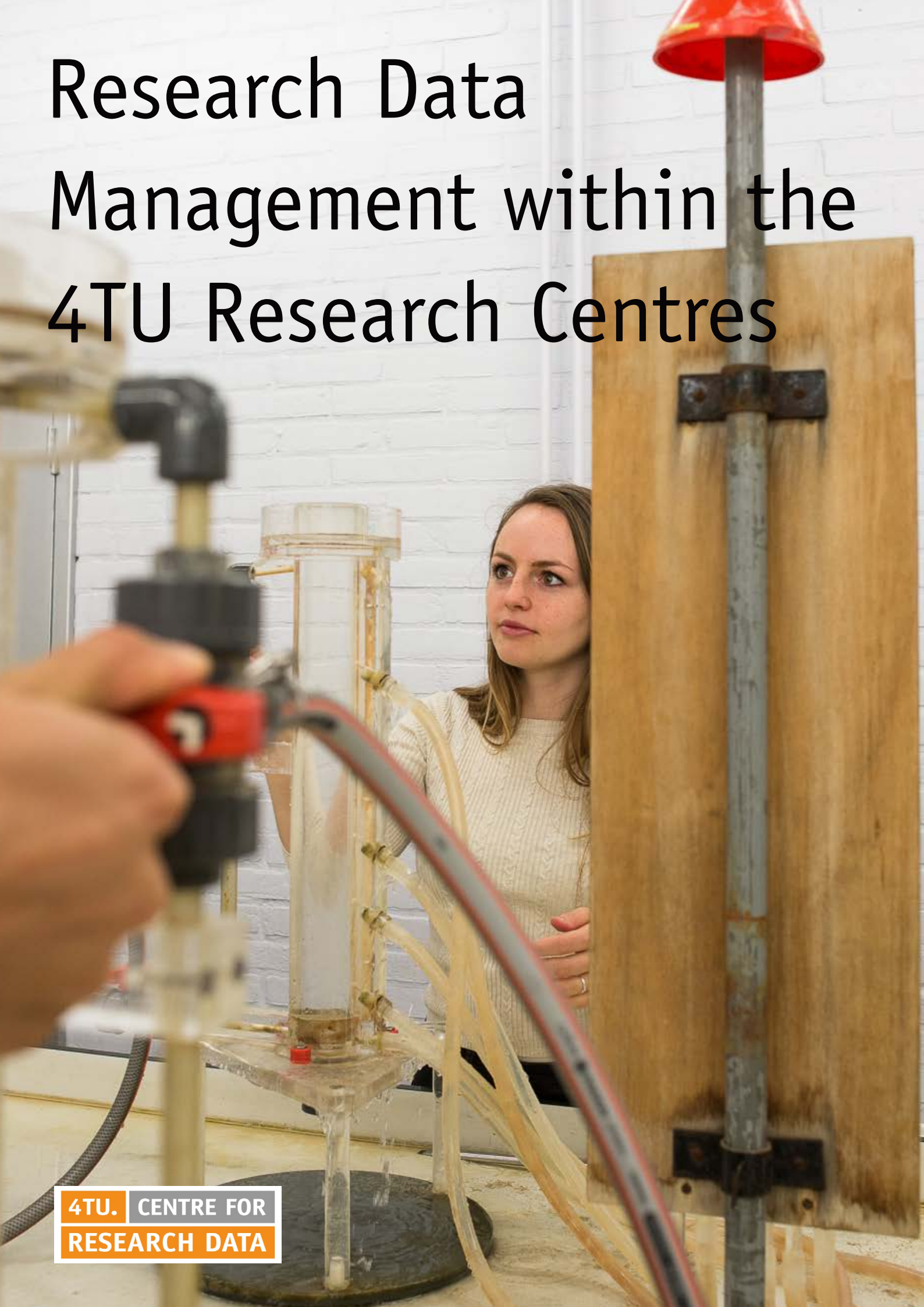
Research Data Management within the 4TU Research Centres



4TU. CENTRE FOR RESEARCH DATA

Contents

The 4TU.Federation and its Research Centres 2

The 4TU.Centre for Research Data 2

Aims and scope of this report 2

General findings 3

Possible next steps 3

Detailed findings: research data management attitudes and practices 4

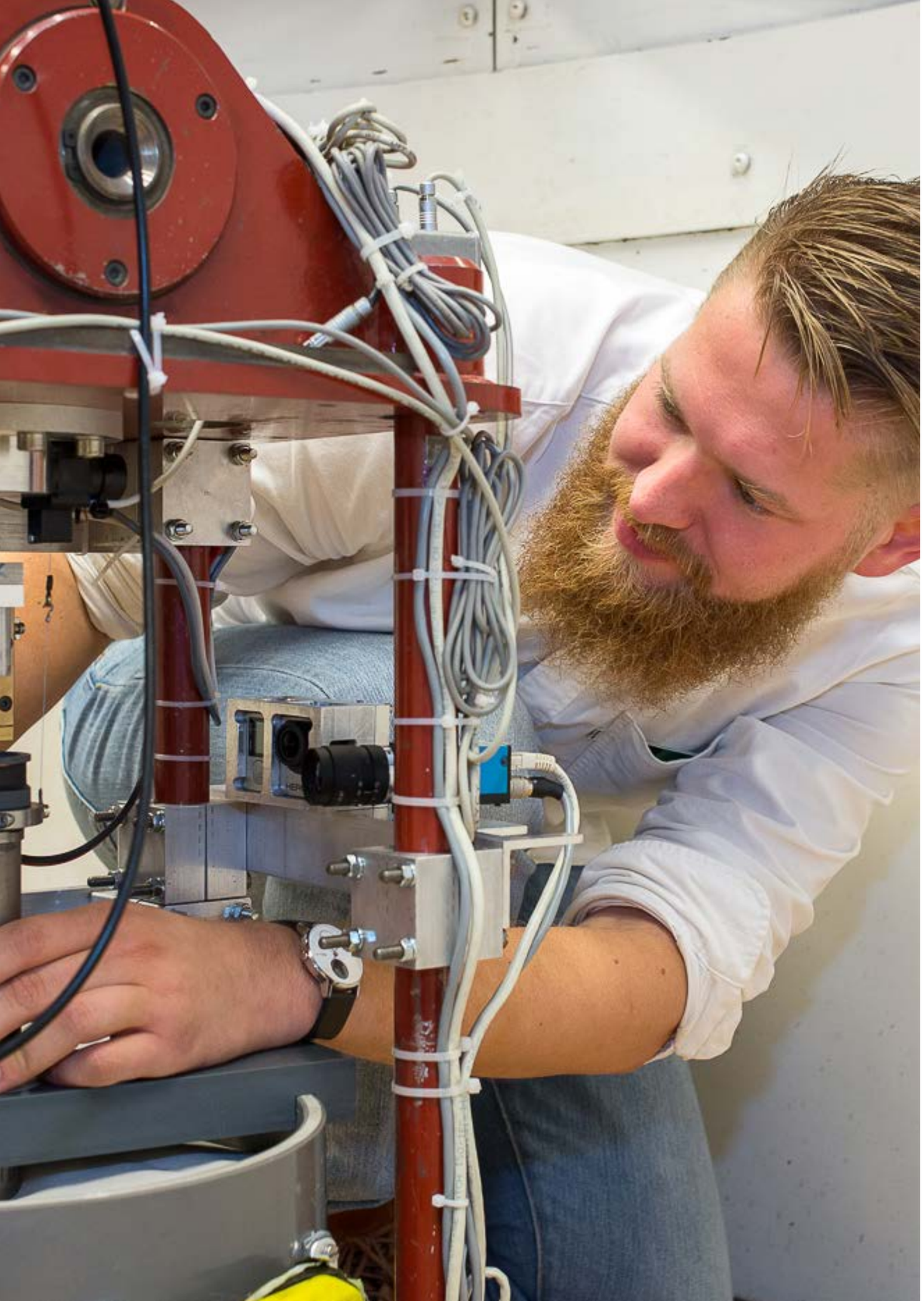
- Responsibility for the stewardship of research data 4
- Data storage and long-term preservation 4
- Data publishing 6
- Software and software sustainability 6
- Commercial, private, and sensitive data 7
- Funder requirements on research data management 8
- Data sharing: extrinsic requirement or intrinsic motivation? 8

Challenges and opportunities 9

Appendix: How we worked 11

By Maria Cruz and Alastair Dunning - 4TU.Centre for Research Data - March 2018

All full page pictures by Annemiek van der Kuil | PhotoA.nl



The 4TU.Federation and its Research Centres

Founded in 2007 as the 3TU.Federation, the 4TU.Federation comprises the four technical universities in The Netherlands: Delft University of Technology, Eindhoven University of Technology, University of Twente, and Wageningen University & Research. Together these universities include close to 50,000 students and 3,000 (full time equivalent) scientific staff[1], representing a large combined source of expertise in engineering and the technical sciences.

Creating links between the four universities, and by doing that, strengthening their research and teaching, and encouraging innovation are the core values and mission of the Federation. Cooperation in the areas of research, teaching, and valorisation has led to the creation of the nine 4TU Research Centres[2]:

- High-Tech Systems (4TU.HTS)
- Netherlands Institute of Research on ICT (4TU.NIRICT)
- Fluid and Solid Mechanics (4TU.FSM)
- Applied Mathematics Institute (4TU.AMI)
- Ethics & Technology (4TU.Ethics)
- Built Environment (4TU.BE)
- Design United (4TU.DU)
- High-Tech Materials (4TU.HTM)
- Human-Technology Interaction (4TU.HTI)

4TU.Centre for Research Data

4TU.Centre for Research Data (hereafter 4TU.ResearchData) began as a project in 2008 (at the time it was named 3TU.Datacentrum) as a "state-of-the-art facility for carefully storing scientific research data and making it permanently accessible." By 2010, it was up and running as a fully operational data archive. Now comprising just over 7500 (and growing) datasets with Digital Object Identifiers, 4TU.ResearchData has evolved to become a certified and trusted repository for the technical sciences. Funded by Delft University of Technology, Eindhoven University of Technology, and the University of Twente, the Centre offers researchers knowledge, experience, and tools for sharing and safely preserving scientific research data.

Aims and scope of this report

Although 4TU.ResearchData is part of the 4TU.Federation and has almost the same age as the Federation itself, its existence and services may not be widely known among the researchers connected to the 4TU Research Centres. To introduce- or re-introduce, as the case may be – 4TU.ResearchData, and to explore how the 4TU Research Centres are managing their research data, in the summer of 2017, we contacted the Scientific Director of each of the nine 4TU Research Centres with an invitation to meet individually with 4TU.ResearchData for a structured interview[3]. The aim of these interviews was to learn about the Centres' focus and activities and to get a sense of their research data management practices and attitudes. This Report highlights the results from those contacts and interviews, starting with general findings and related possible next steps, followed by detailed findings and challenges and opportunities.

1 As per https://www.4tu.nl/en/about_4tu/facts-figures/.
2 Financial support for 4TU.HTS and 4TU.FSM was discontinued at the end of 2017.

3 For details about whom we contacted, whom we met and when we met, and what we discussed, see the Appendix - How we Worked at the end of this document.

General findings

1. Community building and networking are the main focus of activity for those who work for the Centres on a day-to-day basis. Research activities are on the whole the responsibility of individual research groups connected to the Research Centres. Research data management is not addressed at a strategic level, but left to individual research groups or to individual researchers.

2. The Centres include a broad range of disciplines and research topics, even within one Centre. There is, therefore, a broad range of attitudes towards data and a broad range of data types and characteristics, including large datasets; commercially sensitive datasets; privacy and ethical concerns regarding data; software and its sustainability.

3. Software sustainability is an important and much discussed topic, particularly in the computer science and applied mathematics communities, but also in materials science. There are currently no standards or systematic way of looking after software. Deciding which software should be sustained and how it should be maintained are important questions that are still being considered.

4. Research on human subjects and datasets including personally identifiable information or sensitive personal information are more prominent than might be expected in engineering and the technical sciences. Computer scientists and applied mathematicians, in particular, often work with these types of data. Lack of transparency and reproducibility of scientific results can be an issue in these areas because the underlying datasets are often not available. There is no easy solution to these problems, but more sophisticated sharing of output data could play a role.

Possible next steps

1. Maintain the interaction between 4TU.ResearchData and the 4TU Research Centres, and ensure that the Centres are aware and communicate 4TU.ResearchData's core services to their communities: Archiving, Strategic Guidance, and Training[4].

2. Support and work with winners of the 4TU research programmes 2018-2021 on the development of their data management plans.

3. Software Sustainability: Communicate best practice findings from 4TU universities (e.g. TU Delft Sandbox Sessions Open ICT) on Software Sustainability. Organise potential workshop amongst interested parties on training, archiving, and career recognition related to research software.

4. Data privacy and ethics: Communicate best practice findings from 4TU universities on data privacy and ethics. Organise potential workshop amongst interested parties on informed consent, sharing of sensitive data, and tools for anonymisation.

4 As a result of recent interactions with 4TU.ResearchData, 4TU.HTM now includes information about 4TU.ResearchData on their website: https://www.4tu.nl/htm/en/research/4tu-centre-for-research-data/.

Detailed findings: research data management attitudes and practices

Responsibilities for the stewardship of research data

By and large, the Scientific Directors felt that the researchers or research groups within their Centres were individually responsible for the data they collected, produced, processed, or analysed. This means that, on the whole, the Centres do not address research data management at a strategic level. As the Scientific Director of 4TU.Ethics put it in an email: "We currently do not see an active role for the 4TU.Ethics management to devise our own policies or practices in this regard. Most data that is produced is produced in the context of funded research projects. We trust that these projects adhere to data management policies of the institutions at which they take place." In the case of this Centre, it is fair to say that they do not produce much data. 4TU.Ethics researchers engage mostly in philosophical research, and this is a conceptual and normative discipline, rather than an empirical discipline. However, the statement from 4TU.Ethics could perfectly summarise the attitude towards data stewardship encountered in the meetings with the Scientific Directors of the other Centres.

Data storage and long-term preservation

Data management is left to individual research groups or individual researchers, such as PhD candidates and postdocs. If a researcher loses data – because, for example, there wasn't a storage or backup strategy in place – this is considered to be their own problem. As one of the Scientific Directors put it: "It comes down to the survival of the fittest."

A common problem – mentioned by 4TU.BE, 4TU.AMI, 4TU.HTM, and 4TU.HTI – is that PhD students often depart without leaving their dissertation data behind. At times, this happens because there is confusion about who owns the data. More often than not though, it's because students are busy at the end of their theses and there is no good protocol for data sharing at the end of the project. Sometimes the data can become available upon request, but this can be difficult to achieve, let alone enforce. The data are usually stored on students' individual laptops or external disks. Students often move to another country, or can change careers, and may find it difficult to retrieve the data. Even when students leave data behind, the files might be in formats that are not conducive to long-term preservation, or there isn't enough documentation for others to be able to understand and work with the data.

Researcher mobility may also prevent scientists from taking advantage of institutionally provided solutions for data storage and preservation. Some researchers feel that it is safer and more practical to take care of the data they handle, collect, or produce – rather than relying on local ICT provision – because they will need to move to different institutions, often in different countries, as they progress in their academic careers. More rarely, there is simply lack of trust because of previous institutional failures.

Data publishing

For some of the research communities served by the Centres, data publishing is still not an habit. In materials science, for example, researchers will publish the results of their data analysis in a scientific paper, and may sometimes provide extra information about the data in the Supplementary Materials section, but they will very rarely share the raw data. In few instances, the datasets are simply too large to be easily shared with others. For example, materials scientists who collect data at the European Synchrotron Radiation Facility in Grenoble often come back with a couple of terabytes of raw data to process. Such volumes of data are difficult to store, manage, and preserve. The researchers bring the data back from the Facility in their own hard drives and use external hard drives for backup. Long-term preservation is a concern, but there are no good solutions available.

Software and software sustainability

Computer scientists and applied mathematicians consider themselves mainly as data users rather than data producers. Therefore, data stewardship and data management were concepts that didn't resonate much with the communities served by 4TU.AMI and 4TU.NIRICT. In general, these communities produce mathematical models, algorithms, and computer code to analyse third-party data. Software is also extensively used and produced by the communities served by 4TU.HTM for data analysis and mathematical modelling of physical processes and phenomena.

Software sustainability is an important and much discussed topic in these communities; however, there are currently no standards or systematic way of looking after software. Deciding which software should be sustained and how it should be maintained are questions that are still being grappled with – for example, by the Netherlands eScience Center, of which NIRICT's Scientific Director is an advisory board member. Software sustainability is a tough task to accomplish even for such an organisation, which has software sustainability in its mission and dedicated research software engineers.

There is a big push for open source software, and a lot of what is produced by the computer science community is open source – shared through platforms such as GitHub. In general, the community is willing to share software, but that software may not always be easily findable or accessible. Sometimes it is a matter of asking researchers directly. Even when software is available – shared as part of a journal article, for example – it isn't always the case that it will be reusable. Frequently not all the pieces that are required for the reproducibility of computational results are in place. The primary data used to test the algorithms are often not available. One of the Scientific Directors felt though that all the building blocks for software sustainability and sharing already exist; it's the connections between them that are missing.

Commercial, private, and sensitive data

Computer scientists and applied mathematicians (4TU.AMI and 4TU.NIRICT) often work with (but generally do not produce) sensitive, private data, and commercial data. In the field of cyber-security, for example, researchers work together with government agencies whose data cannot be shared. Bioinformatics relies heavily on patient data collected in hospitals; these data are also restricted. There is also a lot of work done in collaboration with commercial partners that may be willing to share the data privately with the researchers, but do not want to have that data publicly shared. Commercially sensitive seismic data, for example, is sometimes redacted before it is shared with researchers. Although there are some partners who do not present these constraints, there is a recognised problem with the transparency and reproducibility of scientific results in computer science and applied mathematics, because the underlying datasets are often not available. This is a problem that is not easy to solve. Sharing of output data could play a role in transparency.

In the case of 4TU.HTI, which was initially part of NIRICT, researchers deal with both highly sensitive data and very large data volumes. They regularly record multimodal corpora, which can be defined as "digitized collections of audio- and video-recorded instances of human communication connected with transcriptions of the talk and/or gestures in the recordings."[5] These data are used in psychology, artificial intelligence, robotics, signal processing, computational linguistics, and human-computer interaction.

For researchers at 4TU.HTI, the biggest challenge is to share these data with international colleagues during the duration of a project and to preserve these data for the long term. There are routines and protocols, but not formal plans, for how to share data during research projects, for getting ethics committee approval and participant consent, and for data anonymisation. Data are usually stored in password-protected servers at the researchers' institutes and accessible only to the researchers involved in the project; data are transferred to other servers if cleared for sharing (if anonymised, for example). However, there seem to be no good solutions (infrastructure wise) for sharing highly sensitive data with international colleagues during a research project[6]. Long-term preservation is also a challenge because of the large volumes of data involved.

There are also open datasets in the field of human-technology interaction. There are community challenges, where groups of researchers have access to the same dataset and each group tries to come up with the same solution – say a speech recognition algorithm. Some datasets are open with restrictions, where users have to sign a form (or license). Usually, the research paper will point to a URL containing access instructions and a contact point.

5 See Allwood, Jens (2008). Multimodal Corpora. In: Lüdeling, A. & Kytö, M. (eds) Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter. p 207-225.

6 Sensitive data services and their integration with European e-infrastructures was discussed at the Digital Infrastructures for Research 2017 conference held in Brussels 30 November - 1 December 2017. It was acknowledged under this discussion topic that "there is still the need to implement policies and the corresponding technologies that effectively allow cross border, inter-disciplinary research on personal sensitive data."

Funder requirements on research data management

The Scientific Directors were aware that funding bodies, such as NWO and the European Commission, now require data management planning and data sharing for funded projects; in general, they viewed these developments positively. For 4TU.HTM, for example, "a statement on fulfilling these requirements is important in H2020 project proposals, although training of researchers may be necessary for optimal implementation in practice." Tailored training would be most useful because of the large variation in how scientists do research.

4TU.BE used to run a successful funding scheme of their own, called the Lighthouse Projects. There were no specific requirements regarding data management and data sharing in this funding scheme. "This was not in the picture as we started," explained the Scientific Director of 4TU.BE. The data stayed with the individual researchers, and it was their responsibility and choice whether to preserve and share the data – or not. However, these were small-scale, fast-track, high-risk projects to be developed further if they successfully delivered tangible results (e.g. a prototype or a test environment) after one year. The best projects secured funding from NWO for further development. Although data archiving is not common practice in this area, "if that's a requirement, we will do it" was the reaction from this community to NWO's new data management protocols.

Data sharing: extrinsic requirement or intrinsic motivation?

For many researchers, data sharing and research data management are extrinsic requirements – from funding bodies, research institutions, and publishers. Researchers may appreciate and agree with the higher goals of transparency, good science, reproducibility, and so forth, but at an individual level, there are, or there seem to be, no immediate payoffs. As put by NIRICT's Scientific Director, "others may prove you wrong by using your data, which is good from the perspective of scientific integrity, but that won't help you get your PhD." On the other hand, as also pointed out by NIRICT's Scientific Director, providing data to others – especially if the data are unique, well documented, and citable – can help researchers get more credit for their research (e.g. citations[7]). Formulated this way, research data management can become an intrinsic motivation.

7 For a recent brief review of the literature showing that data sharing confers a citation advantage, see McKiernan et al. eLife 2016;5:e16800 DOI: 10.7554/eLife.16800. Look for the section "Resource management and sharing" and the paragraph titled "Gain more citations and visibility by sharing data" within that section.

Challanges and opportunities

In October 2017, the 4TU.Federation issued a call for proposals for the 4TU research programmes 2018-2021. Funded programmes will need to develop a Research Data Management Plan with the support from the 4TU.Centre for Research Data in the first months of the research programme. 4TU.ResearchData welcomes this deepening of its links to the 4TU Research Centres. The hope is that, together with this Report, this is the beginning of a closer and fruitful collaboration.

Research data management is increasingly considered an important part of high-quality research. International and national funding bodies have in the last couple of years started to mandate institutions and researchers to make data available. Data sharing is predicated on good research data management and has the potential to make scientific research more transparent, open, and efficient. In view of these principles and developments, could the 4TU Research Centres take a more strategic view towards data management? How could 4TU.ResearchData help with that?

It is clear that there are many areas where 4TU.Research Data could immediately offer help and expertise. For example: provide tailored training and advice on research data management; help with transforming routines and protocols, where those already exist, into formal data management plans; help create catalogs for data that cannot be shared.

There are other areas where 4TU.ResearchData and the 4U Research Centres could collaborate. Software sustainability, for example, is an area of common interest. 4TU.ResearchData's mission and core values are to offer high-quality advice and training on research data management and a long-term trusted digital archive for storing scientific research data, including software. How can 4TU.ResearchData contribute towards better standards and protocols for software storage, maintenance, and preservation? How to ensure that 4TU.ResearchData provides a state-of-the-art facility for storing and making scientific research software permanently accessible in a useful and sustainable way? How can 4TU.ResearchData provide good quality advice and training in this area? The same for sensitive, private data and ethics related to research data. Could 4TU.ResearchData provide better services and guidance in this area?

Regardless of the exact form of any future potential collaboration, 4TU.ResearchData would like to keep in touch with the Centres and to meet up regularly to exchange ideas and to keep each other informed. Could there be a contact person or a liaison for research data at each Centre?

Appendix: How we worked

In mid-August 2017, we sent the Scientific Director of each of the nine 4TU Research Centres an invitation to meet individually with 4TU.ResearchData to discuss research data management and to introduce 4TU.ResearchData. We received responses from all the Centres bar 4TU.DU[8]. Three Centres preferred not to meet with us: 4TU.FSM, 4TU.HTS and 4TU.Ethics. However, 4TU.Ethics sent their views on research data management via email. Meetings with the five remaining Centres took place in Delft throughout September 2017 (see the Table below). Maria Cruz, Alastair Dunning, or both participated in these meetings. The meetings were also sometimes attended by other staff from the TU Delft Research Data Services and in one case by one of the TU Delft Data Stewards. The Scientific Directors were informed about the aims and scope of the project and that the results would be published as a Report to the 4TU.General Board. The meetings were structured, although not always strictly, around the following topics: the history and focus of the Centre and its activities; the research data (including software) collected, produced, processed, or analysed by the communities served by the Centre; data management practices and attitudes; data storage, archiving, and long-term preservation; funder requirements; training needs; help and support 4TU.ResearchData could provide. All Directors, including those we did not meet in person, were given the chance to read and to comment on the first draft of this Report between 15 December 2017 and 2 March 2018.

8 The Scientific Director of 4TU.DU later responded to a call for comments on the first draft of this Report.

4TU Centre	Meeting date	Participants: 4TU.ResearchData & TU Delft	Participants: 4TU Research Centre
4TU.NIRICT	6 September 2017	Alastair Dunning, Maria Cruz	Prof. Inald Lagendijk, Drs. Eveline Vreede
4TU.HTM	12 September 2017	Alastair Dunning	Prof. Jilt Sietsma, Drs. Reina Boerrigter
4TU.BE	13 September 2017	Maria Cruz, Marta Teperek	Prof. Ulrich Knaack
4TU.AMI	19 September 2017	Maria Cruz, Jasper van Dijck	Prof. Kees Vuik

4TU.Centre for Research Data
TU Delft Library
Prometheusplein 1
2628 ZC Delft
T +31 (0)15 27 88 600
E researchdata@4tu.nl
@4TU.ResearchData