H Subject-Specific Requirements for Open Access Infrastructure – Attempt at a Synthesis

Christian Meier zu Verl and Wolfram Horstmann

1 Introduction

This study addresses the question how to characterise subject-specific requirements for research infrastructure with a focus on the influences of Open Access (OA), in the general sense covering open access to literature, open data and open science. The introduction – which is assumed to have been read before this synthesis – specified the following.

We refer to the scope of OA in terms of the Berlin Declaration:

Establishing open access as a worthwhile procedure ideally requires the active commitment of each and every individual producer of scientific knowledge and holder of cultural heritage. Open access contributions include original scientific research results, raw data and meta data, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

We refer to the definition of Open Access in slightly modified terms of the Budapest Declaration:

By open access, we mean its immediate, free availability on the public internet, permitting any users to read, download, copy, distribute, [export], search or link to the [materials], crawl them for indexing, pass them as data to software or use them for any other lawful purpose.

By research infrastructure we mean the entirety of production and service, which includes instruments like large sensors, satellites, laboratories, and many more facilities, like digital services and virtual research environments. The research process within that refers to all facilitating processes: the researcher and his or her behaviour is not part of the infrastructure.

The chapters in this study present subject-specific views on OA infrastructure for research by analysing research workflows as well as researcher behaviours. They specifically take into account two aspects, namely (i) working with literature and (ii) working with data. Throughout the preceding chapters and throughout this chapter, the topic of OA infrastructure is centred on digital resources. Even though there are many transitions between physical and digital resources mentioned – for example, between the human researcher and the computer or a digital resource, and the physical, experimental as well computational facilities – these transitions will not be addressed explicitly in most of the cases for the sake of lingual simplicity.

The following sections will discuss commonalities of and differences between the different presented views on OA infrastructure and formulate recommendations for supporting the development of infrastructure (e.g. through funding initiatives) under specific consideration of the question how principles of "openness" or OA can be applied. In line with the qualitative approach of this whole study, the synthesis will be provided as an interpretative account.

When comparing the chapters, the most obvious observation can be summarised in one word: diversity. The archaeologist in a desert excavation has different requirements from a climate researcher crunching observational satellite data or an engineer building a biologically inspired robot hand. On the first view, this diversity may appear to be the natural enemy of infrastructure, since infrastructure is about commonalities in terms of global standards, joint facilities and shared resources rather than about differences between diverse subject-specific requirements. Simultaneously, it is obvious that research must be extremely diverse in terms of thematic and methodological specialisation in order to tackle the ever-more specific challenges of the world. **Thus, any roadmap for OA infrastructure must address this natural tension between diversity and infrastructure.** This study chose the approach of addressing this tension directly by providing an account of diversity and then reflect this diversity in specific aspects of OA infrastructure such as OA to literature and OA to data. It is not expected that the study will provide a complete picture and a detailed plan for the next decades: rather it is expected that the reader will gather impressions of diversity and develop a (maybe sometimes tacit) understanding of how diversity can be managed within research infrastructure development in a way that leaves research with sufficient degrees of freedom for self-organised developments while supporting the emergence of synergies between those developments through shared resources that apply principles of openness.

Attempting to provide a synthesis, the following sections will consequently analyse the commonalities and differences. This is done first on a high conceptual level and then on a detailed, systematic case-by-case basis. Thus, the resulting qualitative, rather than quantitative, account shall inform strategic decisions for future developments with respect to conceptual rather than procedural aspects. The specific measures, programmes or plans are assumed to be the result of these strategic decisions.

2 Methodological reflections

The approach taken in this study is unusual – or even extremely particular. Rather than analysing the principles of research practice through large-scale, representative questionnaire exercises, a small selection of partners provide individual and often descriptive accounts of specific subject areas. Rather than mapping the world of research with a broad account aiming at comprehensiveness, the analyses in the specific subject areas dive deep from the institutional and departmental level to the individual researcher and even research project.

2.1 Localising the study in the world of research

The world of research represents the most specialised activities in human behaviour. Being always on the verge of the unknown – things that never have been experienced and discovered before – researchers have to develop extremely resourceful, creative and swift capabilities in order to "squeeze" novel knowledge out of their minds and the world. Additionally, considering how much knowledge has already been generated through research in the last centuries and decades, the questions posed and methods used are becoming ever more capillary. At the same time, the phenomena analysed by research are becoming ever more complex and significant. Topics such as cancer, climate, consciousness or terrorism require many researchers of different subject areas to join forces.

The question of how to characterise research in a comprehensive sense is the subject of specialised research (e.g. philosophy of science or science studies) and goes far beyond the scope of this study. This study rather shall provide an explorative account of a very specific aspect of research practice, namely OA infrastructure. Thus, this study deliberately did not attempt to provide a representative account of research. Instead a pragmatic approach was taken: six partners (institutions, organisations) were chosen to provide their subjective view on OA infrastructure. The selection of partners originally referred to funding areas of the European Commission (EC), which were chosen as pilot areas for implementing the OA policy of the EC. These partners are considered as exemplars of infrastructure institutions in a given subject area (Table H.1): they not only perform research in a given subject area but also provide some sort of infrastructure for their subject area. Thereby, the analysis of both aspects – subject-specific requirements and infrastructure – should be made possible.

Table H.1 Pairings of partners and subject areas

Partner	Subject area (corresponding to EC funding)
CGIAR	Environment (health)
CITEC	ICT – cognitive interaction and robotics
CNR/NKUA	ICT/capacities – e-Infrastructure
DANS	Science and society
EBI	Health
WDCC/DKRZ	Environment

CGIAR, Italian Consultative Group on International Agricultural Research and Bioversity International; CITEC, Cognitive Interaction Technology – Center of Excellence; CNR/NKUA, Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie dell'Informazione and the Department of Informatics and Telecommunications of the National Kapodistrian University of Athens; DANS, Data Archiving and Networked Services; EBI, European Molecular Biology Laboratory/European Bioinformatics Institute; WDCC/DKRZ, World Data Center for Climate/Deutsches Klima Rechenzentrum.

The six subject areas and corresponding institutions definitely do not represent the complete world of research. However, they are spread across different domains of research, such as (natural) science, the social sciences and the humanities. Many of the partners are themselves highly interdisciplinarily organised, sometimes bridging between sciences and humanities (e.g. CITEC and CGIAR) and often showing overlaps in their constituent disciplines with other partners participating in this study, For example, both CGIAR and WDCC include environmental science, and both CITEC and EBI include computer science. In a sense, the approach taken in this study tries to provide vertical “drilling cores” into the world of research infrastructure rather than represent research infrastructure with a horizontal coverage.

2.2 The process of writing the chapters

Each exemplar partner appointed one or more chapter authors. In addition to all participants receiving written briefings from the editors, all chapter authors physically met three times to discuss concepts and progress. It was a deliberate decision not to provide too strict methodologies and structures for the chapter authors in the writing process. The reason for this liberal methodological approach was to provide a degree of freedom that could elucidate obvious but also subtle differences between the subject areas and inform future studies about possible approaches. The authors are experts in their fields and it was assumed that they know best themselves how to characterise their subject area. The reader should be provided with an expert subjective view of the given subject area with a taste of the sometimes implicit principles of thinking and working in that subject area rather than a normalised account constrained by too many pre-fabricated assumptions. In this sense, the free choice of chapter authors in how to characterise their subject areas is part of the design of this study, since the individual methodology chosen by authors to characterise their subject area also informs the reader about the self-perception within that subject area. Further, comparing the different methodologies applied in the chapters provides in itself an account of diversity between the subject areas.

2.3 Observations during the writing process

Why is the understanding of diversity so important for the future development of research infrastructure? Why did this study not try to focus on uniformity? It became immediately clear in the discussions at the meetings and the written correspondence that chapter authors were highlighting differences rather than commonalities. Everybody pointed very much to the "special character of their case" and that it is "not comparable to other cases". Since it can be assumed that this attitude would become prevalent in any measure that is aimed at implementing infrastructure on a broader scale, it was decided to address this diversity directly. Thus, the diversity of requirements has to be studied, understood and respected with the greatest possible care.

An obvious observation with respect to diversity is that almost every partner emphasised that it is impossible to provide a single typical research workflow, even within the work scope of a given institution. Thus, a generic model of research workflows applying to all subject areas was not feasible. Accordingly, almost each partner subdivided the corresponding chapter into several sections, describing different typical researchers, research groups, disciplines or a number of different research workflows, which defined their very specific requirements for the infrastructure services. The methodologies used to characterise these typical research workflows varied from descriptive, observational accounts to semi-structured expert interviews and systematic questionnaires. This variation in methodology shows that the authors found different methods appropriate to characterising their subject area and supports, again, the presence of strong differences within and between subject areas.

However, the constituent disciplines in one subject area show overlaps between partners in a non-systematic fashion, with almost no overlaps in the subject-specific infrastructure services described, even in cases where the same discipline is involved in two subject areas. In other words, the infrastructure services for one subject area, say biology, provided by two different institutions, say EMBL-EBI and CGIAR, serve very different functions in the scientific community, even though they may be used by the very same researcher. But it has to mentioned that the descriptions of OA to literature show much more homogeneity with respect to infrastructure services than the descriptions of OA to data.

In sum, these unsystematic overlaps between research practices and infrastructure services show that not even a partial Cartesian "map of research" can be produced by analysing and comparing the different chapters. Rather than a traditional disciplinary division, say infrastructure for biology vs. infrastructure for geology, **specific research problems and their corresponding research projects performed by collaborative interdisciplinary organised groups can be identified as the drivers of research infrastructure**. Thus, a multidimensional organisation of research infrastructure – a network model – appears to be the appropriate model for describing research. both a layer cake model, in which a subject research layer is based on a data layer, in turn based on an ICT layer, or a hierarchical matrix model in which layers are pervaded by subject-specific "columns", seem too simple to catch the subtleties of research infrastructure.

2.4 Initial observations summarised

The analyses provided by the exemplars are "drilling cores" that characterise research infrastructure in a given subject area. Initial observations about these drilling cores can be summarised in a first coarse approximation as follows:

i. Each institution or organisation provides research infrastructure specific to the subject area in terms of multiple and focused requirement satisfactions, defined by the constituent subject-specific research processes.
ii. Institutions or organisations, although considering themselves subject-specific, do not have the self-perception for serving a single subject area. Rather, they serve a multitude of disciplines, with the tendency of becoming even broader in their constituent disciplines.
iii. Infrastructure provided by the institutions or organisations is designed to support the sharing of resources and collaborative research with a multitude of different services, such as databases, repositories, analytic software or communication tools. Those tools seem to be modular to serve the diverse needs of the researchers involved rather than providing an integrated virtual research environment for one subject area.

iv. OA to literature is described as a relevant phenomenon in each different subject area. The degree to which OA to data is established in a given subject area varies.
v. OA to data is characterised as much less established than OA to literature and often accompanied by enumerations of obstacles that prevent OA to data.

3 General assumptions throughout the chapters: the benefits and obstacles of OA infrastructure

Even more prominent than the question of characterising research in the different subject areas, the current state and perspectives of OA infrastructure was addressed by this study. Before providing a more detailed account in the later sections of this summary, a general interpretation of assumptions regarding OA infrastructure is given here first.

3.1 Benefits of OA infrastructure

OA infrastructure is a complex concept determined by multiple aspects, most obviously by the two aspects infrastructure and OA. These two aspects will now be characterised separately in terms of their benefits, as can be concluded from the chapters.

3.1.1 Benefits of infrastructure

Cost considerations As the predominant benefit, cost considerations can be easily identified as a benefit of providing infrastructure. It is generally assumed to be more efficient when a given service, say a database, is provided once to a research community rather than providing the service twice or multiple times in different locations. Today's digital services easily allow remote access to a single shared service from different locations for different users, wherever they are and whatever their particular research interest.

Enabling research Another benefit is providing researchers with access to resources that would otherwise be not accessible, to enable research processes that would otherwise be not possible. Examples are access to licensed literature for which the individual researcher has no access rights and access to research results (e.g. personalised surveys) that are only accessible on special conditions (e.g. highly secured workstations) or expensive experimental facilities.

Transparency and comparability Good research practice, for example in terms of reproducibility of research results, dictates the comparability of research results in order to verify and falsify them. When researchers use the same infrastructure, say again a database, the research processes are more likely to be comparable than when differing infrastructures are used: file formats, metadata standards and statistical methods tend to be similar in an integrated infrastructure. The emergence and the application of standards is thus a very important implication with respect to transparency and comparability.

Synergies Providing infrastructure is a way of sharing resources among researchers. Synergies emerge through sharing when the research process can develop a novel quality that would not be possible without sharing. A prominent example is the Human Genome Project, in which joint infrastructure and standards were used to collaboratively build a resource of research results that could practically only be achieved in a global and collaborative manner.

3.1.2 Benefits of OA

Cost considerations Any barrier to resources for research causes costs. In a simple case, licensing access to electronic literature requires researchers or institutions to work with registration or accreditation obstacles (e.g. logins or IP-checks, digital rights management) and payments (e.g. invoice processing, bank transfers). In a more complex case, missing access to primary research data can force research funders to finance the same experimental projects several times. In general, the innovation capacity and creativity of research is limited wherever research resources are kept behind barriers. Thus, anything that is OA can help to reduce the effort and costs incurred when dealing with barriers.

Enabling research OA can even play a more crucial role when a given research project is simply not possible without OA, i.e. situations in which a researcher is endowed with access to resources specifically because of their open character. This is seen, for example, when a researcher grounds a project on data that have to be open in order to be re-used, say for an application that performs runtime public transport monitoring.

Transparency and comparability It almost goes without saying that OA enhances the possibilities for researchers to use, analyse, assess and check the work of their peers. The recent trend of data publishing as a supplement to research literature corroborates this observation.

Synergies Intensified peer communication and collaboration through OA resources is instrumental to effective division of labour and complementary, rather than redundant, research projects. OA can enhance the information flow between otherwise isolated research activities and is therefore crucial for performing collaborative, interdisciplinary research projects.

Summary The benefits of infrastructure and OA considered separately reveals a strong relation between these two main aspects addressed in this study. Even though it might seem trivial, it should be noted at this point that OA and infrastructure are two completely different phenomena: OA is a mode of communication while infrastructure refers to facilities. However, the benefits of both can be characterised referring to the same aspects of research: cost considerations, enabling research, transparency, comparability and synergies. The most obvious reason is that **both infrastructure and OA imply a notion of sharing**

This study shows that there are general assumptions underlying the analyses of OA infrastructure. In summary, infrastructure is an essential prerequisite of research that:

- reduces costs by providing shared resources instead of building multiple local solutions,
- enables research that is otherwise not possible,
- enhances comparability by providing joint standards and methodological frameworks,
- creates synergies between researchers, groups or disciplines by sharing the same resources.

If infrastructure is operated according to OA principles, all benefits of infrastructures are boosted because the degree to which the sharing of resources can be exploited is maximised.

3.2 Obstacles to OA infrastructure

The obstacles mentioned in the chapters are so manifold that such an analysis would justify a dedicated study on these obstacles. Consider only one example within one specific chapter, namely data collected at a archaeological excavation site: the necessity of barriers to excavation data is explained by the protection of the data against the possible abuse by treasure hunters and the possible abuse by political activists. Treasure hunters or political activism are rather surprising in the context of research resources! It would be interesting to collect all such examples throughout the chapters, but that would not emphasise the obvious observations with respect to the obstacles for OA, namely:

i. There may be good reasons against a completely open research infrastructure, particularly when they are grounded in the research processes themselves, for example needing time to exploit the results before someone else (competition among colleagues or industries), protection of privacy (medical records, surveys) and risk of abuse (dangerous technology).
ii. These good reasons apply to a much lesser degree to the aspect of OA to literature than to the aspect of OA to data, since literature is localised at the end of the research process, where many processing and refinement steps on the results to prepare them for publication have already been performed.
iii. Obstacles to OA infrastructure vary so dramatically across subjects that they cannot be foreseen in a general OA policy. Therefore, a procedure for allowing exceptions from a general OA policy is required. Exceptions can be justified and assessed on a case-by-case basis for each research project, particularly with respect to the question of OA to data.

4 Comparative analysis

The following sections provide a comparative account of the main aspects addressed in the chapters, namely the characteristics of the research lifecycles as a whole and its constituent aspects of literature management and data management.

4.1 Research lifecycles

Each subject area is organised in different ways as a result of differing research lifecycles. Even individual fields in one subject area can be organised differently. Therefore, it is necessary to compare parts of research practices of locally situated units belonging to an entire research field. Depending on the subject of research itself, it is possible to find concordances of data management between research fields (e.g. between climate research and ICT). Thus, the purpose of this comparison is to find commonalities and differences in research workflows and to emphasise the research steps described as at the core of research activity. All considerations below are grounded on rather abstract and minimal descriptions of observable workflows, which were described in each subject chapter. We will discuss each research workflow by pointing at essential steps of research practices. In the end of this section we will highlight common steps and main research activities of each presented case by comparing all research lifecycles.

CGIAR/Bioversity International observed four projects to define requirements on agricultural research by focusing on typical groups within this field. The research field of agriculture is highly interdisciplinary and includes economics, geography, geology and climate research as well as biology. Most observed work groups are collaborating internationally to study agricultural developments in different parts of the world. Therefore, these projects need simple and stable instruments to measure, for example, developments of plants or behaviour of farmers. Generic steps of workflows within the field of agricultural research are (i) data collection, (ii) cleaning, (iii) archiving, (iv) use and (v) dissemination. It is not possible to locate the steps of the workflow to which researchers pay more attention, but the effort in collecting data is huge, which suggests that data collection and use are the steps with most activity.

CITEC categorised four different research areas within the institute. The categorisation runs along the following four sections:(natural) science, social science and the humanities, computer science, and robotics and engineering. Groups within one section behave similarly for data and literature management and also conduct research in similar ways. But there are major differences between these sections on research objectives, methods used and infrastructures that influence the entire way of conducting research. CITEC performs highly interdisciplinary research on ICT and each working group is well engineered. The common steps of research are: for (natural) science (i) data collection, (ii) processing, (iii) enrichment and (iv) re-use; for social science and the humanities (i) data collecting, (ii) processing, (iii) archiving and (iv) enrichment; for computer science (i) data collection or re-use, (ii) processing, (iii) archiving; and for robotics and engineering (i) data collection, (ii) enrichment, (iii) processing, (iv) archiving and (v) re-use. Beyond this interdisciplinary cooperation of groups, CITEC cooperates with international researchers and companies all over the world. It is not possible to locate main research activities but all groups basically have three steps in common: data collection, data processing and data archiving. So these steps are typical and most important for CITEC as an exemplar of ICT research.

CNR/NKUA describe six research workflows within the field of e-Infrastructure. The scope includes public and commercial research institutes and three of the cases have the following workflow: (i) requirement analysis, (ii) design, (iii) development, (iv) documentation and (v) testing and deployment. All six research lifecycles have standard phases in common, such as : (i) requirement analysis, (ii) designing and (iii) implementation. Depending on the research objective itself, e-Infrastructure research is heavily engineered and uses a vast amount of computing hard- and software. International collaborations are common to all research groups.

DANS describes a research workflow with five steps within the field of the humanities and social science, which is based on the workflow of the large-scale activity Digital Research Infrastructure for the Arts and Humanities (DARIAH). There are the following steps: (i) search/discovery, (ii) gather, (iii) analysis/experiment, (iv) publish/disseminate and (v) store/archive. Collaboration and sharing of current research results with the public or internal working groups is possible in all steps. It is not possible to identify steps in this research lifecycle that are more prominent than others.

EMBL-EBI describes five cases within the research field of health and life science. All cases have different objectives, such as examining genomic sequences, mechanics and dynamics of cell divisions, imaging brains, simulating neuronal cell signals and developing databases for mouse embryonic models. The common research lifecycle mentioned by EBI has seven steps: (i) data collection or re-use, (ii) processing, (iii) analysis, (iv) enrichment, (v) archiving, (vi) dissemination and (vii) publication of literature. All five cases use other methods to explore their subjects but it is impossible to make general statements about any kind of emphasis of a specific activity. Only research modelling is primarily related to data processing (for modelling) and archiving. The rest of research workflows are equally distributed in their activities throughout the complete research lifecycle.

WDCC/DKRZ describes five different cases of research institutes within the field of climate research. Most types of data are observational data such as images or sheets of numbers. A common research workflow includes four steps: (i) data collection or re-use, (ii) processing, (iii) enrichment and (iv) archiving and re-use. Climate research is very well engineered and uses a vast amount of computing hard- and software and technical equipments such as satellites, airplanes and observation stations. Researchers share their facilities internationally to constantly use these expensive instruments. Therefore, data sharing with colleagues and/or the public is commonly established at the climate research community. Some institutes are more specialised in data collection and archiving than others, and some researchers spend more time on data collection, archiving and disseminating than researchers who have no access to data or access to data-collecting facilities only for a limited period of time.

Summary Comparing these descriptions of research workflows from different subject areas and cases, it becomes obvious that some workflow steps are generic to conduct research beyond disciplinary and institutional boundaries. Even if we sample different research fields we can observe five steps that emerge in nearly every workflow. These five steps are (i) data collection (as direct or indirect collection by searching through databases), (ii) processing, (iii) enriching, (iv) archiving and (v) re-using. These steps are rather abstract and you can discover more differences by focusing in-depth on a single step. For example, collecting data enforces other research practices and facilities in climate research as in the field of social science.

Another aspect of the research workflow is the order of individual steps. Hence, it is instructive to have a look at the ordinal dimension of research workflows. Most descriptions of workflows start with the collection of data but some start with the re-use of data or requirement analysis, for example.

Also, the combination of steps is different, even in within one institute, and depends mostly on the research subjects, applied methods, technologies and collaborations. WDCC and CITEC have one arrangement of workflow in common. All other observed research workflows are different in combination. Most research workflows are workflows with four, and sometimes three, main steps. Sometimes, even the understanding of what can be count as an autonomous step of research workflow varies from case to case. But a common understanding of steps necessary to conduct research or to build an entire research workflow is observed. This is generic for all analysed data-driven research practices. As we mentioned, subjects, approaches and applied methods diverge at more complex levels; therefore, generic infrastructure has to be highly configurable in combination (such as modules to rebuild individual research rhythms) and suitable for different research environments by adapting the modules.

We conclude that there are five generic steps of workflow, even if these steps are always very specific on closer consideration. The arrangement of steps depends for the most part on the objectives, applied methods, technologies and collaborations. Therefore, any approach to generic infrastructure has to be highly configurable.

4.2 Literature management

A common final good of all research is literature. Almost all significant research knowledge is transformed into literature at a certain point and to a certain extent. The most obvious advantage of literature as container of knowledge is the way it supports understanding and dissemination of insights through time and space. Of course literature is indexical and written in different terminologies (with which one has to be familiar) but it is more durable and reaches more recipients than talk and more generic than data. Hence, management of literature is a generic task beyond disciplinary boundaries to reach large audiences. We differentiate three dimensions of literature management: (i) production, (ii) organising and (iii) dissemination. Nowadays, researchers explore new ways to present their literature. E-publishing, social media, OA and data publishing are only a few aspects depicting the current change of research publication. These upcoming developments influence all of the dimensions mentioned above and even restructure the principles of research. To serve new needs, it is necessary to analyse the management of literature in different research fields. How is literature managed through different disciplines? Which reasons can be observed for differing literature management? Where are the most progressive developments of literature management? How are these new developments organised? In the following section, we will look at each individual chapter one by one to finally compare all of the approaches and discuss commonalities and differences.

CGIAR activities are massively dominated by data management so that literature management is characterised concisely. There are several branches, which show established practices of OA publishing (gold) and CGIAR manages 14 OA repositories spread over all partner institutions. Since 2006 CGIAR provides a virtual library which gives access internal and external research literature on agriculture, hunger, poverty and the environment. This is a shared, integrated service that allows users to tap into leading agricultural information databases, including the online libraries of all 15 CGIAR Centers.

CITEC describes several ways to manage literature. Self-written literature can be presented through the central service PUB which is provided by the Bielefeld University Library. This service manages the bibliographic information as a generic service for all departments of Bielefeld University, which is locally configured to specific needs. Future developments by CITEC are semantic enrichment which allows formal representation of literature and the relations between them. Beyond this generic literature management, CITEC has four different groups which diverge because they are using different tools to write, manage and publish literature. The BehNatNeur group uses Endnote, Mendeley and Reference Manager to manage non-self-written literature. Data and literature can be published together. There are two forms of publishing which are preferred within the group: printed versions and electronic versions (accessible via the Internet) and 34% of published literature is OA (followed the green way). The SocHum group uses BiTeX, Citavi, Zotero and Mendeley. For collaborative writing they use Google Docs and Subversion. Data and literature are usually not published as a compound object and 57% of published literature is OA (followed the green way). The CompSci group uses BibTeX, Mendeley and Drupal (for metadata management of literature and for the literature itself, with modifications). Collaborative writing is managed by Subversion. It is not possible to publish data and literature together. The RobEng group uses Drupal (for metadata management of literature and for the literature itself, with modifications), Endnote, BibTeX and Subversion to manage literature. Both forms are established to publish as a printed version and as an electronic version (via the Internet), and 68% of published literature is OA (followed the green way).

CNR/NKUA stores most of the research literature locally on personal computers and manages and shares via e-mail or with software tools such as Google Docs and Dropbox. Literature writing is often realised with online tools like Google Docs to produce texts cooperatively, but each interviewed group behaves in a slightly different way. The D-Lib group searches to find literature via Google, Google Scholar, Wikipedia and DRIVER. If they write literature collaboratively, they use Google Docs or share their file via Dropbox or BSCW. The Agro-know group uses Google Scholar to find literature and to manage literature via Mendeley. They write collaboratively in many different ways, for example via e-mail, BSCW, Dropbox, Google Docs and Wiki (only if they collaborate with external researchers). Publishing data and literature together is not established. The group prefers to publish their literature OA. The researchers within the National Documentation Center search for literature via Google Scholar and Scopus. Literature is managed with CitUlike. They write documents collaboratively with SVN. They prefer to publish OA. The Greek Research & Technology Network uses Google to search for literature. The researchers manage literature with Mendeley and publish in journals and conference proceedings. A combination of literature and data publishing is not established. They do not prefer to publish OA. The MADGIK group searches for literature with Citeseer, Google Scholar and also with DRIVER. They collaboratively write documents via Google Docs but mainly they exchange documents via e-mail. In general, the common literature lifecycle is: (i) survey, (ii) analysis of literature, (iii) drafting and (iv) publishing. OA publishing is not desired by researchers within one mentioned organisation but by all the others.

DANS facilitates a self-archiving system called EASY (Electronical Archiving SYstem) which can archive both literature and data. There are four interviewed researchers, who come from different disciplines and manage their literature in different ways. (i) By using eDNA it is possible for archaeologists to conduct desk-based research with access to literature and data of other excavations. OA journals are not highly rated within the field of archaeology, so therefore they are not preferred. Some, but not every, researchers have an online list which shows his or her record of publications. Normally publications from excavations are published as reports under institutional copyright. These reports are necessary to understand the datasets in a better way. (ii) The historians described collaboratively written author literature for historical demography data. The specific role sharing depends on the difficulties in handling these demography data. Therefore, some historians prepare the datasets and the others interpret the datasets. (iii) The social scientist remarked that there is nowadays an inflation of publications. Literature writing focuses on articles, which are the major form of publication. There is no high-quality OA journal within the field of social science and publishing OA is not preferred. (iv) For linguists, literature is not only literature for reading but also data for research; therefore there is a clear tendency to OA with literature which can be used in both directions. The world of linguistic publications shifts towards enhanced forms, which make it possible to publish literature with data together. Some publishers embargo the literature for a period of time before it can be distributed OA (green).

EBI mentioned that journal articles are the primary output of life science. Most journals are published by commercial publishers, medical charities, learned societies, medical institutions, universities and research institutes. There are extensive, subject-based repositories of OA literature which are a well-established and integral part of the life science community. OA publication ratio varies between disciplines from 5% to 16%.

WDCC mentioned that online access is established in most subject-specific journals. One interviewed climate researcher reports that publishers support the publications of literature and data together. But only some formats of data are published, for example it is not possible to publish video data within one document. One climate points out that the German national license (covered by the DFG) provides access to the most relevant publication repositories for climate research. The interviewed researcher of the Climate Service Center mentioned that some publishers enable the exchange of data within literature. OA is established within the Climate Service Center. The climate researcher at the Karlsruhe Institute of Technology mentioned that they use Zotero to manage their collections, citations and sharing of literature. They prefer printed forms of literature. The interviewed researcher at the Alfred Wegener Institute for Polar and Marine Research mentioned that they prefer printed forms of literature and that OA is established in some extent.

Summary The landscape of research literature includes six fields of research which are currently similarly organised. If you compare these literature management descriptions, it is obvious that there are various tools to manage, write, publish and find literature. Some tools are common and you can find them all over research fields. These tools serve generic needs going beyond each disciplinary requirement. This is particularly the case for all literature management tools. Even though there are many tools like BibTeX, Citavi, CitUlike, Drupal (with modifications), Endnote, Mendeley and Zotero, they serve the same needs with slightly different modifications.

The infrastructure for literature is well established. The choice may depend on personal or institutional reasons. There are some tools which serve the form of writing via the Internet. Google Docs, BSCW, Dropbox, SVN and e-mail are the mentioned tools to write and share within the writing process documents. The common way is to write one document with many different versions which have to be merged by someone. Google Docs and Wikis serve the function to edit one document through different authors without document exchange. This can be done at the same time and the document will be stored at the server.

The use of OA publications is different between subject areas. Life science and climate research have well-established OA repositories. In the field of social science and the humanities, there are no high rated (golden) OA publication options. Most of the fields prefer to publish articles. Only two fields mentioned publishing books or using websites.

In sum, there is a common ground of literature management on which a generic infrastructure can build to manage the metadata and the literature itself. With respect to the publishing of data together with literature, it is obvious that there is no generic way to do this yet, but there are emerging techniques such as standardised forms of "enhanced publications".

We can conclude that there exists a generic and specific infrastructure which serves the needs of researcher at different research fields. On the one hand, management, discovery and writing literature is organised with the same tools and is not heavily dependent on subject-specific requirements. On the other hand, publication locations are organised in a subject-specific manner through different publishers, journals and OA repositories. OA is well established in life science and climate research and partly established in ICT and agricultural research, but to a lesser extent in social science and humanities.

4.3 Data management

In describing the data lifecycle and how data management is organised among research fields, we first describe the data lifecycle or parts of the cycle worth considering for comparisons. Then we compare these different subject-specific ways of managing data. It should be noted that research data management is but one part of the research lifecycle workflow and does not cover the complete lifecycle.

CGIAR research is data intensive, just as agricultural research is generally characterised. Therefore, the CGIAR chapter focuses on the openness of data sources and not on data management practices in general. Common steps are (i) data collection, (ii) cleaning, (iii) storage, (iv) use and (v) release. Data collection includes, for example, researchers installing portable laboratories in undeveloped landscapes to study agricultural processes. Therefore, these researchers need robust, simple and user-friendly instruments. Data down- and upload can be organised via a cell-phone Internet connection. For cleaning steps, software and manpower were used to describe the structuring process for collected data in order to decide which parts of the data have to be archived and which parts can be deleted. Afterwards, the cleaned data will be stored at a server. Storage also goes beyond the backup data in that these data will be reviewed according to formal standards for data archiving. After reviewing, the data are used and analysed for reports or publications. The data itself will be prepared for release after the publication of the research outputs. Describing metadata follows the standards in the field of agricultural research. CGIAR has several technical solutions for data management which depend on the research objectives. Dataverse is one example for a technical environment of data management and is used for water and agricultural research. Dataverse is a data repository run by Harvard University which provides metadata storage, file format conversion, collection management and customisation of display.

CITEC research is very data intensive. Between the three common steps of data management within the CITEC (data collection; processing, enrichment and analysis; and archiving), there can be additional steps, and especially the last step is not generalised. Archiving is performed by different groups in different ways. For example, there is no common server that archives everything. After archiving, the question of data exchange is important to all researchers. Currently, most researchers exchange their data by personal request and only a few data (e.g. Open Source software) are freely accessible without asking for permission. Open Source software is archived and distributed on a dedicated Open Source server and repository that manages software developments and data. This shall serve as an example for establishing data management on a broader scale. Currently, each group is managing their research data on their own based on a common internal infrastructure with local policies on group level.

CNR/NKUA describes several ways to manage data, from local storage up to Cloud or Grid storage. Different SVN and CMS solutions are used in e-Infrastructure research to manage and disseminate data. Research data are often stored locally; only software as a special kind of data are often stored and found at software sources on the web. These sources are well known within the e-Infrastructure community. Within e-Infrastructure, many kinds of data are produced, processed and archived, but there are no common standards for metadata to simplify data exchange: within one project or organisation, data exchange is well established but there are obstacles to exchanging data with the entire community or the public. This is true for nearly all kinds of data – software, again, being an exception. Reports, technical descriptions and system logs are shared with more access restrictions.

DANS is developing different data management solutions for different research fields. But there is one generic national data management system for the entire field of social science and the humanities which serves demands such as archiving data, curation and publication of data by DANS' staff. Data management is based on a research lifecycle model and supports archiving and exchange of data. In general, data management is integrating the following research steps: (i) discovery, (ii) collection, (iii) annotation and enrichment and (iv) publishing. First of all, data corpora have to be discoverable. Second, data are collected and generated with different kinds of tools. Most data are digital but sometimes digitising artefacts of archaeological excavations is time consuming. Third, annotation and enrichment of data is mostly necessary for all researchers to understand and interpret data correctly. Fourth, publishing data accompanied by literature is not well established within social science and the humanities. There are problems such as no standards for referencing data and less rewards for publishing data than literature.

EBI describes data management as different challenges for different subparts within the field of health and life science. All subjects within the field have databases that store and disseminate data to researchers and the public in general. Data publication is well established in the life sciences as long as the collected and published data do not touch personal rights. All other kinds of data are mostly archived in databases that are accessible for the scientific community. In many cases, there are standards, formats and ontologies that support data exchange.

WDCC describes data management as a major objective in climate research. Hence, there are international projects to organise data storage and dissemination to climate researchers and a broader public. Standards for data exchange and archiving are established within the field of climate research: most research facilities are expensive and therefore data are shared by big collaborative working groups distributed all over the world. The Coupled Model Intercomparison Project 3 and 5 are two large projects which are part of the current infrastructure of storage and exchange of data. Collection, quality control, annotation with metadata and publication of data is well established within the field of climate research.

All institutions and research fields analysed are managing large amounts of diverse data. There are many differences: some fields are more data driven than other fields. Looking at the technical basis of data, such as data types, formats, standards and metadata, it is obvious that data management is organised in many different ways but it can be observed that many fields use similar types of data. Building on similar data types, it could be possible to construct generic research infrastructure, for example managing image data across social science, the humanities, health, life science, and climate research. It also becomes obvious that data management becomes more restricted whenever privacy issues are involved.

Summary The comparative analysis provided descriptions of the main aspects addressed in the chapters, namely the characteristics of the research lifecycles as a whole and its constituent aspects of literature management and data management. Research lifecycles show common steps: (i) data collection, (ii) processing, (iii) enriching, (iv) archiving, and (v) re-using. However, the variance in the descriptions appears stronger than these rather abstract commonalities. Literature management shows strong commonalities in tooling but strong differences in publishing practices: data management shows a large variance in both tooling and data management practice. The step models for the different aspects provided in the chapters indicate the presence of systematic infrastructural services, but the variance of the step models indicates that each infrastructural service is built around a very specific research question or project. Corroborating the general observations in the beginning of this chapter, the comparative analysis shows that OA to literature is a growing or established practice in the subject areas studies but is not yet fully developed. OA to data is considered an important future activity.

5 Conclusions

The general observations on the writing process of all subject-specific chapters and the overall impressions as well as the comparative analysis point to one key challenge: developing research infrastructure that operates in an open mode and thereby supports the diversity of research practices. In a way, infrastructure is an opponent to diversity since infrastructure is not only an essential prerequisite but also a collection of rigid conditions or constraints: it is an inherent property and explicit objective of infrastructure to make research uniform. Openness, however, is a way to maximise the permeability of research resources (literature and data) within research infrastructure so that the collaborative, interdisciplinary and international research activities needed to tackle the next given challenge can emerge.

Measures to support infrastructure developments (e.g. funding programmes) should therefore take into account the following observations, interpreted on the basis of the subject-specific requirement descriptions throughout this volume.

i. Digital literature and data resources are an essential precondition of research. The provision of digital literature and data resources through infrastructural services are perceived as a matter of course (or implicitness) and are not questioned unless they are obviously missing. Thus, knowledge infrastructure, as the entirety of resources and processes related to digital literature and data resources used in research, is not conceived as an explicit facility but rather as an invisible capacity.

ii. OA is described as a *modus operandi* for working with digital literature and data resources, rather than as an end in itself or an ethical principle.

iii. OA to literature and OA to data refer to very different parts of the research process. While literature shows universally generic characteristics, data are much more related to subject-specific methodologies and facilities. Even though the benefits are the same for literature and data – namely cost considerations, enabling research otherwise not possible, transparency, comparability and synergy – the obstacles vary broadly and require that OA to literature and OA to data are treated separately in policy and infrastructure development.

iv. Due to the universally generic role of text-based resources in research, OA to literature can be regarded as a general prerequisite for efficient and effective as well as innovative research and should be mandated uniformly over all subject areas – even if the specific implementation of OA to literature is left at the discretion of the subject areas (e.g. through subject-specific repositories) – and should be arranged in the grant conditions. For non-subsidised research results, organisations should strive for access as open as possible.

v. OA to data has (yet) to be reflected in a fully subject-specific way in policy and research infrastructure development. The emerging practice of mandatory project-specific data management plans that address the question of OA to data could be sharpened by asking the question: “Are data open and if not, why not?” Also, OA in data management plans could be supported by providing a generic Open Data policy with subject-specific *addendi* to such a generic policy. A given subject-specific addendum to a generic Open Data policy may well be mandatory in a given subject area.

vi. The difference between OA to literature and OA to data may be transient as more and more systematic connections between literature and data are made. In many cases, the literature is the data: text-mining and text-annotation enrichment treat text as data and therefore contribute to provide a continuum of semantically connected knowledge resources on the long run. Explorations towards infrastructural linkage between literature and data (e.g. enhanced publications) should be intensified.

vii. The provision of research infrastructure services by institutions and organisations is requirement driven and depends on the research context – even within a smaller subject area – but supports collaboration among researchers from various disciplines. The development scheme in practice tends to be incremental and evolutionary and based on prototypes and working solutions rather than applying theoretical frameworks and capacious facilities.

viii. The layer cake model of research infrastructure does not reflect the complex organisation of research infrastructure. The distinction between horizontal developments, based on generic research processes and ICT standards, and vertical developments, based on subject-specific research questions, is helpful since it breaks up the layer cake model and suggests a hierarchical matrix model. However, a network model of research infrastructure, consisting of a multitude of subject-specific nodes that apply common local design principles (e.g. metadata standards, exchange protocols) in order to communicate with one another and share resources amongst other nodes, reflects best this study's descriptions of research infrastructure and is assumed to be the most promising approach for designing future research infrastructure developments.

Future research infrastructure developments should consider the following principles in order to reflect the diversity of research as the key challenge:

i. Support subject-specific developments that are research driven, incremental and evolutionary in order to match and adapt to the established situated practices.

ii. In a separate strand, support the development of generic infrastructural services and standards applicable in local subject-specific nodes. Services and standards should obviously be maintained by institutions and organisations with long-term commitment.

iii. Provide systematic cross-talk between the subject-specific and generic developments by:

    a. providing research and development programmes that explicitly address the question of how to link subject-specific and generic developments. Examples for activities are science and technology studies, networking events and focused infrastructure projects,

    b. installing advisory boards or oversight groups for projects and funding programmes that have representations of both subject specialists and infrastructure specialists, and

    c. enforcing the mutual participation of subject specialists and infrastructure specialists in assessments and reviews.

iv. Apply OA as a *modus operandi* in all activities. It should be mandatory for literature and is recommended for data. Appropriate exceptions for specific subjects can be considered.