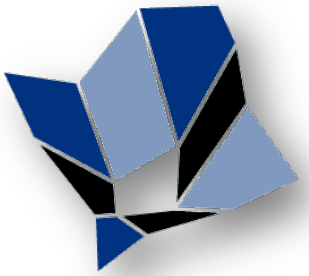

Project “RADIESCHEN” (Project RADISH)

Framework conditions for an inter-disciplinary research data infrastructure

“Synthesis” report

**“Final report of the project and
roadmap for the development of a research data infrastructure
in Germany”**

Content

1. Introduction

Around the world scientific disciplines are increasingly facing the challenge of a burgeoning volume of research data. This data avalanche consists of a stream of data generated from sensors and scientific instruments, digital recordings, social-science surveys or drawn from the World Wide Web.

All areas of the scientific economy are affected by this rapid growth in data, from the logging of digs in archaeology, telescope data with observations of distant galaxies in astrophysics or data from polls and surveys in the social sciences. The challenge for science is not only to process the data through analysis, reduction and visualisation, but also to set up infrastructures for provisioning and storing the data.

The following image shows how the use of the term "big data" has grown on the Internet. "Big data" became a buzz-word around 2012 and is currently a hot topic.

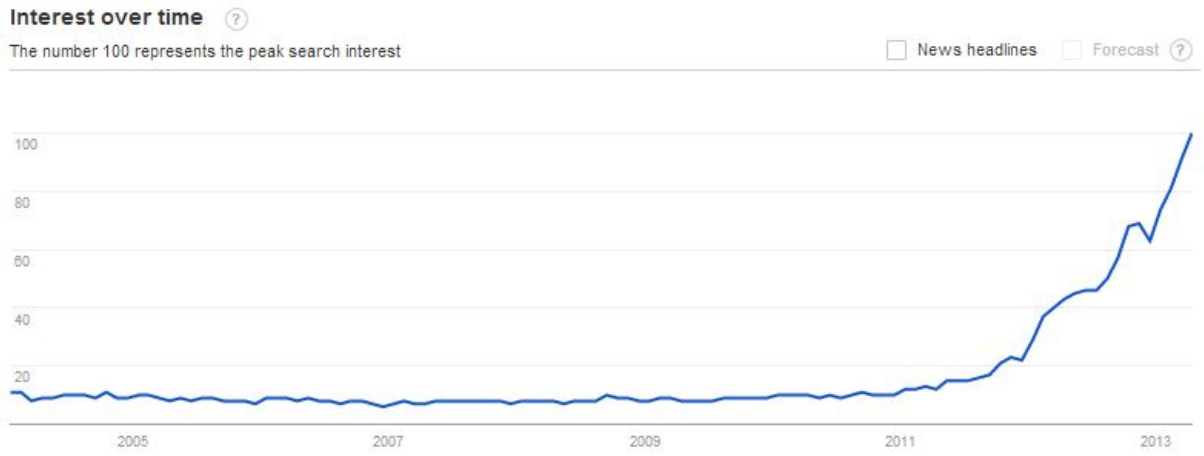

Fig. 1.: Google Trends – the use of the term "big data" in Internet searches from 2004 to today. The curve covers all searches around the world and exhibits a steep rise since the beginning of 2012[1].

However, research data and "big data" are not automatically the same thing. Smaller data sets such as the daily observations from a local weather station, are also research data that also deserve to be analysed, stored and annotated in case they are used further. In fact, these small data sets make up the lion's share of available research data. This means that research data infrastructures and their associated services and tools vary greatly and are designed to handle these discipline-specific data sets.

The project "Framework conditions for an inter-disciplinary research data infrastructure (Project "Radieschen", meaning "radish")" now tackles the issue of what demands are placed on generic components of an infrastructure and the intermeshing of that with discipline-specific components. This is based on a review of existing systems and infrastructures and an analysis of them in terms of organisational structure, technology used and costs incurred. Cross-disciplinary topics, such as the value system of scientific publications or the role of social media in science, play just as much a role as the trend towards the outsourcing of services to service facilities and computer centres.

[1] http://www.google.com/trends/

This report provides an overview of the results of the "Radieschen" project. This starts with a classification of the project and the project objectives within the German and European research landscape (chapter 2). The next chapter deals with future scenarios that describe how the scientific world in Germany could develop by 2020. Chapter 4 summarises the outcomes of the work packages of costs, organisation and technology, shows recommendations for action and provides an outlook on the topic in question.

An important component of the Radieschen project was the interaction with the research data community. Chapter 5 summarises the outcomes of the interviews performed during the review and provides an overview of the discussions that took place in the course of the Radieschen workshops and the research data symposium (FDI 2013). Chapter 6 looks at interdisciplinary topics that, although not a main component of the investigation, were repeatedly mentioned and were discussed during the course of the project. The report closes with an outlook and provides recommendations for a further development of research data infrastructures.

2. Classification

E-science, e-infrastructure, research data management, virtual research environments – these buzz words are often heard in the context of research data and its infrastructures. There are a corresponding range of projects and institutions around the world that deal with these topics.

Fig. 2 shows an overview of **large-scale research infrastructures**, funded by the EU ESFRI initiative.[2] This takes a very broad view of the term "research infrastructures" and includes in it institutions, equipment and associated services that are available to the scientific community from a wide range of disciplines. An example here is Géant,[3] a high-speed network that is intended to facilitate cooperation as well as the sharing of knowledge and resources between researchers. Géant is a project of the e-Infrastructures Initiative of the EU Commission. Project Radieschen looks at a section of this broad topic area and focuses specifically on research data and the associated infrastructures required in Germany. This enables a more precise view of the actual conditions and needs of researchers on the ground in Germany.

[2] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what
[3] http://www.geant.net/Pages/default.aspx

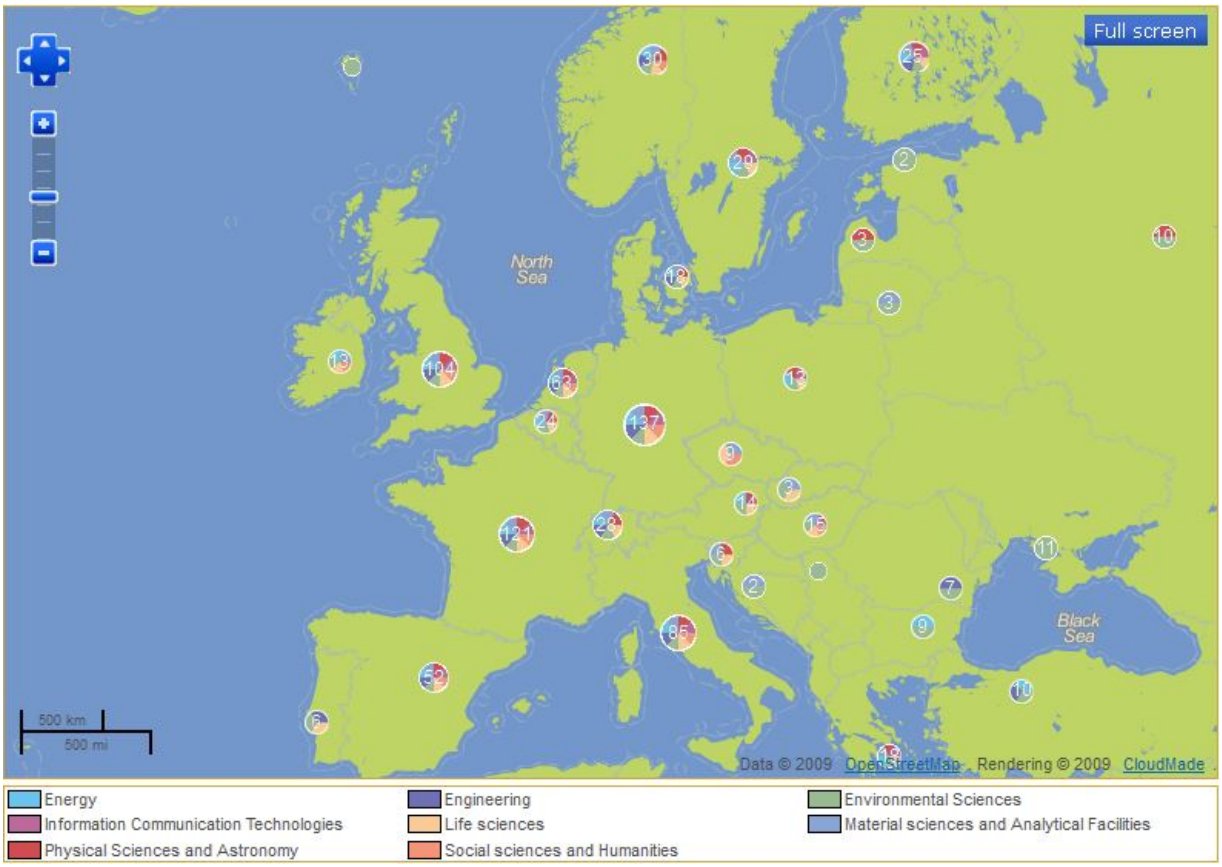

Fig. 2: Overview of all current large-scale research infrastructures in Europe[4]

The orientation of the **EU EUDAT**[5] **project** is similar to that of project Radieschen. It, too, focuses on the creation of a "cross-disciplinary data service". Key areas of research and development at EUDAT are the creation of a pan-European data service, which is aimed at supporting many different scientific communities.

EUDAT sees the heterogeneity of the data as the starting point, but at the same time also takes into account the integration of the data by means of solutions and services that can be used in common. The collaborative data infrastructure (CDI) to be created requires an abstract architecture that permits existing data solutions to be integrated with data centres that support common data services. Similar to Radieschen, EUDAT deals with the reprocessing of data, metadata solutions, persistent identifiers and special solutions for all types of data, including what is known as "small data".[6]

In contrast to EUDAT, Radieschen not only examines the technical opportunities, but also considers key aspects such as the organisational circumstances and the costs of setting-up and operating a research-data infrastructure. Further, Radieschen is specifically orientated towards the circumstances of the German research landscape.

[4] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=mapri
[5] http://www.eudat.eu/
[6] http://www.isgtw.org/feature/towards-collaborative-data-infrastructure-science

The aim of the **Research Data Alliance (RDA)**[7] is to accelerate innovations and discoveries related to research data at the international level, to increase their benefit through reprocessing, to harmonise standards and to increase the retrievability of research data. The initiative seeks to develop and promote the acceptance of infrastructures, associated policies, recommendations for action and the development of standards. The RDA is a newly-established organisation (August 2012). The alliance operates globally with partners in Europe, Australia and America. Radieschen focuses on developments in Germany, thus making a national contribution to the goals of the alliance's work.

The **DFG** is facing up to the challenges of ever-increasing volumes of data for scientists, universities and research funding bodies with its call for proposals "Information Infrastructures for Research Data". Within the total of 27 projects funded by these projects, referred to below as **"FD" projects**, scientists and information specialists are developing infrastructures for research data that are tailored to the respective subject area (as at May 2011). This not only looks at medium to long-term archiving, but also includes subject-appropriate metadata as well as questions of combining publications with research data or quality control.[8] The 27 projects of the call for proposals asked to give an interview in the scope of the Radieschen review.

Further, since 2007 it has been possible to apply for service projects from the DFG in the scope of the DFG-funded Collaborative Research Centres (CRCs) "that deal with the set-up of information infrastructure for the research project. These projects, too, bring scientists together with information specialists from libraries and computer centres. To date the DFG is funding 27 such INF projects within the 232 CRCs that are currently active".[9] The review also examined these projects using interviews. Feedback with the INF community took place through the INF project workshop held by Radieschen.

Both the FD projects and the INF projects have a practical focus and are aimed at the realisation or further expansion of research data infrastructures. In this context Radieschen acts as an observation and roadmap project and is not allocated to any of the specified funding lines.

The **Helmholtz initiative "Large Scale Data Management and Analysis" (LSDMA)** offers the research centres of the Helmholtz community in Germany a data service with community-specific data life cycle laboratories (DLCL).[10]

The DLCLs work in close cooperation with the scientists. Their aim is the processing, management and analysis of data during the entire data life cycle. The joint research activities in the DLCLs produce in community-specific tools and methods. The DLCLs are enhanced by a Data Services Integration Team (DSIT). This team offers generic technologies and infrastructures for use in the various research communities based on research and development in the areas of data management, data security, storage technologies and long-term archiving of data.

The Helmholtz LSDMAs specialise in the handling of large to very large data volumes. However, this is not applicable to every research discipline. For example, the research data archive of EarthChem[11] is

[7] http://rd-alliance.org/

[8] Making Scientific Research Data Accessible: Current Trends and Perspectives in Germany, information workshop in Washington DC, 21 June 2011, http://www.dfg.de/dfg_profil/geschaeftsstelle/dfg_praesenz_ausland/nordamerika/berichte/2011/110621_informationsworkshop_washington/index.jsp

[9] Cf. Effertz and Schoch (2013)

[10] http://www.helmholtz-lsdma.de/

only 8 GB in size and yet still contains the research results of past decades in more than 300,000 data sets. The aim of project Radieschen is also to take these relatively small data sets, known as “small data”, into account in the development of research data infrastructures and their tools.

Research data and the set-up of research data infrastructures are not local challenges, but instead affect all scientific disciplines and research institutions. Whether a global, European or local solution is needed depends on the object of the research. Not every discipline works with petabytes of data and not every discipline requires constant access to the data across national boundaries. However, because researchers and scientists predominantly operate internationally, international or at least European solutions are to be favoured. This also applies in respect of knowledge transfer and increasing researcher mobility.

Common to all projects considered here, however, is a trend that is moving away from the “silo” solution of individual data collections and towards a solution in which the services and structures of special service providers, such as computer centres or special repositories, are used. The reports from the Radieschen project provide the requisite background information in this respect.

[11] http://www.earthchem.org/

3. Future scenarios

The rise of new technologies and developments also poses new challenges for the actors in the area of research data infrastructures. Libraries, as one of the actors, enable access to digital media and support the publication of research data and its long-term archiving. Digital media and research data, however, introduce new aspects into the libraries' range of activities. How are we to imagine the library of the future? The library as an interface to the computer centres? Will library and computer centre fuse into a new service unit? What role will scientific publishers play in future? Currently the traditional form of publications still carry greater weight – articles for conferences and journals. But will this still be the case in future? New forms of publication are already making their presence felt. The tasks of the computer centres may also change. Yesterday their remit was provisioning of rapid hardware, whereas now everything revolves around the topic of data.

This gives rise to the question of what tools are required to locate and pursue the correct course in a networked world. One tool from the area of innovation management is the **scenario technique.**[12] Following Kurt Sontheimer,[13] the scenario technique is less about predicting the future and more about thinking ahead of the future. Scenarios describe possible future situations such as the future development of Germany as a business location in which the project would be positioned.

Future scenarios are based on a networked system of influencing factors in which several conceivable future development opportunities are computed for each influencing factor. A key objective of the scenario technique is to identify future opportunities and threats in order to take strategic decisions.

Fig. 3 shows a projection of various potential future scenarios. The current starting position is marked by the blue circle on the left. The X axis shows changes over time. The Y axis shows the spectrum of possible scenarios in various large circles. The circle around

- "Possible" describes developments that may occur. The prediction is based on extrapolated knowledge, for example projections.
- "Plausible" describes scenarios that may occur. The prediction is based on current knowledge.
- "Probable" represents a situation that will probably occur. The prediction is based on current trends.
- "Preferable" describes a situation that is hoped for based on the reasoned evaluations of the current situation.

The method enables factors to be included that are otherwise difficult to record, such as new insights about the future, a deep-seated value shift or new rules and innovations.

[12] Gausemeier, J., Stoll, K., Wenzelmann, C. (2007)
[13] Sontheimer, K. (1970)

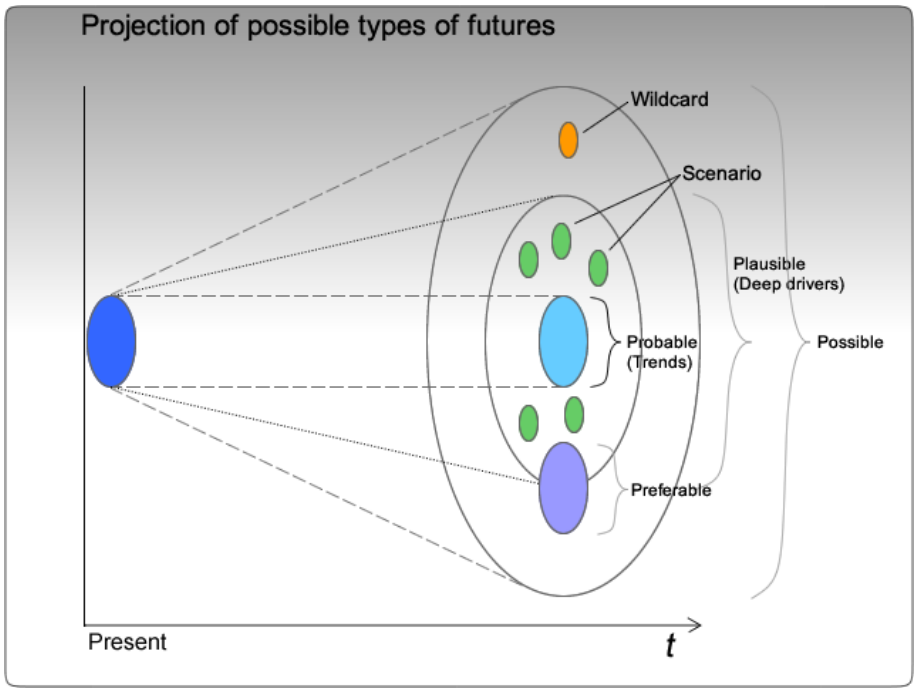

Fig. 3: The diagram[14] shows a projection of possible future scenarios. The blue dot on the left represents the starting position. The green dots on the right show possible scenarios. The circles of the funnel specify whether the respective scenario is in the preferable, probable, plausible or plausible range.

The following visions of the future describe possible developments of the scientific world in Germany in 2020 (or later). The situations are presented in an exaggerated manner to illustrate trends and determine potential development steps. The scenarios describe extreme situations. It is not expected that the situations described will occur precisely as presented.

The diagrams related to the scenarios illustrate the positions of the actors as compared to the situation today. The starting position, the situation today, is located exactly on the cross in the middle. The actors presented are the scientists (S), the scientific libraries (L), the scientific computer centres (C) and the data scientists (DS) as the incarnation of a new professional profile amongst knowledge workers.

[14] Image source: http://www.quesucede.com/page/show/id/scenario_planning

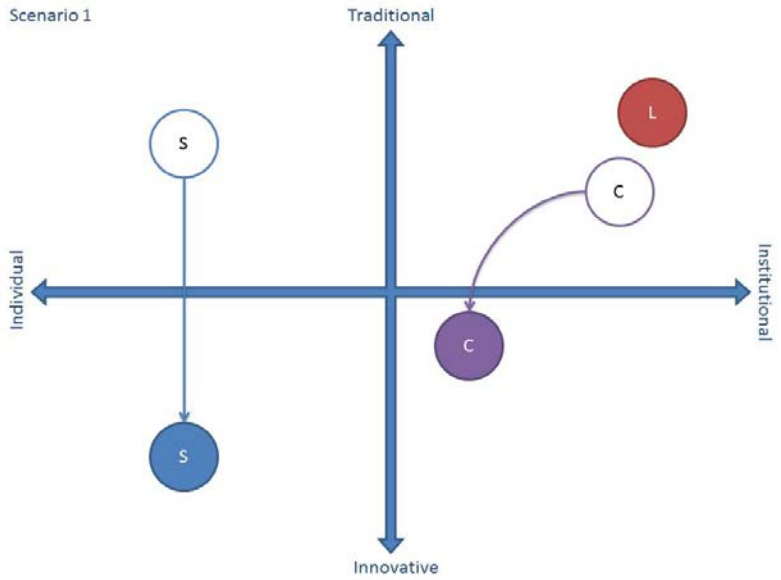

Fig. 4: Scenario 1 – New performance identifiers in science

Scenario 1 – New performance identifiers in science

Lori is a successful scientist in the earth sciences. She has just returned from a lecture tour in South Africa and is told that her software publication in the open access journal "Earth Science & Computing" has been accepted. She has already published substantial data in data journals but this is her first software publication. Lori is particularly excited because this publication finally allows her to apply for the highly sought-after position of lead researcher in Australia. Prerequisites for applicants for these hotly-contested positions in data-intensive disciplines like the earth sciences are no longer just the citation index, but now the triad of peer-reviewed publication, data publication and software publication, because only this combination of publications is considered to be complete and a substantial contribution to science.

Her colleague Matthis enters the room. He is also happy because, as the co-author of the software publication, he is credited with valuable European research credit points (ERC points) that he can now use to book coveted measurement time on a mass spectrometer. Although there is a mass spectrometer close to the GFZ labs in Potsdam, you are only granted measuring time after you have amassed a minimum number of ERC points. The ERC points used quickly pay dividends because the planned measurements are sure to generate new insights that Matthis can use for his next journal and data publication. This means he can continue his research with Lori and her team and hopefully complete his PhD soon.

Main aspects of the scenario:

- Mere counting of publications and citations as an evaluation of academic performance is replaced by a combination of peer-reviewed publications, data publications and software publications.
- A scoring system is established and governs access to resources.

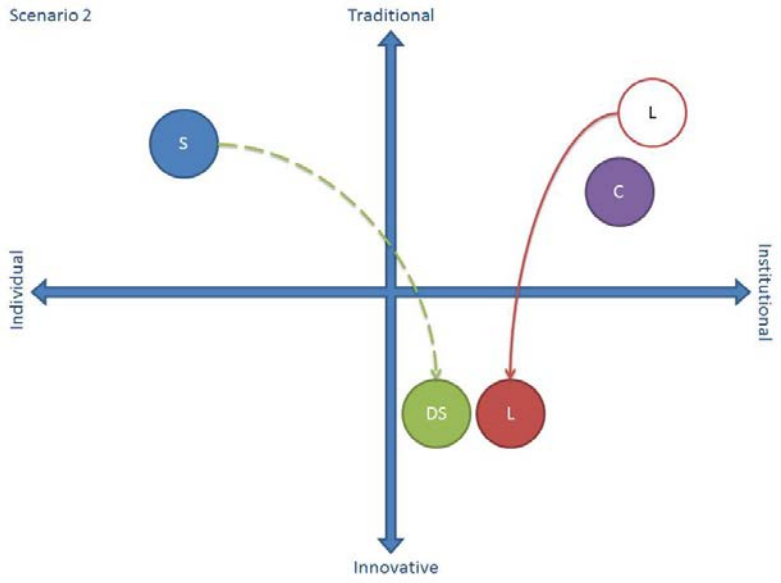

Fig. 5: Scenario 2 – Libraries are the future

Scenario 2 – Libraries are the future

Robert is looking around a redesigned library, a library affiliated with the Union of German Libraries for Science and Technology (UGL-ST). The image is characterised by a pleasant ambience, meeting rooms, discussion areas, places for training and discussion, and is adorned with displays showing current data streams. Wireless gigabit network access goes without saying. The library's databases are linked nationally with the databases of other university and research libraries in the UGL-ST. After the dissolution of the traditional library associations, the foundation of the UGL-ST enabled Germany to keep pace with the rapid development of research libraries around the world. Long gone are the times when solitary institutional libraries held catalogues, books and journals, and data lay hidden deep in the computer centres and workstation computers. Books and journals are still around, albeit only a few in printed form. The library is no longer a "paper museum", but rather it has developed into an information service provider that provides researchers with data and information – and not only text-based media.

Robert's colleagues include some highly educated data scientists. These sift the incoming data streams, perform quality checks and initial assessments of a potential follow-up use of the data. The remit of the UGL-ST libraries, however, goes beyond the archiving and cataloguing of data. As a consequence of the crisis in the journal sector at the beginning of the century, the libraries sprang into action and forced the scientific publishers, and with them the traditional subscription model for journals, out of the market with their own online publications. A globally-operating open access publishing house, the German Science Press under the umbrella of the UGL-ST, set the ball rolling. Although the scientific publishers had long had a leading position in the field of the traditional publication of journals and books, they did not stand a chance when it came to software and data publication due to their lack of capacity and willingness to reform. They lost the battle for position and were confined to niches or were taken over by the large academic publishers.

The interior of the new UGL-ST-libraries therefore only vaguely resembles the old university libraries. Instead information and competence centres, with their core units of library and computer centre, are an indispensable part of today's research infrastructure. In order to keep abreast of the times

and to hold one's own in the fierce international competition, it was necessary to develop new offerings independently. The UGL-ST had set up a dedicated research department in which highly-qualified data scientists developed software and researched new techniques for data analysis and visualisation in order to keep pace with the rapid changes arising from new forms of scientific communication and data-driven research. The strength of the libraries in the Union of German Libraries for Science and Technology lies in developing services for science jointly and centrally, providing them with a high level of reliability and at the same time providing competent advice and services on the ground, close to where the scientists need them.

Main aspects of the scenario:

- Libraries continue to develop into innovative, networked information and competence centres.
- Data scientists, highly-qualified experts in data, work in libraries in areas such as curation, quality assurance or archiving.
- Libraries take on the role of today's academic publishers.

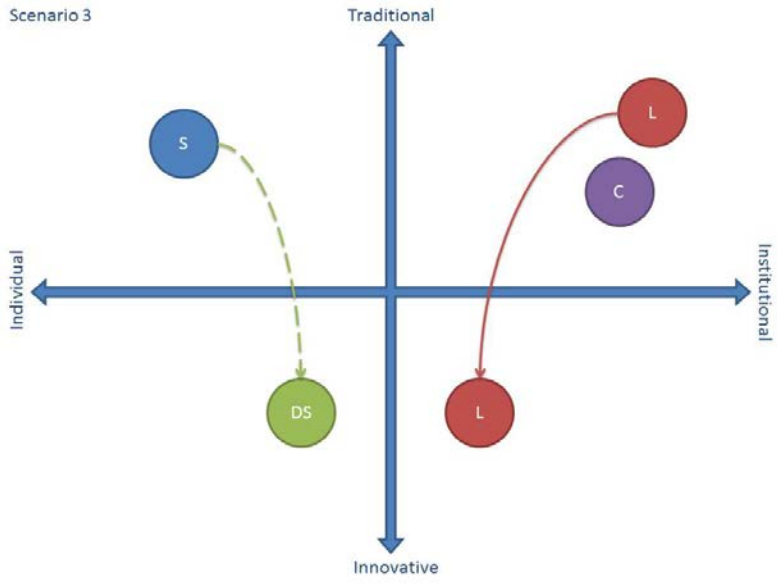

Fig. 6: Scenario 3 – Data scientists, the stars of a new generation

Scenario 3 – Data scientists, the stars of a new generation

Tom is a data scientist specialising in research data. After graduating in science and completing an additional qualification in research data, he was able to choose the most interesting position for him from a large number of offers. Despite lucrative offers from Great Britain and the USA, he opted for a position in Germany at the German National Library for Science and Technology, GNL-ST for short. This newly-established library was equipped with the latest technology and therefore offered the best conditions for the start of his career. The tasks for data scientists at GNL-ST also include the development of algorithms for classifying and annotating data, because they realised that an unstructured deluge of data cannot be meaningfully analysed. Their tasks also include quality assurance, review of the data and support with any follow-on use of the data.

For Tom the most exciting part is viewing the content of the data and analysing potential cross-references. Human-computer interfaces enable the GNL-ST data scientists to interact with the data in a 3D space via non-text-based input devices (e.g. data gloves, gesture control) and spatial displays. Immersive data analysis is a particularly well-liked process for data-driven research in high-dimensional data rooms. For example, by simply combining an analysis of satellite data on the state of the ionosphere with an analysis of oceanographic data, he could determine the risk of a possible tsunami in the Indian Ocean. The scientists on the ground gratefully received his tips and immediately incorporated them into their early-warning system. This enables people to react quickly and more precisely to a warning signal and the population to be evacuated if required.

For the future Tom is hoping for the development of further innovative interaction technologies that enable him to represent even very large data volumes abstractly and to sift for the desired piece of information in no time at all.

Main aspects of the scenario:

- The professional profile of the data scientist is developing and is also establishing itself in the academic world.
- Data scientists work with state-of-the-art academic information service providers that have evolved from the traditional science libraries.
- Their tasks include services such as ingest and archiving for scientists as well as research in the area of data analysis.

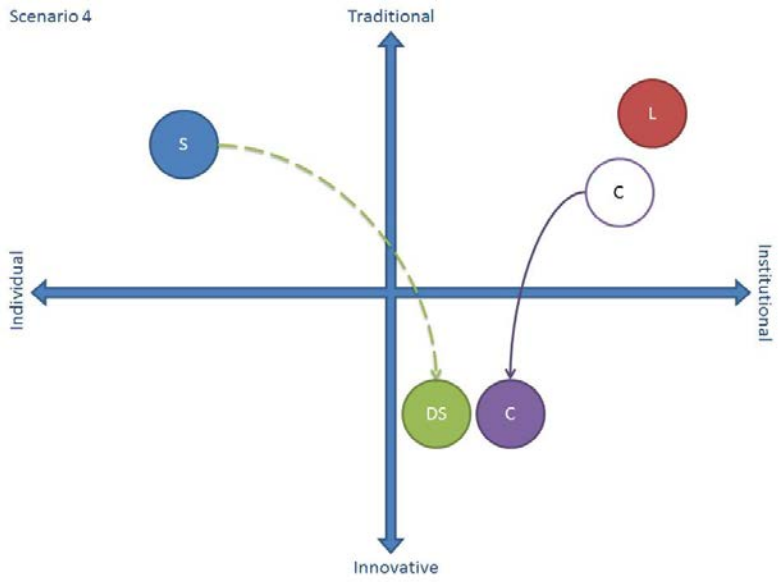

Fig. 7: Scenario 4 – Data centres take on a new role

Scenario 4 – Data centres take on a new role

Gone are the times when computer centres in science were merely conservative service centres that provided storage space and servers on demand. The cliché of the computer centre in the mind of most researchers were places where peculiarly-dressed figures ran between computers in large halls with roaring ventilation systems. Storage was the order of the day and computer centre staff tended to be sceptical about new developments because somewhere yet another server had given up the ghost. The ambitious computer centres were involved in high-performance computing. However, at

some point there was a quantum leap and some of the academic computer centres developed into data centres.

Today data centres are the natural point of contact for data management, software services as well as traditional publications. The data centres took over the latter task from the libraries and publishers that could no longer deal with the growing data deluge as the call for combined software and data publications became ever stronger. And the next logical step was to offer their own online journals for data and software publications. They had the capacity after all. Now the former library was affiliated as a department of the data centre and the academic publishers with their paper and subscription-based offering all but died out.

Data scientists now populated rooms that were equipped with the latest interactive technology and computers with the latest analytical software. Of course, the servers and high-performance computers still existed – at a central location and well-secured against unauthorised access, power outages and other disasters. The data centres remained service institutions; however, their tasks extended well beyond the level of Internet hosting and cloud offering. The focus was on offerings for virtual research environments (VREs) and research data engines (RDEs). For the scientists the details of the administration and the software installation remained hidden. They could use a toolbox to assemble their own working environment for all of their projects in which they could easily collaborate with other researchers from their teams or community. It was now also possible to publish research results, data and software directly from this environment.

Main aspects of the scenario:

- Computer centres develop further into data centres that serve researchers as the primary point of contact both for data management and software services and also publications of all kinds.
- In the data centres data scientists work on the provision of the various services (VREs, RDEs) for the communities.

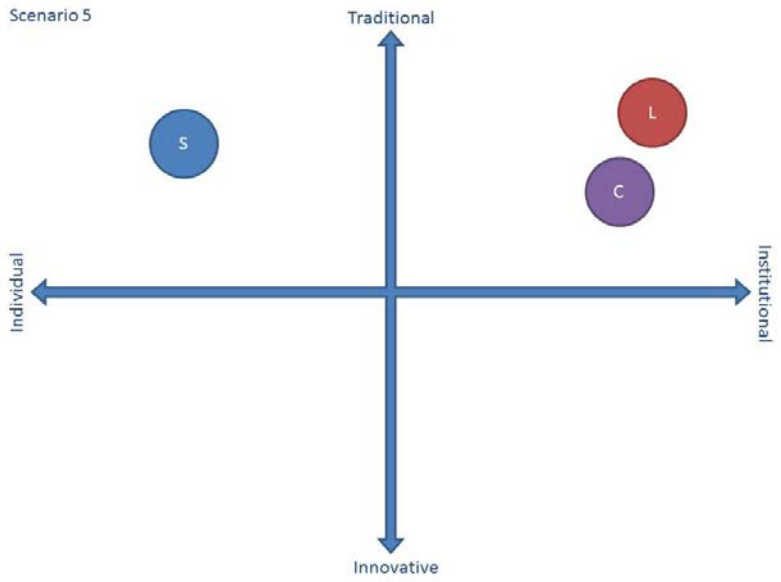

Fig. 8: Scenario 5 - Sticking with tried-and-tested methods

Scenario 5 - Sticking with tried-and-tested methods

Peter sits in front of his screen sorting the data from his project "Benedikt" into his database. These are unique data sets that enable an almost complete analysis of icon painting of the 14th century. The data is in part non-reproducible and thus particularly valuable for him and his colleagues. Peter initially stores his data on his external hard drive. Later he intends to transfer it to a repository for art-historical data, but first of all he plans to finish analysing his data and publish his results – in one of the remaining academic publishers. He is wary of the new offerings from the online publishers after one of his colleagues was accused of plagiarism. Plagiarism researchers claimed that his colleague had used third-party data from one of the online databases without acknowledging its source. To prevent that happening to him, he only uses his own data sets and stores them securely on his hard drive. What a sorry state of affairs it would be if other scientists reaped fame and glory on the basis of his work?

He is not the only one in Germany to adopt this position. In recent years Germany has increasingly developed into an enclave of tried-and-tested methods in a changing world. In Germany you can still find the traditional academic publishers with their paper publications as well as computer centres that cater directly for the needs of researchers on the ground and provide them with tailor-made working environments. In other European countries everything is already virtual – virtual research communities (VRCs), research data engines, data management plans. A whole host of things he considers to be no more than a waste of time for researchers.

The downside, however, was that he and his team were cut off from the global research activities in art history. These days collaboration and the exchange of information is done via VRCs in many cases. Data exchange was possible via the large online databases, but relied on mutuality. Further, a carefully-cultivated publication list appeared to be increasingly worthless. The interesting job offers from overseas stopped a long time ago. There increasing value was placed on the triad of publications of data, software and methodology. Maybe he should take a closer look at these opportunities after all? The prospects for advancement as a scientist in Germany alone seemed very limited to him. He will have a long, hard think about it later. Before that, analysing his data was the priority. The funding body wanted a report and the deadline was approaching rapidly.

Main aspects of the scenario:

- Attempts at innovations are refused for a wide range of reasons.
- Germany falls ever further behind in the international comparison. The scientists are increasingly isolated.

Conclusion

An optimum outcome can only be achieved if the different actors interact with one another and are prepared to rethink and alter their current position.

The scientific world is dynamic and is always changing. It is impossible to say what direction this development will take. The scenarios presented show possible developments – both positive and negative. It is now up to the actors themselves to define their own position in this context, to rethink it and consider steps that can achieve a positive development for the future.

4. Synthesis of the results from the work packages of costs, organisation and technology

The core aim of this project Radieschen is to develop three reports on the topics of technology, organisation and costs. In terms of content, these reports focus on the analysis of the technical components of the infrastructure (technology report), the analysis of processes and workflows in the lifecycle of research data and the observation of organisational aspects (organisation report) and the examination of cost structures for operating research data infrastructures (costs report). This chapter draws a conclusion for the three reports, shows recommended action for the respective topics and provides an overview of potential further development in the near future.

Technology

The evaluation of the interviews and the analysis of the materials examined produced the following striking results:

- The projects and specialist disciplines prefer bespoke developments to be created and used. However, generic tools at the lower technical level (e.g. file systems, databases) are used as the basis for the bespoke developments. The use of these basic tools is widespread across disciplines.
- In terms of hardware, hard drive and tape systems are primarily used as storage media.
- If existing software solutions are used, sustainability plays a pivotal role in the selection process, i.e. the software solutions must offer ongoing prospects in terms of support and maintenance. Arguments against using generic components frequently cite their lack of customisability to today's working environments and requirements. Commercial services, in contrast, are frequently shining examples of user-friendliness and the potential for integration into the private working environments.
- The analysis of the workflows of the various disciplines and institutions surveyed indicated a range of data-related working steps that occur almost everywhere. These are both discipline-specific and discipline-independent processes. Examples of discipline-specific working steps are data collation, quality control of the data and discipline-specific metadata. Examples of discipline-independent processes are the general generation/enhancement of metadata/data and metadata storage, data transfer, data replication, (long-term) archiving and web-based access to the data.

The analysis of the technical systems enabled the following recommended action to be determined:

- Interdisciplinary methods and tools can be used for a range of discipline-independent working steps. These include: data formats, metadata generation/enhancement for technical and contextual metadata, data and metadata storage, data transfer, data replication, backup, (long-term) archiving, web-based access to the data.

- Research data management tools should be user-friendly and enable integration into the scientific working environment.

- The development of standards and interfaces is important so that individual obsolete components can be replaced relatively simply in the face of ongoing technological change.

- Persistent identifiers should be increasingly used. Easy-to-use systems with the requisite technical and organisational background already exist.

- Intersubject exchange on the topic of data management should be intensified as existing and potentially reusable solutions are not known in many cases. Knowledge of technical solutions and organisational aspects could be disseminated for instance through topic-based competence centres that would be established.

In general the trend indicates a move towards outsourcing technical services to service institutions and computer centres. As a result of the deluge of data and the short shelf-life of information stored electronically, such outsourcing appears to be essential, especially for professional curating. In future it will be increasingly important to have available and to use professional services for managing research data. Naturally, certain data-management and data-processing steps are specific to a specialist discipline. However, generic solutions can be developed for a range of services, which, however, need to demonstrate long-term prospects and sustainability.

Organisation

The focus of the third work package is on the examination of the structures in which research data management takes place in Germany and the organisations that are involved in them. The focus is on the DFG-funded projects related to this topic. The focus on the DFG is not least due to the fact that – until just a few weeks ago – the topic of digital research data management was barely visible as a prominent funding issue for the other funding bodies, despite the GWK Committee for the Future of Information Infrastructure (*Kommission Zukunft der Informationsinfrastruktur*, KII),[15] the alliance for research data and other initiatives. Further, the scientific value of the data itself was underestimated for far too long and only established in the area of the classical "memory" institutions due to the traditional orientation on scientific publications.

It is clear that the acceptance of the results of previous projects on data management in the individual communities (formats, standards, workflows, infrastructure, metadata etc.) is in need of improvement. The use of this preliminary work by new projects should therefore be given a high priority in the funding process:

- The consideration of existing standards should be a requirement in the approval of new applications.
- In the infrastructure field greater use should be made of the competence of the different committees in the specialist disciplines.

Taking the domain model into consideration, it is clear that the progress regarding the management of research data primarily needs to be made in the area of the transitions between private, group and permanent domains. This is the precise starting point for the INF instrument of the INF projects in the CRCs of the DFG. For that reason this funding instrument should be further developed in a targeted manner:

[15] http://www.gwk-bonn.de/index.php?id=205

- Each CRC should have an INF project.[16] The scientists should be involved at all stages of INF conception and execution.
- Each INF project should be able, through targeted research, to apply the standards and tools that have been established in the specialist disciplines. If this consensus is not given within the discipline, targeted steps (not from INF itself) should be taken to create such consensus.
- Criteria such as novelty or uniqueness are not decisive for INF projects, but rather efficiency and integration into an infrastructure context.
- The INF projects should not include the attempt to create the structures of the entire discipline to start with. Further, the use of generic components should be favoured over the development of special solutions.
- In order to enable a long-term perspective, the INF subprojects require corresponding infrastructure measures from the IT infrastructure providers in the sciences field into which their work can be integrated.

The range of approaches to research data management in the individual disciplines shows that the development of tools and standards for data management remains dynamic:

- There are no clear analyses of the basic workflows in most disciplines. Only by concentrating on these existing workflows can efficient tools be selected or developed that are then actually accepted by the respective community.
- The (further) development of subject-specific metadata is intended to provide a clear focus on retrievability and traceability (provenance). More sophisticated systems than this with their excessively specialised vocabulary lose their reusability and relevance too quickly. Metadata should also be provided to a large extent by the tools themselves.

The field of IT infrastructure also exhibits a further need for development:

- IT infrastructure and its providers concentrate too strongly on scientific computing and to date have not adequately tackled the problem of data management. In terms of better support of research data management, however, a flexibilisation of current funding structures is necessary. The D-Grid[17] was such an initiative. However, it has been shown that there are still substantial problems in flexibilising the provision of IT resources. In this context it is not technical, but rather legal, funding-political and organisational questions that are key.

Although the most urgent issues in the area of research data management are still to be found within the respective discipline, interdisciplinary approaches to solutions are desirable in the organisational area:

- in order to use the results of the many completed projects, topic-related competence networks should be set up that provide support and consultation to new projects. It was clear that there is a strong demand for information about existing tools and standards from researchers and projects that cannot, however, currently be adequately met.

[16] Information management and information infrastructure programme element in the DFG Collaborative Research Centres, also known as CRCs (see also http://www.dfg.de/foerderung/programme/koordinierte_programme/sfb/programmelemente/programmelement_inf/index.html)
[17] http://www.d-grid.de/

Over the coming years progress towards sustainable research data management will still be made primarily within the specialist disciplines. This will be achieved in particular by means of subject-specific standardisation processes and implementation and use of subject-specific workflows. These developments need to be supported by a flexibilisation of the IT infrastructures. It is only on this basis that IT tools developed across disciplines, for instance data-mining tools, can be successfully deployed. That notwithstanding, collaboration, and in particular more intensive communication, is desirable and worth aiming at. Further, common organisational structures are expedient, provided corresponding structures within the disciplines are not neglected.

Costs

The risk of data loss is very high if it is not archived. Even if the data is still around somewhere, locating and recompiling it entails a lot of effort and expense. For that reason at least all important and non-restorable data should be archived.

Primary archives, i.e. those research data archives that take the data directly from the researcher, are the most labour-intensive at ingest. Things are different in secondary archives, i.e. those that do not take the data from the researcher, but exclusively from other archives. For both secondary archives surveyed in the course of Radieschen, selection and ingest taken together require the fewest working steps as compared with storage plus curation and provisioning.

In terms of costs, the following recommendations for action may be made:

- More extensive automation helps to reduce costs. That applies above all to ingest as the most labour-intensive step. An interdisciplinary ingest software must be flexible enough to meet the requirements of a range of subjects but still be easy to handle and be maintained in the long run. Only then will the required effect of an interdisciplinary and sustainable cost reduction be felt. Software projects such as PubFlow[18] need to enable further development and maintenance through follow-up projects.
- Uniform data structures, vocabularies and metadata standards simplify software development, quality control, data maintenance and the search in metadata and data. That also helps to lower costs. Despite that, the attempts to agree standards in many disciplines are still underdeveloped.
- In any event, putting a price on research data services is problematic. Charging for access can make it difficult to perform preliminary studies, for instance.
- If a price is put on research data services, such prices need to be known at a sufficiently early stage to enable researchers to apply for the adequate level of funding.
- The costs incurred should be openly disclosed and transparent. An affine price function is recommended to start with. This consists of a constant base sum that covers costs that are not dependent on volumes and a sum of linear subfunctions. The price should increase in a linear fashion with the number of data sets. The number of data sets may be replaced by a different factor that reflects the number of the logical data units. If data volumes are above average, the data volume should be included in the price function as the second variable, also in the form of a linear subfunction. An affine price function can be easily adapted to recorded or estimated costs and has the further advantage of being transparent and traceable for the customers.

[18] http://www.pubflow.uni-kiel.de/

- It will not work without any improvement in funding for the research data infrastructure. Current data volumes cannot be managed by means of efficiency increases in terms of the archives alone.
- Accompanying business management projects would be expedient to look into the costs in detail. Costs can be recorded more precisely if they are recorded in parallel to the project and not retrospectively.
- Little is known about the costs of the pre-ingest phase and the private domain. Further research is needed in this respect.

According to data from the German Federal Statistical Office (*Statistisches Bundesamt*) in 2011 the state and private not-for-profit institutions spent €11 billion and the tertiary education sector spent a further €13 billion on research and development. If just 1% of this €11 billion could be reallocated to an improved research data infrastructure, approximately 25 further research data centres with the capacity of the DAS (Data Archive for the Social Sciences at GESIS) and approximately ten further virtual research environments with data access such as TextGrid or C3Grid could be operated.

5. Analysis of the discussion with the community

Insights from the interviews

The results of project Radieschen are largely based on the interviews conducted in the course of the project. The interviews were performed with representatives from a variety of scientific communities. The Radieschen project team performed a total of more than 28 interviews. The majority of the interviews were performed in person. Some interviews were performed by telephone for scheduling reasons. The comprehensive catalogue of questions included both a general part as well as questions from the areas of technology, organisation and costs.

It is no easy task to capture the essence of the interviews in words. However, some clear trends can be identified. The personalised interview format made a substantial contribution to the understanding of the data, workflows and general particularities of the individual projects.

The requirements regarding hardware are probably the most common factor between the different projects. All projects named a relatively **standardised hardware set-up**: blade servers for web and databases, high-end hard disk, but otherwise no particularly exotic configurations. Several participants run their own systems, whilst others relied on the hosting facilities of their own data centres. Virtualisation is often used and generally seen as a positive development. Some of the respondents reported that a low-end cluster was used (Hadoop[19]).

Many projects extensively use **free and open source software**, in particular on the server side. If bespoke software was developed, this is often published under GPL or a similar free licence.

Resource repositories are a popular choice for storing data and literature. However, these are often not adequate or specific enough to store entire datasets. Alternatively, **peripheral data storage systems** are used.

[19] http://en.wikipedia.org/wiki/Apache_Hadoop

In terms of **metadata**, on the other hand, many variations are apparent: these range from abstract text fragments in the life sciences through to more detailed schemata (DDI in the social sciences, ABC in biology). Most interviewees specified that compiling a high quality of metadata took serious effort.

Most respondents sympathise with **open access policies** and attempt to treat their data as openly as possible. Exceptions exist in the context of privacy (e.g. data from individual households), legal restrictions on the part of the publishers, high-risk material (e.g. the locations of rare species). In some cases, however, publication may be possible through the anonymisation or the addition of what is termed "noise data".

A frequently-cited concept, especially in the socio-economic context, is to **visit the data-storing facility** itself to work with sensitive data there. However, a number of attempts have already been made to develop a more location-independent solution.

The **workflows** for data archiving are specific to each scientific community. Apart from very abstract steps, commonalities are therefore very difficult to define.

All respondents were aware of the relevance of **long-term archiving** (of data and software) as a topic. However, the issue was not always explicitly addressed because this requires a cost estimate and embedding in the organisation and its working processes.

In general it can be said that **there is no such thing as the average user**. Even within a community, users' affinity for IT varies considerably. However, IT knowledge is more widespread in some areas such as astronomy than in other areas such as the social sciences.

As a whole, the **number of people** who work on a project **cannot be easily defined**. Most organisations assign employees to several projects in order to leverage synergies. Many of the respondents also stated that it was difficult to impossible to calculate the costs for the ingest of a data set.

In general there are **significant differences between the surveyed projects** in terms of size, development status, future prospects, number of partnerships etc. As one aim of performing the interviews was to cover as wide a bandwidth as possible, this does not count as an unusual result. The interviews should therefore be considered to be **indicators of the breadth of the topic area** and not examined too closely for possible generalisations.

Workshop and symposium results

In the scope of the Radieschen project one expert workshop and one symposium were organised. Around 85 people took part in the workshop (April 2012) and some 135 in the symposium (January 2013). The participants mainly came from the German-speaking area and represented a wide range of research disciplines and subject areas – from veterinary science and earth sciences through to computer centres and scientific libraries. The expert workshop was divided into working groups with the topics

- WS1: Policies and incentives: what are meaningful and necessary guidelines in handling research data?

- WS2: Integration into the research process: does the data come to the infrastructure or the infrastructure to the data?
- WS3: Generic vs. discipline-specific services: what are the factors for the success of interdisciplinary services?
- WS4: Opportunities and limits to the outsourcing and centralising of services

The symposium consisted of a lecture part and smaller expert discussions in the supporting programme. The topics of the expert discussions ranged from data management through policy up to virtual research environments. The common aim of both events was to bring the research data community together, to network and to promote the exchange of experiences and discussion between the participants.

In the course of the **expert workshop** the scientists agreed that the question of whether the services are offered centrally or decentrally is determined to a substantial extent by the volume of the data. The discussion should centre around “heterogeneous vs. homogeneous data” and be less about the question of whether we are dealing with “big data” or “small data”. The most urgent question is whether infrastructures exist that can present heterogeneous, highly-complex data in a simple manner and whether infrastructures exist to extract information from very large data volumes. There are currently a range of formats for both large and small, heterogeneous and homogeneous datasets, which is difficult to administer.

A further consequence of the discussion from the expert workshop was that tools for research data management need to reach a level that is comparable with commercial tools. Additionally, policies, infrastructures and incentive systems need to be treated on an equal footing another and developed further.

The discussion demonstrated that a cultural shift is also possible in evaluating a more systematic handling of research data and sustainable, user-friendly applications need to be developed that can be seamlessly integrated into scientific working processes.

The large number of participants in the **symposium on research data infrastructures (FDI 2013)** in January 2013 showed that the topic of research data is regarded as important. In the discussions it could be seen that the development of research data infrastructures continues to progress in a heterogeneous manner. However, the discussion also showed that the topic of research data infrastructures has reached a level of conceptual maturity that is comparable with concepts in other European countries, the USA or Australia.

In particular, a certain level of convergence can be observed in technical and conceptual development. Alongside that there are areas that lack conceptual clarity which, however, need further research, for example into the topics of “quality”, “trust" and “cost and price models”. There are also gaps in terms of data management tools and their integration into working processes of the scientists. The participants criticised the fact that many actors involved in data management were not sufficiently well known and also that networking within the community could be optimised. The demand for offerings in the areas of qualification and consultation was also particularly pronounced.

It became clear at the event that an improvement in handling research data is not only a question of the supporting technical infrastructure, but requires a cultural shift in science. The cultural shift in respect of research data is moving towards more open handling of this part of the scientific narrative,

as proposed in the report “Science as an Open Enterprise” by the Royal Society (2012). The degree of openness is determined by the tension between trust in one’s “peers” and control of one’s own “work”. Alongside the social norms, the legal framework for research data also needs to be developed further.

If the situation is to improve, the creation of data, software and infrastructures as a contribution to the scientific value system needs to be established alongside the former standard of literature publications. This requires the benefit of a research data infrastructure to be more apparent to researchers. Data policies and the consideration of these services in institutional evaluation systems could help support this shift.

The participants positively evaluated the opportunity to learn about new, exemplary solutions for handling research data. Some projects and promising results were presented. Evaluating the results and exchanging experiences can initiate a development in this context that goes beyond the experimentation stage. What is important is that the exchange between the actors continues and also includes the communication of practical approaches to extend the circle of actors.

Both the participants and the organisers of the symposium evaluated the event as being extremely helpful for the exchange of ideas, the generation of new impulses and the networking of actors amongst themselves. Follow-on events are already under consideration.

Outcomes of the discussion of the INF workshop

On 11 April 2013 the Radieschen project held a workshop of the CRC-INF project at SUB Göttingen with the aim of forming a community. The workshop was very well attended with 40 participants from 21 different CRC-INF projects.

With some €561 million, the CRCs receive around 20% of the research funding approved by DFG. After funding of individual projects (around 35%), this represents the second-largest portion of the DFG funding programme.[20] Currently 232 CRCs are being funded, including 27 CRCs with an INF subproject.

The workshop met with great interest in the circle of the CRC-INF-projects. Very broad coverage was achieved. During the event a high demand for discussion was apparent. A range of different questions were raised and discussed.

It was seen that the CRC-INF projects are either located at the ***libraries*** or the ***computer centres*** of the respective locations. Several locations, such as the universities of Bielefeld, Freiburg, Trier, Kiel etc., are using the CRC-INF projects to set up location-wide solutions for a research data infrastructure.

The typical activities of the INF projects that crystallised in the run-up to the workshop through a poll of the respondents were confirmed once again in the workshop. In this survey 18 CRC-INF projects specified the following as typical activities:

- Provision of a platform for the central storage and exchange of the data, e.g. database, repository, file server (13 responses)
- Provision of a collaborative working environment, e.g. project management or portal software, web portals developed in-house with integrated tools (ten responses)
- Consultation and support, in part also training, e.g. data preparation, data analysis, metadata, policy development (seven responses)
- Development, implementation and provision of tools, e.g. computer-linguistic tools (seven responses)
- Publication (six responses)
- Four responses and fewer: archiving, administration, development of standards and formats etc.

Issues that generated particularly intensive discussion were the questions of long-term archiving and sustainability. Some of the CRC-INF projects are aimed at retaining the data produced over a longer period, whilst other participants warned against overloading the CRC-INF activities with such demands. The set-up of a suitable and long-term/sustainably available research data infrastructure is better done outside the projects. The INF projects are no long-term and comprehensive replacement for a missing research data infrastructure. In this context the DFG expert who attended made clear reference to the responsibility of the universities as it is they who apply for the CRC and therefore also need to provide the necessary (digital) infrastructure. This also includes long-term archiving and availability.

[20] The specified figures refer to 2011 (cf. Effertz und Schoch 2013)

A further point that was intensively discussed was the question of acceptance. Experience in respect of the acceptance of the INF projects by the other subprojects varies greatly. In one case (CRC 649) the PI of the INF projects is identical to the spokesman of the whole CRC; this CRC-INF project had no challenges to report in terms of its acceptance. The following experiences were reported by the other CRC-INF projects. Obtaining the acceptance of the other subprojects is a lengthy and time-consuming process, which will succeed above all if the CRC-INF projects respond very specifically to the needs and processes in the working methods of the other scientists.

In the outlook part some very specific wishes for future activities were formulated. In particular the wish to hold further CRC-INF workshops in future was expressed. It was stressed that there is particular demand for a highly focused workshop approach on specific topics. The specified topics include for example:

- Acceptance, or how this can be increased
- Tools used, technologies, collaborative working environments (follow-on use)
- Heterogeneity (of data, formats, requirements, technologies etc.)
- Policies, e.g. the research data policy of a CRC, and also the research data management of the subprojects

6. Interdisciplinary topics

Cost allocation between the actors within an organisation remains unclear. Although a rough cost estimate is possible of which costs are incurred through ingest, preparation of the data and its storage, at the same time it is not usually possible to allocate these costs to the individual actors such as employees or departments of an institute (see also the “costs” report). In most cases there is no clear allocation to a cost centre.

A further open question is the distribution of the costs between the data-producing project (e.g. project “Radieschen”), the institutes where the project is established (e.g. the GFZ-CeGIT department) and the institution (here the Helmholtz Zentrum Potsdam Deutsches Georesearch Zentrum GFZ) that is managing the project. To date there are no corresponding guidelines in place in the large science institutions. As a consequence, the future existence and the maintenance of existing research data repositories is uncertain. Funding bodies fund projects to resolve current challenges. Existing repositories are seen as inventory and thus fall into the remit of the umbrella institution. The umbrella institution in turn refers to the acquisition of funds for the further operation of such repositories. The introduction of a special levy as part of the approval of projects, similar to the German solidarity surcharge, would be one option for improving the current situation. This surcharge would be borne in equal measure by the umbrella institution and the funding body. Further, the topic of sustainability should be given greater consideration when applying for future projects, for example by drawing up the sustainability plan as part of the project work.

A constantly recurring topic of the discussion is the **value system for recognising scientific work in its various forms**. (See also the Knowledge Exchange report on “The Value of Research Data – Metrics for datasets from a cultural and technical point of view”[21]). To date the metrics only take into account the traditional forms of publications (H index,[22] journal impact factor (JIF)[23]). Before now

[21] http://www.knowledge-exchange.info/datametrics
[22] http://de.wikipedia.org/wiki/H-Index

data or software publications have had no or only little relevance in these indices. A recognition of such publications would considerably boost the awareness of the relevance of research data and its publication and thus also drive forward the further development of research data infrastructures. The same applies to the publication of software. What would be helpful in this context would be the development of new categories in the generation of KPIs in scientific evaluation systems that give equal weighting to different forms of publications as well as increased support for and promotion of the new forms of (open access) journals on data and software publication by the publishers, researchers and also the funding body.

Many scientists tend to have reservations about modern forms of communication such as **social media**. This also applies to scientists of Generation Y, who represent science's new blood, but who do not count as "digital natives".[24] Social media and online forums in research are not usually regarded as legitimate research tools.[25]

New web-based tools and other innovative applications, however, can make a contribution to the international networking of scientists, and also be used for data exchange and further data collation. For instance, a small group of scientists is increasingly using open data, exploring crowdsourcing mechanisms and using citizen science[26] to support their work (ESA AstroDrone project[27]). This unconventional approach harbours substantial potential and should be supported accordingly.

7. Outlook and recommendations

This synthesis report describes the status quo of the development of research data infrastructures in Germany. It describes the technology currently in use, provides an overview of organisational structures and examines cost allocation. Further, the individual reports point to gaps in development and sketch out required developments in the near future.

Yet how will the development of research data infrastructures look in the distant future? Will developments proceed in a straight line, i.e. can we imagine what innovations will arise over time? Can major upheavals be expected? Are there already any signs of this?

The history of technology teaches us that innovations whose effects were initially considered to be minimal actually developed the potential to drive tried-and-tested technologies out the market and take their place. Examples are the triumph of the telephone, which replaced telegraphy or Wikipedia, which took a leading position as an online encyclopaedia and thus knocked long-established works such as the Encyclopaedia Britannica off the top spot. Some of what we today take for granted today we did not even imagine ten years ago. In the context of research data the concepts "grid" and "cloud" are good examples of this development. At the end of the 1990s the term "grid" emerged as the promise to pull unlimited IT resources pretty much "out of the socket". That appeared irrelevant for private users as the grid concept was initially aimed at user communities. Commercial applications that had also been conceived for funding by the German Federal Ministry of Education

[23] http://de.wikipedia.org/wiki/Impact-Faktor
[24] http://de.wikipedia.org/wiki/Digital_Native
[25] The British Library and JISC (2012)
[26] The Royal Society (2012)
[27] ESA project "AstroDrone" http://www.esa.int/ger/ESA_in_your_country/Germany/Smartphone-App_verwandelt_Spielzeug-Drohne_in_Raumsonde

and Research BMBF were never developed because the targeted companies had no faith in the security of the grid applications.[28]

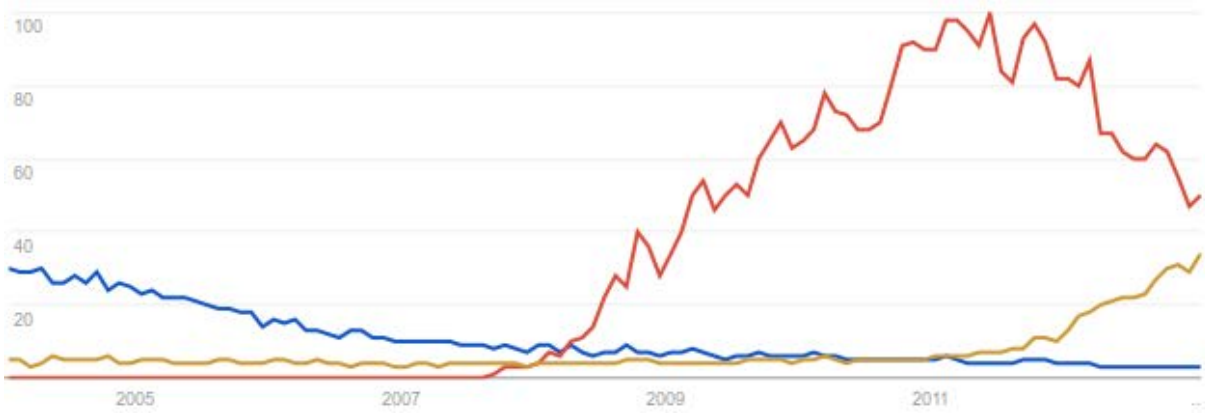

Fig. 9 Histogram of Google searches for the terms "grid computing" (blue), "cloud computing" (red) and "big data" (yellow) in January 2013. Source: Google Trends.

The rise of the concept of "cloud" rendered the "grid" concept practically obsolete (Fig. 9).

What does that mean for technologies and services in terms of dealing with research data? A look at the trends sketched out above shows the dynamics of development and demonstrates the difficulty in predicting how research data will be handled over the next ten years. It is impossible to predict what technical solutions will be available. Also, trends can only be identified to a limited extent because development remains largely influenced by *disruptive innovation* patterns, which in itself represents a trend in further development.

What is actually a trend and what is just hype? Is "big data" actually a significant trend in science? As the scientists of tomorrow, will the "digital native Internet, use these technologies differently and more freely? In this case it helps to take a step back from the technologies and to ask which processes need to be technically supported.

2003 saws the publication of the influential article "e-Science and its implications"[29] by Hey and Trefethen, which introduced the term "data deluge" into the discussion. At that time the discussion was still primarily concerned with the anticipated data volumes. Technical progress, however, brought further aspects into view, namely the possibility of formulating and testing new hypotheses through exploratory analysis of the data. In this context scientific progress depends directly on the availability of the data and the opportunity of processing it, which is called "*data intensive science*".[30]

Disruptive innovation[31] can be seen in many further examples of the history of technology. What they all have in common is that the innovation starts off in a niche market unnoticed or unheeded by the market leaders. A disruptive innovation may consist of a new technology, a novel product or an innovative service. Fig. 10 shows the course of a disruptive innovation, referred to here as disruptive

[28] Klump, 2008.
[29] Hey and Trefethen, (2003).
[30] McNally et al., (2012)
[31] http://en.wikipedia.org/wiki/Disruptive_innovation

technology. The technology starts off being subordinate to established products. After acceptance in the lower market segment (low quality use) the technology improves and higher-level markets are conquered (medium quality use – high quality use) until ultimately a fully-fledged product unseats established market leaders (most demanding use).

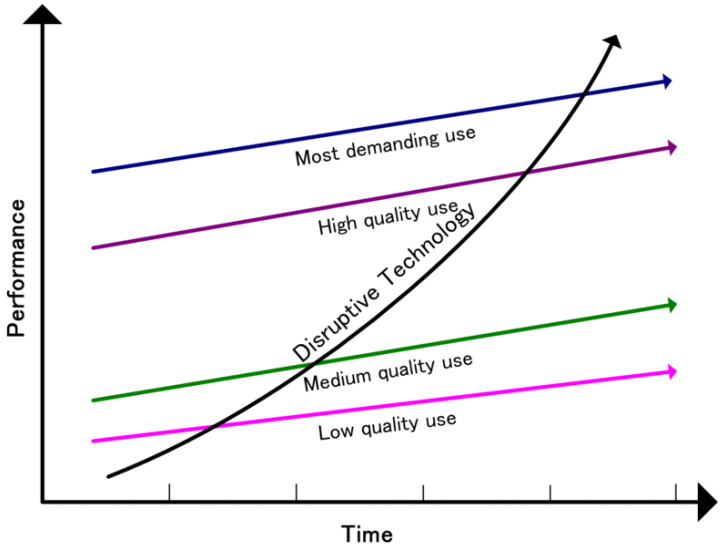

Fig. 10: Course of a disruptive innovation (source: Wikipedia)

Also, further innovations are to be anticipated in the field of research data and research data infrastructures, both at technological level (academic cloud computing, data-driven research), as well as at social level (value system, publication metrics). Not all of these new developments can be classed as disruptive innovations. However, even the area of research data infrastructures should be open to innovations and trends should be observed and new developments promoted. Fig. 11 shows a view of the Gartner Hype Cycle for emerging technologies. The Hype Cycle shows new technologies from their initial emergence in the area of "technology triggers", the start of the hype through the "valley of disillusionment" through to a widespread acceptance of the technology and implementation in marketable products.

As a conclusion it may therefore be interesting for funding bodies to **integrate agile components into their programme management** to be able to react quickly and flexibly to new developments. It may also be conceivable to fund products with a think-tank character or projects as a kind of sounding board for new technologies and developments.

A **project that coordinates the interaction of the research projects of this funding area** is also desirable in the area of research data infrastructures. Such a project should also focus on the performance of networking activities, the exchange and distribution of results and the building up of a joint knowledge base of the projects.

Apart from a couple of exceptions, research data infrastructures are located in the generally more conservatively-orientated environment of scientific libraries and the classical sciences. These focus more on researching issues and provisioning services and less on the search for innovations. A project on the topic of innovation management and research data infrastructures or **open innovation**

in the research data context could be conducive to the further development of research data infrastructures and opening them up to new trends and innovations.

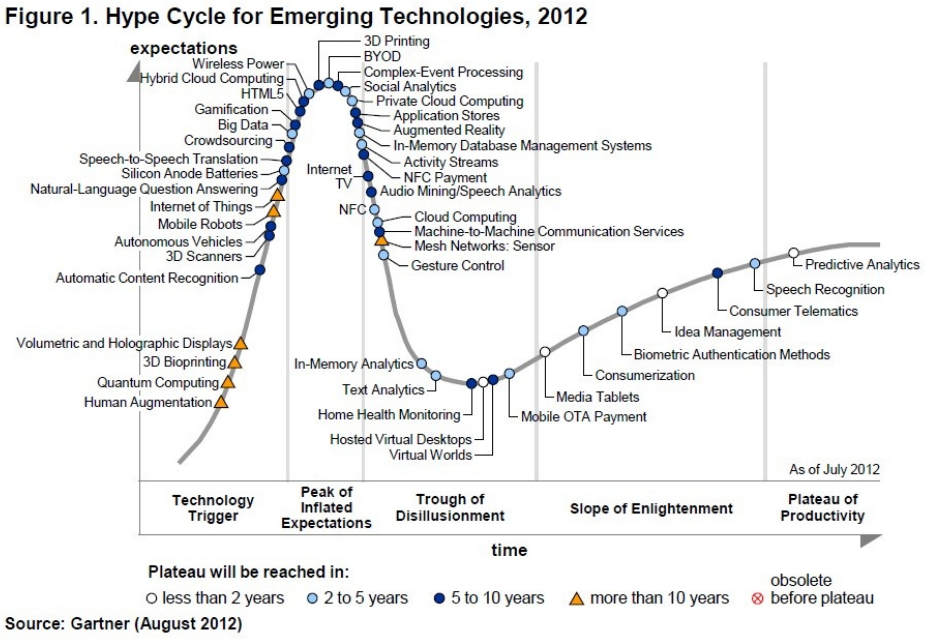

Fig. The Gartner Hype Cycle offers a graphic representation of the status of the maturity and the acceptance of new technologies and applications and shows their potential relevance for resolving real business problems and exploring new opportunities. The "emerging technologies" Hype Cycle focuses on developments with a broad, cross-industry significance and a high potential for transformation.

However, innovations in information technology are also often short-lived. Many of the services offered for example by Google have an average life of around three years,[32] which roughly corresponds to the length of innovation cycle in information technology. The example of Google shows that not even a global company can predict the success of one of its services, so the way to deal with this challenge is to develop a portfolio of services on a common platform. If individual services are unsuccessful, they can be switched off again without impairing the operation of the platform as a whole. A successful strategy for infrastructure facilities could be to develop a **modular portfolio of services** that is based on a common platform. This strategy would enable the facilities to adapt the services flexibly to the constantly changing needs of research, whilst the underlying platform can be continually developed further. In this way the contradiction between the infrastructure's need for stability and the demands of flexible, and potentially short-lived, applications can be bridged.

The general development of research data infrastructures in Germany is characterised by a decidedly active research data community. However, what is needed is **increased cooperation with projects at the European level** as well as an opening up to other disciplines such as the aforementioned innovation management. Attention should also be paid to the **parallel development in the economy** in dealing with large data volumes ("big data") and, in this context, in the **creation of the profession of the data scientist**. Despite different orientations and objectives there are clear parallels here.[33] It

[32] Arthur, C. (2013)
[33] Biesdorf, S. Court, D. and Willmott, P. (2013)

is certainly conceivable for these developments to be correlated, or at least observed, and to take from them suggestions for the further development of research data infrastructures.

8. Bibliography

Arthur, C. (2013), Google Keep? It'll probably be with us until March 2017 - on average, The Guardian, 22 March. [online] Available from: http://www.guardian.co.uk/technology/2013/mar/22/google-keep-services-closed

Biesdorf, S.; Court, D. and Willmott, P. (2013), "Big Data: What's your plan?", McKinsey & Company, McKinsey Quarterly - Insights & Publications, March 2013, http://www.mckinsey.com/insights/business_technology/big_data_whats_your_plan

Effertz, E.; Schoch, K. (2013), Teilprojekte zur Informationsinfrastruktur in Sonderforschungsbereichen. "INF-Projekte". (Presentation at of the joint workshop of the CRC-INF projects on 11 April 2013 in Göttingen)

Gausemeier, J., Stoll, K., Wenzelmann, C. (2007), Szenario-Technik und Wissensmanagement in der strategischen Planung.
In: Gausemeier, J. (Ed.) Vorausschau and Technologieplanung, 1st ed., p.3-30, HNI, Paderborn, 2007

Sontheimer, K. (1970), Voraussage als Ziel und Problem moderner Sozialwissenschaft. In: Klages, H.: Möglichkeiten und Grenzen der Zukunftsforschung. Herder, Vienna, Freiburg, 1970

Hey, T., and A. Trefethen (2003), e-Science and its implications, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 361(1809), 1809–1825, doi:10.1098/rsta.2003.1224.

Klump, J. (2008), Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten, nestor-Materialien, Kompetenznetzwerk Langzeitarchivierung (nestor), Frankfurt (Main), Germany. [online] Available from: http://nbn-resolving.de/urn:nbn:de:0008-2008040103.

McNally, R., A. Mackenzie, A. Hui, und J. Tomomitsu (2012), Understanding the "Intensive" in "Data Intensive Research": Data Flows in Next Generation Sequencing and Environmental Networked Sensors, IJDC, 7(1), 81–94, doi:10.2218/ijdc.v7i1.216.

The British Library and JISC (2012), Researchers of Tomorrow: the research behavior of Generation Y students, June 2012, http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow.aspx

The Royal Society (2012), Science as an open enterprise, June 2012, http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf