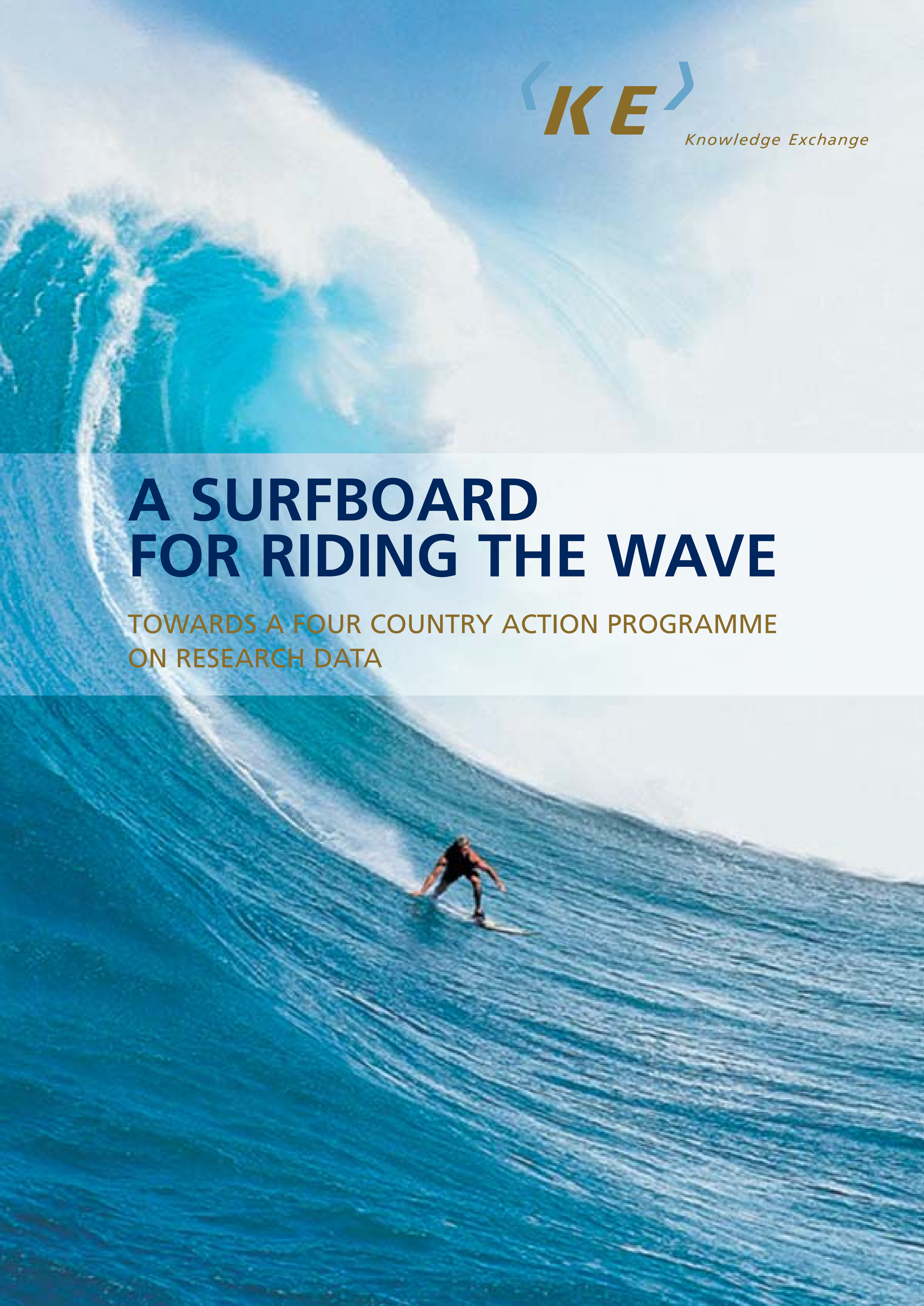
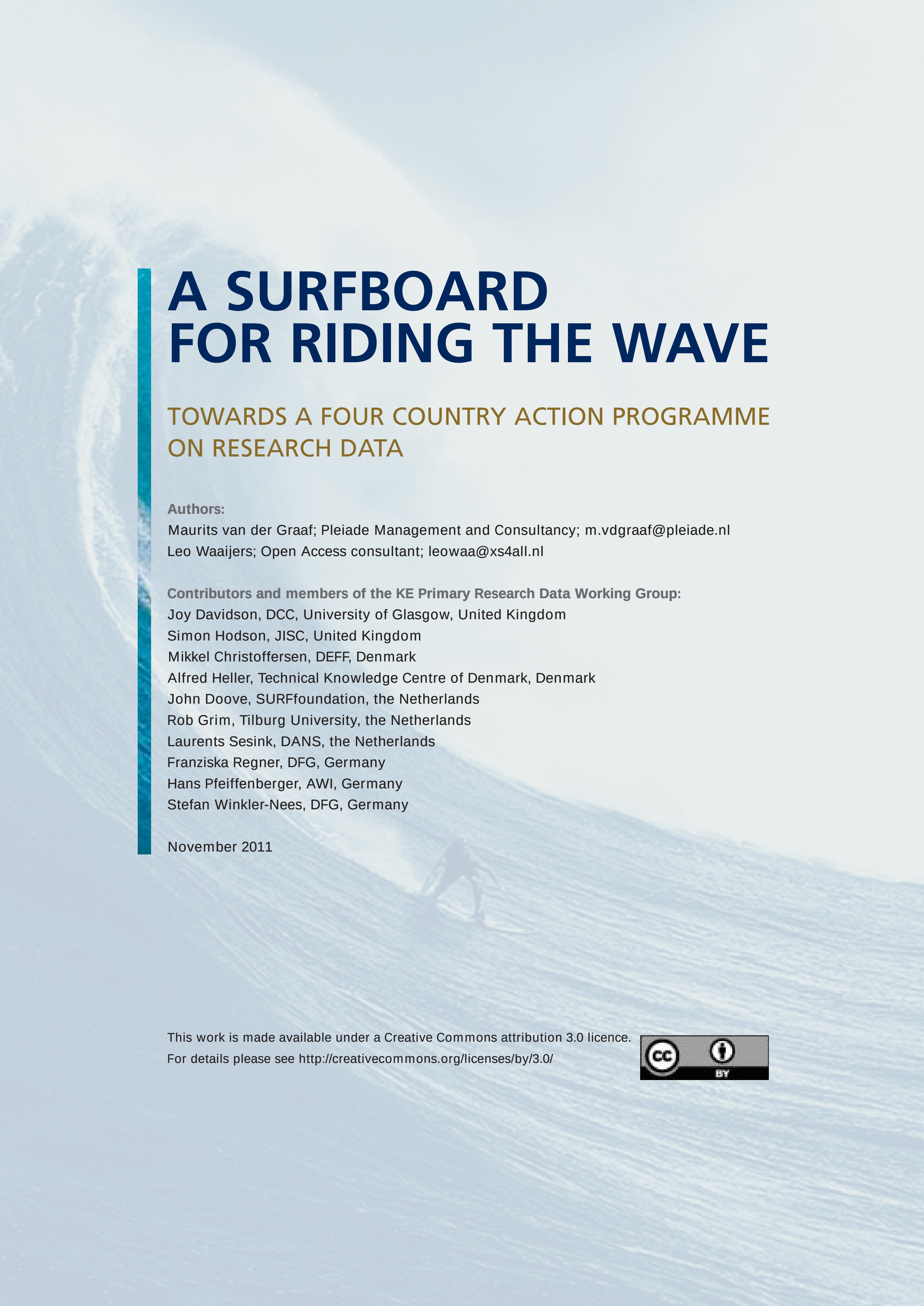


Knowledge Exchange

A SURFBOARD FOR RIDING THE WAVE

TOWARDS A FOUR COUNTRY ACTION PROGRAMME ON RESEARCH DATA



A SURFBOARD FOR RIDING THE WAVE

TOWARDS A FOUR COUNTRY ACTION PROGRAMME ON RESEARCH DATA

Authors:
Maurits van der Graaf; Pleiade Management and Consultancy; m.vdgraaf@pleiade.nl
Leo Waaijers; Open Access consultant; leowaa@xs4all.nl

Contributors and members of the KE Primary Research Data Working Group:
Joy Davidson, DCC, University of Glasgow, United Kingdom
Simon Hodson, JISC, United Kingdom
Mikkel Christoffersen, DEFF, Denmark
Alfred Heller, Technical Knowledge Centre of Denmark, Denmark
John Doove, SURFfoundation, the Netherlands
Rob Grim, Tilburg University, the Netherlands
Laurents Sesink, DANS, the Netherlands
Franziska Regner, DFG, Germany
Hans Pfeiffenberger, AWI, Germany
Stefan Winkler-Nees, DFG, Germany

November 2011

CONTENTS

EXECUTIVE SUMMARY:
TAKING STOCK AND GOING AHEAD

Scope

The *Riding the Wave* report calls for 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. The Knowledge Exchange partners have embraced this vision. This paper presents an overview of the present situation with regard to research data in Denmark, Germany, the Netherlands and the United Kingdom and offers broad outlines for a possible action programme for the four countries in realising the envisaged collaborative data infrastructure. An action programme at the level of four countries needs the involvement of all stakeholders from the scientific community. We identified four key drivers:

- incentives and
- training in relation to researchers in their role as data producers and users of information infrastructures
- infrastructure and
- funding of the infrastructure in relation to further developments in data logistics.

Incentives

For researchers in their role as data producers, we identified four main areas of incentives to share and publish their datasets: (a) re-use and recognition, (b) principles of science, reflected in rules and codes of conduct, (c) requirements by funding organisations and (d) journal data availability policies. Several initiatives in the four partner countries enable both the citing and publication of datasets. Some science organisations have published a code of conduct for data sharing while some science funding organisations have set requirements for grant applicants with regard to data management during the research project and data sharing after the research project. There appears to be a rising number of scientific journals with a data availability policy.

Training

In the data-intensive scientific world, new skills are needed for creating, handling, manipulating, analysing, and making available large amounts of data for re-use by others. We distinguish three actors in this process: (1) researchers, who should have basic skills with regard to data handling (2) a newly emerging professional role with the label ‘data scientist’, who will be responsible for computing facilities, storage and access in their discipline and (3) another newly emerging professional role labelled ‘data librarian’, who will be responsible for data curation, preservation and archiving. The current situation in the four KE countries is rather diverse and very much in development.

Data infrastructure and its funding

This paper distinguishes institutional data infrastructures from disciplinary (inter)national infrastructures, describing the situation in the four KE partner countries in this respect.
We highlight two challenges: gaps in the present data infrastructure and connectivity issues.
We also address the funding of data infrastructure, reporting the results of cost benefit studies and describing the present situation of funding of data archives and data centres.

Toward a four country action programme

Based on the overview of the present situation in the four Knowledge Exchange partner countries, we have formulated three long-term strategic goals:

- Data sharing will be part of the academic culture
- Data logistics will be an integral component of academic professional life
- Data infrastructure will be sound, both operationally and financially.

Focused on achieving these three long-term strategic goals, this report presents the broad outlines of an action programme at the level of the four KE countries, departing from the current situation and advancing towards the realisation of the envisaged collaborative data infrastructure for research.

1. INTRODUCTION

Focus of the Knowledge Exchange overview report

The *Riding the Wave* report (1) calls for 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 scholarly and scientific research and society. In this vision, research data are seen as an integral part of the research infrastructure and are as important and necessary as for example networks and computing facilities. This vision is widely embraced and enthusiastically supported by many scientific organisations in Europe, including the Knowledge Exchange partners operating in Denmark, Germany, the Netherlands and the United Kingdom.

What is the present situation in the four Knowledge Exchange countries with regard to research data? What concrete steps should the partners take next to realise this vision of a collaborative science data infrastructure? That is the central focus of this Knowledge Exchange overview.

Introducing Knowledge Exchange

Knowledge Exchange (KE) is a co-operative effort that supports the development and use of ICT infrastructure for higher education and research. The KE partners (see textbox) share a common vision based on their four national strategies: to make a layer of scholarly and scientific content openly available on the internet. Research data is a key element of this vision. Working together on a common approach will lead to greater availability and re-use of research data, which will benefit research, education, society and business. The KE working group on primary research data has explored this topic and collected data for an assessment of the current state in the four countries, resulting in this overview.

Knowledge Exchange partners

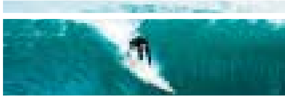
- Denmark's Electronic Research Library (DEFF)
- German Research Foundation (DFG)
- Joint Information Systems Committee (JISC) in the United Kingdom
- SURFfoundation in the Netherlands

Key drivers: researchers and infrastructure

At this stage of development, four key drivers for realising the collaborative science data infrastructure have been identified:
(a) two key drivers in relation to researchers in their role as data producers and users of information infrastructures (incentives and training)
(b) two key drivers in relation to the technical infrastructure (the next steps in development of the infrastructure and the funding of the infrastructure).

This paper reports the results of an overview, assembled by the KE working group on primary research data. After reviewing the international initiatives in science data infrastructure (chapter 2), the report presents the current situation with regard to the four key drivers in the four countries (chapters 3 to 6) and formulates broad outlines of an action programme for the four countries (chapter 7).

2. DATA INFRASTRUCTURE FOR RESEARCH: AN INTERNATIONAL REVIEW



Riding the Wave

Scientific and scholarly research nowadays results not only in publications but also increasingly in research data. Subsequently or parallel to the actual publications, research data sets are starting to have a life of their own as independent sources of information and analysis for further research. This data-intensive model of research has been described as the Fourth Paradigm of Science (2). To facilitate this, research data sets need to be discoverable and accessible in similar ways as publications are for purposes of validation and re-use in meta-analyses, simulation models and other types of studies.

An important milestone in the thinking about this phenomenon is the recent report entitled Riding the Wave by the High Level Expert Group on Scientific Data, which developed an impressive vision for the year 2030 on the issue of how Europe can gain from the rising tide of scientific data. Praised and embraced by many stakeholders including Knowledge Exchange, this vision encompasses a scientific e-infrastructure that supports seamless access, use, re-use and trust of data. The envisaged collaborative data infrastructure should function not only as a valuable asset for technology, but also for the economy and the society as a whole. Figure 1 outlines this infrastructure.

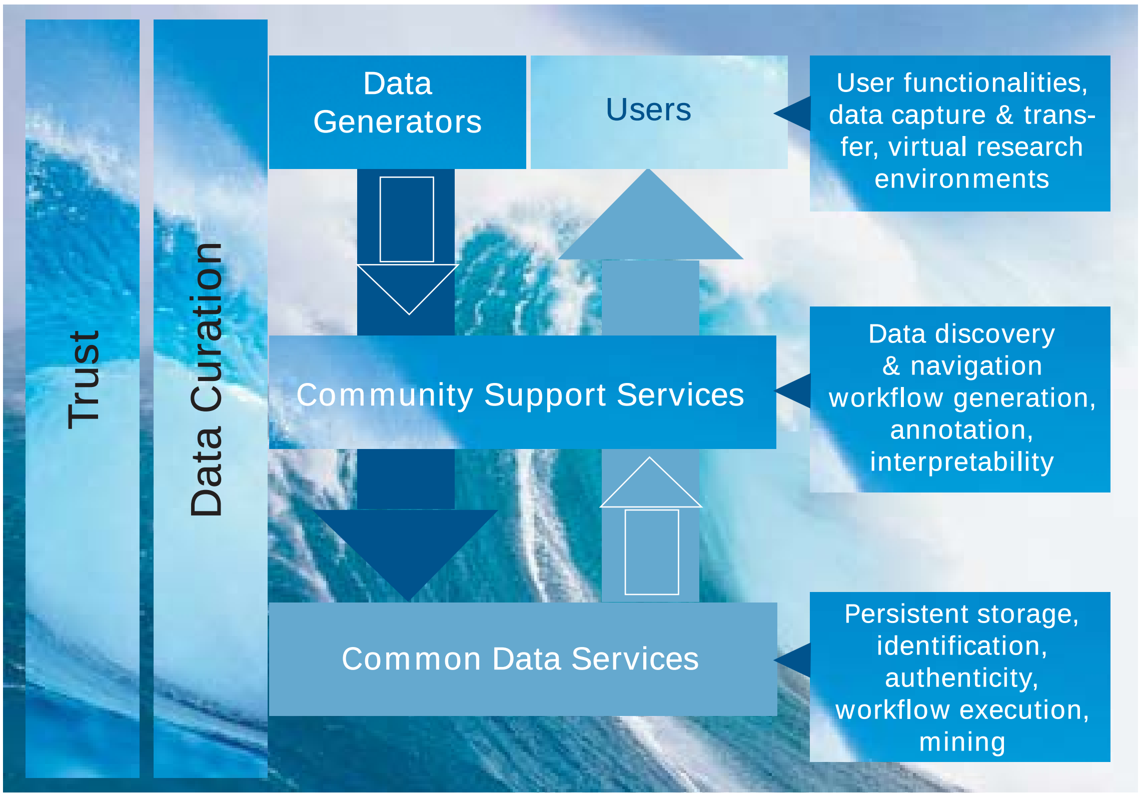

Trust
Data Curation
Data Generators
Users
User functionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation workflow generation, annotation, interpretability
Common Data Services
Persistent storage, identification, authenticity, workflow execution, mining

Figure 1. The collaborative data infrastructure - a framework for the future; from *Riding the Wave*, p. 31

This overview focuses on four key drivers that support the science data infrastructure:

- **Incentives** for researchers to make their research data accessible and re-usable by other researchers and **training** these researchers in data management skills and practices both facilitate the roles of data producers and users that form important target groups of the collaborative data infrastructure (see Figure 1).
- Both **funding** and **infrastructure** enable common data services and community support services that form the backbone of the collaborative data infrastructure.

Present practices in data sharing

How widespread is data sharing among researchers now?

A survey of UK researchers (3) found that 22% of the respondents from science, 37% from arts and humanities and 45% from social sciences shared their research data with many others. A large international survey of researchers by PARSE.Insight found that 25% of the researchers make their data openly available to everyone (4). Legal issues and possible misuse of their data was most frequently cited as a barrier for data sharing, while the possibility of re-analysis of existing data was seen as the most important driver for the preservation of research data (91% of the respondents in this survey thought this important).

In a survey of senior Dutch university researchers across all disciplines, 70% indicated they were data producers, 60% had shared (some of) their own datasets with others and 50% re-use others' datasets in their own scientific work (5) (6).

An international survey by Tenopir et al. (7) reports similar findings: only 36% of the respondents agree that others can find their data easily, although three-quarters share them with others. The authors conclude that there is a willingness to share data, but note a number of thresholds for researchers in doing so. In this survey, insufficient time and lack of funding came out on top of the reasons for researchers not making the data electronically available.

These results show that data sharing is happening now, but the practice is still limited to a minority of researchers and a limited number of datasets and certainly not a common practice. There is broad agreement that data sharing should be elevated to a higher, more systematic and organised level so that it becomes a standard practice throughout science.

International initiatives

As research becomes more data-intensive, research datasets increase in number and size. Re-using (combinations of) research datasets produced by researchers in the same discipline or from different disciplines brings about novel approaches, such as data exploration, simulation and modelling, system level science, and transdisciplinary research.

Not only access but also interconnectivity and interoperability of the various datasets and systems are essential to enable these new approaches. This realisation has led to several international initiatives and policy statements, among which the publication of the *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007) can be seen as a landmark. These guidelines aim to promote a culture of openness and sharing of research data among the public research communities of the OECD countries based on the principle that a data infrastructure should be part of the international research infrastructure (8).

The European Union defined the 'Fifth Freedom' – free movement of knowledge – for the European Research Area and sees research data sets as an integral part of this (9). The research infrastructure in Europe presently consists of around 300 research facilities. Increasingly, a data infrastructure is seen as an essential part of the international research infrastructure. The ESFRI roadmaps (2010, 2008 and 2006) set out a strategy for building on the European research infrastructure and include data infrastructure elements {(10) (11) see also: (12) (13)}. The Alliance for Permanent Access has carried out projects (PARSE.Insight, ODE) to define the functional specifications. The European research funders, combined in EUROHORCS and the ESF, have also committed themselves to promote and ensure permanent access to research data generated with their funding (14).

Similar initiatives have been taken in the USA, such as the Blue Ribbon Task Force report on sustainable data infrastructure (15), the DataNet programme and the recently re-enforced policy on data management and dissemination by the National Science Foundation. In Australia the Australian National Data Service was set up. In the so-called Brussels Declaration of the International Association of Scientific, Technical and Medical Publishers, international journal publishers stated that research data are not included in the copyright of publishers and should be as openly available as possible (16).

In conclusion, a number of major international stakeholders are committed to developing (international elements of) a data infrastructure.

3. INCENTIVES

3.1 Introduction: four types of incentives and one major challenge

For researchers in their role as data producers, the KE working group on primary research data identified four areas of incentives to share and publish their datasets: (a) re-use and recognition, (b) principles of science, reflected in rules and codes of conduct, (c) requirements by funding organisations and (d) journal data availability policies. This chapter explores each area of incentives in more detail and assesses the situation in the four KE countries.

However, it is important to note that there are also disincentives for researchers. Important disincentives lie in the area of risks of publishing data sets, such as possible abuse and ethical or legal issues. Knowledge Exchange has just published a study on the legal status of research data in the four partner countries[1], identifying flaws and obstacles to the access to research data and singling out preconditions for making data openly available. Another area of disincentives for researchers consist of the extra efforts and costs needed to create documentation and metadata for the dataset {see also (4) and (7)}. These disincentives should be understood as qualifiers when devising incentives.

To make the balance of risks and rewards for the researcher tilt towards data publishing, it is important to make the rewards more attractive and compelling while possibly minimize the disincentives. This is the major challenge in this area: to make sharing datasets an accepted and integrated part of the academic culture.

3.2 Re-use and recognition

In the current conventional system of recognition and reward, data do not have an adequate place. Recognition for researchers is still mainly based on publishing in high-quality journals and/or citation metrics of their articles. Published datasets should count in the academic record of the data producer as well. In general, metrics (based on citations, usage statistics etc.) can play a role here, but a distinction should be made between the requirements of funders – seeking impact for the projects or institutes they finance – and researchers, looking for recognition by their peers.

A novel method to achieve the latter was pioneered by the environmental scientists David Carlson and Hans Pfeiffenberger. They started a peer-reviewed journal for data publications, that is, articles describing research datasets. This gives data producers the opportunity to publish a peer-reviewed journal article on their datasets (alongside publishing the datasets themselves) and thus make it count in their academic record (see Box 1). So far, only a few scientific journals are dedicated to publishing data publications.[2]

[1] The legal study is available at: http://www.knowledge-exchange.info/Default.aspx?ID=461

[2] The other journals known to us are *Acta Crystallographica E*, *Ecological Archives*, *GigaScience* and *International Journal of Robotics Research*. However, there are initiatives to start up more publication channels for these data publications (or data papers); see press release of 2011/06/03 at datadryad.org

Box 1. Earth System Science Data

The journal *Earth System Science Data* (ESSD) provides reliability to re-usable scientific data as well as an incentive for their creators to publish them in the first place. ESSD articles ‘wrap’ data with proof of quality-related assertions and provide an object target for – almost – classical peer review. The reviewers do not just scrutinize the article text but also (and even more so) the data themselves.

ESSD, www.earth-system-science-data.net, is an international, interdisciplinary journal for the publication of articles on original research data (sets), furthering the re-use of reference quality data of benefit to Earth System Sciences. The editors encourage submissions on original data or data collections, which are of sufficient quality and potential impact to contribute to these aims.

ESSD has an innovative two-stage publication process involving the scientific discussion forum Earth System Science Data Discussions (ESSDD). It is designed to foster scientific discussion and maximise the effectiveness and transparency of scientific quality assurance.

In the first stage, after a rapid access peer-review, articles are immediately published on the ESSDD website. They are then subject to interactive public discussion, during which the referees’ comments, additional short comments by other members of the scientific community and the authors’ replies are also published.
In the second stage the final revised papers, if accepted, are published in ESSD. To ensure publication precedence for authors, and to provide a lasting record of scientific discussion, both ESSDD and ESSD are ISSN-registered, permanently archived and fully citable.

Pfeiffenberger, H. & Carlson, D., 2011: “Earth System Science Data” (ESSD) – a peer reviewed journal for publication of data. D-Lib Magazine 17 (1/2). doi: 10.1045/january2011-pfeiffenberger

Another method is to stimulate citing of published datasets in the journal literature. Citing (re-used) datasets is not yet incorporated in the habits of most researchers and – more importantly – not yet standardised. DataCite, an organisation started by TIB Hannover in Germany, and other persistent identifier solutions are addressing this latter problem by developing methods for citing datasets. The DataCite method, building on existing methods from publication citation and reference styles, has been rapidly embraced by data centres and data archives throughout the world enabling users to cite a dataset in a recognizable, standard manner.

The ODE report *Integration of Data and Publication* (17) explores several options for linking datasets and journal articles with an eye on best availability, retrievability, interpretability and usability of the datasets. A bidirectional link between journal articles and data sets in public archives came out as the best option, together with data publications. As a next step on the technology and service level, special software for analysis and visualisation tools on the publisher’s website will give readers of online journal articles interactive access to (parts of) the underlying dataset residing in the data archive.

As the first steps in linking datasets and journal articles are made by the collaborative efforts of publishers and data archives, logically the next steps should include developing citation metrics for datasets based on DataCite or equivalent persistent identifier standards: an easy-to-understand citation score that can be automatically generated by simple tools on the internet. A dataset impact factor, using among others the bibliometric indicators of articles that have cited the datasets, will almost certainly provide an enormous stimulus for data producers to publish their datasets. Already, there are indications that biomedical articles with publicly available data are cited more often than articles without the availability of underlying datasets (18). The academic record of the data producer will thus benefit twice: the journal article will be cited more, and there will be additional citation metrics for the published dataset.

Another important stimulant for data producers would be to include published datasets in the same manner as publications in research evaluation exercises. In short, published datasets should give the data producer recognition in a similar way as publications do now.

With regard to the situation in the KE countries, many initiatives in this essentially international domain were generated in Germany (ESSD, DataCite). With regard to possible next steps, it is important to strive for the inclusion of published datasets in research evaluation exercises and for initiating projects to develop citation metrics for datasets.

3.3 Rules and codes of conduct

Another way to stimulate data sharing and publishing is to have important national and international scientific organisations issue codes of conduct on research data management, sharing and publishing. Such codes of conduct on datasharing are seen as important stimulants by researchers, especially from life sciences and social sciences (5) (6).

Table 1. Codes of Conduct issued in KE countries

Generic codes of conduct for sharing research data			
Denmark	Germany	The Netherlands	UK
	DFG: Recommendations for the secure storage and availability of digital primary research data (2009) (21)	VSNU: Code of Conduct for science (2004) (34)	RCUK: Common Principles on Data Policy (2011) (19)
	Alliance of German Science Organisations: Principles for handling research data (2010) (20)		UK Research Integrity Office Code of Practice (2009) (35)

Table 1 gives an overview of the main recent activities in the KE countries. Recently, the joint Research Councils in the UK (RCUK) issued their *Common Principles on Data Policy* (19). These principles:

- affirm the Open Access principle for publicly funded research data
- prescribe proper policies and practices for data management and meta data
- emphasise the need of attributing the creator of the original dataset by the re-using researchers
- state that it is appropriate to use public funds to support the management and sharing of publicly funded research data.

Similar statements on these elements have been issued in Germany by the Alliance of German Science Organisations in 2010 (20), and by the KE partner DFG in 2009 (21). Codes of conduct issued previously include statements on the retention of research data for a number of years for purposes of validation, but lack the above-mentioned elements. More recent code of conduct-like statements on data sharing from national scientific organisations in Denmark and the Netherlands are unknown to us.

3.4 Requirements by funding organisations

The UK research funding organisations appear to be leaders in setting requirements on research data for research grants. All seven research councils and the Wellcome Trust stipulate requirements on data management. Key elements of the fundes' requirements include:

- **data plan**: a requirement to consider data creation, management or sharing in the grant application

- **access/sharing**: promotion of data sharing or re-use. Some research councils also require all research publications to include a statement on how the supporting data can be accessed

- **long-term curation**: stipulations on long-term maintenance and preservation of research outputs. What is meant by long-term preservation varies per research council: expected periods for preservation range from three to more than 10 years

- **monitoring**: whether compliance is monitored or action is taken, such as withholding funds. Two research councils can withhold final grant payments if data are not deposited

- **guidance**: to what extent does the research funding organisation provide guidance to its grant holders on research data management and sharing? It varies from best practice guides and toolkits to professional support from designated data centres

- **costs**: a willingness to meet data management and sharing costs: four research councils and the Wellcome Trust state that these costs can be included in the grant proposal.

In the UK there are also more specific mandates, defining requirements from the individual funders[3] and the beginning of explicit policies in universities[4].

[3] http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

[4] http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies

Table 2. Research funding requirements in the KE countries[5]

Funding organisations requirements	Data plan	Access/ sharing	Long-term curation	Monitoring	Guidance	Costs
UK						
AHRC	+	+	+/-	-	+	-
BBSRC	+	+	+	+	+	+
CRUK	+	+	+	+	-	-
EPSRC	+	+	+	+	-	+
ESRC	+	+	+	+	+	+
MRC	+	+	+	-	+/-	-
NERC	+	+	+	+	+	-
STFC	-	+	-	-	+/-	-
Wellcome Trust	+	+	+	+	+	+
Germany						
Deutsche Forschungs Gemeinschaft	(+)	(+)	(+)			(+)
Denmark						
Council for Independent Research	-	-	-	-	-	-
Council for Strategic Research	-	-	-	-	-	-
The Netherlands						
NWO: arts and humanities social sciences	+	+	+	+	+	+
NWO: other scientific disciplines	-	-	-	-	-	-
STW	-	-	-	-	-	-
Senter Novem	-	-	-	-	-	-

In the Netherlands the main scientific funding organisation NWO claims co-ownership of the research data of projects they fund and as such the right to have a say in making the data available after the projects. Other funders in the Netherlands have no explicit requirements on data sharing and research data management[6].

In 2010, the German DFG added an item to its guidelines requesting grant applicants to state what they plan to do with the research data during and after the proposed research project. This 'light-touch' requirement asks grant applicants to address the issues of data management, access and sharing and long-term curation in their proposals, but has no mandatory components. The requirement is supposed to raise awareness on data sharing and data management amongst applicants. Additionally, the statements made by the proponents are part of the review process. Scientific reviewers will reflect on the willingness to share data and to use potentially existing data repositories or may stimulate the development of such infrastructures where necessary (see Chapter 6).

The funding organisations in Denmark have no requirements on datasharing or research management for research grants to date.

Although UK research funding bodies pay far more attention to research data management and datasharing, one might question the extent to which these requirements are complied with in practice. A better understanding of the findings of monitoring activity is needed. At

[5] Data for the UK research funders are based on the overview at the DCC website http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

[6] DANS (Data Archiving and Network Services), the Dutch data centre for social sciences and humanities, also has agreements with ministries and government bodies on storing data resulting from policy-oriented research

the same time, it is important to ensure the attention to research data management planning does not become an ineffective box-ticking exercise consuming the time of researchers, research assessment panels, host institutions and research council review panels to little overall benefit. The real challenge here lies in achieving a greater level of compliance with both the letter and the spirit of data management and datasharing policies.

For the other KE countries, the first step would be to encourage the research funding organisations to implement and/or strengthen the requirements on data for grant holders along the above-mentioned lines. An important challenge will be to create a greater engagement with these issues, and the development of policies and supporting mechanisms, on the part of research institutions, universities and other stakeholders.

3.5 Data availability policies of journals

Most peer-reviewed journals in science, technology and medicine have files where authors can add supplementary data related to the article, such as data sets, multimedia files, large tables, animations, and protocols. In a longitudinal study of 28 high impact medical journals, the percentage of articles that contained supplementary material (online only) increased from 7% in 2003 to 25% in 2009 (22). Most supplementary material consisted of tables, figures and videos. The trend is towards ‘enhanced’ publications, where journal articles are supplemented with various types of data.

Data sets underlying journal articles are called replication datasets because they can be a subset of a larger dataset that could be used for more than one publication. Editorial boards of scientific journals are increasingly pressing authors to offer access to the underlying (replication) data sets in combination with the journal article. Such data availability policies have prompted the UK Data Archive (UKDA) to offer a special service. The UKDA store enables researchers to deposit their (replication) dataset themselves, thereby bypassing the ingesting procedures of the data archive itself (23).

Although most journals (over 90% according to the PARSE.Insight survey) offer the possibility to deposit supplementary materials with the journal article, there are often limitations to the size and format of the files (thus hampering re-usability) and long-term preservation is not guaranteed in many cases. Data are usually not curated in a professional way.

A number of journals have mandated their data availability policy. Leading journals, such as *PLoS One* and *Science* have these policies in place. As the landscape of public data archives is patchy (see Chapter 5), several scientific areas lack an appropriate public data archive or repository. To fill the gap, a consortium of journals in bioscience has implemented a collaborative data repository, Dryad (www.datadryad.org). A joint declaration in the *American Naturalist*, *Evolution*, *Journal of Evolutionary Biology*, *Molecular Ecology*, *Heredity* and other key journals in evolution and ecology implemented the following mandatory data availability policy[7]:

> “This journal requires as a condition for publication that data supporting the results in the paper should be archived in an appropriate public archive.”

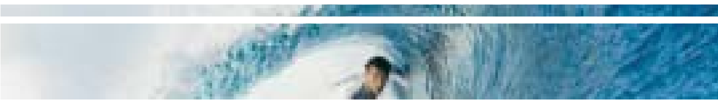
The statement continues by naming data repositories for specific data types (e.g. Genbank) and the Dryad generic repository for all other data.

[7] http://www.datadryad.org/jdap

Furthering data availability policies among journals is mainly in the domain of editorial boards and publishers. However, the KE partners can assert their influence here. Recently, JISC funded the Dryad-UK project that is expanding the Dryad initiatives into new research areas (primarily infectious diseases), establishing new partnerships with journals (including BMJ, BMC and PLoS titles[8]) and developing a robust business model for this data repository. Next to furthering open access to research articles a second line of funder action could be to stimulate a data availability policy for all journals, both OA and subscription based.

[8] http://blog.datadryad.org/2011/06/27/bmj-open-a-new-partner-and-an-expanded-scope/

4. TRAINING



4.1 Introduction

In this fast developing data-intensive world in scientific and scholarly research, what kind of skills are needed for creating, handling, manipulating, analysing and storing for re-use of large amounts of data by others? In a landmark study (24), Swan and Brown made an inventory of the skills and needs of data scientists, data managers and data librarians now and in the future. Their report observes that some researchers in data-intensive research areas have acquired considerable skills in handling and managing data themselves or have a colleague who has these skills, but in other areas researchers turn to the institutional IT services or library for assistance and advice. The report also observes that data scientists usually ended up in their roles accidentally as formal education hardly exists.

The report distinguishes several specialist roles: **data scientists** working as part of a team of researchers or in close collaboration with them, who are responsible for computing facilities, storage and access, and **data librarians** from the library community who are specialised in the curation, preservation and archiving of data.

The main recommendations of the Swan and Brown study are:

- develop data skills and data science in the research domains by postgraduate training courses on the fundamentals of data management for researchers and develop career options for data scientists
- develop data skills in libraries of research institutes by training
- develop curricula for data librarians.

The report observes the difference between ‘big science’ and ‘small science’.
Big science – large research facilities with data centres – employs data scientists but has no system in place for professionalisation, career structures and recognition. In small science – smaller research programmes and projects that are run most often at universities – the responsibility for data management is in the realm of the institutes. There, gaps in the handling of research data exist in terms of skills and specialised personnel. Libraries are trying to fill some of these gaps by creating new positions for data librarians.

4.2 Data skills in research domains

Based on the Swan and Brown study, several projects have been initiated in the UK. The Digital Curation Centre runs training programmes (DCC 101 and Tools of the Trade) that directly target researchers and a ‘train the trainer’ programme on these subjects.
Summarising findings from a ‘Research Data Management Forum’ organised by the DCC, an article by Pryor and Donnelly made the case that ‘data skills should be made a core academic competency’ and that ‘data handling [should be] embedded in the curriculum’ (25). In response to this, JISC funded a set of five projects to embed research data management training in postgraduate academic curricula. The aim of these projects is to create discipline-focused postgraduate training units, which can be re-used by other institutions in order to stimulate curriculum. Most projects are complete or are about to complete at the time of writing[9].

[9] http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain.aspx

Another approach in the UK is formed by the Researcher Development Framework, developed by Vitae[10]. This framework describes competences needed by researchers over their whole career, including knowledge and competences in the area of data management. The Research Information Network's Working Group on Information Handling is developing practical guidance for this through the Data Management Skills and Support Initiative (DaMSSI[11]) which is jointly funded by JISC.

In the Netherlands, DANS (Data Archiving and Network Services) annually provides 15 to 25 workshops, training courses and guest lectures for data managers of archaeological institutes, for researchers of the institutes of the Royal Netherlands Academy of Arts and Sciences (KNAW) and for master's students in humanities and social sciences at various Dutch universities. The other KE countries appear to have fewer activities in this area. In Denmark, DEFF is starting a project aimed at mapping the institutions' data policies and ultimately raising awareness among researchers and other stakeholders. Germany has no systematic nation-wide research data management training activities targeting researchers.

4.3 Data librarians

Many research libraries are strategically repositioning themselves within their institute and seek a role in supporting research by setting up repositories for open access publications (24) and creating discovery services for datasets[12]. Along the same lines, libraries are taking responsibility for facilitating research dataset publishing and archiving (see also Chapter 7). As for skills and competences, the relatively new position of the data librarian has been proposed and in some cases has already been created.

In the US data librarian positions and data libraries were established some time ago. American library units acquire datasets from third parties (often governmental agencies or commercial parties) and make them accessible for their academic communities, thus creating a library of datasets. In the UK, an example of such a data library can be found at the University of Edinburgh. Most data libraries are now also supporting researchers with their research datasets by offering services such as a data repository (23).

Generally, the role of data librarians in the new setting of supporting researchers in data publishing and datasharing has yet to be developed. University libraries in the Netherlands are in the process of adding this role to the official job description of information specialist. SURFfoundation has set up a discussion forum on research data that includes data scientists and members from the library community to raise awareness (*Onderzoeksdata Forum*).

Formal training programmes or curricula for librarians to acquire the appropriate skills and competences for this new position appear to be both scarce and scattered. However, there are a few examples: in Germany, the universities for applied sciences ("Fachhochschulen") in Potsdam and in Cologne have developed Bachelor and Master Programmes in "Library and Information Sciences" that include research data management. This seems to reflect the general rising awareness of the future importance of data librarianship.

[10] http://www.vitae.ac.uk/rdf

[11] http://www.dcc.ac.uk/training/data-management-courses-and-training/skills-frameworks

[12] British Library: http://www.bl.uk/reshelp/experthelp/science/sciencetechnologymedicinecollections/researchdatasets/datasets.html

4.4 Challenges

The longer-term challenge is to make best practices in research data management skills a core and fundamental component of all disciplines provided to researchers in their undergraduate and postgraduate training. The Research Development Framework in the UK could provide a vehicle to realise this[13]. However, it seems clear that far more effort is needed to bring about this change in the relevant curricula of universities (see Box 2). It also seems that specialised curricula to educate future data scientists are needed. However, it is not yet clear what kind of career opportunities will arise for data scientists.

In the short term, the main challenges are to set up learning frameworks that offer progressive training options for professional development over the course of a career. This would enable researchers to develop themselves.

As for training data librarians, there are issues to be solved in the curricula for librarians and information professionals. Is it feasible to train information science students in the more (discipline-specific) technical aspects of data management and curation? There are also a number of issues regarding boundaries and workflow: what is the overlap in skills between researchers trained in the basics of data management and specialised data scientists and data librarians? What would a seamless workflow look like from data creation to data management to longer-term curation and re-use, including the three roles? Finally, to what extent can training be generic versus discipline-specific?

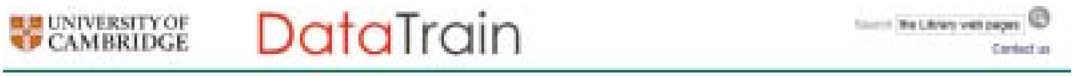
Box 2. The Data Train project

To help build data management capacity among its postgraduate students and early career researchers, the University of Cambridge is working with the Archaeological Data Service (ADS) to develop discipline-specific data management modules for both archaeology and social anthropology. These departments are closely associated with students across the departments attending the same introductory courses. Both departments currently run courses in related topics such as computing and research methods but data management is currently not well covered.

In close collaboration with staff and students from the participating departments, the University Library, and the ADS, modules on data management planning, data creation, selection, long term preservation, access management, use and reuse and rights issues are being adapted to fit alongside existing course modules on research methods. The new course modules will be piloted as part of the departments' research methods training in spring 2011 and will be continued within the departments beyond the life of the project in collaboration with staff from DSpace@Cambridge.

In order to ensure that training resources are aligned with relevant standards, staff from the ADS are involved in the project to provide guidance and support, feedback has been sought from the UKDA on the Social Anthropology modules. The ADS will also serve as a dissemination point for the training resources to help them reach the wider archaeological community.

http://www.lib.cam.ac.uk/preservation/datatrain/



[13] http://www.vitae.ac.uk/policy-practice/234381/RDF-overview.html

5. DATA INFRASTRUCTURE

5.1 Introduction

> “Data can be equated with money that has value only if it is used and circulated. As the different currencies can be stored in the globally interrelated bank infrastructures, we need persistent, highly available and compatible data infrastructures where data from various disciplines can be stored and fetched from.”
> PARADE - Partnership for Accessing Data in Europe[14].

Infrastructure is a broad notion. It may have technical, legal, organisational and sometimes cultural or political connotations. For an action programme choices have to be made in relation to available resources and the mission of the actor. Given the nature of Knowledge Exchange as a co-operative effort of four national catalyst organisations, a focus on interoperability seems most appropriate, especially as there is still a world to be won in this respect.

5.2 Heuristic view of the landscape

In big science, data sets generated by large research facilities are often stored in their own data centres. An example is LOFAR, a radio telescope generating large datasets. Researchers can submit a request for an observation. When granted, the researcher has privileged access to the observation data for a period. Then the datasets can be re-used by other researchers as well under the sole condition of mandatory acknowledgement of LOFAR.

In small science, datasets are often stored by the individual researcher or research group during the research project. After the research project is completed, the data can be stored for some time in a similar manner – no transfer takes place. The dataset or part of the dataset can be a replication dataset linked to a journal article. In these cases, the dataset is mostly transferred to a repository or a data archive by the institute or by the journal publisher.

The resulting infrastructure is diverse, fragmented, in flux and organised differently across various disciplines in different countries. Figure 2, presented in a study by SURFfoundation gives an impression of the diversity related to origin and storage of datasets (23). Later stages of the data lifecycle reflect this diversity when it comes to aggregation services and data discovery services. Further, sometimes re-use is enabled by legal frameworks and codes of conduct. For the underlying infrastructure the application of standards, both technical and semantic (e.g. ontologies), is critical. Even the notion of quality is affected by the provenance and purpose of the data.
By working towards interoperability all these issues will pop up automatically and are to be ‘solved’ i.e. to be dealt with in a practical, non-academic way. There is less use in tackling them in isolation.

[14] http://www.csc.fi/english/pages/parade

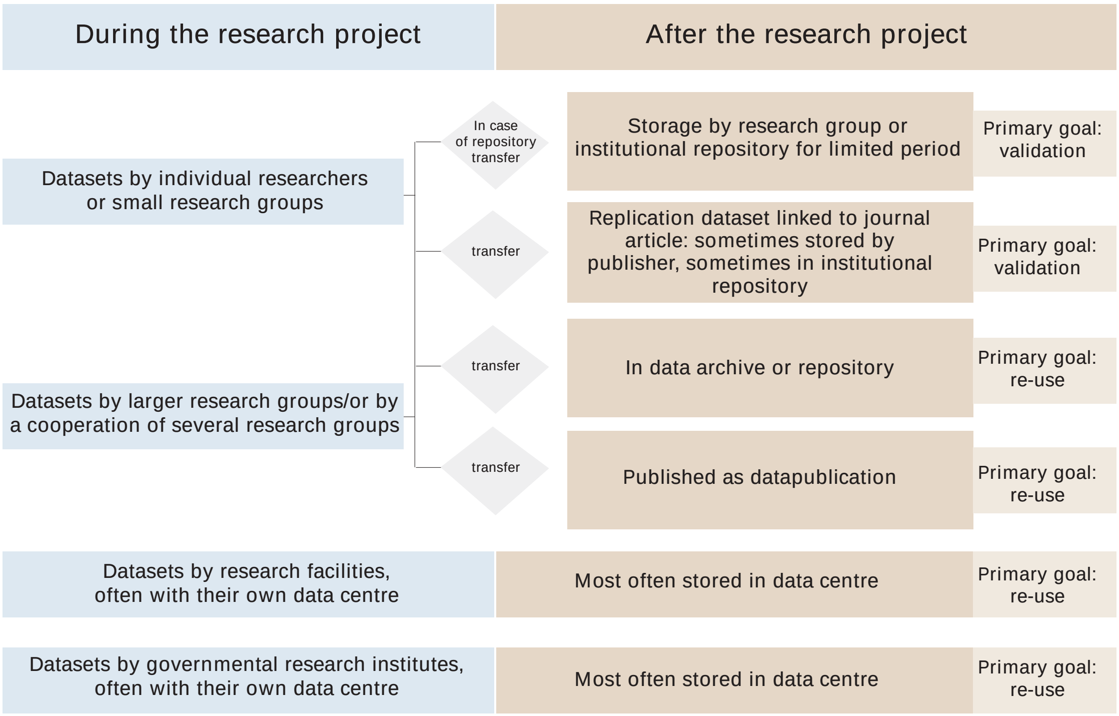

During the research project
After the research project
Datasets by individual researchers or small research groups
Datasets by larger research groups/or by a cooperation of several research groups
In case of repository transfer
transfer
transfer
transfer
Storage by research group or institutional repository for limited period
Primary goal: validation
Replication dataset linked to journal article: sometimes stored by publisher, sometimes in institutional repository
Primary goal: validation
In data archive or repository
Primary goal: re-use
Published as datapublication
Primary goal: re-use
Datasets by research facilities, often with their own data centre
Most often stored in data centre
Primary goal: re-use
Datasets by governmental research institutes, often with their own data centre
Most often stored in data centre
Primary goal: re-use

Figure 2. Origin and storage of research data sets; from: SURFfoundation [15]

Nevertheless, two different levels can be roughly distinguished:

1. **Institutional data infrastructure**: Development of data management procedures and processes in research institutes is very much in flux. On the one hand there are virtual research environments – a sort of collaborative 'electronic workbenches' – that enable researchers to work together and create, collect and process research data. On the other hand, there are institutional data repositories, where the datasets resulting from research projects can be stored and shared with other researchers. Because both parts of the institutional data infrastructure are in their infancy, and more often absent or very scattered than omnipresent, there is a wide variety of approaches to these issues among the various institutions.

2. **Disciplinary or (inter)national infrastructure**. The ecology of data centres and data archives organised along disciplinary, national and international lines is diverse.

   Disciplinary data centres are developed by scientific communities according to their own needs. Examples are the world data centres in the geo and environmental sciences (see Box 3), genome databases, crystallography databases, and the International Virtual Observatory Alliance in astronomy.

   With regard to internationally organised data centres, the European Union is building on an international research infrastructure for the European Research Area (10), which increasingly includes units focused on research data. Related to this effort, the MERIL project will make an inventory of the European research landscape including research data

[15] Like the Riding the Wave report, this paper focuses on data created in academic research. Commercial data and data from governmental agencies are not included

infrastructure elements. This project will lead to a portal that will probably be publicly available in 2012. In this context the European legal framework for developing European research infrastructure consortia (ERIC)[16] and the Vision for Global Research Data Infrastructures (GRDI project) should also be mentioned as they cover important organisational aspects to enhance the European research infrastructure.

Box 3. World data centres in the geo and environmental sciences

The World Data Centre System
The World Data Centre (WDC) system includes 52 centres in 12 countries. Its holdings include a wide range of solar, geophysical, environmental, and human dimensions data. These data cover timescales ranging from seconds to millennia and provide baseline information for research in many disciplines focused on monitoring changes in the geosphere and biosphere – gradual or sudden, foreseen or unexpected, natural or synthetic. WDCs are funded and maintained by their host countries on behalf of the international science community. They accept data from national and international scientific or monitoring programs as resources permit. All data held in WDCs are available for no more than the cost of copying and sending the requested information.
http://www.ICSU-wds.org

World Data Centre for Geomagnetism
The World Data Centre for Geomagnetism, http://web.dmi.dk/projects/wdcc1/, situated in Copenhagen, DK, has collected analogue and digital geomagnetic data from a worldwide network of magnetic observation. The data and services are available for researchers and organisations without restriction. Data are exchanged based on common guidelines enabling sharing and re-using, together with online publication and visualisation, and are available through an online catalogue.

5.3 Institutional data infrastructures in the Knowledge Exchange countries

In the UK, JISC has funded a number of projects under the banner Research Data Management Infrastructure. Some projects are directly focused on setting up an institutional infrastructure for data. For example, the Institutional Data Management Blueprint (IDMB) aims to create a practical institutional framework for managing research data that facilitates ambitious national and international e-research practices, encompassing a whole institution, exemplified by the University of Southampton. Practices are based on an analysis of current data management requirements for a representative group of disciplines with a range of different data. The results of the IDMB project are expected during 2011.

The more recent Shared Services and the Cloud Programme resourced by the University Modernisation Fund are concerned with developing a national shared infrastructure for research data management and will set up a virtual server infrastructure (a ‘cloud’) to offer cost effective data management and storage services to higher education institutions in England. Complementing this shared infrastructure, four projects have been funded to develop software as service applications for managing research data. Roughly £3.5 million has been invested in this aspect of the shared IT infrastructure programme[17].

In Germany, the DFG recently launched a call for proposals entitled “Information Infrastructures for Research Data”. The programme text states, “A nationally addressable organisational structure is urgently required, for both the humanities and the natural and life sciences.” The text leaves open how this is to be done. It merely suggests “by discipline, by institution, or in national repositories for research data”. Each initiative was required to establish a close cooperation between information facilities (primarily libraries, but also scientific data centres) and stakeholders in research. The intention was to match the

[16] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric
[17] http://www.jisc.ac.uk/whatwedo/programmes/umf.aspx

researchers' requirements with the expertise of the information professionals in developing the required information infrastructures. In spring 2011, 27 project grants were approved in this programme.

In the Netherlands, three technical universities joined forces a few years ago to set up the 3TU Datacentre, which can be seen as a multi-institutional data repository. At Tilburg University, the library is actively supporting researchers in data management and has developed best practices in supporting research data management in the fields of finance and experimental economics. Other Dutch universities are experimenting with DataVerse, an open source application for publishing research data from social sciences. Discussions are now taking place about options to broaden the scope of the 3TU Datacentre to develop it into a data repository for all Dutch universities and possibly merge it with the national data archive DANS (see Section 5.4/Table 3).

In Denmark a prototype is under development at the Technical University of Denmark on the basis of Fedora-Commons software. DataVerse is applied at other libraries, the Danish Data Archive uses its own software. The coming e-Science Centre in Denmark will have to address this issue.

5.4 Disciplinary, national and international data infrastructure

Table 3 (below) presents a, probably incomplete, overview of data archives, data centres and data repositories in the four KE countries[18]. With 17 entries, the UK data infrastructure seems the most widespread, covering the most disciplines. In Germany, there is presently no comprehensive overview on the data repositories and data archives. To solve this, there are concrete plans to develop a portal to access various German data archives and repositories. Table 3 lists only nine data centres based in Germany but this list is almost certainly incomplete. The Danish national data infrastructure is very limited with only the Danish Data Archive for social sciences as its national data archive. In the Netherlands, DANS covers the arts, humanities, and social sciences and is initiating services for other scientific areas. The 3TU Datacentre – currently serving three universities – was discussed in Section 5.3, under the institutional infrastructure.

LOFAR site at Effelsberg - see table 3. Source: www.lofar.org

[18] The aforementioned MERIL project may produce a more comprehensive overview. The project will lead to a portal that will probably be publicly available in 2012

Table 3. 'Impressionistic view' of data archives, data centres and data repositories in KE countries [19]

Denmark	Germany	Netherlands	UK
Danish Data Archive	World Data Centre for earth and environmental sciences PANGAEA	DANS (Data Archiving and Network Services)	Archaeology Data Service
European environment agency data service	GFZ Seismological Data Archive	World Data Centre for soils	Biological Records Centre
World Data Centre for Geomagnetism (shared with the BGS in Edinburgh/UK)	World Data Centre for climate	3TU.Datacentre	British Atmospheric Data Centre
	World Data Centre for remote sensing of the atmosphere	Max Planck Institute language archive	British Oceanographic Data Centre
	GESIS Data Archive for the social sciences	European Directory of Marine Environmental Data (EDMED)	Chemical Database Service
	German Satellite Data Archive	LOFAR (radio telescope)	eCrystals/Crystallography Data Service
	CellFinder	KNMI (meteorology)	Edinburgh DataShare
	DNA Bank Network		Environmental Information Data Centre
	ZPID Forschungsdatenzentrum für die Psychologie		European Bio- Informatics Institute
			MarLIN / DASSH
			National Biodiversity Network (NBN) Gateway
			National Cancer Research Initiative / Information Network
			National Digital Archive of Datasets (NDAD)
			National Geophysical Data Centre (NGDC)
			National Geoscience Data Centre
			NERC Earth Observation Data Centre
			NERC Environmental Bioinformatics Data Centre
			Oxford Text Archive (OTA)
			Polar Data Centre
			ShareGeo (EDINA)
			The UK Solar System Data Centre
			UK Data Archive
			UK National Air Quality Archive
			Visual Arts Data Service
			World Data Centre for Glaciology and Geocryology

[19] Based on www.datacite.org/repolist with additions by the KE working group on primary research data and from the SURFfoundation studies (5) and (21)

5.5 Challenges in realising an ecosystem of data repositories

There are a number of challenging issues in the further development of the proposed ecosystem of data repositories. These issues include, but are not limited to:

- Gaps in the present data infrastructure: at the institutional level as well as the disciplinary/(inter)national level. As a result data sets may be 'homeless' i.e. even if the creator is prepared to deposit and share them an adequate repository is missing.

- Connectivity issues: connectivity issues play a role within institutions, where the issues are concerned with the connection between institutional data infrastructures and the workflow of the researchers. Also, connectivity issues play a role with regard to the connection between institutional data infrastructures and the national data infrastructure. These issues are part of the broader problem of how to make the research data infrastructure interoperable – not only within a scientific discipline, but also across disciplines. Basically, this concerns an international challenge to set technical standards for software, data models and protocols and includes semantic aspects (12) (26).

The KE partners can initiate further developments in both areas by identifying gaps in their national data infrastructure (see Box 4) and support projects that deal with connectivity issues.

Box 4. An example of a gap in a national data infrastructure

The **Galathea expeditions** comprise a series of three Danish ship-based scientific research expeditions in the 19th, 20th and 21st centuries, carried out with material assistance from the Royal Danish Navy and, with regard to the second and third expeditions, under the auspices of the Danish Expedition Foundation. All three expeditions circumnavigated the world from west to east and followed similar routes.

The first two historical expeditions (1845-1847 and 1950-1952) gathered large collections of research data on paper. The third expedition was carried out in 2006 and 2007 and collected a great deal of research data in digital form. Now, however, only a few data collections from the third expedition can be preserved. The remaining research data are dispersed over local hard disks with little or no chance of long-term preservation. As a result, this third modern-day expedition could leave less research data behind than the two earlier ones. This dramatic case shows the importance of a establishing a research data management plan at the start of a research project that includes an approach for long-term preservation.

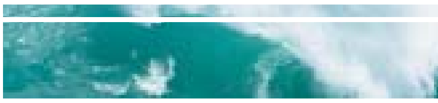
HRH The Crown Prince is the royal patron of the Danish Expedition Foundation.

Scientific research of high quality provides the foundations of the Galathea 3-expedition.

Read the latest status report from selected research projects.

You can see pictures from the different stages of the expedition in the photo gallery.

6. FUNDING THE DATA INFRASTRUCTURE



6.1 Introduction

After presenting the vision of a collaborative science data infrastructure, the *Riding the Wave* vision document states bluntly, “This is expensive.” But how expensive will it actually be? What are the present insights into the costs and benefits of such a data infrastructure? And, the ultimate question: who will bear the costs?

6.2 Overview of the main research funding organisations in the four countries

The national research infrastructures and the funding organisations in the four KE countries differ considerably in size and structure. A broad outline for each country is presented in Table 4.

Table 4. Broad outlines of the research funding infrastructure in each partner country

Country overview	Denmark	Germany	Netherlands	UK
No. of universities (full member of EUA)[20]	7	75	14	66

Other important research performers[21]	Governmental research institutes are merged with universities since 2007	Max Planck Society (80 institutes);
		Fraunhofer Society;
		Helmholtz Association;
		Leibniz Society
			KNAW Research institutes (19);
			NWO Research institutes (9),
			Large technological institutes at TNO (4)
				Main actors in public sector research are higher education institutes, mostly universities

Some major funding organisations	Council for Independent Research (five research councils);
	Council for Strategic Research (policy-oriented research)
		Deutsche Forschungsgemeinschaft DFG (German Research Foundation);
		Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)
			Research council: NWO;
			STW (technology foundation);
			SenterNovem (innovation agency) and ZONmw (health sciences)
				Research Councils (7 per disciplinary area);
				The Wellcome Trust (private funder)

Germany and the UK have the largest national research infrastructures, each with around 70 research universities. In the UK, most research takes place in universities, whereas Germany also has important public research organisations, including the Max Planck Society, the Fraunhofer Society, the Helmholtz Association and the Leibniz Society.

In Germany, the German Research Foundation is the main channel for public funding of projects for basic research. The Federal Ministry of Education and Research supports programme oriented research and provides large scale support for initiatives of national relevance.

The UK has seven research councils, each covering a research discipline and an important charitable funding organisation (the Wellcome Foundation).

[20] From the website European University Association, checked at 2011-06-18; http://www.eua.be. Not all universities are member of the EUA, but these numbers give an overall indication of the number of the larger research universities

[21] From the ERAWATCH Country report 2009 series, JRC-IPTS; Germany, the Netherlands, United Kingdom, Denmark

Denmark and the Netherlands have much smaller public science infrastructures with seven and 14 universities respectively. Denmark has two councils funding research while in the Netherlands, the main funding body is the research council NWO, there are also two smaller funding organisations for technical and applied sciences and ZONmw for health research and development (see Table 1).

6.3 Costs and benefits

In an ESFRI document, the overall cost of the research data infrastructure is estimated to be 10-15% of the overall cost of the research infrastructure (27). JISC has funded a number of projects to investigate this in detail. The project entitled Keeping Research Data Safe (28) resulted in the following findings:

- An institutional data repository is likely to cost a factor of 2.5 to 4 more than an institutional repository for publications. More personnel are needed (typically 2.5 to 4 FTE) and equipment costs are much higher (several tens of thousands pound sterling on an annual basis).
- The costs are distributed as follows: approximately 55% on outreach/acquisition/ingest, 31% on access and 15% on archival storage and preservation.
- Relatively high preservation costs in the early years diminish substantially over time.

These findings led to the following recommendations:

- Potential efficiency effects will come mostly from future tool development that supports the automation of ingest and access.
- Repositories should take advantage of economies of scale. This recommendation is supported by the UKRDS feasibility study that advises universities to share a data repository (29).

Subsequent RDMI projects funded by JISC investigated in detail the benefits of an institutional data infrastructure (30). Some of the main benefits are:

- **Innovation**: new research funding and research opportunities, inspiration for new research, stimulating new networks and collaborations.
- **Effectiveness**: reliable citations to data, no loss of access to data as a result of postdoc turnover, guidance and training for researchers embedded in the schools.
- **Efficiency**: rapid access to research data and derived data, time and efficiency savings, enhanced finding and organising of data, no recreation of existing data.

A recent article in *Nature* highlighted the benefits of data archiving from the perspective of research funders. The underlying study looked into the number of papers generated by re-using data from the Gene Expression Database and found that more than 1100 articles – published by authors other than the original data producers – re-used one or more of 2700 datasets that had been deposited up to three years earlier. Translated into investment terms, it was estimated that the annual investment of £400,000 in the Dryad repositories could contribute to more than 1000 papers within four years. This compares favourably with an estimated 16 papers from the same amount of money invested in original research. The authors of the *Nature* article conclude, “Public data archiving can generate important new results for a small fraction of the currently accepted cost of doing science. To maximise the impact of the support they provide to individual investigators, research funders should include the maintenance of data archives as an integral part of their investment portfolios” (31).

In conclusion, a well-organised data infrastructure at the institutional level gives the institute and its researchers a competitive edge because of increased effectiveness and efficiency, and new opportunities for novel research approaches. Similarly, this also appears to be valid for a disciplinary/(inter)national data infrastructure. Thus, there appears to be a clear business case for setting up data infrastructures at the institutional and disciplinary/(inter)national levels. These arguments are corroborated by the findings of a recent RIN/JISC study on data centres showing a high usage of data centres (thousands of researchers with millions of downloads each year) and showing widely-perceived benefits of research efficiency and research quality (32).

6.4 Funding

The roadmap presented by PARSE.Insight (33), distinguishes three stages for data infrastructure: prototypes, emerging infrastructures used by early adopters, and long-term infrastructures. The funding for creating prototypes and developing emerging infrastructures is and will be carried out by research project funding, for which national and international funding opportunities are available. The PARSE.Insight roadmap states that developing a business model for the long-term infrastructure is ‘difficult’.

Research funding organisations appear to have taken on a responsibility to support elements of the long-term disciplinary/(inter)national data infrastructure. In theory, this is made clear by funding organisations stating that it is appropriate to allocate public funding to data structure elements (14) (19) (20) (21). In practice, some research funders in the four KE countries do indeed fund data infrastructures: some UK Research Councils are funding data archives (NERC, ESRC), the Dutch NWO is partly funding DANS, and the Danish government is funding the Danish Data Archive. In addition, JISC and SURFfoundation have various programmes in place for stimulating the development of a collaborative data infrastructure, as mentioned in various paragraphs in this report.

Germany’s DFG is restricted by its statutes to funding on project basis. DFG can fund development and pilot projects for data infrastructures, but long-term funding is supposed to be taken up by the host institutions, such as universities or research institutes. There are a several examples of this, including the Psychology Data Archive (PsychData) run by the Leibniz-Institute for Psychology Information, and the PANGAEA system run by the Helmholtz Association. On the whole, funding support for elements of the data infrastructure appears rather patchy and uncoordinated at a national level.

The institutional data infrastructure and development projects mentioned in this report are funded by national and international organisations. Long-term funding of data infrastructures is an issue since institutional budgets are under great pressure. Additional structural costs for a relatively novel infrastructure can be expected to meet strong resistance in the governing boards of universities and institutes. However, as part of a strategic realignment of research libraries, a number of libraries are taking on a new role in organising and maintaining a data infrastructure. Some of these libraries absorb a part of the costs of data infrastructure by their existing library budgets in the course of this strategic realignment.

6.5 Challenges in funding the data infrastructure

In terms of boundary issues between institutional and disciplinary/(inter)national and data infrastructures, it is generally thought that data archives would be better organised along disciplinary lines because of the expertise needed to preserve and curate the datasets. Following a similar line of thinking, short term storage of data is seen as a task for institutes, while long-term preservation – again due to the expertise needed – is seen as a task for data archives (see Box 5). However, these lines of thought might become out-dated as a result of upcoming technical developments in virtualisation and federated data repositories. These developments might make it more feasible for institutes to join forces in setting up a data infrastructure. In short, the ‘cloud’ might allow the boundaries between institutional and national data infrastructures become permeable in the longer term. This will make the challenge to create criteria for funding the data infrastructure even more pressing – who funds what and why and what are the evaluation criteria? Another important challenge will be to minimize the costs of the data infrastructure while maximising the benefits for the researchers.

Box 5. An example of a funding model: the e-depot of Dutch archaeology

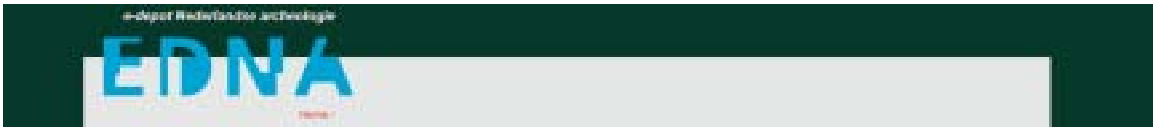
“Digital archaeology requires a digital memory”
This slogan was used to bring care for digital data to the attention of Dutch archaeologists during the EDNA pilot project that set up the e-depot for Dutch archaeology. In 2007, it was backed up by EDNA II, the retrospective archiving project. In the years that followed, the e-depot continued to grow, from 5,000 to 10,000 deposited datasets by 2009, and reached 15,000 at the beginning of 2011.

The archaeology e-depot is located at DANS. The e-depot stores digital files of research data from Dutch archaeologists. These files contain primary data on excavations, regional explorations and material studies. Notably, they concern completed and published research results, in which the authors have made the basic data accessible to other scientists. The e-depot ensures durable archiving and access to all the digital documentation from archaeological research. Research descriptions and data can all be downloaded via the EASY archiving system.

Agreements to this end have been laid down in the quality standard for Dutch archaeology. Dutch archaeologists accepted that obligation together, based partly on their good experience with the usability of EASY. EDNA is a collaboration between DANS and the Cultural Heritage Agency (RCE). DANS is willing to invest in EDNA and is supported by the Dutch Ministry of Education, Culture and Science on the condition that in due course the archaeological field will be self-supporting in financing the digital archiving of research data.

To safeguard the continuity of the e-depot for Dutch archaeology it is important to ensure adequate funding for the longer term. Project grants are limited and this means that a new cost model is necessary. DANS will continue to support the discipline of archaeology and has the expertise and digital infrastructure that the e-depot needs. Structural financing of the costs of archiving focuses mainly on personnel costs to process and control data, as well as the conversion into the correct sustainable format and display of data sets. There are both one-time ingesting costs and structural archiving costs as well as additional charges, extra services and overhead costs. The proposal is a cost model where funding bodies, scientific researchers and commercial archaeological researchers can take into account fixed rates to deposit datasets, which are known in advance.

The e-depot is investigating whether there is enough support for this model. Commercial archaeological companies can pass the digital deposit costs onto the client who contracted them. This way, at the end of a research project, digital information will be transferred to the e-depot and DANS can guarantee the long-term archiving.
http://www.edna.nl



7. TOWARDS AN ACTION PROGRAMME ON RESEARCH DATA FOR THE KNOWLEDGE EXCHANGE PARTNER COUNTRIES

7.1 Introduction

The starting point of this paper is the vision, articulated by *Riding the Wave*, of a collaborative data infrastructure that enables researchers and other stakeholders from education, society and business to use and re-use research data. The focus lies on the key drivers of this development: researchers and infrastructure. The current situation for those key drivers in Denmark, Germany, the Netherlands and the United Kingdom was surveyed and analysed in the international context regarding:

- incentives for researchers to publish datasets
- training in data management for researchers, data scientists and data librarians
- data infrastructure at the institutional level and the disciplinary and national and international levels
- data infrastructure funding

In previous chapters this report described and analysed the current situation in the four countries for each of the above areas. The recommendations of *Riding the Wave* with regard to those key drivers are the following 'first steps':

- develop an international framework for a collaborative data infrastructure
- earmark additional funds for scientific e-infrastructure
- develop and use new ways to measure data value, and reward those who contribute to it
- train a new generation of data scientists [and broaden public understanding].

Following the analysis of the situation in the KE countries and seeking to translate the recommendations of *Riding the Wave* into concrete options, this chapter recommends actions in each area. Together the recommendations outline an action programme at the level of the four KE countries that will facilitate the realisation of the envisaged collaborative data infrastructure. We want to 'Take stock and go ahead'.

7.2 Incentives for researchers

For researchers as data producers, there are four main areas of incentives to share and publish their datasets:

1. **Re-use and recognition**: currently, publishing datasets is of little account to the academic record of researchers. Several initiatives are trying to change this: to enable dataset citation and data publications (through peer-reviewed journals specialised in this type of article). Data centres in the four KE countries have implemented persistent identifiers such as the DataCite method to facilitate dataset citation. However, as yet there is no standard bibliographic format for citing datasets and there are only a handful of specialised peer-reviewed journals for data publications. It is generally assumed that if published datasets counted in the academic records of the dataset-producing researchers, this would provide a powerful incentive for researchers to make the effort to publish them.

2. **Rules and codes of conducts**: in the UK and Germany, several important scientific organisations have issued codes of conduct or similar statements emphasising data management and data sharing issues. Such statements have an impact on researchers and can be seen as paving the way for more data sharing.

3. **Requirements by funding organisations**: several research funding organisations in the UK, Germany and Netherlands have implemented requirements with regard to data management and data sharing for research grant applicants. This is seen as a powerful incentive for researchers.

4. **Journal data availability policies**: increasingly, editorial boards of scientific journals are pressing authors to offer access to the underlying datasets in combination with the journal article. Sometimes, these policies are mandatory.

Increasing incentives for data producing researchers will be a cornerstone in any action programme to make data sharing and data publishing an integrated part of the academic culture. The following table (page 32) presents the long-term strategic goal and primary stakeholders, with mid-term objectives and suggestions for possible actions.

Table 5. Possible actions to increase researchers' incentives for datasharing

Long-term strategic goal:

Data sharing will be part of the academic culture

Primary stakeholders to be involved in this part of the action programme*:

Data centres/data archives; academic institutions/professional bodies/learned societies; research funders; editorial boards/journal publishers

Mid-term objective	Possible actions
Standardise data set citation using persistent identifiers such as DataCite	Set up a committee to develop standards for dataset citations
Have considerably more journals for data publications	• Grant seed money for bottom-up initiatives to start peer-reviewed data publication journals • Set up a Community of Practice for national initiatives
Develop citation metrics for datasets	Conduct a feasibility analysis to develop citation metrics for published datasets
Make published data sets and citation metrics count in research assessment exercises	Advocate making published data sets and citation metrics count in research assessment exercises in the four countries including, as a preliminary step, the registration of datasets in the annual reports of research institutes and universities
Define and issue codes of conduct on data sharing on institutional or disciplinary/(inter)national levels	• Conduct awareness campaign among academic institutions, professional bodies/learned societies • Develop appropriate educational modules for early career researchers and ongoing professional development training courses for research staff
Set requirements for data sharing and data management in grant applications and show willingness to meet costs (This is relevant to Denmark, Netherlands and Germany. UK funding bodies have already developed these)	Develop requirements and policies on: • Data management plan • Guidance and support • Mandatory depositing in data archive/data centre • Monitoring compliance
Have considerably more journals with data availability policies	• Convince editorial boards of journals to have a data availability policy with workshops/seminars • Create a website with an overview of data availability policies of different journals (along the lines of the SHERPA/RoMEO website) • Data availability policy for funding OA journals

* KE and its partners are conscious that other parties are also active in this field. We would therefore like to take up these actions together with other stakeholders.

7.3 Training

Training in data management and data sharing can be distinguished in two categories: first, training researchers to improve data skills within research domains and second, training librarians so that they can function as data librarians. DCC in the UK and DANS in the Netherlands are carrying out initiatives to improve the data skills of researchers and research support staff. Similar initiatives could be taken in Germany and Denmark. As the role of data librarians is in development and only a few data librarians will be needed in each of the four countries, a supranational effort to define a curriculum for training data librarians could be part of the action programme.

Incorporating data management in the curricula of researchers and possibly setting up specialised curricula for data scientists should be encouraged at universities and in scientific fields. The next table outlines a possible action programme on the issue.

Table 6. Possible actions to facilitate data logistics (data sharing/management)

Long-term strategic goal:

Data logistics will be an integral component of academic professional life

Primary stakeholders to be involved in this part of the action programme*:

universities, learned societies, library schools

Mid-term objective	**Possible actions**
Develop data management training courses targeting data librarians	• Define a curriculum • Develop benchmarks for assessing course content • Provide infrastructure for international internships
Incorporate data management training in the curricula of researchers	Conduct an awareness campaign among academic institutions and learned societies with regard to training of data skills (and with regard to rules and codes of conduct on data sharing)
Develop curricula for data scientists	Develop means for assessing researchers' data management skills; seek informal and formal accreditation from professional bodies, learned societies and industry.

* KE and its partners are conscious that other parties are also active in this field. We would therefore like to take up these actions together with other stakeholders.

7.4 Data infrastructure and funding

The report described and analysed the data infrastructure of each KE country at the institutional level and at the disciplinary/(inter)national level. At the institutional level, all four countries have undertaken initiatives, however, the institutional infrastructures are still in development and have not yet crystallised. At the national level, the UK seems to have the most widespread data infrastructure with 17 data centres and archives. Germany identifies nine data centres but lacks a comprehensive overview. The Netherlands lists seven data centres, while Denmark lists three. The three main challenges in developing an ecosystem of data repositories are (1) gaps in the present data infrastructure and (2) connectivity issues (between the workflow of researchers and the institutional data infrastructure and between institutional and national data infrastructures) and (3) long-term financial basis. According to these results, the proposed action programme should focus on these three challenges. This could include several actions as presented in the table below.

Table 7. Possible actions for developing a sound data infrastructure

Long-term strategic goal:

Data infrastructure will be sound, both operationally and financially

Primary stakeholders to be involved in this part of the action programme*:

research funders, universities and research institutes, data centres/data archives

Mid-term objective	**Possible actions**
Improve institutional data infrastructure	• Initiate and support projects for the development of institutional data infrastructure • Periodical webinars where project participants can exchange practical experiences and knowledge
Improve coverage of disciplinary and (inter)national data	• Identify gaps (‘homeless’ data sets) with a KE survey infrastructure • Coordinate national data infrastructure elements and investigate whether mutual opening up of facilities could fill the gaps using cloud technology
Clarify the basics of data infrastructure funding (who pays for what and why?) by establishing relevant funding criteria for the various stakeholders	• Initiate a study to investigate the principles of funding data infrastructure elements
Understand costs and benefits of data sharing and its infrastructure with the aim of minimizing the financial burdens	• Develop a benchmarking model of costs for data infrastructure so that ensuing cost studies will have comparable results and make the exchange of ‘lessons learned’ possible • Initiate studies into the benefits and costs of re-use, publishing and archiving of datasets • Initiate/support projects developing automatic ingest tools for datasets (as the most important cost driver)

* KE and its partners are conscious that other parties are also active in this field. We would therefore like to take up these actions together with other stakeholders.

7.5 Direct role for Knowledge Exchange: Quick wins

The previous tables present an ambitious four country action programme. The feasibility of this action programme relies on the concerted effort of a number of key stakeholders. Although Knowledge Exchange is in an excellent position to oversee the status quo in the four countries and develop the programme, it will certainly take time to inform, convince and involve the potential partners.

Meanwhile, however Knowledge Exchange can take certain actions under its own steam. Unsurprisingly, these actions refer to its core competence: knowledge exchange. Here are four examples of such actions. Giving them a certain priority could bring Knowledge Exchange some quick wins, thus contributing not only to the status of KE itself but also to the authority of the action programme as a whole.

Table 8. Possible concrete steps for Knowledge Exchange partners

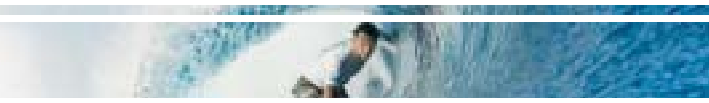
Possible actions	Possible concrete steps
Regular activities for the exchange of experiences and knowledge (i.e. working group meetings, roundtables, seminars and workshops)	Organised by Knowledge Exchange
Identify gaps (‘homeless’ datasets) by carrying out a four country survey	Wide survey of researchers asking: “If you were prepared to share your data, would you know where to deposit them safely?” Survey data centres asking, “Are you prepared to foster ‘homeless’ datasets from other KE countries (possibly on a swap basis)?”
Influence editorial boards of journals to have a data availability policy	List OA journals in KE countries, benchmark against a ‘standard’ data availability policy. Publish yearly e.g. in RoMEO
Awareness raising campaign among academic institutions and learned societies on rules and codes of conduct	Collect existing codes of conduct in KE countries, analyse them and compile a (discipline-specific) model. Make this the basis of an awareness raising campaign among academic institutions

7.6 Arriving at a cohesive and comprehensive action programme

As a collaboration of four partner organisations in Denmark, Germany, the UK and the Netherlands, Knowledge Exchange has already achieved several successes in the field of open access in its relatively short lifetime. Based on the status quo survey conducted by the KE working group on primary research data and the challenges related to the availability of research data, this report has outlined an action programme including a series of possible actions and associated concrete steps.

The aims of this fast developing field of research align completely with the common vision of the KE partners to make a layer of scholarly and scientific content openly available on the internet. Initiatives, coordination and exchange of knowledge would greatly contribute to the development of the data infrastructure for scholarly and scientific research. Therefore, the Knowledge Exchange partners should bring together the various stakeholders in the four partner countries to develop a cohesive and comprehensive action programme. The aim should be to initiate a concerted effort that will speed up the development of the desired data infrastructure and ensure that KE national infrastructures will be embedded in the future international research data infrastructure.

INFORMATION SOURCES



1. **High level expert group on scientific data**; *Riding the Wave: how Europe can gain from the rising tide of scientific data*; European Union, 2010; http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

2. **T. Hey, S. Tansley, K. Tolle (eds)**; *The Fourth Paradigm: Data-intensive Scientific Discovery*; Microsoft Research, 2009; ISBN 978-0-9825442-0-4; http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf

3. **N. Beagrie, R. Beagrie, L. Rowlands**; *Research Data Preservation and Access: the views of researchers*; Ariadne, 2009, nr. 60; http://www.ariadne.ac.uk/issue60/beagrie-et-al/

4. **PARSE.INSIGHT**; *Insight into digital preservation of research output in Europe*; 2009; http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

5. **L. Waaijers, M. van der Graaf**; *Quality of research data, an operational approach*; D-Lib magazine; Vol.17; nr.1/2; http://dx.doi.org/doi:10.1045/january2011-waaijers

6. **M. van der Graaf, L. Waaijers**; *Over kwaliteit van onderzoeksdata*; SURFshare report, 2010; http://www.surffoundation.nl/nl/publicaties/Documents/SURFshare_Organisatorische%20aspecten%20van%20duurzame%20opslag%20en%20beschikbaarstelling%20onderzoekdata.pdf

7. **C. Tenopir, S. Allard, K. Douglass, A.U.Aydinoglu, L. Wu, E. Read, M. Manoff, M. Frame**; *Data Sharing by Scientists: Practices and Perceptions*; PLoS ONE ; Vol. 2011; http://dx.doi.org/doi:10.1371/journal.pone.0021101

8. **OECD**, *OECD principles and guidelines for access to research data from public funding*; 2007; http://www.oecd.org/dataoecd/9/61/38500813.pdf

9. *European Research Area Vision 2020*; http://ec.europa.eu/research/era/pdf/era_vision_2020_en.pdf

10. **ESFRI**; *Strategy report on research infrastructures, roadmap 2010*; European Union, 2011; ISBN 978-92-79-16828-4; http://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf

11. **ESFRI**; *European roadmap for research infrastructures, roadmap 2008*; European communities, 2008; ISBN 978-92-79-10117-5; http://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf

12. **e-IRG Data Management Task Force**; *Report on data management*; 2009; http://www.e-irg.eu/images/stories/publ/task_force_reports/dmtfjointreport.pdf

13. **K. Koski, C. Gheller, S. Heinzel, A, Kennedy, A. Streit, P. Wittenburg**; *Strategy for a European data infrastructure*; PARADE: Partnership for Accessing Data in Europe, 2009; http://www.csc.fi/english/pages/parade

14. **EUROHORCS and ESF**; *The EUROHORCS and ESF vision on a globally competitive ERA and their roadmap for actions to help build it*. 2008; http://www.eurohorcs.org/SiteCollectionDocuments/EUROHORCs_ESF_ERA_RoadMap.pdf

15. **Blue Ribbon Task Force**; *Sustainable economics for the digital planet: ensuring long-term access to digital information*; 2010; http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

16. **STM**; *Brussels Declaration*; http://www.stm-assoc.org/brussels-declaration

17. **S. Reilly, W. Schaller, S. Schrimpf, E. Smit, M. Wilkinson**; *Integration of Data and Publication*; ODE report: Opportunities in Data Exchange, to be published autumn 2011

18. **H.A. Piwowar, R.S. Day, D.B. Fridsma**; *Sharing detailed research data is associated with increased citation rate*; PLoS ONE 2(3): e308; http://dx.doi.org/doi:10.1371/journal.pone.0000308

19. **RCUK**; *Common principles on data policy*; 2011; www.rcuk.ac.ul/research/Pages/DataPolicy.aspx

20. **Alliance of German Science Organisations**; *Principles for the Handling of Research Data*; 2010; http://www.allianzinitiative.de/en/core_activities/research_data/principles/

21. **Deutsche Forschungsgemeinschaft**; *Recommendations for secure storage and availability of digital primary research data*; 2009; http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen _200901_en.pdf

22. **D.L. Schriger, A.C. Chehrazi, M.M. Rashida, D.G. Altman**; *Use of the Internet by Print Medical Journals in 2003 to 2009: a longitudinal observational study*; Annals of Emergency Medicine; 2011; 57;2;153-160. http://dx.doi.org/doi:10.1016/j.annemergmed.2010.10.008

23. **M. van der Graaf**; *Organisatorische aspecten duurzame opslag en beschikbaarstellingen onderzoeksdata*; SURFshare report, 2010; http://www.surffoundation.nl/nl/publicaties/Documents/SURFshare_Organisatorische%20aspecten%20van%20duurzame%20opslag%20en%20beschikbaarstelling%20 onderzoekdata.pdf

24. **A. Swan, S. Brown**; *The skills, role and career structure of data scientist and curators: an assessment of current practice and future needs*; JISC, 2008; http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareers finalreport.pdf

25. **G. Pryor, M. Donnelly**; *Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?*; The International Journal of Digital Curation, 2009, 2, Vol. 4, 158-170; http://www.ijdc.net/index.php/ijdc/article/view/126http://www.ijdc.net/index.php/ijdc/article/viewFile/126/133

26. **GRDI2020**; *Towards a10-year vision for global research data infrastructures*; 2011; http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=acfd704c-3cfb-436c-ba82-3f5a22c7e152

27. **C. Rizzuto**; *Research Infrastructures and the Europe 2020 strategy*; ESFRI, 2010; http://ec.europa.eu/research/infrastructures/pdf/esfri/publications/esfri_inspiring_excellence.pdf

28. **C. Beagrie**; *Keeping Research Data Safe fact sheet*; 2010; www.beagrie.com/KRDS_Factsheet_0910.pdf

29. *The UK research data servers feasibility study, report and recommendations to HEFCE*; December 2008; www.ukrds.ac.uk/resources/download/id/16

30. **N. Beagrie**; *Benefits from the infrastructure projects in the JISC Managing Research Data Programme*; JISC, September 2011; http://www.jisc.ac.uk/media/documents/programmes/mrd/RDM_Benefits_FinalReport-Sept.pdf

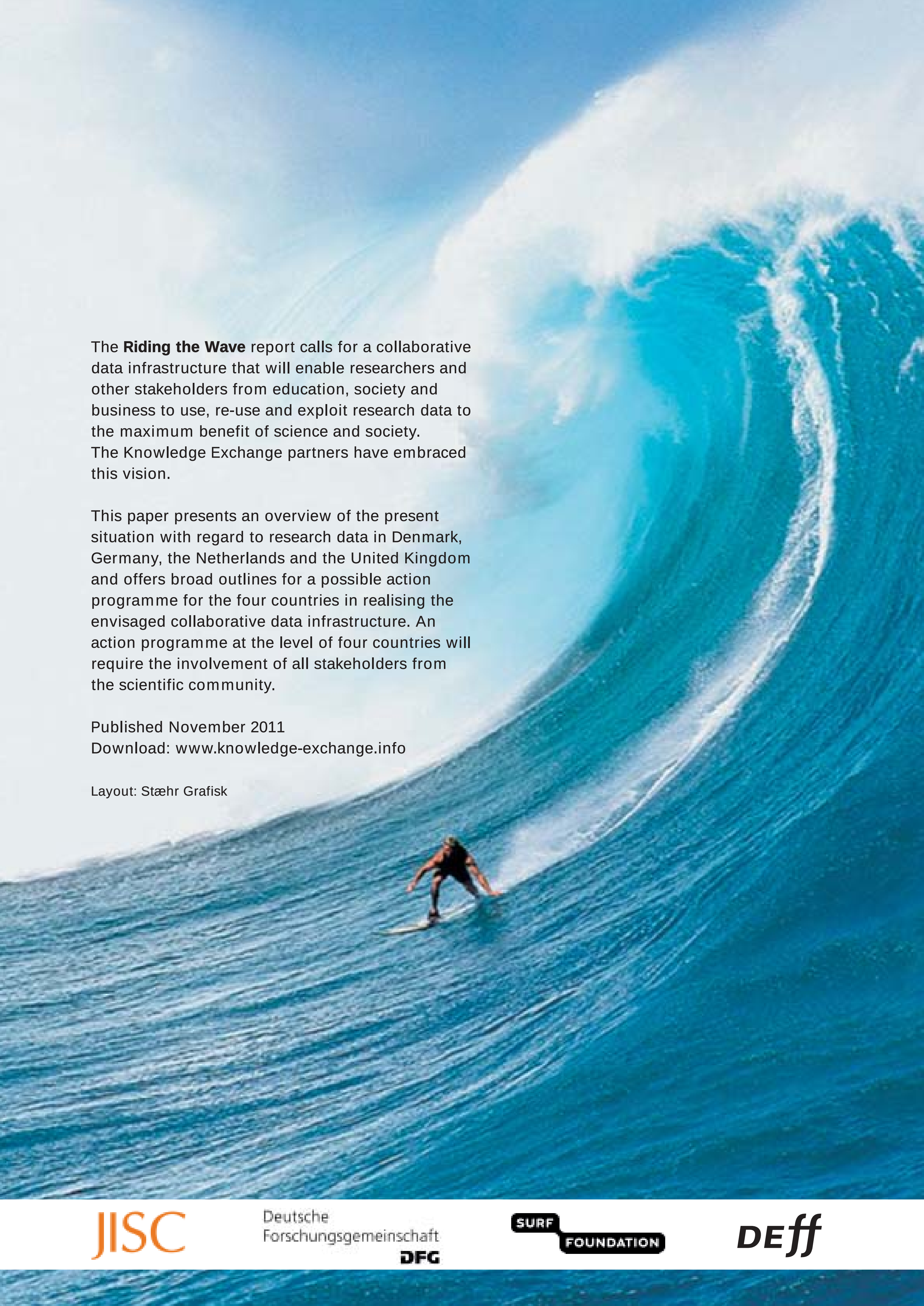
31. **H.A. Piwowar, T.J. Vision, M.C. Whitlock**; *Data archiving is a good investment*; Nature, Vol. 473, 285/285; http://dx.doi.org/doi:10.1038/473285a

32. **Research Information Network**; *Data centres: their use, value and impact*. 2011. www.rin.ac.uk/data-centres

33. **PARSE.Insight**; *Science Data Infrastructure roadmap*; 2010; http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

34. **VSNU**; *Nederlandse Gedragscode Wetenschapsbeoefening*; 2004; http://www.vsnu.nl/Media-item/Nederlandse-Gedragscode-Wetenschapsbeoefening.htm

35. **UK Research Integrity Office**; *Code of Good Practice for Research*; 2009; http://asset.ukrio.org/ukR10htre/UKRIO-Code-of-Practice-for-Research.pdf



The **Riding the Wave** report calls for 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.
The Knowledge Exchange partners have embraced this vision.

This paper presents an overview of the present situation with regard to research data in Denmark, Germany, the Netherlands and the United Kingdom and offers broad outlines for a possible action programme for the four countries in realising the envisaged collaborative data infrastructure. An action programme at the level of four countries will require the involvement of all stakeholders from the scientific community.

Published November 2011
Download: www.knowledge-exchange.info

Layout: Stæhr Grafisk

JISC
Deutsche
Forschungsgemeinschaft
DFG
SURF
FOUNDATION
DEff