C Information and Communication Technology

Dennis Spohr and Philipp Cimiano

This chapter describes the case study carried out at the Centre of Excellence Cognitive Interaction Technology (CITEC) at Universität Bielefeld. The aim is to provide a representative example of a research institution in the wider field of information and communications technology (ICT), with a specific focus on cognitive interaction and robotics engineering. After a brief introduction to the general structure and mission of CITEC, we will discuss the general scope of the case study, as well as how the methods presented in the introduction to this book have been applied in detail.

1 History, structure and mission

CITEC is a research institution founded at Universität Bielefeld as part of the Excellence Initiative of the German federal government and the state governments in 2007. According to its statutes,[1] CITEC is a competence centre for fundamental research and technology transfer and cultivates an international network of cooperation with industrial and scientific institutions. This includes industrial partners like Miele & Cie KG and Honda Research Institute Europe GmbH, as well as members from internal and external institutions, such as the Research Institute for Cognition and Robotics (CoR-Lab) and the Centre for Interdisciplinary Research (ZIF) at Universität Bielefeld. The network of external researchers is integrated into the so-called *Virtual Faculty*, which includes renowned experts in research fields related to cognitive interaction technology from all over the globe. Finally, CITEC maintains an international and multidisciplinary graduate school offering scholarships for PhD students.

[1] http://www.cit-ec.de/sites/www.cit-ec.de/files/CITEC_Satzung_0.pdf.

CITEC consists of 37 research groups[2] – 11 of which newly funded with financial support from the excellence cluster, 24 that were part of different departments of the university before the foundation of CITEC and two senior professorships – comprising overall more than 250 scientific staff. The groups come from a variety of scientific backgrounds, such as neurobiology, linguistics or computer science. What all of these groups share as part of their mission within CITEC, however, is the common goal to obtain a better understanding of cognitive interaction, as well as its implementation in technical systems. Within CITEC, they are organised in four major research areas, namely:

1 **Motion intelligence:** This area investigates how perception and action can be combined in a way that allows robots to operate autonomously in unpredictable environments and situations. This is approached by investigating animals and humans performing different cognitive tasks, from various perspectives (such as biological, psychological or physical), in order to arrive at a comparable level of sensorimotor capabilities in robotic systems.
2 **Attentive systems:** The primary aim of this area is to combine experimental and empirical methods as well as engineering approaches in order to identify the mechanisms that enable artificial systems to understand and actively focus on what is important and align their processing resources with their human partner accordingly.
3 **Situated communication:** This research area focuses on how language, perception and action can be coordinated in a way that enables efficient cooperation between humans and technical systems. As such, this involves research of linguistic and psychological phenomena in communication, as well as the computational aspects of their implementation in artificial systems.
4 **Memory and learning:** The focus of this area is to find technical solutions to cognitive issues like memorising in order to arrive at architectures which enable an artificial system to acquire, store and retrieve knowledge, as well as to improve its capabilities by learning. This is achieved by combining experimental research into biological brains with the development of new algorithms for learning and memorising knowledge.

As can be derived from the above description, each of these research areas combines aspects from engineering with aspects of other scientific disciplines, such as life sciences and computer sciences, indicating a high degree of in-

[2] Being a very young institution, the number of research groups at CITEC is still constantly changing. At the time this study was carried out, it consisted of only 32 research groups.

terdisciplinary cooperation. In fact, the CITEC website[3] states the overall vision and goals as follows:

The vision of the CITEC scientists are interactive tools that can be operated easily and intuitively, ranging from everyday objects to fully-blown humanoid robots. The future technology should adapt itself to its human users instead of forcing us humans to adjust to the often cumbersome operation of the current equipment. Just as every human being automatically adapts his speech and actions to the addressee in order to be understood, technological systems should adjust their behaviour to their interaction partner. In order to interact naturally with humans and to flexibly adapt to changing conditions, a system needs to be endowed with the corresponding cognitive abilities. Consequently, the study of the fundamental architectural principles of cognitive interaction – be it between humans or human–machine interaction – is the necessary pioneering work. It is supplemented by new possibilities of technological application, which need to be designed, constructed, and tested. We believe that this dual goal of combining basic research with technological application in order to significantly advance our understanding of cognition itself can only be realized through ***intense interdisciplinary cooperations****.*

As the previous paragraphs have shown, CITEC is a very young research institution that is heterogeneous along several dimensions. On the one hand, it comprises newly created research groups as well as groups that existed at Universität Bielefeld long before CITEC was founded – some of which that had not been in cooperations before. On the other hand, cognitive interaction technology is in itself a very heterogeneous field of research, requiring a high degree of interaction between researchers from various disciplines. Moreover, many research questions require empirical and experimental insights into biological and behavioural aspects of cognition in animals and humans (e.g. when interacting with artificial devices) as well as investigations from engineering and computational perspectives. Finally, CITEC is very internationally networked, including academic as well as industrial partners all over the globe (such as Finland, Ireland, Japan, Spain and the USA).

This heterogeneity poses high demands both on the infrastructure that needs to be in place in order to support such interconnectivity, as well as on the specification of common policies for the management and exchange of both data and literature, which may differ considerably between disciplines, due to very different traditions and historical backgrounds, as well as between academic and industrial partners, in particular with respect to legal issues. These aspects will be discussed later in this chapter, after an introduction to the methodology and the case narratives.

[3] http://www.cit-ec.de.

2 Methodology

In this section, we specify how the methods explained in the introduction to this book have been applied in this case study, as well as the scope of each method in terms of the number of research groups which participated. As was further mentioned in the introduction to this book, this case study had been preceded by a preliminary study on a small subset of research groups, covering, however, all of the research branches introduced above (see 1 History, structure and mission). This study put different methods to the test and thus helped to estimate the qualitative and quantitative impact of each of the methods applied. In addition to this, the results of each method were then integrated into the actual case study itself, thereby obtaining a detailed and representative picture of research infrastructure, literature and data management at CITEC. The methods that have been applied in this case study are described below.

2.1 Introductory interview

Interviews were held with the leader(s) of a research group in order to get a general understanding of the research topics dealt with in a group and to determine the applicability of the other methods. These interviews, which lasted between 10 and 40 minutes each, were first recorded on audio and later protocoled and analysed.

2.2 Observation

Observations of experiments were carried out as part of the typical research agenda of a group (i.e. the experiments were scheduled independently of the observation). An experiment was observed only if it was ensured that the observation would not interfere with the workflow of the experiment. Each observation lasted between 45 and 60 minutes and was recorded on audio and video and later protocoled and analysed.

2.3 Questionnaire

Based on the introductory interview and observation, a questionnaire was developed in cooperation with the Universitätsbibliothek Bielefeld, containing questions believed to cover the most important aspects of research infrastructure. This questionnaire was used to guide the semi-structured interview and was circulated among all research groups which had not participated in a semi-structured interview.

2.4 Semi-structured interview

The questionnaire was used to guide semi-structured interviews with a selection of groups and lasted between 30 and 60 minutes. As with the introductory interview, the semi-structured interview was audio recorded, protocoled and analysed.

2.5 Website analysis of publication behaviour

The above methods were complemented by an empirical analysis of a selection of group websites in order to investigate publication behaviour. If such websites were available, the publication section of a group's website was inspected by applying the heuristics in Figure C.1 to each of the publications listed there.

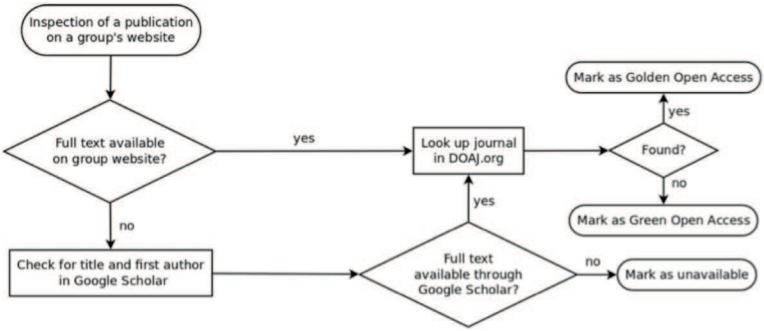

Figure C.1 Flowchart of the literature analysis process

The analysis was limited to the first 100 publications listed on the website, starting with the latest ones. In addition, some pre-processing was applied before consulting Google Scholar (e.g. in case the bibliographic information on the website contained typographical errors). More often than not, several variations of titles were queried (e.g. subphrases enclosed within double quotes), as Google Scholar frequently returned inaccurate results for titles containing hyphens. Moreover, because groups frequently publish articles in cooperation, the chance of counting a particular publication more than once is quite high. In order to exclude such cases, at least to some extent, the analysis was carried out without emptying the cache memory of the web browser. In other words, when analysing a particular publication, it was immediately visible to the investigator whether the publication had been accessed at a previous point in time. If this was the case, it was not counted a second time. Finally, it should be noted that this analysis did not consider the university-

wide publication repository PUB available at Universität Bielefeld because it was still under development at the time the analysis had been carried out and was still in a transition state at the time of writing.

This analysis included only groups that have a group website on which they provide bibliographic information about their publications, since an exhaustive investigation – which would have required identifying the researchers of a particular group and then navigating to their (potentially external) homepages in order to apply the aforementioned process – did not seem feasible. Therefore, in order to cover such cases as well, the questionnaire contained a series of questions dealing with literature management and publication behaviour, in particular with respect to Open Access. Moreover, as will be discussed below (see 3.4 Robotics and engineering (RobEng) and 6.2 Results of empirical website analysis), a representative collection of the aforementioned branches is covered by this analysis.

3 Case narratives

In order to be able to interpret the results of this case study beyond the level of individual groups, some form of classification is needed. The research areas of CITEC presented above, however, are not suitable for such a classification, as they do not partition the set of research groups into disjoint sets. As a result, it would not be feasible to attribute the findings within a particular group to one particular research area. In addition to this, CITEC is a highly interdisciplinary research centre and even the boundaries between groups are not always clear-cut. For instance, some researchers are affiliated with more than one research group. Moreover, since almost all groups have a cognitive and a computational component, a strict classification in terms of "traditional" scientific disciplines can at best be approximated.

In order to arrive at a meaningful classification nevertheless, we present below four case narratives that try to approximate a classification of groups in terms of traditional scientific disciplines. Each case narrative gives a brief description of the groups associated with the respective disciplines, as well as the common practices observed with respect to research data management, literature management and research workflows. However, we restrict the discussion to those groups which participated in the study.

3.1 Behavioural sciences, natural sciences and neuroscience (BehNatNeur)

Research group profiles Members of this branch of research either belong to one of the traditional natural sciences (e.g. biology or physics) or are

characterised by a high degree of experimental work involving living beings like humans or animals in order to investigate neural or psychological aspects of cognition. In Table C.1 and the following, we give brief descriptions of each of the groups which we classify as belonging to this research branch, along with a brief description of their approximate size, main research objectives and primary research instruments.

Table C.1 Flowchart of the literature analysis process

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Co-operations internal/ external
Active Sensing	Sensory capacities of electrical fish, neural mechanisms of parallel processing, as well as hydrodynamics	4	7/1		1/3
Biological Cybernetics	Sensory control of behaviour, esp. motion sequences	15	15/1 (and 10 set-up PCs)	Three motion capture set-ups (one Vicon MX10 with eight cameras and two custom-made ones with three Basler-A602 cameras each, objectives and ringflash system)	3/3
Neurobiology	Visual information processing in the brains of flying insects	19		Flymax	
Neurocognition & Action Biomechanics	Human perception and sensomotorics, cognitive representation and motion intelligence in humans and robots, cognitive biomechanics and augmented reality, neuromotion and neurosimulation	25	35/0	Virtual and augmented reality, motion tracking, electroencephalography, electromyography	5/15

Neuro-cognitive Psychol-ogy	Attentional control of visual perception and of sensori-motor actions	9		Motion capture set-ups, eye-tracking	
Physio-logical Psychol-ogy	Memory and memory deficits in humans	11	16/0		3/5

Empty table cells indicate that no information has been provided.

Research data Table C.2 shows the different types of data created by a typical series of experiments on a specific research topic in the behavioural sciences, natural sciences and neuroscience. As can be seen, video data especially arise in large quantities and sizes (more than 1000 files and in the terabyte range), followed by "other types of data". Here, groups indicated mainly Vicon[4] data files, electrophysiological measurement files, electroencephalographic data and spectra, as well as other binary data.

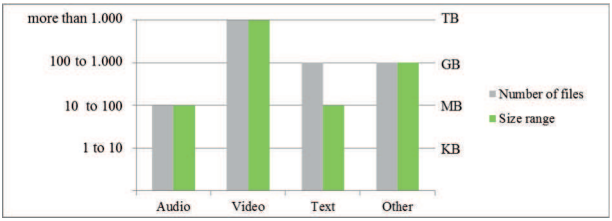

Figure C.2 Data types in terms of number of files and sizes in BehNatNeur

Three out of four groups further indicated that there are established standard formats in their field and that they use them either very frequently or with rare exceptions. With respect to metadata, two groups indicated that they annotate their data with metadata, with the other two stating that they do not.

Research data lifecycle Groups indicated the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1 **Data collection**
All groups stated that they begin with the collection of data in experiments, with two groups indicating that the data collection process may take up to several months.

[4] Vicon is a widely used motion capture system; see http://www.vicon.com.

2 **Archiving**
Only one group mentioned an archiving step of primary data on a file server accessible to several institutes (Bioserver).

3 **Processing**
All groups indicated that they process the primary data, either by means of manual analysis in statistics programs like Microsoft Excel or SPSS or computationally using Matlab, for example, or performing video and cluster analyses.

4 **Archiving**
Three groups indicated a backup step involving DVDs, external hard drives, file servers or individual PCs.

5 **Re-use/enrichment**
Two groups indicated that the data are re-used at a later stage, for example in the context of new studies, or processed and analysed further.

6 **Archiving**
One group indicated a final archiving step of the analysed data, again on the Bioserver and individual PCs.

Research data and Open Access As was mentioned at the beginning of this section, research in behavioural sciences, natural sciences and neuroscience involves a considerable amount of data gathered in experiments with humans. As a result, primary data are expected to be problematic with respect to the application of Open Access principles. This is supported by the results of the questionnaire, which shows that only secondary data could conceivably be shared with the public, albeit to a limited extent (Table C.3). For primary data, one group indicated that they would not even share data with close colleagues.

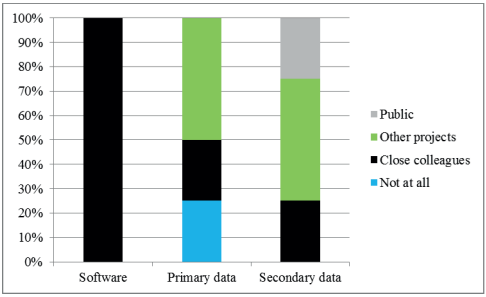

Figure C.3 Willingness to share software and primary and secondary data in BehNatNeur

Research literature Research groups in behavioural sciences, natural sciences and neuroscience at CITEC primarily use Endnote, Mendeley and Reference Manager as well as the university-wide publication repository PUB[5] for managing their scientific literature. With respect to publication preferences, groups stated that both print and electronic publication media are preferred and established in the field.

Literature and Open Access Three out of four groups indicated that Open Access is hardly established in their field, and one group even indicated that Open Access is not established at all. This is supported by the empirical website analysis, which revealed that, of the 231 publications analysed, only 16 (6.93%) were Golden Open Access publications and 79 (34.20%) were Green Open Access publications. The remaining 136 publications were unavailable following the strategy explained above (see 2.5 Website analysis of publication behaviour).

Table C.2 Research groups in behavioural sciences, natural sciences and neuroscience

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Cooperations internal/ external
Applied Computer Linguistics	Dialogue systems, dialogue, conversation, language interaction	3	3/1	Motion tracking lab, audio recording studio	1/10
Clinical Linguistics	Language, cognition, interaction (basic functions and dysfunctions)	10	15/0		3/5
Emergentist Semantics	Interaction of children between 3 months and 6 years of age, mothers and fathers, human–machine interaction	12	20/0	Four high-definition cameras, several camcorders	2/1

[5] http://pub.uni-bielefeld.de.

Gender and Emotion in Cognitive Interaction	Emotions and gender stereotypes and their role in human–machine interaction	8	10/1		3/7
Language and Cognition	Language understanding, influence of visual context on language understanding	10	15/1	Eye-tracking, video cameras, reaction time PCs	2/5
Phonetics and Phonology	Prosody, human–machine communication, speech synthesis	6	10/1	Audio recording studio, electroencephalography, video cameras	3/6
Psycholinguistics	Interaction, gesture, priming	8	10/1	Two video cameras, audio recording equipment	1/2

Empty table cells indicate that no information has been provided.

Linking research literature and data All groups indicated that it is currently possible to publish literature and data together.

3.2 Social sciences and humanities (SocHum)

Research group profiles Members of this branch (Table C.2) either belong to social sciences or humanities in the traditional sense, such as social anthropology and language studies or they focus on the immediate application of such studies in a computational context. As with BehNeurNat, members of this research branch carry out experiments involving humans.

Research data Figure C.4 shows the different types of data arising in the course of investigating a typical research topic in the social sciences and humanities. Similar to the findings in BehNatNeur, video data constitute the largest part of the data (between 100 and 1000 files and in the terabyte range), followed by audio and text data in the gigabyte range. No other data types were specified in the questionnaire, although the above list of research instruments suggest that at least eye-tracking data and electroencephalographic data arise.

Five out of six groups mentioned that they annotate their data with metadata, although only two groups stated that they are aware of existing standard formats in their field.

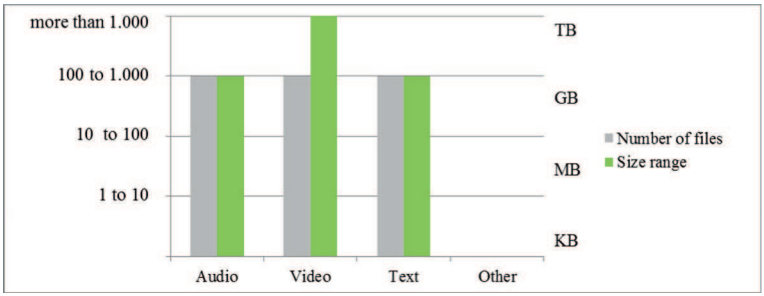

Figure C.4 Data types in terms of number of files and sizes in SocHum

Research data lifecycle Groups identified the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1 **Data collection**
As was mentioned above, groups collect primarily video and audio data in experiments involving humans.
2 **Archiving**
Two groups indicated an intermediate archiving step.
3 **Processing**
The collected data are, in some cases, analysed statistically, whereas some groups mentioned post-processing steps like cutting and compressing.
4 **Archiving**
At least one archiving step is involved in all data lifecycles described.
5 **Enrichment**
Three groups mentioned time-consuming manual annotation and transcription steps, as well as semi-automatic annotation using tools such as ELAN and Praat.
6 **Re-use**
Two groups stated that they re-use the data, for example to generate natural stimuli on the basis of their transcribed and annotated data.

Research data and Open Access Similar to BehNatNeur, research in social sciences and humanities involves experiments with humans. In contrast to BehNatNeur, however, groups in SocHum seem to be more willing to share their data (Table C.5). In general, the majority of groups are willing to share software and primary and secondary data beyond the level of close colleagues. However, only secondary data could, in principle, be made publicly available.

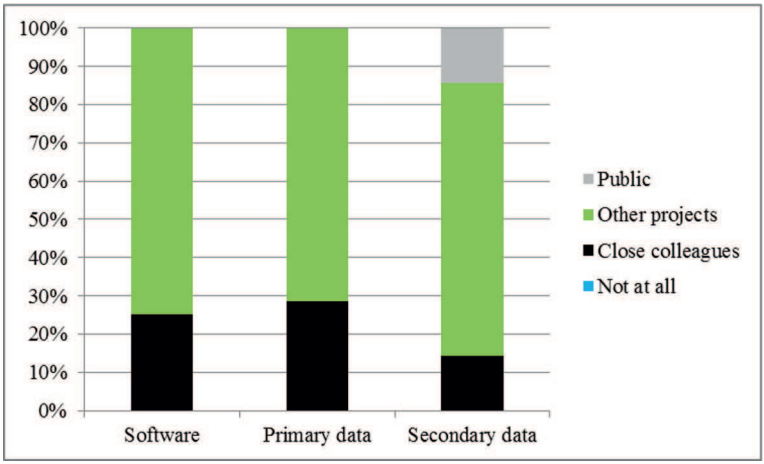

Figure C.5 Willingness to share software and primary and secondary data in SocHum

Research literature Research groups in the social sciences and humanities at CITEC primarily use BibTeX to manage their publication metadata and Citavi, Zotero and Mendeley for managing their scientific literature. They typically create their own publications collaboratively using Google Docs and Subversion. As with BehNatNeur, most groups stated that both print and electronic publication media are generally preferred and established in the field, although one group indicated that only the print medium is established.

Literature and Open Access Two out of seven groups stated that Open Access is not established in their field and four indicated that it is hardly established. Again, this is supported by the empirical website analysis, which showed that of the 38[6] publications analysed, only one (2.63%) was a Golden Open Access publication and 22 (57.90%) were Green Open Access publications. The remaining 15 publications were unavailable.

Linking research literature and data Only two groups mentioned that they are aware of solutions for publishing literature and data together. However, all groups but one agreed that it would be a reasonable and desirable development.

[6] The number of analysed publications of SocHum groups is so low because – as of February 2011 – most of these groups either do not have a group website or do not list their publications.

3.3 Theoretical and applied computer science (CompSci)

Research group profiles Members of this branch (Table C.3) deal primarily with the development of algorithms and computational models and have software as primary output of their research. In contrast to the areas discussed above, experimental studies generally do not involve humans or animals.

Research data Table C.6 shows the different types of data arising in a typical study in theoretical and applied computer science. In contrast to the findings for the previous research areas, text data make up the largest part of the data (more than 1000 files and in the terabyte range), although it can generally be said that all types of data arise in large amounts in this research area. Other types of data indicated in the questionnaires and interviews comprise dialogue data and spectra.

One out of three groups indicated that they annotate their data with metadata, with three groups stating that they are unaware of established standard formats.

Table C.3 Data types in terms of number of files and sizes in BehNatNeur

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Cooperations internal/ external
Computer Graphics and Geometry Processing	Acquisition, modelling, optimisation and animation of virtual 3D objects or characters	10	20/1	3D scanners, motion tracking systems	3/6
Genome Informatics	Theoretical and algorithmic bioinformatics with applications in genome research	15	1/0	Use of the computer infrastructure at the partner institution CeBiTec	0/15
Semantic Computing	Knowledge representation and management, Semantic Web, information retrieval	8	12/2		5/10

Theo-retical Comput-er Science	Neural computation and methods of computational intelligence, esp. prototype-based learning approaches as well as learning theory and self-organisation	5	15/0		1/10

Empty table cells indicate that no information has been provided.

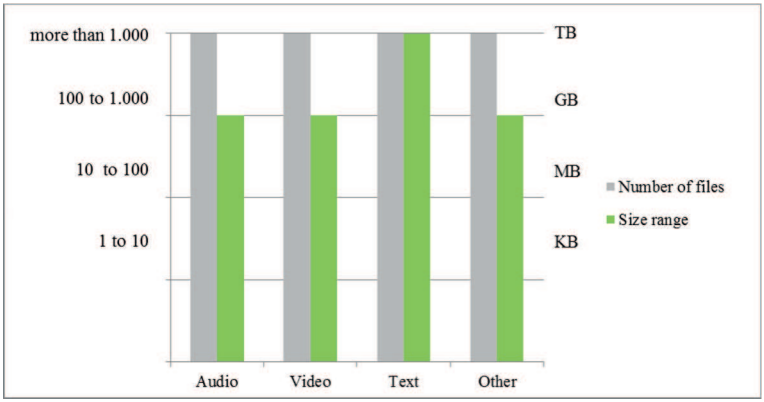

Figure C.6 Data types in terms of number of files and sizes in CompSci

Research data lifecycle Groups identified the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1 **Data collection/re-use**
As was mentioned above, in contrast to the previously discussed research areas, the data collection step does typically not involve experimental work. In fact, groups tend to start by re-using existing data, such as those available on the World Wide Web. However, these data are typically not research data as such (i.e. this step should not be considered as re-using research data), but rather data from social media like Twitter or Flickr.

2 **Enrichment**
One group indicated that they first annotate the initial data set with further information before they start processing the data algorithmically.

3 **Processing**
In CompSci, this step typically marks the central stage in the research workflow, as it is concerned with the application of the algorithms developed by the researchers.

4 **Archiving**
As with the previous research areas, at least one archiving step is involved in the data lifecycles observed.

5 **Re-use**
One group indicated that the archived data are typically re-used at a later stage for testing and comparing the performance of newly developed algorithms.

Research data and Open Access In contrast to BehNatNeur and SocHum, groups in CompSci are generally more open when it comes to sharing data. In particular, all data types could conceivably be shared beyond the level of close colleagues, with public availability being accepted by the majority of groups both with respect to software and secondary data (Table C.7). This shows that Open Access (or Open Source in terms of software) is a well-established practice in CompSci.

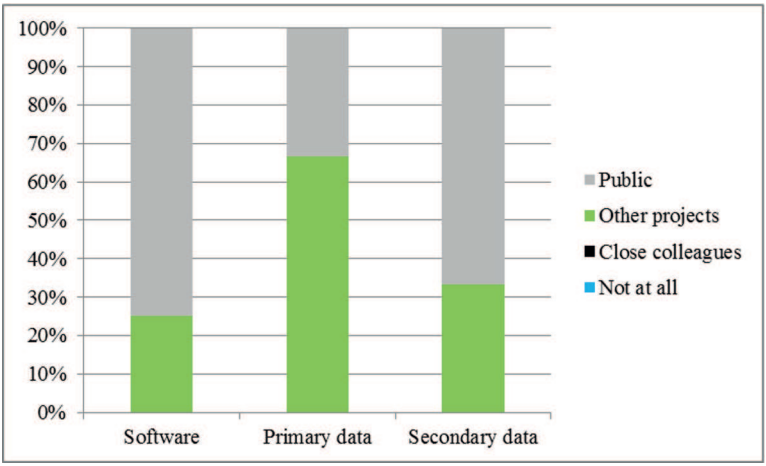

Figure C.7 Willingness to share software and primary and secondary data in CompSci

Research literature Research groups in theoretical and applied computer science at CITEC primarily use BibTeX to manage publication metadata, as

well as the Drupal content management system for managing metadata and the publications themselves. Besides this, Mendeley and the university's PUB system were mentioned, as well as Subversion for collaboratively creating literature. In terms of publication media, electronic publications seem to be preferred over publications in printed media.

Literature and Open Access The openness with respect to the willingness to share research data is also reflected in the status of Open Access to literature, as only one out of five groups considers Open Access not to be established in the field. However, it needs to be said that the actual status of Open Access to literature as suggested by the empirical website analysis is still very similar to the BehNatNeur and SocHum. In fact, only one out of 100 publications was a Golden Open Access publication. Green Open Access publications, however, made up the vast majority of the publications, namely 78%. The remaining 21 publications were unavailable.

Linking research literature and data Two out of three groups indicated that they are unaware of possibilities for publishing literature and data together, agreeing, however, that it would be desirable.

3.4 Robotics and engineering (RobEng)

Research group profiles Members of this branch (Table C.4), if not concerned with the actual engineering of machines, typically have software and models as one of their primary research outputs as well. In contrast to the previous branch, however, work in robotics and engineering at CITEC typically involves experimental studies with humans and robots, and the software and models are generally directly applied to robots or other engineered systems.[7]

[7] It should be noted that the research focus of the Applied Informatics group has changed towards robotics in recent years, and a considerable amount of research deals with experimental studies involving, for example, the interaction between humans and robots. Therefore, although being traditionally rooted in the computer science field, it has been classified as belonging to the robotics and engineering branch.

Table C.4 Willingness to share software and primary and secondary data in BehNatNeur

Name	Research topics	Members	Computers PCs/ servers	Further instruments	Cooperations internal/ external
Applied Informatics	Pattern recognition, computer vision, software engineering, evaluation of cognitive systems, human-inspired memory, social robotics	40	50/2	Robotic platforms, 3D cameras, headmounted displays)	10/8
Cognitive Robotics and Learning	Neural learning methods, esp. recurring reservoir networks, transfer of other machine learning approaches to interactive scenarios	8	8/2	Several robot platforms	2/2
Cognitive Systems Engineering	Motion generation using dynamic systems, software and systems engineering for cognitive robotics, architecture of intelligent systems	10	10/1	iCub (humanoid robot), several special hardware platforms, usage of the general CoR-Lab infrastructure	3/1
Cognitronics and Sensor Systems	Cognitronics, microelectronics, CPU design, sensor systems	20	40/10	High-performance measuring instruments with network connection, research platform "tele-workbench" with network cameras and video and data servers, special hardware platforms for rapid prototyping of microelectronic switches based on FPGAs	3/15

Neuroinformatics	Data mining, brain–computer interfaces, evolutionary computation, complex systems integration	32	60/1	Brain–computer interface system, eye-tracking devices, three cybergloves, depth cameras, in-house developed tactile sensors, robot set-up (2 PA10 arms, two shadow hands), robot set-up (two Kuka lightweight arms, Schunkhand), manual intelligence lab (14 Vicon cameras)	10/10
Sociable Agents	Development of systems for intuitive and natural human–machine interaction	9	20	3D camera, two time-of-flight cameras, eye-tracking, two cybergloves, headmounted display, two 60" displays	

Empty table cells indicate that no information has been provided.

Research data As with the previous disciplines, data arise in large quantities also in the robotics and engineering groups. Figure C.8 shows that video and other data types constitute the largest portions of the research data (both in the terabyte range), where eye-tracking, motion capturing, tactile, simulation and design data, as well as binary data, were named among "other data types".

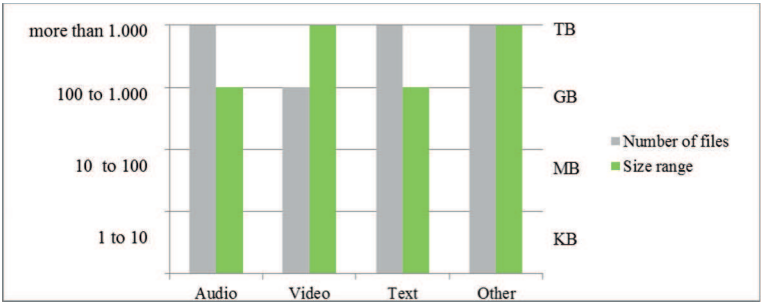

Figure C.8 Data types in terms of number of files and sizes RobEng

As with the CompSci groups, half of the groups stated that they annotate their data with metadata and three of the four groups stated that they are unaware of established standards in the field.

Research data lifecycle Groups indicated the following stages in the data lifecycle, with bold stages being those shared by at least half of the groups.

1 **Data collection**
All groups begin by collecting data in experiments, typically involving the interaction between humans and machines, humans performing cognitive tasks or autonomous robots performing tasks.
2 **Processing**
One group indicated a processing step consisting of post-processing the data recorded by the Vicon system, as well as compressing the recorded videos.
3 **Enrichment**
Two groups indicated that they annotate their data, using a tool like Anvil, for example.
4 **Processing/analysis**
All groups analyse their data, for example by applying different machine learning algorithms to them.
5 **Archiving**
As with the previous research areas, one archiving step is part of the data lifecycle.
6 **Re-use**
Two groups stated that they re-use their data and experimental set-ups at later stages.

Research data and Open Access Similar to the observations for the CompSci groups, research in robotics and engineering is rather open with respect to the willingness to share data. As can be seen in Figure C.9, software and primary data are mostly considered to be made publicly available. One group indicated, however, that in some projects the amount of generated data is so large that it is considered to be too much for being shared.

Research literature Research groups in robotics and engineering at CITEC primarily use Drupal, Endnote, BibTeX and Subversion for handling their publications. There seems to be no preference with respect to publication media, with printed and electronic publications being well established.

Literature and Open Access Two of the three groups which answered the respective question in the questionnaire stated that Open Access is hardly

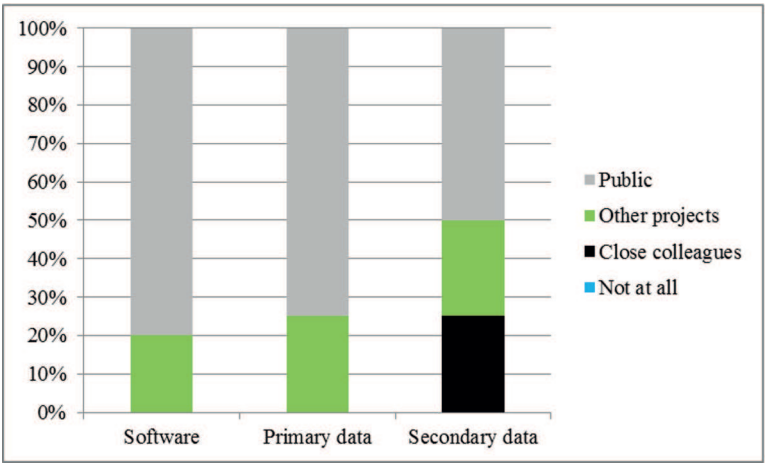

Figure C.9 Willingness to share software and primary and secondary data RobEng

established in their field. As with the other research areas, the empirical website analysis confirmed that only a tiny fraction of the analysed publications are Golden Open Access publications. In particular, five (1.17%) of the 428 publications were Golden Open Access publications and 292 (68.22%) were Green Open Access publications. The remaining 131 publications were unavailable following the strategy explained above (see 2.5 Website analysis of publication behaviour).

Linking research literature and data Two groups indicated that it is currently possible to publish literature and data together, and the two groups which were not aware of such possibilities stated that it would be a desirable development.

4 Representativeness of this case study

The methods presented above (see 2 Methodology) were applied to a number of research groups in each of the research branches just introduced. Whenever possible, it was attempted to apply each method to at least one representative of each branch, which succeeded for almost all methods.[8] However, since all branches are well covered in the questionnaire that forms the basis of the

[8] One of the reasons for not having observed experiments in CompSci is due to the fact that such are rarely carried out.

semi-structured interview, we do believe that this study gives a representative description of the entire case CITEC nonetheless. The overall participation according to methods and research branches is given in Figure C.10, and can be seen as an attempt to quantify the representativeness of this study. It should be noted, however, that representativeness refers to "representative of the institution" in this context, not to "representative of the field of ICT". We are aware of the fact that while this chapter gives a representative account of CITEC, it describes only one of many possible examples within the field of ICT.

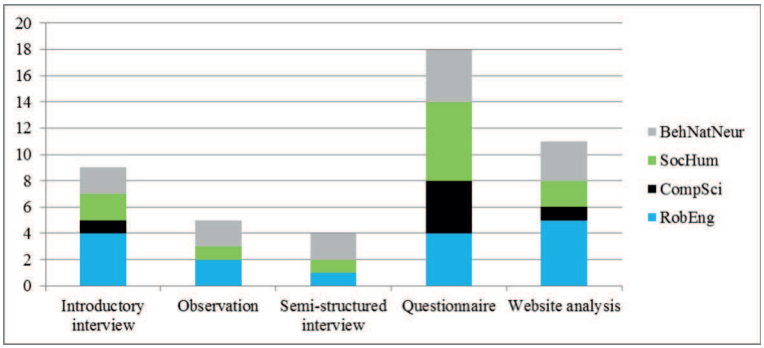

Figure C.10 Overall participation in the case study according to methods and research branches

As was mentioned above, because the semi-structured interview had been guided on the basis of the questions in the questionnaire, a group either participated in the semi-structured interview or completed a questionnaire. As such, 21 out of 32 research groups (65.63%) answered detailed questions (i.e. either participated in a semi-structured interview or completed a questionnaire) and 23 groups (71.88%) were covered by this study in some way or another. Figure C.11 summarises the participation according to research branches and overall.

5 Current status of research infrastructure

This section presents a detailed and consolidated view on the current research infrastructure at CITEC, focussing first on available infrastructural facilities and services. Afterwards, we will discuss the types and amounts of data dealt with at CITEC, as well as the various stages they pass through.

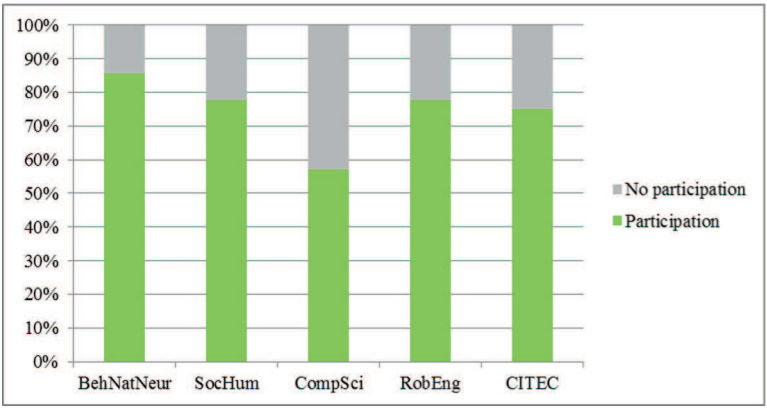

Figure C.11 Overall participation in the case study

5.1 Infrastructural facilities and services

5.1.1 Computing and network infrastructure

CITEC hosts a computational infrastructure that is maintained in large parts by the *Rechnerbetriebsgruppe* (RBG; IT services group) of the Faculty of Technology at Universität Bielefeld, which operates the computational infrastructure at the Faculty of Technology. Based on the figures given above (see 3 Case narratives), a research group has on average 18.67 desktop PCs or notebooks, the majority of which run on Unix-based operating systems like Mac OS X or Linux. This is supported by a survey carried out in a different context prior to this case study, which revealed that the use of non-Unix operating systems (such as Microsoft Windows XP) was well below 10%. Although this figure cannot be taken at face value, it nonetheless gives a hint as to the actual distribution of operating systems at CITEC, at least in technical disciplines. In the BehNatNeut groups, however, Windows seems to be the predominant operating system. In fact, the interviews as well as observation sessions have shown that groups are sometimes forced to resort to a non-Unix operating system in case they use special hardware or commercial software which depends on such.[9] These systems are, however, in most cases maintained by the researchers of the groups, have restricted network access and are thus largely independent of the infrastructure operated by the RBG. RBG further operates a *Subversion* revision control system[10] server that can

[9] The importance and use of special hardware and commercial software in the data lifecycle is discussed in more detail below (see 5.1.3 Research instruments and 5.2 Overview of the data lifecycle).

[10] http://subversion.apache.org.

be used to store group-related or project-related data. Finally, the CITEC Compute Cluster (C3) provides powerful computational support and is operated by *Central Labs Facilities* (CLF) and connected to the network operated by the RBG.

In addition to the facilities offered and maintained by the RBG, there is a wide range of further services available at CITEC. Although these are still hosted on servers provided by the RBG, the services themselves are generally offered by CLF and consist of, for example, collaborative research environments like the *CITEC Social Network*, as well as central repositories like the *CITEC OpenSource Server*. These will be introduced in the following subsections.

5.1.2 Social network and collaborative services

The *CITEC Social Network*[11] is a platform based on the OpenSource social networking engine Elgg,[12] where researchers at CITEC can create and join groups in order to exchange opinions, discuss particular research topics or upload documents. One of these groups is concerned with, for example, the CITEC Software Round Table, a regular strategic meeting aiming to discuss issues regarding the creation of a cognitive interaction toolkit, as well as exchange experiences and best practices regarding software and frameworks used. Members of this group can participate in these discussions and have access to the documents presented at sessions of the regular colloquium. In addition to this, every member of CITEC has access to an *instant messaging service*, further supporting exchange and collaboration among researchers. As of February 2011, the *CITEC Social Network* has more than 400 members and around 30 different groups.

The *CITEC Project Development Platform*[13] is a collaborative development environment which builds on the Redmine[14] project management web application and features an integrated Wiki and discussion forum. For each project, the platform provides a file space for work documents and project deliverables, as well as issue tracking and milestone management. Moreover, the platform is connected to a *project repository farm*, a revision control system that provides a dedicated Subversion repository for each project.[15] Access to a project repository requires a login and is determined on the basis of the role that the respective person has in the project. Moreover, projects in the

[11] https://social.cit-ec.uni-bielefeld.de.
[12] http://www.elgg.org.
[13] https://projects.cit-ec.uni-bielefeld.de.
[14] http://www.redmine.org/.
[15] https://projects.cit-ec.uni-bielefeld.de/svn/<project-id>.

farm are arranged by topic, which allows for collaboration beyond the level of research groups or institutes.

The *CITEC OpenSource Server*[16] provides a central repository for depositing and obtaining open-source software. While primarily aiming at storing software developed at CITEC, the *CITEC OpenSource Server* openly invites researchers from other institutions to not only use the software, but to contribute and collaborate on software development as well, with the goal of creating an open library of software related to cognitive interaction technology. Similar to the Project Development Platform, the OpenSource Server is connected to revision control system repositories. However, both the OpenSource Server and Repository Farm are publicly accessible. Figure C.12 summarises the collaborative research infrastructure available at CITEC. As can be seen there, all components access a *directory service*, an LDAP server hosting a directory with information on all CITEC members and associates, such as contact details and affiliations.

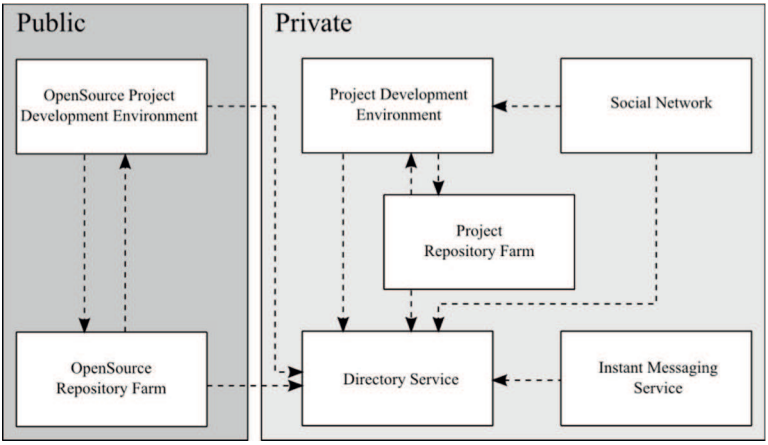

Figure C.12 Summary of the collaborative research infrastructure at CITEC[17]

5.1.3 Research instruments

As was mentioned above (see 1 History, structure and mission), research in cognitive interaction involves a considerable amount of empirical investigation of humans and animals, focusing in particular on the way they handle certain

[16] http://opensource.cit-ec.de.

[17] This image is based on a figure created by Thilo-Paul Stueve presented at the CITEC Software Round Table colloquium.

cognitive tasks. In the following, we will illustrate the variety of instruments being used in such studies by means of a concrete example. In particular, we discuss a collaborative experiment between the groups Neurocognitive Psychology and Neuroinformatics that aims at analysing learning, interaction and automatisation in speedstacking. Speedstacking consists in stacking and destacking ten plastic mugs in several predefined formations in the shortest possible amount of time.[18] At the beginning of one speedstacking exercise, a previously untrained participant is asked to rest his or her hands on a hand timer, a device measuring the time needed to complete the exercise. The exercise starts with the participant taking his or her hands off the timer. The participant then performs the task and puts his or her hands back on the hand timer as soon as the task has been completed. In particular, the participant tries to complete as many individual speedstacking task iterations as possible within 3 minutes. This is repeated for five times and makes up one complete experiment.

In the time between each experiment, the participant is asked to practice the speedstacking task for at least 45 minutes per day. In these experiments, the previously mentioned goals are investigated along different lines. For example, the focus of the eyes of the participant is recorded by means of a so-called eye-tracker, in order to examine whether the position changes over time (e.g. at a certain point in time and 1 week later). Moreover, measuring the progress that the previously untrained participant makes after a certain amount of training is taken as a learning indicator, where progress is measured primarily on the basis of the decrease in the time needed to perform the task over time. Finally, the hand movements of the participant are recorded by means of special markers attached to them, as well as special cameras tracking those markers. This is done to have three-dimensional trajectories of the movements of the hand in digital form, which means that they can be transferred to an artificial system, such as a robot, at a later time. Due to these different aspects, each experiment involves a number of different instruments serving different purposes. In addition to the ones just mentioned, for example, irregular speedstacking completions (e.g. falling mugs or any other kind of disruption) are manually annotated as containing mistakes, in order to ensure that the times measured during such iterations is not taken into account when analysing the overall learning curve and thereby ensuring a certain level of quality of the data observed within an experiment. This annotation is done in a different computer than the one used, for example, to record the eye-tracking data. Figure C.13 shows the complete set-up of instruments used in the particular experiment that was observed.

[18] https://www.cit-ec.de/research/ALIAS.

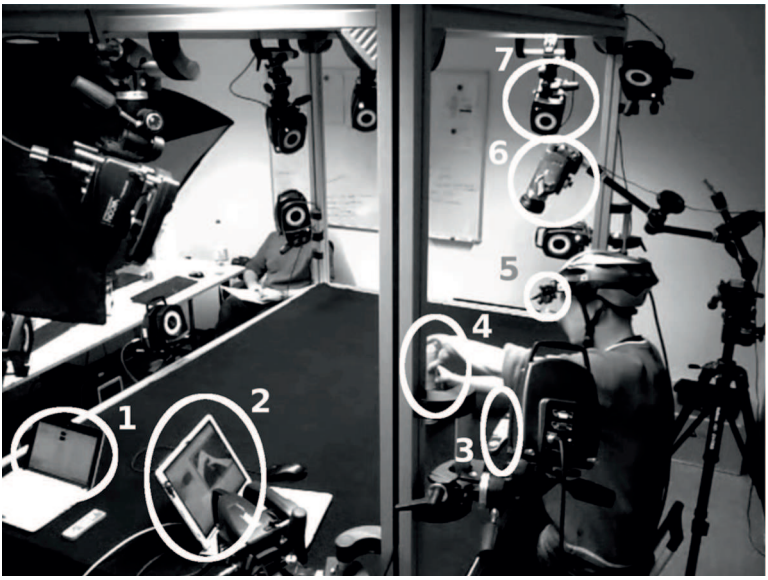

Figure C.13 Set-up of a collaborative experiment of the groups Neurocognitive Psychology and Neuroinformatics

In particular, the following instruments have been used, according with their purpose:

1 **Laptop "MacBook Pro":** for recording the amount of time needed to complete one speedstacking iteration and to manually annotate whether there has been a mistake or disruption in this iteration.
2 **Laptop "Windows XP":** for handling the data recorded by an infrared camera (no. 5) that is attached in front of the eyes of the individual, as well as of a further head camera. The recorded videos are displayed on the laptop in real-time.
3 **Hand timer:** recording the amount of time needed for stacking.
4 **Six special markers:** (three per hand) allowing a camera (no. 7) to track the movements of the participant's hands in 3D.
5 **Head camera and eye infrared camera:** for recording the view the participant has, as well as where the participant is looking.
6 **Scene camera:** to have a further perspective of the stacking scene. Here, all mugs are always in sight, which is not necessarily the case with the images recorded by the other cameras.
7 **Fourteen Vicon cameras**: recording the 3D coordinates of the markers

8 **Computers "Windows XP":** (not shown in Figure C.13) with special commercial software processing the images recorded by the Vicon cameras.

Table C.5 below summarises the other research instruments used at CITEC, as observed in other experiments or indicated by answers in the questionnaire (see 3 Case narratives).

Table C.5 Research groups in social sciences and humanities

Instrument/platform
3D scanners
Audio recording studio
Brain–computer interface system
Data gloves
Depth cameras
Electroencephalography
Electromyography
Electrophysiology (at least five set-ups including binoculars, intra- and extracellular amplifiers, analogue-digital converters, PCs, oscilloscopes, micromanipulators, stimulators, frequency generators and electrode pullers)
Eye-tracking
fMRT
High-performance measuring instruments with network connection
Hydrodynamics
iCub
Microsoft Kinect
Motion capture (at least four set-ups, e.g. Vicon MX10 with 8–13 cameras, two self-built set-ups with three Basler-A602 cameras each, objectives and ringflash system)
Research platform "Tele-workbench" with network cameras and video and data servers
Robot platforms
Special hardware platforms for rapid prototyping of microelectronic switches based on FPGAs
Tactile sensors
Video cameras
Virtual and augmented reality

5.1.4 Data management

As mentioned above (see 5.1.1 Computing and network infrastructure), CLF provides central revision control repositories for archiving project-related data. However, the overall analysis of the various interviews and questionnaires clearly suggests that there is no general data management strategy

that is followed by all groups. In addition to this, however, the interviews especially revealed that there is no general archiving strategy within a group, but that it rather depends on the particular project as well as the partners involved in a project. For example, data created in EU-wide projects are frequently stored in external repositories which are hosted by one of the participating project partners. Some CITEC-internal projects make use of the collaborative research environment operated by CLF, whereas others make use of the storage infrastructure provided by project X1 of the Collaborative Research Centre 673 "Alignment in Communication". Finally, projects involving either a rather small amount of people, such as PhD projects, for example, or group-internal projects seem to be primarily stored on group-internal – or even personal – storage devices. Figure C.14 summarises the overall data management strategy at CITEC, represented by the answers given to the question which devices the group uses for storing their data.

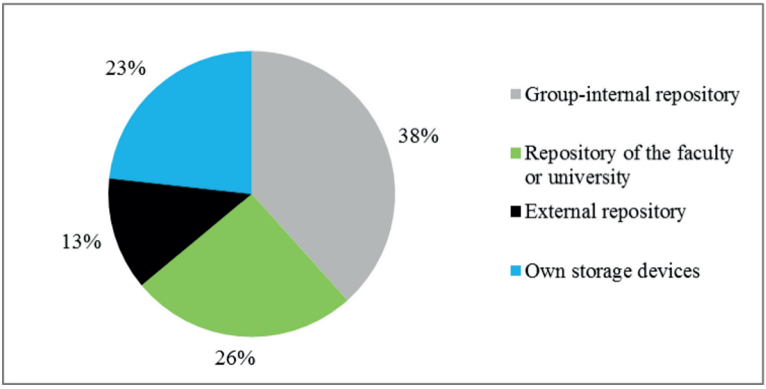

Figure C.14 Overall data management at CITEC

As can be seen in the figure, despite the availability of a central data management infrastructure, there is no homogeneous data management strategy building on this infrastructure, since only 26% of all research groups make use of it. On the one hand, this is certainly in part due to the reasons mentioned in the introduction to this chapter, namely that CITEC is a very young institution involving previously existing research groups, some of which adhere to the management procedures they had previously established. On the other hand, however, this may in part be because groups do not have designated personnel for dealing with questions of data management and may therefore be not perfectly well informed about the available infrastructure, unable to make use of the infrastructure, possibly due to a lack of sophisticated technical background knowledge, or have their own independent infrastructure.

This is supported by the answers given by groups as to whether they have personnel in charge of data management questions. The results are given in Figure C.15.

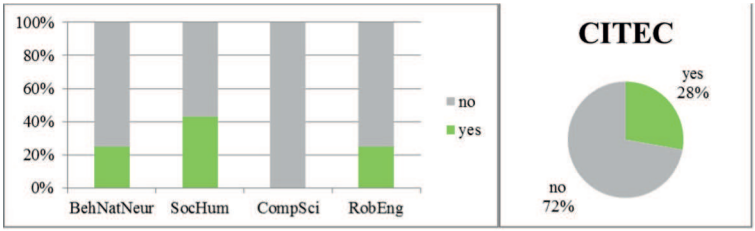

Figure C.15 Groups having a person in charge of data management

The results show that only 28% of the research groups at CITEC have a person in charge of data management. A follow-up question revealed that of the 72% which do not have such personnel, 69% would like to have such personnel. The distribution according to research branch is given in Figure C.16.

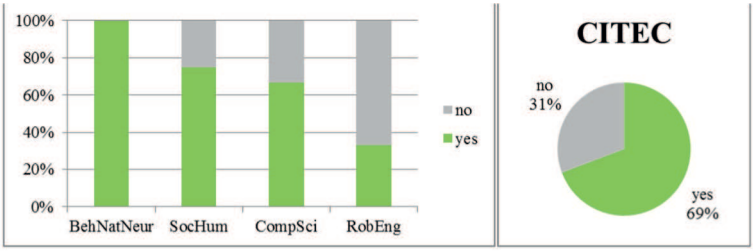

Figure C.16 Groups wishing to have a person in charge of data management

In case groups motivated their choices, the main reasons indicated in favour of having a person in charge of data management are the large amount of data being dealt with and the possibility to obtain a better general overview of the data being managed in a group. This would in turn increase the sustainability of the data, allowing for better reuse and thus comparability. Main reasons against having a person in charge of data management were that it was not a primary task area and as such not is financeable or that it works as it is – either due to an easily manageable amount of data or because researchers themselves are well grounded in data management issues. Especially the latter seems to be the case in CompSci and RobEng groups. While this is certainly true, however, it clearly explains the heterogeneous distribution

shown in Figure C.14 above, since data management tasks are transferred to the researchers themselves. Therefore, there is the strong requirement to have budget assigned to the task of managing research data. This does not mean, however, that each group needs to be given a new member that takes over this task – although in some cases this may certainly be reasonable – but rather that groups are provided with budget to be able to appoint and finance a specific member of the group to take over data management tasks, such as participating in strategic meetings of CLF in order to be informed and to agree on global data management practices at CITEC.

In addition to the resource-based aspects just discussed, the data management strategy taken strongly depends on the diversity of data types that are to be managed. For primary data obtained in experiments involving humans, for example, it may in many cases not even be permitted to store them in external repositories, because other people may have access to them. This means that a more fine-grained analysis of data management is necessary, which is given as part of the discussion of the data life cycle (see 5.2 Overview of the data lifecycle).

5.1.5 Publication management

With regard to publication management, groups were asked which tools they use for managing internal and external literature (with multiple answers allowed). Here, 26% of all groups use BibTeX to store publication metadata and one-fifth use the Drupal CMS to manage internal publications, including metadata – which can in turn be exported in BibTeX format – and, in some cases, the publications themselves. The most frequently used tools are shown in Figure C.17. It should be noted that the vast majority of tools used are freely available, with Endnote being the only exception. Moreover, the figure suggests that publication management is done on a group-internal basis. This is to say that no group answered that they use a group-external repository for depositing publications. This is obviously because no such repository was available at the time of writing. However, as will be discussed later in this chapter (see 9.1 Literature management), current developments are clearly headed into this direction.

5.2 Overview of the data lifecycle

In the following, we will discuss the data lifecycle extrapolated from the descriptions given by researchers in the questionnaire. Figure C.18 shows the general stages that can be identified (see 3 Case narratives for a description of the data lifecycle in the various disciplines).

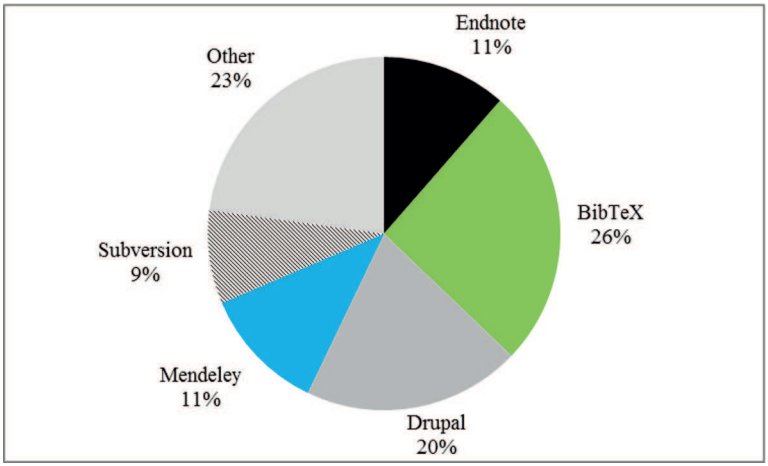

Figure C.17 Overview of publication management tools used at CITEC

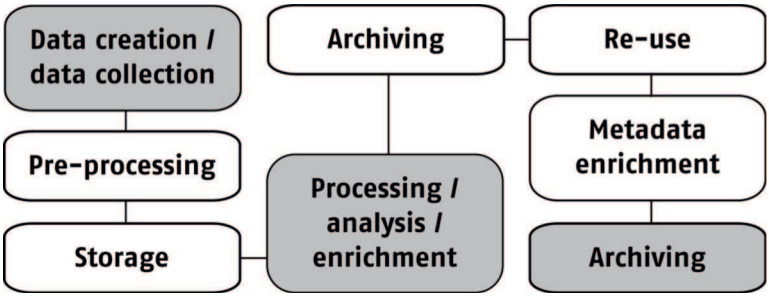

Figure C.18 Different stages in the data lifecycle

The grey boxes in the figure represent those stages which could be identified by (almost[19]) all groups participating in the survey. In the following subsections, we will discuss each stage in more detail, starting with data creation/data collection in the top left-hand corner.

5.2.1 Data creation and collection

The variety of research instruments being used in a single experimental study, as well as the other instruments that are typically used at CITEC, has been illustrated earlier in this chapter (see 5.1.3 Research instruments). These in-

[19] In case one or two groups did not list a particular stage, we still considered it as representative and marked the respective box in grey.

struments do, of course, produce very different types of data. In the following, we will discuss the different types of data that arise, as well as the scale (in terms of storage requirements) at which they are created. In addition to the primary data created this way, the following also includes cases of collecting primary data. This is, for example, the case in which the primary data of a group consist in material found on the web, such as images or texts on general websites.

Data types and scales In order to get an overall picture of the types of data being created at CITEC, we asked research groups to specify the types of data typically arising in the investigation of a particular research object, as well as a rough estimate of the quantity in terms of number of files and memory requirements. The distribution according to types of data is shown in Figure C.19 and Figure C.20.

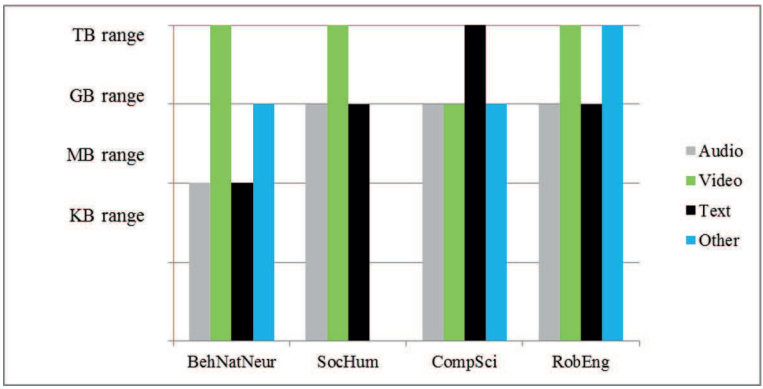

Figure C.19 Data types occuring in different disciplines (memory requirements)

The figures show that video data arise in all research areas in considerably large quantities, posing very high storage demands in almost all of them. For example, the Neurobiology group reported behavioural experiments which involve high-speed cameras recording 500 images per second, each of them with a resolution of 1 megapixel. With up to three high-speed cameras in an experimental set-up, each second recorded thus requires around 1.5 GB of disk space, and a currently running PhD project has thus created around 9 TB so far. In addition to audio, video and textual data, several groups specified other types of data occurring in considerable quantities and size, such as eye-tracking data, electroencephalograms or nuclear magnetic resonance spectra. Overall, it can thus be said that all research areas at CITEC pose high

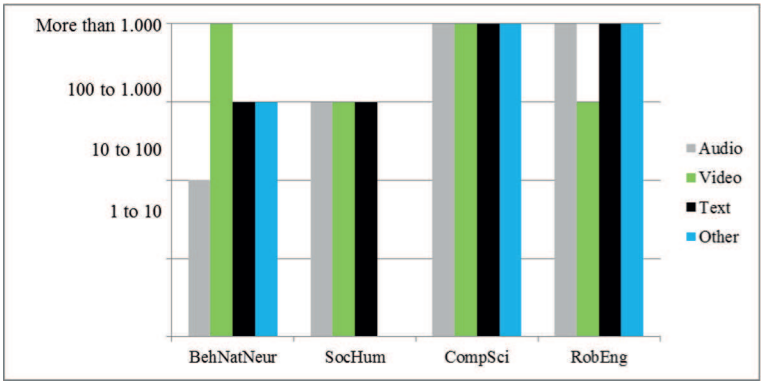

Figure C.20 Data types occuring in different disciplines (number of files)

demands on storage infrastructure. How these are managed will be discussed below (see 5.2.2 Pre-processing).

Software In addition to the different types of data arising in the different disciplines, groups were asked to specify whether they rely on commercial software in order to obtain primary data. Here, Figure C.21 clearly shows that more than half of the groups rely on commercial software at least to a considerable extent, with 11% relying (almost) entirely on commercial software. This further suggests that groups depend on the hardware required by the software in order to run successfully.

5.2.2 Pre-processing

Two groups indicated a pre-processing step taking place in the data lifecycle. For example, this was the case for the experiment described above (see 5.1.3 Research instruments). Here, the data recorded by the Vicon cameras (i.e. the trajectories of the markers which are attached to the participant's hands) are directly reviewed in the Nexus[20] tool after the experiment, before the data are transferred to other storage media. This is done in order to make sure that the cameras were able to trace the hand movements correctly. In cases where this is not the case, the data can be corrected manually in the tool. Other groups mentioned processing steps like digitisation, cutting of audio and video data or data compression. In all cases that include such a processing step, the answers indicate that raw primary data are not used in later stages of the workflow.

[20] http://www.vicon.com/products/nexus.html.

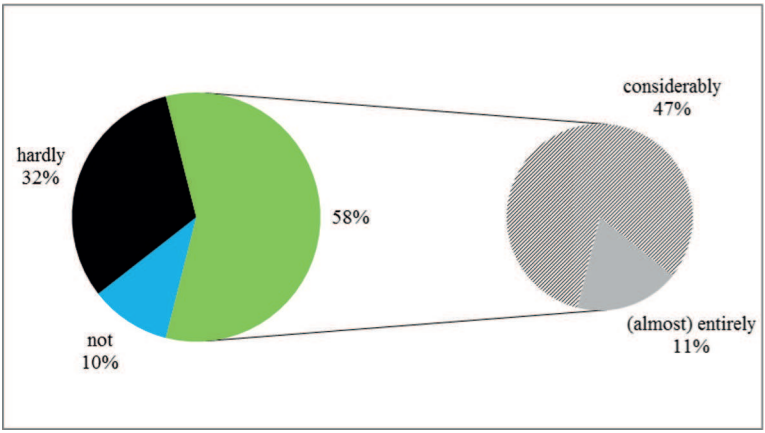

Figure C.21 Overall dependence on commercial software for generating primary data

5.2.3 Storage and transmission

After the data creation (or collection) and a possible pre-processing step, many groups indicated a first archiving step. However, since most groups did not mention this step, we do not consider it to be a crucial step in the data lifecycle. For those groups which did indicate it, this step consists in archiving either the storage media on which the data had been recorded, such as digital audio tapes, or other storage media to which these data had been transmitted, such as DVDs or external hard disks. In other cases, this step simply consists transferring data to the hard disk of the respective researcher.

5.2.4 Analysis and enrichment

All groups which create or collect primary data mentioned an analysis step, which mainly consists in analysing the primary data by means of statistical tools like SPSS or programming languages like R or MatLab. In many cases, this includes an enrichment step, in which secondary data are obtained by annotating the primary data with tools like ELAN or Praat. Since in some cases enrichment precedes analysis and in others vice-versa, we grouped these two steps together in one stage.

5.2.5 Archiving

In most groups, this intermediate archiving step does not exist. It is, however, an integral part of those groups involved in the creation of the so-called

Manual Interaction Database,[21] an effort to create a strong empirical basis for investigating research questions in the area of *Motion Intelligence* (see 1 History, structure and mission). Here, groups store the post-processed data from their experiments in SQL databases in order to allow for future re-use.

5.2.6 Re-use

Re-using, for example, algorithms, methodologies, models or annotated data either for follow-up experiments or – in the case of software components – even in other contexts is a very common step in the data lifecycle. In addition to the re-use of manual interaction data just mentioned, models for classifying data, as typically created by machine learning approaches, are frequently applied to data sets other than those on the basis of which they had been obtained, primarily in order to measure their performance on these previously unseen data sets. Similarly, physics-based models representing the physical properties of a human hand are re-used in robotic systems in order to achieve comparable behaviour in a robotic hand. Finally, annotated data are frequently used in other settings to investigate other research questions, such as linguistic phenomena in the case of text corpora. For a number of research groups, however, data re-use – aside from publishing the findings obtained on their basis – is not a common procedure, with one group indicating that data are in fact re-used too rarely.

5.2.7 Metadata enrichment

Most groups annotate their data with metadata, which is shown in Figure C.22 below.

A follow-up question asked whether existing standards are used for this annotation or whether groups use custom formats. All of the groups which annotate their data with metadata use existing standards do so either quite frequently (57%) or almost always (43%). Reasons for deviating from standard formats were the lack of metadata fields for annotating proband information. Of all groups, 59% indicated that there are no established metadata standards in their field.

5.2.8 Archiving

As was discussed above (see 5.1.4 Data management), different data management strategies are followed at CITEC. The strategy chosen depends on different factors, one of which is the type of data that is being managed. In line with general practice, we distinguish between primary and secondary

[21] http://www.cit-ec.de/research/MINDA.

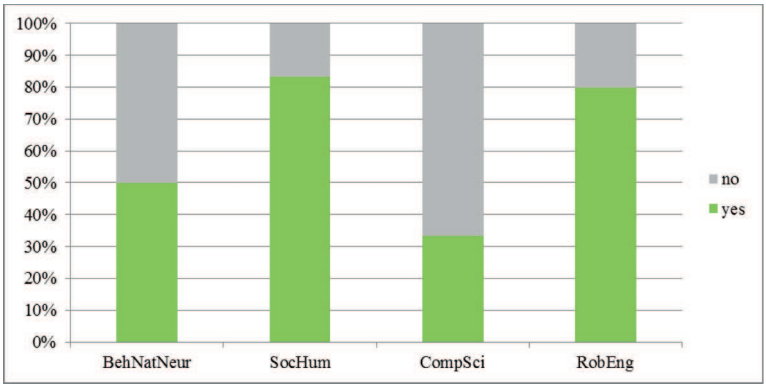

Figure C.22 Metadata enrichment according to research branches

data. Although software could be considered a special kind of secondary data, we treated it separately in this study – mainly due to its importance in a computationally oriented research field and the expected difference in how it is managed in contrast to primary or secondary data. Figure C.23 summarises the different archiving strategies according to types of data.

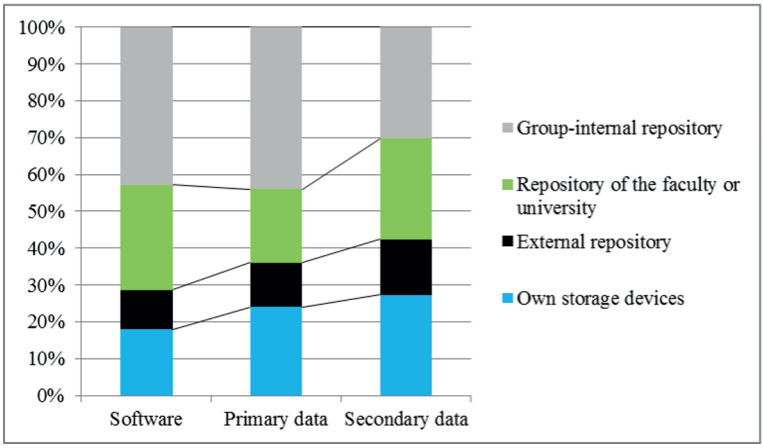

Figure C.23 Archiving strategies according to data types

As can be seen in Figure C.23, it is far more common to use repositories provided by the faculty or university for storing software and secondary data than it is for storing primary data. The figure suggests that the latter are typically stored on own storage devices or those provided internally by the

group. In general, it can be clearly seen that using external repositories for storing data is rather uncommon for all kinds of data, though slightly more common for secondary data.

As far as backup strategies are concerned, most groups mention that they generally perform regular backups with standard backup systems, on file servers and/or individual computers. In addition to this, indirect backups are in many cases achieved by using a revision control system like Subversion, since people working with the data stored there usually have a local version of the respective repository. However, it is generally the case that no complete snapshots are backed up this way, which means that this strategy cannot be considered a standard backup procedure. In contrast to this, groups which store their data on repositories of the IT services department of the Faculty of Technology at Universität Bielefeld can make use of the backup policies followed there, which includes backing up complete snapshots of the data stored in the repositories. Finally, some groups cooperating with the Collaborative Research Centre 673 at Universität Bielefeld make use of the server provided by infrastructure project X1 in order to archive their data.[22]

6 Current status of Open Access to literature

The results presented in the following sections are based on literature-related questions in the questionnaire, as well as the empirical website analysis as described earlier in this chapter (see 2.5 Website analysis of publication behaviour), which was carried out in February 2011. A discussion of the results is given below (see 6.3 Discussion of results).

6.1 Results of questionnaire

In the questionnaire, groups were asked to state whether Open Access is established in their group and field of study. The results – grouped according to research branch – are given in Figure C.24.

6.2 Results of empirical website analysis

The following figures illustrate the distribution of Golden and Green Open Access publications, in relation to all publications of a particular group (see 2.5 Website analysis of publication behaviour for classification criteria). Figure C.25 shows the results grouped according to research branch, and Fig-

[22] http://www.sfb673.org/projects/X1.

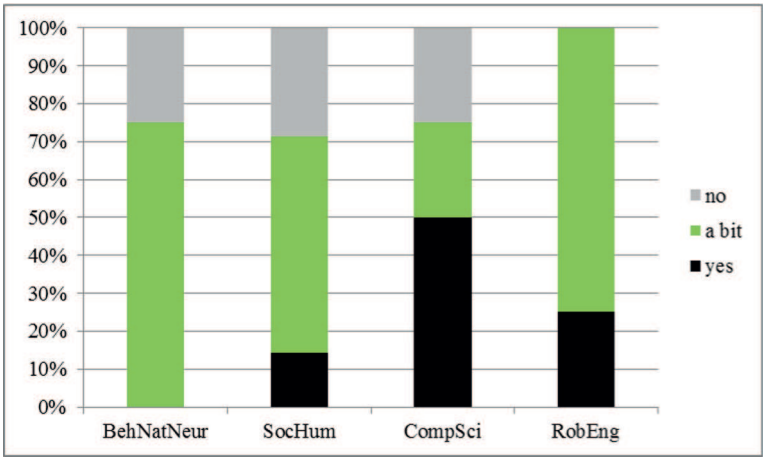

Figure C.24 Groups' replies to whether Open Access is established in their group or field of study

ure C.26 summarises the overall publication behaviour at CITEC. Absolute figures are given in Table C.6.[23]

Table C.6 Data types in terms of number of files and sizes in SocHum

Discipline	Publications analysed	Golden Open Access publications	Green Open Access publications	Unavailable publications
BehNatNeur	231	16	79	136
SocHum	38	1	22	15
CompSci	100	1	78	21
RobEng	428	5	292	131
CITEC	797	23	471	303

6.3 Discussion of results

As the results given above show, Open Access seems to be established in all research groups at least to some extent. In particular, of all groups participating in the questionnaire, 78% answered "yes" (22%) or "a bit" (56%). When looking at the results of the empirical website analysis, it becomes clear that

[23] As was mentioned above (see 3.2, the number of analysed publications of SocHum groups is so low because – as of February 2011 – most of these groups either do not have a group website or do not list their publications.

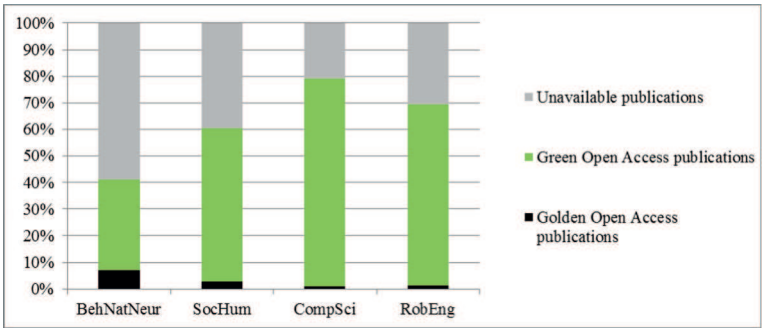

Figure C.25 Golden and Green Open Access publications according to research branch

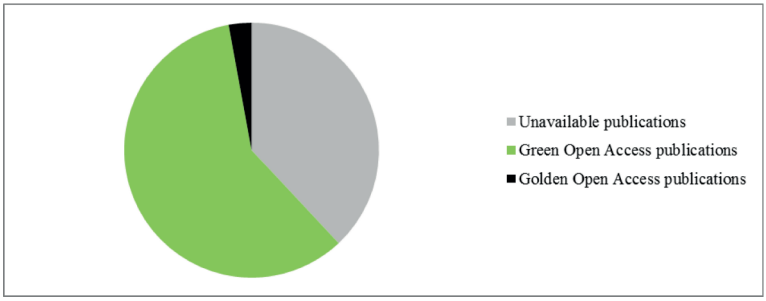

Figure C.26 Overall publication behaviour at CITEC

this can only refer to Green Open Access publications, since only 3% of all publications analysed were Golden Open Access publications, according to the criteria given above (see 2.5). In addition to this, it should be noted that a freely available publication is not necessarily freely available with the authors' knowing about this. For example, the American Physiological Society mentions that articles may be made temporarily free as part of a press release or for other promotional purposes, which means that the actual number of Open Access publications may well be below the one given here. Nonetheless, it can be clearly seen that the majority of publications are available online in some way, with the empirical analysis suggesting an estimate of around 62%.

7 Current status of Open Access to research data

We have given an insight into the existing research infrastructure, as well as the places in which research data are managed (see 5 Current status of research infrastructure. On the one hand, the existing research infrastructure distinguishes between data to which access is restricted to members of the respective project in which the data had been created or collected, and data which can be accessed publicly. On the other hand, the results of the questionnaire have shown that not many groups actually make use of these facilities. In particular, it was shown that archiving of both software and primary and secondary data is mostly done in group-internal repositories (see 5.2.8 Archiving). In the following, we will discuss which policies are followed by the with respect to exchanging data and/or making them publicly available, as well as which kinds of data are generally eligible for exchange.

7.1 Policies and limits

The interviews with the groups revealed that exchange of research data is generally done on a per-request basis. This is to say that researchers from other research institutions who have become aware of a certain data set being created or used in a study, typically by reading a paper describing the respective data, contact the authors of the respective paper to ask for access to the data. Although even this is up to now only very infrequently the case, it does happen occasionally. In such cases, groups generally appoint the person responsible for a particular data set with the task of determining whether the data can be made available to other institutions or not. Here, the general rule for granting access is that the requester acknowledges the cooperation in future publications based on this data set. In addition to this, all groups which gave answers to these questions stated that the general rule is that the set of primary data is believed to have been fully analysed. The reason for this is mainly that the amount of financial and human resources that has gone into the creation of primary data is typically too high to just give the data away "for free". On the other hand, some groups are realistic about the fact that some data sets cannot possibly be fully analysed by a single research group in a reasonable amount of time. Likewise, when asked for conceivable benefits of making primary data available, however, groups tend to see added value in getting additional and even alternative analyses of the same data sets, primarily for reasons of comparability. Only one group (from the SocHum branch) indicated that they make their data available on a public platform, in order to achieve wider (re-)distribution.

Given that the data are believed to have been fully analysed, there is a clear tendency towards providing Open Access to these data. In addition

to this, more concrete plans of developing the necessary infrastructure for providing Open Access to research data will be discussed below (see 9 Future developments). However, since this infrastructure had not been put into place at the time this case study was being carried out, the survey focused on the question as to whether it is generally conceivable for groups to make data available, and if so, which kinds of data and to what degree. In particular, this means that groups were asked to specify whether they would make software and primary or secondary data to close colleagues (only), to other research groups or even to the general public. The following section presents the results of this enquiry.

7.2 Willingness to share data

Figure C.27 to Figure C.29 illustrate the willingness to share primary data, secondary data and software respectively. In addition to this, Figure C.30 shows whether groups which share the software they develop do also make the corresponding source code available.

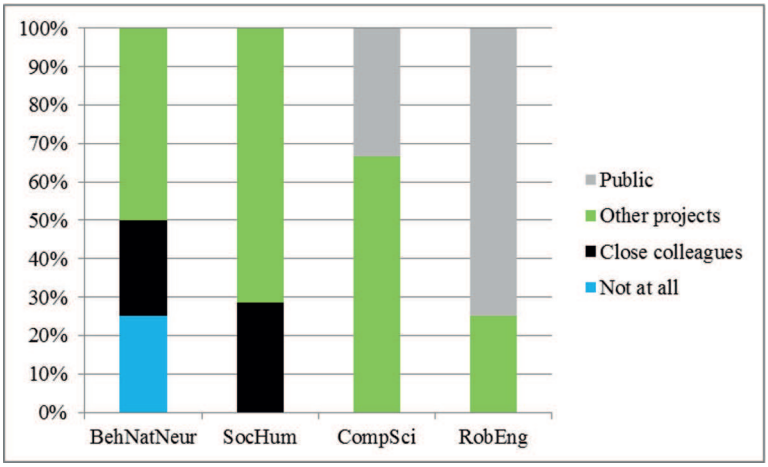

Figure C.27 Willingness to share primary data

7.3 Discussion of results

The results shown above indicate that the willingness to share data beyond a circle of close colleagues differs significantly both between types of data and between types of groups. Starting with primary data, it is clear that groups in behavioural and natural sciences – as well as those in social sciences and

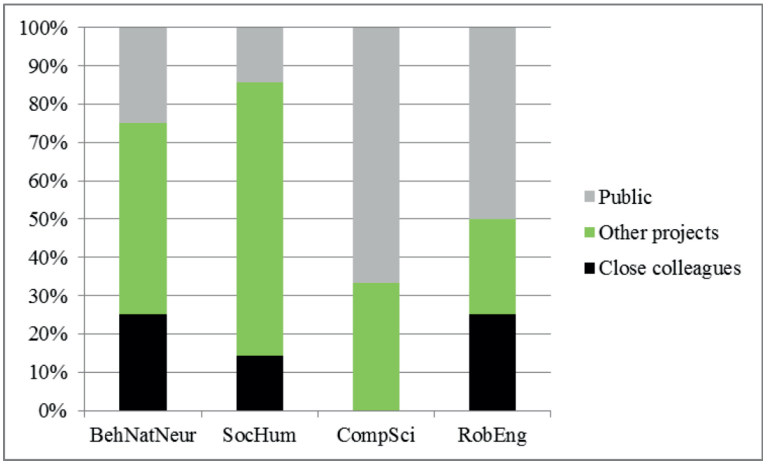

Figure C.28 Willingness to share secondary data

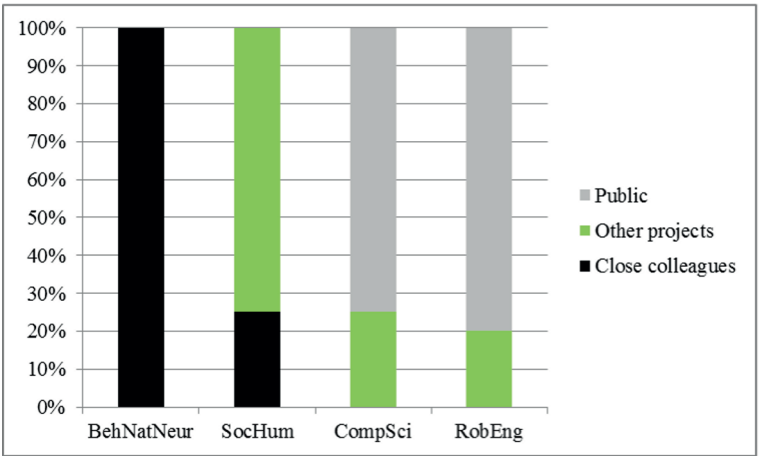

Figure C.29 Willingness to share software

humanities to some extent – are more restrictive than groups in computer science and robotics. In fact, one group in behavioural and natural sciences even indicated that they would not share primary data at all. A straightforward explanation for this is that primary data arising in the former two research branches very often deal with experimental data involving humans. As a result, the free availability of primary data is in many cases not possible from a legal perspective, or at least not desired, which suggests that Open

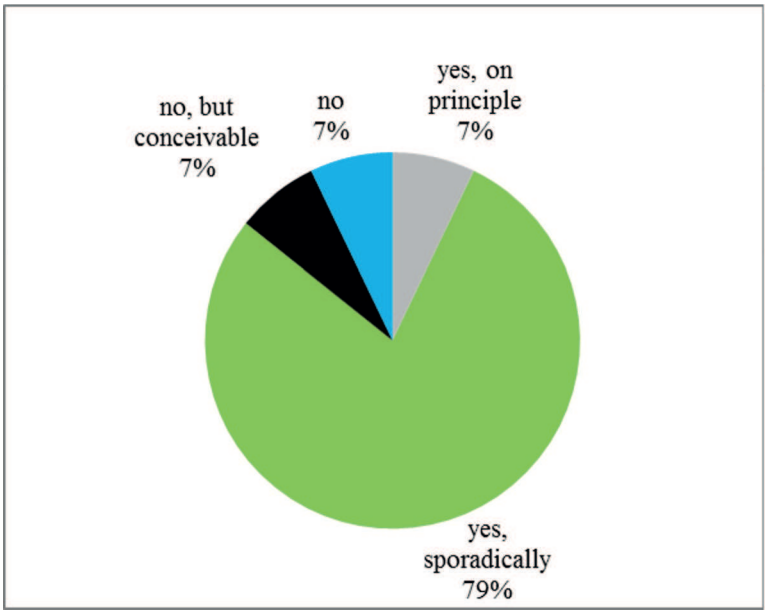

Figure C.30 Willingness to share source code with software

Access to primary data is, at least in these disciplines, a complicated issue. In the latter two research branches, however, primary data are often not created by the group itself – such as in the Semantic Computing group, which frequently uses data available on the World Wide Web – or involves non-human individuals, such as animated characters or robotic devices. As such, legal restrictions like personal rights are less of an obstacle in these disciplines, and the general willingness to share these data is considerably higher. However, potential re-use of the data is limited due to the lack of standard formats in the field. Moreover, one group of the robotics and engineering branch indicated that there are projects whose primary data they would not share at all. The group indicated, however, that this is because the amount of primary data produced exceeds a limit beyond which exchange does not seem reasonable. The situation is slightly different with respect to secondary data. Here, some groups in BehNatNeur and SocHum consider their data to be suitable for being made available to the public, and CompSci groups are less restrictive as to making their data available.

With regard to software, the situation is again very different between the different disciplines. In general, however, the figures seem to suggest that software – at least in those disciplines in which it represents one (if not *the*)

primary research output – could very conceivably be shared with the general public. On the one hand, this could be because well-established platforms for sharing Open Source software exist which enable straightforward sharing of software, such as Sourceforge[24] or Google Code.[25] In addition to this, however, it seems to be the case that software as is generally far less in terms of size than, for example, primary or secondary data. In fact, it seems highly unlikely (in fact almost impossible) that software developed within a single project would ever reach the range of terabytes. Given that software is generally written by humans,[26] this would mean an incredible amount of code being created by hand. While software packages including source code, compiled binaries and the libraries on which the software depends may reach the range of several gigabytes, even this is typically not the case. In line with the above findings for primary data, this suggests that size of data may have an influence on the ease and willingness to share data.

On the basis of these figures, we investigated whether the availability of standardised metadata formats for the description of primary and secondary data is correlated with the willingness to make data available. As was mentioned before, 59% of all groups indicated that there are no standardised metadata formats in their field (see 5.2.7 Metadata enrichment). Therefore, we checked for the remaining 41% whether they could conceive sharing primary and secondary data beyond the level of close colleagues. The results are shown in Table C.7.

Table C.7 Willingness to share software and primary and secondary data in SocHum

	Field has standard format	Conceivable exchange of primary and secondary data
BehNatNeur	3	0
SocHum	2	2
CompSci	0	0
RobEng	2	2
Overall	7	4

At first sight, the figures seem to indicate a rather low correlation, which is in fact 0.23. However, when analysing the figures more closely, it becomes evident that this is due to the difference in the behavioural, natural and neural sciences, where none of the three groups that indicated the availability

[24] http://sourceforge.net.
[25] http://code.google.com.
[26] We are ignoring code generators like Apache Velocity since they are not believed to constitute the main part of software development.

of standard formats are willing to share neither primary nor secondary data beyond the level of close colleagues. In line with what has been discussed above, however, this is mainly because BehNatNeur groups generally tend not make primary data available, due to the reasons mentioned before. In fact, two of the three groups indicated their willingness to share secondary data, which would suggest a correlation of 1.0 using this laxer interpretation of "willingness to share".[27]

Summing up the findings in this section, the figures suggest in general that technical disciplines like CompSci and RobEng are less restrictive when it comes to making data available, be it to other projects or to the general public. As was discussed above, this is in part due to the different extensions of the individual types of data in each discipline, with primary data being a primary concern in BehNatNeur and SocHum. Abstracting from individual research branches, Figure C.31 summarises the overall willingness to share data, according to the types of data. As can be seen there, software and secondary data could far more conceivably be made available to the public, whereas primary data – though being conceivable to be shared with other projects – are either unsuitable for general Open Access or would require very flexible licensing schemes.

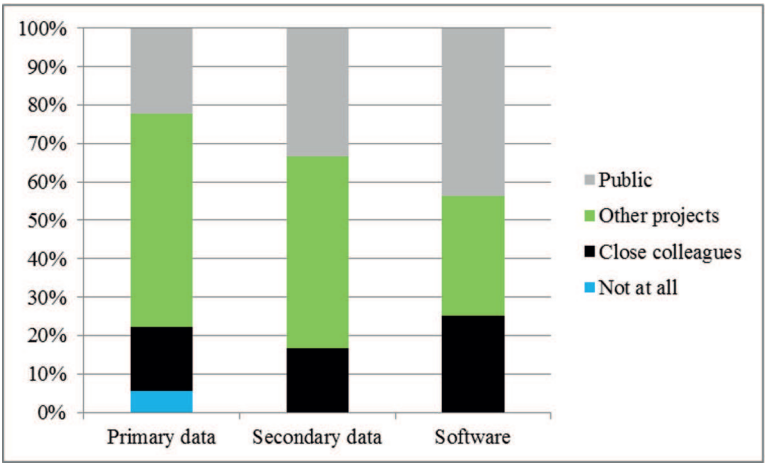

Figure C.31 Overall willingness to share data, according to types of data

[27] Note that we have used a strict interpretation of "willingness to share" – in the sense that willingness to share both primary *and* secondary data was counted as positive evidence only – since the laxer interpretation (i.e. "willingness to share primary *or* secondary data") is true for almost every group.

8 Challenges

In this section, we will discuss the general challenges for an Open Access infrastructure as suggested by the findings above. In general, it has been shown that, in most cases, the data management strategy followed is up to the individual researcher, which results in rather heterogeneous strategies being followed not only between groups, but also within groups. The risk of data sets becoming unavailable due to a researcher leaving the institute is therefore rather high. It should be noted here that this challenge is not solved by creating central repositories alone, since the potential re-usability of a data set is determined by a number of factors. On the one hand, the nature of an experiment or study has a deep impact on re-usability. For example, it seems reasonable to assume that in BehNatNeur, experiments are typically carried out in order to verify a specific hypothesis under very strict conditions. This means that the data collected in such experiments are less likely to be useful in other contexts, which would need to be tested under different conditions. This is certainly different in other disciplines such as CompSci and the data collected there are more likely to be re-used. On the other hand, especially in those disciplines where data are in principle suitable for exchange, it is the degree of documentation of a data set that decides whether it is re-usable at a later point in time or not. In the following, we address different infrastructural challenges with respect to data and publication management.

8.1 Data management

8.1.1 Models for data types, provenance and access rights

Given the variety of data types generated at CITEC, an immediate challenge is to develop models which are capable of representing all aspects of a particular data set. In addition to very general aspects such as type (e.g. audio vs. video), these include the following:

- Given the dependence on proprietary hardware and software identified above (see 5.2.1), it is vital to **document any hardware or software requirements** that need to be fulfilled in order to be able to view or process the data set, as well as other technical aspects like encoding – similar to software package dependencies known, for example, from the popular Linux distribution Debian GNU/Linux.
- Given the guidelines for data sustainability issued by the German Research Foundation, it is necessary to **develop policies and storage infrastructures for short-term, mid-term and long-term archiving** of research data.

- Given the lack of standardised metadata formats in some research areas, it is necessary to **find a reasonable balance with respect to what can be expressed** about a data set, in order to support the re-use of metadata categories whenever possible and thus enhance the interpretability of metadata annotations.
- **Guidelines for ensuring the quality of published data** need to be developed.
- It seems reasonable to **assign data management responsibilities to particular persons** in each group, in order to make sure that all research groups are aware of the available infrastructure.
- It should be possible to **link data sets with publications and vice-versa**, in order to enhance the ways in which both can be explored. Here, it is recommendable to **use technologies and practices developed in the Semantic Web**,[28] in order to ensure that this challenge is addressed in a principled way and achieves appropriate impact.
- Data generated at CITEC poses challenges for storage and backup strategies. For example, given experiments in which 1.5 GB of video data are generated per second, it is vital to have **reasonable backup strategies**. It is understood, however, that this is even more of a requirement in other research areas besides ICT.

The willingness of people to share data with external people, be they researchers involved in other projects or members of the general public, has been discussed above (see 7.2 Willingness to share data). The primary finding was that research in cognitive interaction technology – primarily due to its high degree of experimental work with humans and animals – raises a number of concerns regarding personal rights, and unrestricted Open Access does not seem feasible here. For other cases, excluding those in which access to data is completely impossible due to legal restrictions, it is necessary to have a sound model of access rights to individual data sets – which may even require entirely new licence models, especially with a view on re-usability and modification. Here, it is necessary to encode the provenance of a data set, in order to document its source and development history. As with other challenges mentioned above, this should be approached by making use of available vocabularies as much as possible, in order to achieve interoperability between resources. In addition to this, it is necessary to have a functioning system that implements this model of access rights. As trivial as this aspect may seem, it should be noted that a security leak in the system – or even accidental publishing of confidential data – may have far-reaching legal consequences.

[28] http://semanticweb.org.

8.1.2 Rules, incentives and limits to research data exchange and Open Access

As was just mentioned, Open Access raises legal concerns especially with respect to primary data obtained at CITEC. Moreover, as was mentioned above, the sheer amount of primary data produced may be a limit to exchange in itself (i.e. if the data exceeds an amount at which sharing the data does not seem feasible; see 7.2 Willingness to share data). In addition to this, especially in cooperations with industrial partners, confidentiality agreements have to be signed which restrict the future use of the data in other projects, let alone its publication to the general public. Besides this, however, we have shown that actual data exchange is still performed to a rather limited extent, with exchange upon request, and only after the data are believed to have been fully analysed, being the main policy for data exchange (see 7.1 Policies and limits). It should be noted, however, that researchers admit that it is, in most cases, not possible to say when a specific data set has been fully analysed, while in other cases it is not even possible for a research group to fully analyse a specific data set in a reasonable amount of time. Finally, groups indicated that they expect the amount of maintenance work (e.g. documentation) required to transform a data set into a state in which it can be released to the general public to be very high and the resources that would be needed cannot be allocated – on the one hand due to lack of funding for such tasks and on the other due to lack of scientific reward or appreciation by the community. This is further supported by the analysis presented above (see 5.2.8 Archiving), which showed that only a small number of groups deposit their data on external repositories. Here, a concern was that – given that a data set is, for example, made available to the scientific community but not to the general public – how would it be possible to trace where the data actually end up, after having been downloaded by a large number of people? Finally, it may be possible that experimental approaches will experience a dramatic decrease in the number of probands, because they have to sign very complex data privacy statements. Here, the general trend towards freedom and openness that can be observed on the World Wide Web today faces the desire for more privacy and protection of personal rights at the same time.

On the other hand, many groups expressed the benefits of Open Access to research data. Some of the incentives for Open Access stated by researchers are given below.

- **Increase data transparency**, which would enable researchers, federal agencies or members of the general public, to obtain a better overview of the data generated at a research institution or in a research field.
- **Benchmarking and contrastive analyses** being carried out by different institutions on the same data sets.

– **Support from other institutions** in analysing a particular data set, and thus faster progress in a research field.

8.2 Publication management

In addition to the challenges for data management just discussed, there are a number of requirements on publication management as well. In particular, there is no CITEC-wide publication repository, and publication management is therefore handled very differently not only between research groups, but also within groups (see 5.1.5 Publication management and 5.2.8 Archiving). Therefore, what is needed is, on the one hand, a shared technical infrastructure for depositing publications and, on the other hand, general guidelines and policies regulating deposit and access. In the context of Open Access, we identify the following requirements:

– The interface to the publication deposit process – be it the user interface, application programming interface or web service interface – needs to allow the depositing client to **upload both metadata and the full text of a publication**.
– It should further be possible to **specify the rights (e.g. copyright) that the depositing client is in possession of**, in order to determine whether the client has the permission to set further access rights for the full text of the publication.
– If the client is in possession of the appropriate permissions, the system should allow him or her to **specify the restrictions that possibly apply to the full text**, such as whether it is publicly accessible or only accessible to people belonging to a certain group of users.
– It should further be possible to **determine whether it is permitted to search or crawl the full text and/or metadata of a publication**.
– In order to be able to interoperate with other literature management tools, **metadata should be exportable in several (de-facto) standard formats**, such as BibTeX or Endnote.

Depending on the input by the client on the previous points, the system should then be able to select the appropriate measures for storage and access and allow for flexible search and retrieval. For example, the literature analysis carried out as part of this case study would have been greatly facilitated by being able to search for all downloadable publications of a specific group (or of all CITEC, with results grouped by research group) or for publications which have been written in cooperations between groups. Finally, as was mentioned in the previous section, it should be possible to link publications to research data sets, in order to enhance the information services provided by the system.

9 Future developments

As part of its second funding period, CITEC has very concrete plans for the future development with respect to managing literature and research data, in particular in the direction of linking the two in order to obtain an ecosystem of semantically enriched descriptions of all kinds of research artefacts. First steps into this direction have already been taken and implemented during the course of this case study and the following subsections discuss these current and upcoming developments in more detail.

9.1 Literature management

9.1.1 Interaction with central facilities provided by the university

As was mentioned above (see 5.1.5 Publication management), research groups at CITEC generally take care of literature management themselves, which means that they host and make the descriptions as well as – to some extent – the full texts of the publications authored or edited by members of the group available. Recently, however, the library of Bielefeld University has released the PUB system, a university-wide repository intended to host metadata and full texts of all publications created at Bielefeld University. In order to make use of this repository while still being able to annotate publications with metadata fields not provided by the PUB system, CITEC has developed a module based on the widely used Drupal CMS allowing for a smooth interaction between group-administered publication repositories and PUB. In particular, the module enables the management of a local publication repository which is synchronised with the PUB repository. Here, the module ensures that the local repository always contains at least the group-relevant publications available in PUB, with the possibility of containing additional publications not available in PUB. This concerns, for example, those items which have not been published yet and whose descriptions are therefore not complete yet. Even though PUB provides way for handling such cases as well, authors may prefer not to expose their manuscripts on the university-wide repository until they have been published. In addition to this, the module allows for attaching the aforementioned additional metadata descriptions to a group's publications, which will be described in more detail below.

9.1.2 Semantic enrichment

CITEC is taking concrete steps towards annotating the locally stored publications with additional metadata. On the one hand, this concerns the use of standard schemas for the description of bibliographic entities, such as Dublin

Core.[29] On the other hand, however, CITEC aims at making the descriptions not only useful for human users navigating to a group's website, but also for machines harvesting the website for information. Here, formalisms developed in the context of the currently evolving Semantic Web, such as the *Resource Description Framework* (RDF) or the *Web Ontology Language* (OWL), as well as formal models for representing bibliographic entities by means of these Semantic Web formalisms are of particular interest. In addition to those established, this concerns the analysis and exploration of bibliographic ontologies currently under development, such as the *Semantic Publishing and Referencing* (SPAR) ontologies,[30] which includes Semantic Web versions of established bibliographic models like the *Functional Requirements for Bibliographic Records* (FRBR).[31] Such ontologies are particularly interesting since they provide a rich vocabulary that not only allows for a formal representation of bibliographic entities, but also of the relations between them. Beyond the usual citation relation, this concerns relations such as *usesDataFrom* or *disagreesWith*. It is clear to see that having such relations between entities would greatly enhance the ways in which publications could be queried not only by humans, but also by machines. Here, current development focuses on the integration of such descriptions into the aforementioned module in order to provide such enhanced services.

9.2 Data management

In addition to the management of literature, CITEC has recently launched a research data management task force involving the leaders of several research groups as well as members of the university library and the Collaborative Research Centre 673. The goal of this task force is to design and implement a strategy for achieving sustainability and reusability of all kinds of data created at CITEC. In the first instance, this development is concerned with providing an appropriate framework for storing the data in a way that enables a smooth integration into the existing research infrastructure as explained before (see 5 Current status of research infrastructure) and implements the necessary procedures for enabling Open Access to the data. A second development phase deals with providing suitable vocabularies that allow for a fine-grained description of all aspects of the data, as well as interlinking with other descriptions, such as those of literature already mentioned. The final phase of this development then deals with aspects of making the data available. Here, planning has begun on extending the already existing OpenSource

[29] http://www.dublincore.org/documents/2010/10/11/dces.
[30] http://purl.org/spar.
[31] http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.

server to an OpenData server[32] that on the one hand provides direct access to the data and on the other hand enables access to metadata descriptions by metadata harvesters like CLARIN via the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH). As a result, the vision of research data management at CITEC is ultimately an open one, where all kinds of research artefacts created at the institution – including literature and data which do not affect personal rights – are made available to the research community as well as the general public.

10 Implications for Open Access infrastructure

10.1 Technical implications

- **The diversity of data types** arising even in individual experiments on a single research topic requires mechanisms that allow for linking heterogeneous data types in a way that allows flexible and intuitive exploration.
- **The amount of data** being generated requires the storage infrastructure to be able to deal with data in very large quantities and sizes.
- **The dependence on non-standard formats and proprietary software,** including non-free operating systems needed for the operation of specific research instruments entails a number of issues like backward (in)compatibility, maintenance and licensing that require exact specifications, for example, which software version is needed in order to be able to process the data file in the intended way. These need to be stored and linked with the data in order to make the data re-usable.
- **Privacy issues** of experimental primary data involving humans, as well as data arising in cooperations with industrial partners, pose special requirements on the security of the data, as misuse or accidental release can have far-reaching legal consequences.

10.2 Scholarly implications

- **Fine-grained licensing schemes** regulating access, re-use, linking, manipulation and redistribution of research data need to be developed, as current schemes cannot handle critical cases, for example where anonymised primary data lose their anonymity by other data sets linking to them.

[32] At the time of writing, the CITEC OpenData server has officially gone live at http://opendata.cit-ec.de and published the first freely available data set of manual interaction data.

- **Rewarding and acknowledgement schemes for data creation, curation and publication** need to be developed and established, as these tasks typically take up much more time and effort than, for example, the creation of a scientific article, while they are at the same time not recognised as indicators or measures of quality of research as the latter.
- **Rewarding of golden Open Access publications** in order to establish it as a recognised means of publication.
- **Institutional, disciplinary and/or funder-driven** guidelines and policies for data exchange need to be established in order to provide a framework and incentives for data exchange.
- **Advertising the availability and benefits of the infrastructure** in a way that allows researchers from less technical fields to know what is available and where to find it.
- **Educational support in using the infrastructure** so that researchers not only find available services, but also know how to use them and benefit from them.
- **Funding for designated resources** dealing with data management issues, since data curation is currently done at a subjective level, instead of being a designated part of the general research agenda.

11 Acknowledgements

We would like to thank all research groups who have taken the time to participate in this study, as well as Jochen Steil and Sven Wachsmuth for their helpful comments and suggestions.

12 List of figures

13 List of tables