OECD publishing

BUILDING DIGITAL WORKFORCE CAPACITY AND SKILLS FOR DATA-INTENSIVE SCIENCE

OECD SCIENCE, TECHNOLOGY AND INNOVATION
POLICY PAPERS
July 2020 **No. 90**

OECD
BETTER POLICIES FOR BETTER LIVES

BUILDING DIGITAL WORKFORCE CAPACITY AND SKILLS FOR DATA-INTENSIVE SCIENCE

STI POLICY PAPER

This paper was approved and declassified by the Committee for Scientific and Technological Policy (CSTP) on 19 June 2020 for publication by the OECD Secretariat.

Note to Delegations:
This document is also available on O.N.E. under the reference code:
DSTI/STP/GSF(2020)6/FINAL

Abstract

This report looks at the human resource requirements for data intensive science. The main focus is on research conducted in the public sector and the related challenges and training needs. Digitalisation is, to some extent, being driven by science and at the same time it is affecting all aspects of scientific practice. Open Science, including access to data, is being widely promoted and there is increasing investment in cyber-infrastructures and digital platforms but the skills that are required by researchers and research support professionals to fully exploit these tools are not being given adequate attention. The COVID-19 pandemic, which struck as this report was being finalised, has served to emphasise the critical importance of data intensive science and the need to take a strategic approach to strengthen the digital capacity and skills of the scientific enterprise as whole. This report includes policy recommendations for various actors and good practice examples to support these recommendations.

Foreword

Over the past decade, as part of overall movement towards Open Science, there has been growing policy interest and public investment in enhancing access to data and in data intensive research. This parallels the rapid development of information and communication technologies and digital tools, which are being adopted across the scientific enterprise as a whole. With this comes the promise new scientific breakthroughs, greater transparency and reproducibility in scientific results and increased innovation with overall benefits for science and society. This promise is beginning to be realised but is also limited in many areas by the need for an appropriately skilled scientific workforce.

This report was commissioned by the OECD Global Science Forum to identify: the skills needs for data intensive science; the challenges for building sustainable capacity as these needs evolve; and, the policy actions that can taken by different actors to address these needs. The principle focus is on the requirements of public sector/academic research, with it being recognised that many of the skills that are required for data intensive research are transferable across different sectors, including business and industry. The focus is also on the specific requirements of post-graduate scientific research and related training rather than generic digital skills and the roles of undergraduate and school education. The role of education in relation to digital skills and competencies is addressed in many other reports from OECD and other organisations and this work is not repeated here.

Whilst this report was being finalised, the world underwent a dramatic transformation with the COVID-19 pandemic threatening the lives and livelihoods of millions of people. This has emphasised the increasing dependency of citizens, businesses, public authorities and scientific research on digital tools and data. Those with access to these tools and the skills and capacity to use them are generally better equipped to cope with the massive disruptions that the pandemic is creating. Data-intensive science is proving to be critical for mapping the course of the pandemic and informing policy-making. Integration of data and information from many different scientific domains and new software development are important for modelling and assessing the longer-term socio-economic effects. Digital tools are playing a major role in fundamental research and international collaboration to understand the functioning of the virus and develop diagnostics, therapies and vaccines. At the same time the capacity of some of these tools to collect, link and analyse personal data raises new ethical and legal concerns, with important implications for science and society.

The global research community has responded rapidly to COVID-19 to develop digital platforms that facilitate access to research methods and outputs, not only to other researchers but also to government, industry and the broader community. In as much as these initiatives are new, they also demonstrate that in other contexts the infrastructure required to provide open access to critical research is not yet in place. The success of the scientific response to COVID-19 largely depends on the existence of resources in appropriately curated, interoperable, and preserved states. A skilled workforce is needed to create and utilise these resources. This requires both digitally skilled research support professionals and digitally skilled researchers. Several of the case studies included in this report, including the ELIXIR bioinformatics network, are at the forefront of the research response to COVID-19 and there are important lessons that can be learned from their experiences in strengthening digital skills and capacity.

The COVID-19 pandemic highlights the importance and potential of data intensive science. All countries need to make digital skills and capacity for science a priority and they need to work together internationally to achieve this. To this end, the recommendations in this report are even more pertinent now than they were when they were first drafted in late 2019.

Acknowledgements

An international Expert Group (EG, annex 1), was established through nominations from GSF delegates, to oversee and implement this project. The EG was chaired by Michelle Barker (Australia). This final policy report is the product of that Group's work. It was drafted by the chairwoman, with input from all EG members and support from the OECD-GSF Secretariat, Carthage Smith and Yoshiaki Tamura.

In addition to the EG members, a number of other experts made important contributions. This included the project leaders for the 13 case studies who kindly shared their valuable experience in interviews and provided comments on the final report. Many of these project leaders participated also in the dedicated project workshop, held in Cologne, Germany in October 2019 (see annex 3). The workshop was generously hosted by the Leibniz Institute for the Social Sciences (GESIS) with thanks to Christof Wolf (GESIS President) and Ingvill Mochmann (EG member).

Table of contents

Tables

Figures

Boxes

Executive Summary

As we move into a new era of data-intensive science, expectations are high. Big data analysis and access to new forms of data are already providing important and novel scientific insights and driving innovation. As the ability to combine and analyse data from different domains increases, complex societal challenges, including those embedded in the United Nations' Sustainable Development Goals, are becoming more amenable to scientific analysis. Data-intensive science has the potential to generate the new knowledge that is necessary to inform transformations to more sustainable and prosperous futures..

Digitalisation is transforming all fields of science, as well as other fields of scholarly research. In addition to opening up important new directions and avenues for research, it presents an important opportunity to increase the transparency, rigour, and integrity of research. Digital technologies and access to data are driving change, but human digital capacity and skills are likely to be the critical determinant of scientific success in the future.

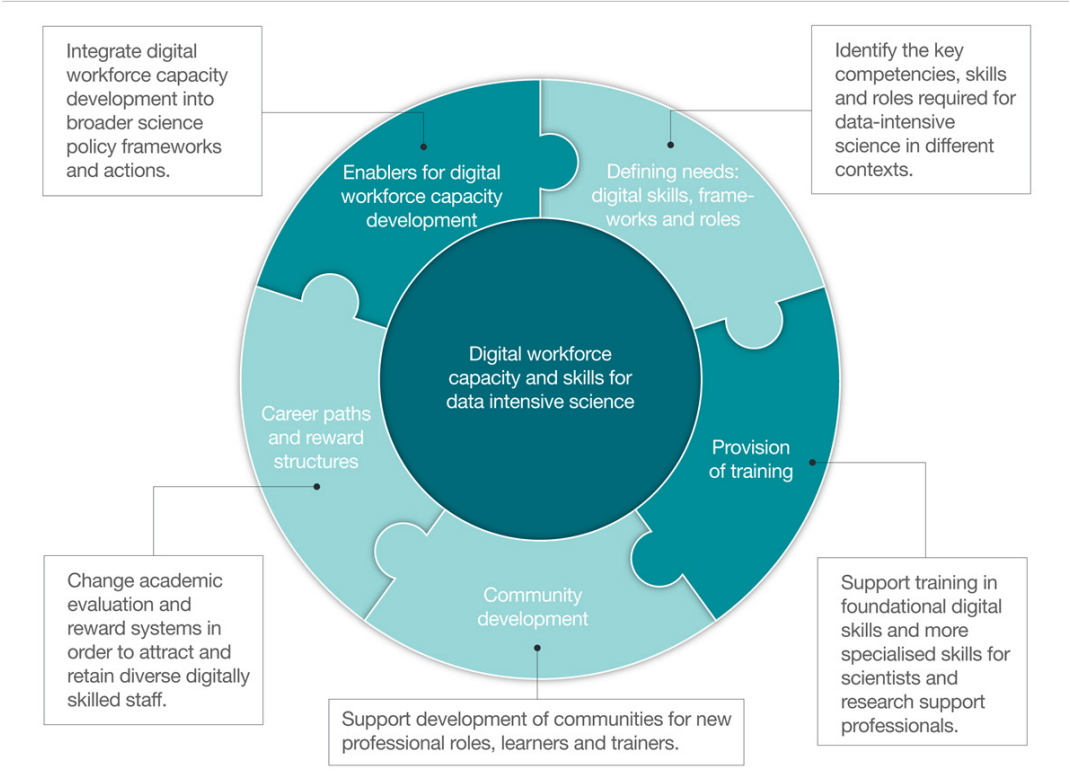
This report builds on recent work from OECD and other organisations that analyses the digital workforce capacity that will be required across different industrial sectors, and examines the specific requirements of data intensive science. This encompasses all fields of science, including social sciences and humanities..

The analysis, which includes thirteen in-depth case studies, explores what actions to date have led to improvements in digital workforce capacity for data intensive research, and what further actions are required. This includes an assessment of how the digital workforce requirements for science differ from other sectors of society and the economy, concluding that there are unique conditions in science that are reflected in specific skills requirements. Some of these requirements are generic for science as a whole and some are specific to disciplines or domains of research.

There is a need for both digitally skilled researchers, who have a common set of foundational digital skills coupled with domain-specific specialised skills, and a variety of professional research support staff, including data stewards and research software engineers. Increasingly research is conducted in teams and the distribution of competencies within a team is highly variable, so it is difficult to be prescriptive as to what should be expected of researchers and what can be best provided by support services. This will vary across different research domains. However, it is increasingly recognised that data intensive science requires not only technical skills but also people-focused skills, such as communication and team working. In many fields there is also a need for ethical and legal expertise, particularly when sensitive data is being used.

There are five key action areas that need to be addressed in parallel in order to build and maintain digital workforce capacity for science (see FIGURE ES1 below). Multiple actors, including governments, research funders, science associations, research institutions, and universities, need to work together across these areas.

Figure ES1. Five key action areas and goals for digital research workforce capacity development



Source: authors' design.

Important actions that national governments can take include:

- Recognising at the policy level the need for a digitally skilled workforce in research, and the importance of strategic planning that integrates the five key areas that are necessary to build and maintain this workforce: defining needs; provision of training; community building; career paths and rewards; and broader enablers.
- Analysis of national digital workforce capacity needs and the status (or preparedness level) of the research ecosystem to provide the training and other actions necessary to meet these needs. In this context, it is important to take account of international as well as local developments, and to consider how government actions can most effectively leverage these.
- Facilitation and coordination of the efforts to build workforce capacity at the speed and scale necessary to optimise the benefits of data intensive science, including monitoring and assessment processes that keep pace with a changing landscape.

Actions that research agencies can take include:

- Ensuring that funding mechanisms support digital workforce capacity development
- Aligning strategies and investment for physical research infrastructure and for digital workforce capacity

There are also a range of broader policy actions that national governments and/or research funders can take, with regard to education, open science, scientific integrity and research

evaluation and assessment which provide an enabling environment for data intensive science and reinforce efforts to strengthen the digitally skilled research workforce.

Universities are the main centres of tertiary education, training, and public research in most countries and hence have a central role to play in building sustainable digital research capacity. Whilst universities have considerable autonomy, they are also responsive to the mandates and incentives that governments and research agencies provide. Universities can take a number of actions in each of the five key areas that need to be addressed to strengthen digital workforce capacity and skills for data-intensive science. These include:

- Provision of training for scientists and research support staff; and
- Development of new career paths with appropriate evaluation, recognition and reward mechanisms.

A more detailed description of actions that universities can take is given in chapter 7. Some of these actions can be built on existing structures, e.g. university libraries can provide a focus for facilitating the development of data management skills and computing departments can help to propagate software and coding skills across the research endeavour. Other actions will require more systemic structural and cultural changes.

A number of other actors, including science associations and academies, research institutes, and research infrastructures have an important role to play, particularly in relation to community building and training provision and more detailed recommendations for all actors are included in Sections 6 and 7 of this report. There is also an important role for private sector actors to play both in the provision of training and in working together with public sector partners to define and address digital research capacity needs.

Many countries and institutions are already implementing some of these recommendations and there are considerable opportunities for mutual learning. A general recommendation for all actors is to engage in international collaboration in this area and share materials and experiences.

Building Digital Workforce Capacity and Skills for Data-Intensive Science

1. Introduction

The digital age is changing the practice of science[1] and all fields of research are increasingly data dependent. At the same time, there is a strong move towards open science, which is being enabled by digitalisation, and new challenges are emerging with regards to ensuring the rigour and integrity of science. These changes are happening rapidly and present a challenge for workforce development, particularly in scientific domains that have historically been less data-rich. Digital workforce capacity is required at multiple levels, including: individual scientists, research teams, data service providers, research infrastructures and institutions. There are different professional roles emerging, including multiple types of "data scientist", some of whom are supporting research and others who are actively involved in conducting research. Different fields of research require different types and levels of digital expertise.

It has been estimated that up to 5% of the scientific research budget needs to be dedicated to the management of FAIR (findable, accessible, interoperable and re-useable) data and that 1 in 20 staff in the research workforce should be digitally skilled research support professionals (Mons, 2020[1]). In Europe alone, this means about 500 000 professionals of various kinds to support researchers through experimental design and data capture, curation, storage, analytics, publication and reuse.

Many activities are taking place in different research domains and different countries to address perceived gaps in digital workforce capacity; however, they are disparate and largely uncoordinated activities, and more strategic approaches may be beneficial. Whilst there has been a lot of focus on digital skills, frameworks and roles, other areas including provision of training, community building, career structures and reward mechanisms are equally important. There has also been a tendency to focus on technical skills, although it is becoming increasingly apparent that data intensive science is a team game and people-focussed skills will be an important element for success in the future. Likewise, in the new era of open science, ethical and legal issues will continue to move from the periphery to the core of research and corresponding capability will need to be developed.

There are separate but related policy discussions on the digital skills and training required for scientists in the age of open science, and the new incentive systems and career paths that will be required for various types of staff. Many training initiatives are already being provided by a variety of public and private actors and steps are being taken to define professional roles and new career paths, but it is not clear that these are adequate to meet immediate and future needs. It is necessary to consider human digital capacity for science more systematically in order to identify gaps and good practices, and to develop effective policy. The emergence of Artificial Intelligence (AI) as an enabling technology that is likely to become pervasive across much of science exemplifies some of the urgent challenges, including the importance of data ethics, and the need for a balance between cooperation and competition between academia and industry for skilled personnel.

This report has been produced by an OECD Global Science Forum (GSF) Expert Group on Digital Skills for Data-Intensive Science (members are listed in Annex 1) and aims to provide information and options for policy makers to strengthen the development of a digitally skilled workforce for science. One of the first tasks was to agree on a common

language and definitions of key terms and concepts as used throughout the report. These are listed in a Glossary at the end of the report. The main report begins with a short description of the digital science ecosystem, the importance of a digitally skilled workforce for data-intensive science and what is needed to build and sustain this. An introduction to the thirteen in-depth case studies that were undertaken to inform this work is then provided as the basis for the analysis that follows. This analysis focuses on five key action areas:

1. Defining needs: Digital skills, frameworks and roles
2. Provision of training
3. Community building
4. Career paths and reward structures
5. Broader enablers of digital workforce capacity

Utilising this analysis, five corresponding goals are identified. Multiple actors have roles to play in achievement of these goals, and specific recommendations are made for national governments, research agencies, professional science associations/academies, research institutes and infrastructures, and universities and libraries.

2. What is known about the digital workforce needs for data-intensive science?

A number of recent studies have examined labour market demand and digital skills. However, this work has largely focused on broader societal needs and not specifically on the needs of science. This section examines the type of broader analysis that is available, including on the economic importance of a digitally skilled research workforce, and details the small amount of specific work undertaken on supply and demand issues in the research sector. It is clear that there are gaps in current needs assessments even taking into account that the rapid evolution of technology and research can make it difficult to predict future need.

The need for a digitally skilled research workforce

There are number of studies that analyse digital skills in the context of the labour market, often in conjunction with the role of education systems. These commonly identify the need for increased digital literacy across many sectors of the workforce, and for both digital and people-focussed skills. An example of the type of broad data that is available is Piatetsky's 2018 analysis of work examining increasing demand based on job advertisements, which noted that IBM claimed "that by 2020 the number of Data Science and Analytics job listings is projected to grow by nearly 364 000 listings to approximately 2 720 000" (Piatetsky, 2018[2]).

With specific regard to science, a recent report from the UK on *Dynamics of data science skills* (The Royal Society, 2019[3]) notes that "with major industry players hiring many of the most experienced data scientists and AI researchers, media reports have suggested that the natural flow of researchers from academia to industry may be reaching unsustainable levels." The report's analysis of job advertisements shows that in 2013, 6.7 million UK-based job postings in relevant digitally skilled roles were listed. In 2017 there were 9.2 million postings (an increase of 36%).

Several recent OECD GSF reports on open science highlight the need for digitally skilled professionals and professions. This includes reports on research data repositories, international research data networks, and open and inclusive collaboration in science (OECD, 2017[4]; Dai, Shin and Smith, 2018[5]; OECD, 2017[6]). These studies recognise the significance of expertise across all scientific domains in key areas such as data curation and

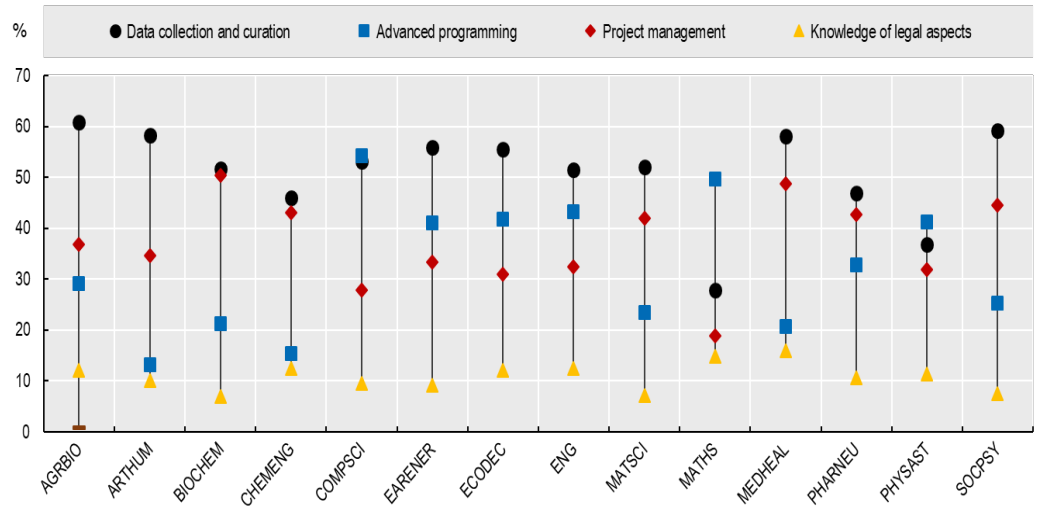
stewardship, data analysis and software development, leadership, teamwork, communication and problem-solving. Similarly, at the European level there have been calls for "an unprecedented effort in training of data scientists as a specialised profession and of broad data literacy to enable more and more users of the data system" (European Strategy Forum on Research Infrastructures, 2018[7]).

Charting the digital transformation of science (Bello and Galindo-Rueda, 2020[8]) includes a detailed analysis of the nature and effects of digitalisation on scientific research. The report is based on an analysis of ~10 000 individual responses to the 2018 OECD International Survey of Scientific Authors (ISSA2), a global online survey of that explored four distinct facets of digitalisation in scientific research:

- Adoption of digital tools and digitally-enabled practices throughout all stages of the scientific process
- Digitally-enabled diffusion of data and code
- Use of advanced and data-intensive digital tools to gain insights and develop predictions
- Development of digital identity and online communication of scientific work.

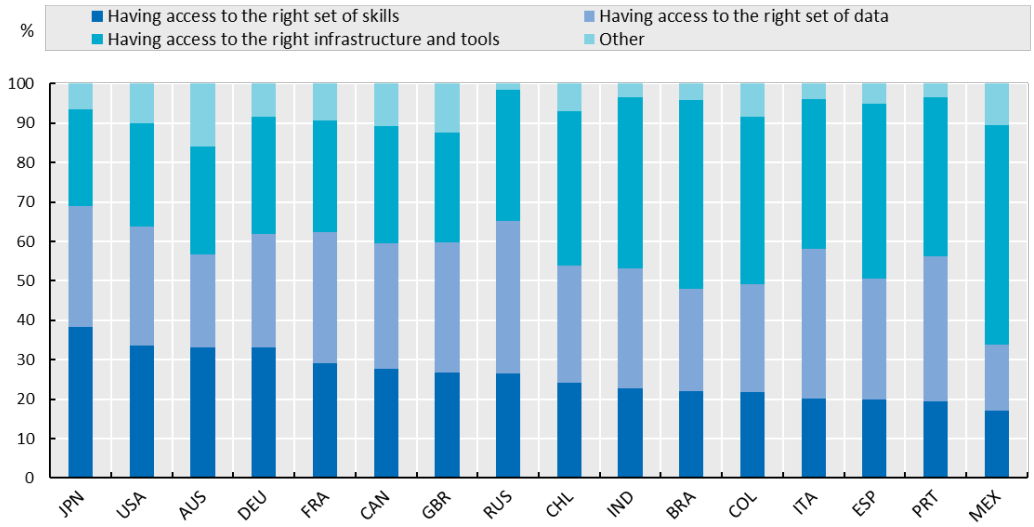
These facets are used to map how different communities, fields, sectors and even countries are responding to the digitalisation of research. The results provide empirical evidence that digital workforce capacity and requirements vary considerably across, fields or disciplines (Figure 1) and countries (Figure 2). It is also apparent from this survey analysis that skills profiles are significantly influenced by factors such as age and gender.

Figure 1. Digital skills requirements as perceived by different disciplines



Most important skills for scientific authors' research work. Percentage of authors in different disciplines, who deem each type of skill as important. Note: Weighted estimates based on sampling weights adjusted for nonresponse.
Source: Bello and Galindo-Rueda (2020), based on the OECD International Survey of Scientific Authors 2018. http://oe.cd/issa.

Figure 2. Challenges for data-intensive research in different countries



Most important challenge faced by scientific authors, by country of residence. Percentage of authors within each field selecting the relevant option. Note: Weighted estimates based on sampling weights adjusted for nonresponse.
Source: Bello and Galindo-Rueda (2020), based on the OECD International Survey of Scientific Authors 2018. http://oe.cd/issa.

There are a number of estimates of the potential economic returns from enhancing data access and use in research (which requires a digitally skilled research workforce). The recent publication, *Cost-benefit analysis for FAIR research data* (European Commission, 2019[9]), estimates that the overall cost to the European economy of not having FAIR research data is EUR 10.2 billion per year in Europe. Earlier Australian studies also indicate the value of increasing skills to maximise research outcomes, with a 2014 study estimating that the value of data in Australia's public research to be at least USD 1.9 billion and possibly up to USD 6 billion a year (Houghton and Gruen, 2014[10]).

Supply and demand issues for digital skills in data-intensive science

Whilst the higher education sector has well-developed processes for analysing supply and demand for some types of skills to inform their offerings, very little work has been undertaken specifically on digital skills for those who remain in the science research sector. Accurately predicting supply and demand figures for the science sector is not straightforward (Ashley, 2016[11]). However, there are a limited number of studies, which all point in the same direction, i.e., that demand far exceeds supply for digital training and skills. Recent studies, conducted in 2106-2018, include:

- An analysis of US National Science Foundation principal investigators in biological sciences, which identified that the most unmet need was training**:** "These needs suggest that NSF, universities, and other institutions have done a fantastic job at providing physical computational resources but haven't provided some of the necessary catalysts for their effective use" (Barone, Williams and Micklos, 2017[12]).
- Australian analysis of bioscience infrastructure elements, which categorised researchers in terms of their data and technique intensity and found that significant changes were expected to occur over the next five years as more bioscience

researchers require data-intensive and bioinformatics-intensive skills (Lonie and Francis, 2017[13]).

- Japanese analysis of research data management skills in research libraries, which revealed that 41% of staff were not familiar with data management planning (Kurata, Matsubayashi and Takeda, 2017[14]).

There is limited data on the numbers of digitally skilled research support professionals in science and reliable data on the demand for these roles in not available, partially reflecting the challenges in defining clear roles and variability in needs across different research domains (discussed ahead 5.1). An analysis of data librarians at R1 universities (doctoral universities with very high research activity) in the US in 2019 concluded: "Around a quarter of R1 universities don't have any dedicated data librarians on staff. Another quarter have only one dedicated staff member. Around a third have a small team of two to three data librarians, while the remaining small proportion of R1 libraries have large research data teams of four to ten. The average number of data librarians per R1 university is a little over two" (Springer, 2019[15]). A recent analysis of the Australian research-IT support workforce (Buchhorn, 2019[16]) calculated the number of equivalent-full-time (EFT) staff providing digital support to researchers as follows:

- 1 EFT per 60-70 researchers for collection and analysis of research data
- 1 EFT per 90 researchers for management or stewardship of research data
- 1 EFT per 250 researchers providing training/advice on research data
- 1 EFT per100 researchers involved in various software-engineering roles
- 1 EFT per 200 researchers providing training and advice on software-engineering.

In the European context it has been proposed that 1 in 20 members of the research workforce should be digitally skilled research support professionals, which supports anecdotal evidence that the present ratios are inadequate. It is suggested that in Europe alone, at least 500 000 professionals of various kinds will be required to support researchers, and that up to 5% of the scientific research budget needs to be dedicated to the management of FAIR data (Mons, 2020[1]).

There is a need for more information on supply versus demand for digitally skilled research staff, both broadly speaking and in specific domains. The importance of this type of information is illustrated by the fact that it was a primary driver for the development of several of the training initiatives that were included as case studies in the current work:

- Canadian company, Element AI, reports that Canada was training >500 graduate students in AI in 2018-2019 in universities (Kiser and Mantha, n.d.[17]). However, it estimated that thousands are needed. Canada's aim to increase the current levels led to the development of the Pan Canadian AI strategy, to address both lack of training and the effects of international mobility on reducing the talent pool. The Canadian Institute for Advanced Research (CIFAR, see ahead, table 1) is leading the implementation this strategy.
- In 2014, the Software Sustainability Institute (SSI) completed a survey of 15 000 randomly chosen researchers from 15 different research-intensive Russell Group universities in the UK. More than 90% of respondents acknowledged software as being important for their own research, and about 70% of researchers said that their research would not be possible without software. This highlighted the need for professional skills in research software development, which became a key focus of SSI.

- The Carpentries training is built on identified need, such as a 2013 Australian bioscience community survey, in which "more than 60% of researchers surveyed said that their greatest need was additional training, compared to a meagre 5% who need access to additional compute power …" This sentiment is shared by bioscience researchers in other countries and has been clearly identified by ELIXIR in Europe" (Teal et al., 2015[18]).

In conclusion, it is important that more nuanced analysis of supply and demand issues for digital skills in data intensive science be undertaken. In this context, *Measuring the Digital Transformation: A Roadmap for the Future* (OECD, 2019[19]) provides suggestions on how new insights into supply and demand issues could be gained, by exploiting and harmonising detailed national surveys on tasks and skills and by working with the business community to define new metrics of skill shortages.

3. A digitally skilled society and a digitally-skilled workforce for science

A key question for this project was whether any issues relating to the digitally skilled workforce are unique to science. This section examines the context around the research sector that creates unique conditions and how these are reflected in specific skills requirements that extend beyond those described in generic digital skills frameworks. Scientific research itself is heterogeneous and the differences between different research domains are explored also.

Public research systems have typically developed in different ways to research in other sectors, creating their own unique conditions. A recent analysis of digital workforce issues in Germany (German Council for Scientific Information Infrastructures (RfII), 2019[20]) observed that, "the scientific labour market is subject to conditions that differ from those for research and development in the digital economy or in industry. Some of these conditions are helpful to promote change and innovative responses to the digital transformation while others tend to hinder them." These conditions will be examined in more detail in Section 5, and include barriers to movement between academia and industry, and academic salaries and career paths that are often not competitive with those in the commercial sector.

Analysis of the thirteen case studies that are described in the next chapter shows that the science endeavour often frames digital workforce issues in the context of new requirements, such as open science, ensuring integrity and reproducibility in data intensive research or the ethical use of algorithms and data. These requirements are sometimes perceived as specific to science, or at least as more important in science than in other sectors, although in practice this is not necessarily the case. For example, whilst requirements for open science may be specific to public sector research, the requirements for both reproducibility and ethical practice are shared by researchers in both the public and private sector.

This project analysed whether different skills are required by science relative to society at large. An evaluation of the competencies encompassed in the European Commission's Digital Competence Framework for Citizens (DigComp)[2] was undertaken to assess the relevance and adequacy of these competencies for science. DigComp is one example of a framework that is considerably broader and more general than those which focus on the research or data-intensive domain. It consists of five areas that categorise the key components of digital competence, and each area contains general competencies that can be usefully applied to research:

- Information and data literacy: Browsing, searching and filtering data; critically evaluating credibility and reliability of data sources; organising and storing data.
- Communication and collaboration: Sharing data; knowing about referencing and attribution practices; using digital tools and technologies for collaborative processes; protecting one's reputation.
- Digital content creation: Creating new, original and relevant content and knowledge; understanding copyrights and licenses; programming and software development.
- Safety: Protecting personal data.
- Problem solving: Customising digital environments to personal needs; using digital tools to create knowledge and innovate processes; identifying digital competence gaps and seeking opportunities for self-improvement.

The DigComp framework provides a sound basis to assess the overall skillset of those engaged in research. However, there are potential missing competencies, or competencies needing extension, in each area that are required in a publicly-funded research context (and possibly also in other research sectors):

- Information and digital literacy: Understanding of statistics to help evaluation and analysis of data; understanding of requirements for reproducibility.
- Communication and collaboration: Following open science principles to share data, information and content, engage in good digital citizenship, and improve collaboration; extend knowledge of referencing and attribution practices to research data and software citation/referencing; protecting academic reputation, both of one's own organisation and that of academic research more generally.
- Digital content creation: Visualisation of data and information to convey knowledge.
- Safety: Protection of sensitive data, understanding of tools and techniques such as delinking, anonymisation and safe havens.

There are also differences in the mix of skills required in research, which is affected to some degree by the discipline/s in which the research occurs, and the size, type and variety of data they produce and consume (Figure 1). Some disciplines have had a strong computational base for decades, resulting in undergraduate courses that produce researchers with a high level of relevant digital skills. High-energy physics, astronomy, bioinformatics and medical informatics are commonly used examples here, but even in these disciplines, researchers may need additional training and access to digitally skilled research support staff. In contrast, some disciplines are in the early stages of integrating computational approaches into their discipline, and these researchers can lack even basic digital research skills. The maturity of a science field in handling large datasets can also affect the size and nature of the digital capacity gap, with fields with a more recent increase in their computational base possibly requiring a substantive investment in basic digital skills. It is important to understand the disciplinary differences that exist, as this affects approaches to generic training (or training in foundational skills) for all researchers.

Between disciplines, some differences in the skills required also emerge due to the nature and scale of data produced by individual research projects, and the degree of data integration necessary to produce datasets which can answer different research questions. High-energy physics typically produces large datasets which present technical challenges in distribution and analysis, but which do not require integration with data from many, if

any, other sources. Clinical research can benefit from integrated analysis of the results of a modest (10 - 100) number of clinical trials, each of which themselves may have had to integrate data from a variety of clinical settings. Bioinformatics will often integrate the result of hundreds of thousands of experiments carried out by thousands of researchers, and the shift to digital astronomy makes large integrated studies possible which could never have been carried out when telescope images resided on photographic plates. The curation of integrated data collections in these latter fields require skills that are less relevant to high-energy physics. How these skills requirements map onto professional roles and composition of research teams is highly variable and is discussed ahead in section 5.1.

In summary, there are unique digital skills or competencies (or levels of competency) that are required for science, although these are mainly an expansion of the basic digital skill set that is required across all sectors of society and industry. These skills relate not only to the performance of data intensive-science but also to other related drivers such as the shift towards open science and need to ensure reproducibility of research. They are not unique to academic/public sector research and many of them are equally relevant in the private sector. However, the mix of required skills varies across scientific disciplines and domains.

4. The science ecosystem and case studies

The challenges for science in developing digital workforce capabilities reflect the ways in which the science ecosystem operates. Countries typically deploy a range of approaches for assessing and building digital research capacity that reflect their unique mix of actors, history and culture. These approaches are themselves situated in the broader research, societal, and economic context, which includes national digital and skills strategies and factors such as internet connectivity. This complexity raises interesting questions for science policy makers regarding how to identify strengths and weaknesses and gaps in their own systems, and how best to support or encourage the most relevant initiatives.

The thirteen case studies that were undertaken as part of this project are introduced in this chapter. The case studies represent initiatives from different parts of the digital science ecosystem and a range of countries. All the case studies were led by public sector actors and focus mainly on academic research needs, although several work closely with private sector actors. The key features of the cases are summarised in Table 1 and each is then described in more detail.

Table 1. Case studies classified by key characteristics

	Structure (remit)	Target audience	Disciplines
Alan Turing Institute (Turing)	Research institution (national)	Doctoral students	All disciplines, some focus on Artificial Intelligence (AI)
American Data Science Alliance (ADSA)	University network (national)	Researchers at all levels and Research Software Engineers (RSEs)	All disciplines
Australian Research Data Commons (ARDC)	Government programme (national)	Researchers and research support professionals	All disciplines
Canadian Institute For Advanced Research (CIFAR)	Government programme (National)	High school to Early Career Researchers (ECRs)	All disciplines, focus on AI
The Carpentries	Community programme (International)	Researchers at all levels	All disciplines
CODATA-Research Data Alliance (RDA) School of Research Data Science	Community programme (International)	ECRs	All disciplines
Delft University of Technology (TU Delft)	University (local)	Researchers at all levels	All disciplines
ELIXIR	Intergovernmental organisation (International)	Researchers and research support professionals	Focus on bioinformatics
GESIS - Leibniz Institute of the Social Sciences (GESIS)	Research infrastructure (national)	Mid-career to advanced researchers	Focus on application in social sciences
Japanese Consortium for Open Access Repositories (JPCOAR)	Research network (national)	Librarians	All disciplines
Millennium Institute of Astrophysics (MAS)	Research institution (national)	ECR and RSEs	Focus on application in astronomy
RSE Association	Professional association (international)	RSEs	All disciplines
Software Sustainability Institute (SSI)	Government program (national)	Researchers at all levels and RSEs	All disciplines, focus on software

More details are provided below on the thirteen case studies, classified as international, national, disciplinary or university/multi-university initiatives.

International initiatives

International initiatives are able to leverage expertise and support from multiple countries and can be a critical adjunct and/or support for developing and implementing national and institutional strategies for digital capacity. The international initiatives included in the case studies encompass both inter-governmental and community-led initiatives.

I. The Carpentries: This international program delivers training in the basics of data analytics and programming for researchers and those working with them. It has transitioned from being a grant-funded activity to a membership subscription model. Materials are all open access, and the role of the central organisation includes managing this material as well as validating instructors. The materials (Library Carpentry, Software Carpentry and Data Carpentry) are generic, but

training events are provided in the context of the learner's discipline and background.

II. CODATA-RDA School of Research Data Science: This is an international initiative delivering two-week training events in data science skills at locations around the world and also producing open, reusable training material. It focusses on the needs of students from low and middle-income countries and course topics are selected through discussions with researchers from these areas. It uses some material from the Carpentries and follows other elements of that model; schools are delivered by a mix of roving tutors and local staff, and students are encouraged to become instructors in future. Content is not specific to low and middle-income countries; demand exists in high-income countries and, in the longer term the aim is to leverage support for events in OECD countries to support delivery elsewhere.

III. ELIXIR: ELIXIR is an international distributed research infrastructure for life sciences data and training is coordinated via one of five technical platforms. Training empowers researchers to use ELIXIR's services and tools through developing skills for managing and exploiting data. ELIXIR provides open materials on a portal, events, a toolkit, and train-the-trainer activities. As well as training development and delivery, ELIXIR builds community capacity by assisting knowledge transfer from more mature national activities to newer entrants.

IV. RSE Association: This is a UK professional membership organisation established by the SSI but now independent of it. It organises regular events to allow RSEs to meet, exchange knowledge and collaborate and works to raise awareness of the role of RSEs and improve career paths for them. The model is now being taken up in other countries to develop an international RSE network.

National initiatives

A number of the case studies are national government-supported initiatives and help to illustrate the range and potential impact of direct government actions in this area.

I. ARDC: ARDC is a national initiative to support Australian research, with skilled workforce development part of one of its five areas of activity (people & policy) As a research infrastructure itself, ARDC recognises that optimising the use of research infrastructures requires skills that not all will have. It facilitates in-person and online training and development of communities of practice. Activities include train-the-trainer programs and leadership of digital skills training initiatives for research software professionals.

II. CIFAR: Training activities delivered by CIFAR, under the Pan-Canadian Artificial Intelligence Strategy, focus on maintaining and improving national strengths in AI in Canada. Training ranges from specialised techniques for able graduate students, to work with high school students. Other activities include workshops and conferences. Many activities concentrate on under-represented groups, including young women, and at least one activity includes students from developing countries. By supporting the research activities of a network of AI institutes and Chairs, the Strategy helps create a rich training environment that promotes attraction, retention and development of AI talent in Canada.

III. JPCOAR: The Research Data Management Training Materials Development Project of the Working Group of Research Data of JPCOAR aims to promote research data management in Japan, in partnership with the National Institute of Informatics (NII). These organisations are working together as part of a broader initiative to promote open science. This work began with a focus on data skills for

librarians but has since broadened its reach. The project develops course materials and runs courses based on them and this involves reuse and translation/adaptation of open source training materials from other sources.

IV. SSI: SSI is supported by research funders in the UK to maximise utility of investment in research software; skills development for RSEs and researchers generally are one part of its activity. Skills development is carried out through a mix of activities, including workshops and a fellowship programme. There is a strong focus on career paths and professional recognition for RSEs, and training addresses the demand from this group for both technical and people-focussed skills (e.g., team leadership).

V. Turing: This national UK initiative on data science is intended to conduct and apply research, develop future leaders and act as a national voice and focus. Most activities are focussed on PhD students, either in partnership with universities who offer a formal place at the Turing or are collaborators. e.g., Centres of Doctoral Training. There is a strong cross-disciplinary focus and trainees are expected to have already acquired foundational skills, with the Turing providing specialised training in areas such as AI that is tailored to specific research questions.

Discipline level initiatives

Discipline level initiatives can be international, regional and/or national in focus. There were three initiatives chosen as case studies that had a specific disciplinary focus [ELIXIR (see III above) can fit also under this classification].

I. GESIS: GESIS has a long-standing national role in archiving social survey data and in research on survey methodology and computer science. Data sources for this type of research are changing – personal devices, smart homes – and GESIS is refocusing several activities, including those related to skills, in recognition of this. GESIS runs on-site training days open to all as well as longer workshops and summer schools in computational social sciences. The GESIS training program focusses on the specialist research areas where it has the greatest expertise/can make most difference. At the same time, GESIS is a member of the Consortium of European Social Science Data Archives (CESSDA), which runs training courses in foundational digital skills for social science.

II. Millennium Institute for Astrophysics (MAS): MAS is a National Chilean initiative focussing on the skills necessary to address specific challenges in astronomy. The skills required in the emerging sub-discipline of astro-informatics include real-time event identification in high-volume data streams and have wider potential application beyond astronomy. MAS runs a mixture of activities including workshops and graduate programmes, with a focus on ECRs. It has its own legal status but has strong links to several national universities and is working also to strengthen private sector links.

University or multi-university level initiatives

Universities have a critical role to play in building capacity for data intensive science and university staff were centrally involved in all the case studies (that is, not only those case studies initiated by universities). Two University-led case studies were included.

I. ADSA: A younger initiative than many of the other case studies, though building on earlier funded work in a small number of US universities. Its plan is to support teaching and use of data-intensive tools and techniques throughout research and education. It has a variety of delivery mechanisms, including hack-weeks, mini-workshops and grants for alumni of the preceding programme (Moore-Sloan Data

Science Environments), typically offered on the campus of an affiliated data science institute or organisation with a University.

II. TU Delft: TU Delft's activities are university-led and focussed on its own data and staff. By funding embedded data stewards and appointing researchers as data champions it spreads skills through peer networks as well as with training events and online learning facilities. The focus is on tasks other than data analysis, such as preparation, curation and management as well as on translating generic policy into discipline-specific actions. Skills developed are thus a mixture of generic and discipline specific.

Key personnel from each of the case studies were interviewed via videoconference by one or two members of the Expert Group utilising the interview questions in Annex 2 B. Interview questions focused on issues, challenges, and good practices that were likely to have policy implications. The case studies also contributed to an international workshop on 28-29 October 2019, hosted by GESIS in Cologne, Germany (Annex 3). This workshop facilitated interactive discussions and exchanges of experiences and perspectives that informed the policy messages and recommendations that are included in this report.

5. What is needed to build a digitally skilled research workforce?

A digitally skilled workforce is required to complement investments in data-intensive scientific infrastructure. An inclusive workforce needs to reflect the diversity of society in order to not only improve science but to ensure it is more responsive to societal needs. This workforce needs to be built, nurtured and maintained over the long-term, and as a goal in itself. To achieve this advancements are needed in five key areas: defining needs; addressing needs; scaling up; ensuring sustainability and facilitating change. Each of these areas is explored in this chapter.

5.1. Defining needs: Digital skills, frameworks and roles

As described in section 2.2, there appears to be a substantial unmet need for training in digital skills and overall digital capacity for science in many countries and research fields. This sub-section delineates approaches to identifying the key digital skills needed for data intensive science, including the use of frameworks for digital competencies and how these link with new emerging roles. Understanding of these has particular value at a policy-making level, allowing appropriate interventions to be more effectively designed and targeted.

Digital skills for science include both specialised and foundational skills. Specialised digital skills needs differ across scientific domains and evolve over time in response to the state of the discipline and to new technologies and policy requirements. At the same time, there is a baseline of more generic foundational digital skills for research that evolves more slowly and should ideally be embedded in undergraduate science education but currently is not. These foundational skills need to incorporate the broad concepts and processes associated with Open Science and the social accountability of research.

The digital skills that are commonly listed as being required for data intensive science are sometimes placed into distinct categories. However, the terms and categories utilised vary considerably. For example, Turning FAIR into reality: *Final report and action plan from the European Commission expert group on FAIR data* (European Commission, 2018[21]), defines the two major areas of competencies as data science and data stewardship, with the following definitions:

- **Data science skills** can be understood as comprising knowledge from computer science, software development, statistics, visualisation and machine learning. This also covers computational infrastructures and knowledge of information modelling, algorithms and information integration.
- **Data stewardship** is a set of skills to ensure data are properly managed, shared and preserved throughout the research lifecycle and in subsequent storage. During the active research process, this could involve data cleaning to remove inconsistencies in data sets, organising and structuring data, adding or checking metadata, and resolving data management issues.

By extrapolation, one might conclude that data scientists and data stewards are needed, and that capacity building efforts should be focussed accordingly but this would be over-simplifying a complex relationship between skill and roles that plays out differently in different parts of the science community. Data science, and its derivative, data scientist, are particularly difficult terms for which to achieve consensus definitions, and their definitions and usage vary in different domains. Because of these ambiguities the term data scientist is of limited use from a policy or planning perspective. It is an umbrella term, but one needs to dissect out its components further to assess the status of digital skills in the science workforce. This will be returned to later with regard to Figure 3.

A number of initiatives have produced lists or frameworks for digital competencies (OECD, 2019[22]; European Commission, 2017[23]; FAIR4S, 2019[24]; Research Data Alliance, 2015[25]; Demchenko, Belloum and Wiktorski, 2017[26]; Molloy, Gow and Konstantelos, 2014[27]). With specific regard to research, the EDISON project was an early attempt build a comprehensive framework of skills and competencies for 'data scientists' (Demchenko, Belloum and Wiktorski, 2017[26]). More recent efforts have built on and extended the EDISON project's work to include competencies related to data stewardship, and as a result are seeing reuse in a wider context.

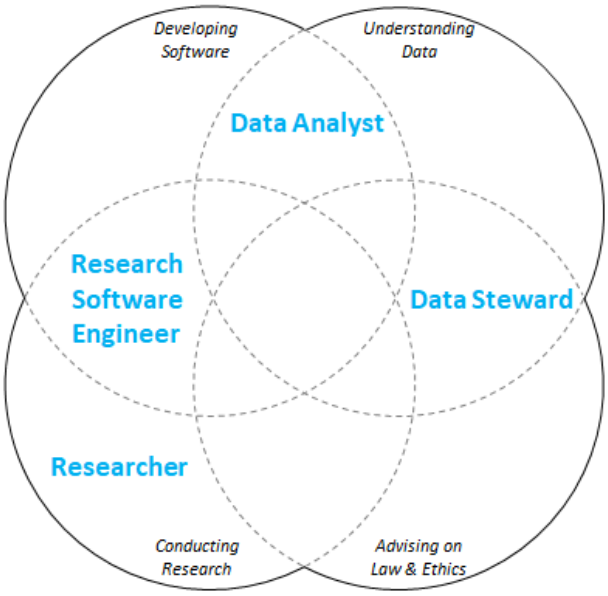
These frameworks vary in their depth and flexibility. More nuanced approaches are apparent in frameworks such as FAIR4S, in particular the recognition that a competency, when applied as a skill, can be done so at a variety of expert levels. In FAIR4S these levels are basic (awareness/comprehension), intermediate (ability to apply) and expert (ability to change practice in this skill.) A role profile then includes both skills and the levels of expertise expected in that role. Work has also been done to identify different types of data stewards in the life sciences, broken down into responsibilities and tasks (Scholtens et al., 2019[28]).

Despite their imperfections, frameworks for digital skills have particular value at a policy-making level. An appropriate framework can help to identify how many people with particular skills at a given level of expertise are required at different scales, from national to local. Some types of frameworks will also indicate how those skills can be acquired - for example, through formal education, self-paced learning or mentoring. This allows appropriate policy interventions to be more effectively targeted. A simplified representation of the main skills and roles that are required for data intensive research was developed at an early stage of this project. This 'meta-framework' was used to clarify the focus of individual case studies and is included with the interview questions in Annex 2. The meta-framework describes a set of roles relevant to data-driven research and activities requiring specific skills or groups of skills at different points in the research lifecycle. Whilst this provided a useful as a way to compare across case studies, the different cases that were interviewed in this work were inclined to modify it in different ways, often wanting to introduce more granularity to reflect the specificities of individual research

fields. At a minimum, this illustrates the role of frameworks in structuring the discussion around digital skills for research.

Skills definition can be used to identify emerging professional roles. Figure 3 illustrates the main areas that need to be addressed and combined for data intensive science (see also the Glossary at the end of this report). These areas include not only data and software but also related legal and ethical issues. The diagram shows four roles and the principle responsibilities of each (rather than the skills required in each). Hence, a data steward is responsible for data and needs to have a thorough understanding of the legal and ethical ramifications of its use, an RSE is responsible for software engineering based on an understanding of the research goals, and a data analyst is responsible for conducting analysis of the data based on an understanding of software. The researcher role is more difficult to demarcate and depends heavily on the research domain and the size and structure of the research team. In some cases, the researcher and data analyst roles may be the same, whereas in larger teams they may be quite distinct. It should be noted that, as for most attempts to categorise human capacities, this diagram is not definitive. The dotted lines within the diagram are used to denote permeability between overlapping roles and responsibilities.

Figure 3. Venn diagram of roles and responsibilities



Source: adapted by authors from an original diagram by Simon Hettrick (SSI) that was developed at the project workshop.

Whilst definition of professional roles can be important, there are limits to how prescriptive role definitions can, or should, be. As the need for a digitally skilled workforce for data-intensive science has continued to evolve, discussions are focusing more on the skills and competencies encompassed, and their placement within a team, rather than attempting to agree on specific job descriptions. In reality, roles such as, data scientist, RSE, data analyst, data steward, data manager, data librarian, or digital curator, encompass a range of competencies but people with these job titles have different skill sets based on their particular speciality. In small research teams, generalists may be needed, whereas larger teams may have more specialist requirements. In some environments, specialist skills may

be delivered by groups or service providers external to the research team, such as might be found in a data repository function in a research institution. Thus, all research groups need access to people with a range of particular skills but not all research groups need to have all skills embedded within them. This is even more true of the lone researcher who will often be dependent on an external technical infrastructure and skills to undertake and disseminate research.

Whilst recognising their limitations, defining skills required and roles (or 'labels') for certain skill groups is important not only for policy development and training but also for the establishment and recognition of professional communities and associated career paths. This is particularly important in academia, which is largely organised around disciplines in a way that is not necessarily compatible with digital workforce structures (see ahead, 5.3 and 5.4). Assessment of the skills needs for data intensive science, can also have broader implications for research governance. For example, the importance of ethics and legal issues has implications not only for researchers and research support professionals but also for the composition of Institutional Review Boards (Grant and Bouskill, 2019[29]; OECD, 2016[30]).

Box 1. Defining digital skills needs

The case studies provide examples of both systematic and informal approaches to defining skills and training needs:

- The Turing collaborates with UK universities to create a program framework that recognises the team approach to science, the need for community development, and the importance of people-focussed skills focused on collaboration. The resulting program brings doctoral students into the Turing community to work in collaborative environments, in alignment with Turing's ambitions to deliver new training and research opportunities, enhance the data science and AI skills agenda nationally, and support the next generation of data science and artificial intelligence leaders.
- CODATA-RDA School of Research Data Science provides an increasing number each year of two week-long courses each attended by 30-60 ECRs in low to middle-income countries. The course content draws heavily on the Carpentries and development has been guided by collaboration with students past and present, some of whom have become course tutors. Course content more strongly incorporates community identified needs than application of top down frameworks.
- In implementing the Pan-Canadian AI strategy, CIFAR emphasises the importance of engagement by all disciplines with Machine Learning (ML) and AI - to the extent that they be considered as part of the foundational digital skills requirement for research. CIFAR cites examples of progress in this direction, such as the University of Toronto's strong ML programme in its business school.[3] CIFAR also supports workshop and projects between researchers and machine learning experts with purpose to identify how AI and ML to could be applied in specific disciplines.
- GESIS, as both a research data infrastructure and research-performing organisation, has good insight into the requirements of researchers needing to use this content, and uses this to continually refine what the training that it offers. It also responds to more substantial changes in the data environment with targeted activity to develop material covering emerging areas such as data from smart devices.

5.2. Provision of training

Provision of appropriate training is essential for building digital workforce capacity. This section examines the variety of teaching and learning modes that can be utilised for training, the challenges in scaling up training efforts, the need for training to encourage diversity in the digitally skilled workforce and the role of the commercial sector in the provision of training.

A commonly cited issue in the case studies was the extent to which available digital skills training fails to meet supply, and the difficulties for organisations in scaling up to meet the ever-increasing demand. The people providing training are usually not certified trainers, and often do not receive recognition for their expertise and contributions in training. Several of the case studies utilise large volunteer workforces to provide training, raising concerns about long-term sustainability. These initiatives usually have limited budgets that

do not extend to paying for trainers, and training is usually offered for free. A recent Australian poll asking 79 research training initiatives on a variety of digital skills how well their supply could meet demand showed that demand exceeded supply for more than 75% of initiatives, with 9% stating that demand was more than 5 times higher than supply.[4]

There are many different teaching and learning modes that can be used to instil and upskill digital skills. Some of the digital upskilling needed for both scientists and research support professionals may eventually be achieved by changes to mainstream undergraduate education pathways. However, even if appropriate foundational skills become embedded in undergraduate science education, digital tools are evolving so rapidly that postgraduate training, both formally within Masters and PhD courses, and on a more responsive ad hoc basis, will surely be required. A range of types of training are available, ranging from more traditional coursework approaches through to Massive Open Online Courses (MOOCs), on-the-job training, hackathons, summer schools, conferences and drop-in sessions such as hacky hours (see case studies, section 4). In the ideal world, formal accredited courses and informal training would complement each other, with the former providing a good grounding in digital skills and the latter ensuring up-skilling as technologies and methods evolve. Given the speed of digital evolution, acquisition of digital skills for research must be supported by lifelong learning, as is being advocated across many sectors of the digital economy.

Scaling up of training is a commonly identified challenge, particularly for the community-led or grassroots organisations that are an important part of the digital research ecosystem. These organisations are often created to progress issues that mainstream organisations are not addressing, or to address these issues at a scale that can transcend organisational and/or national boundaries. Case studies for this project that fit into this category include the Carpentries, CODATA-RDA School of Research Data Science, and the RSE Association.

Capacity building efforts need to create a workforce that reflects the diversity of society in order to ensure that science is not only more productive but also more responsive to societal needs. There is substantial evidence of gender inequality in the digital world, with reports such as *Bridging the Digital Gender Divide* (OECD, 2018[31]) showing how policy interventions can help pave the way to greater inclusion of women. There is a fundamental role for education and training in bridging the (digital) gender divide. This needs to be part of systemic approach that includes: promoting ICT use, skills and learning, and empowering educators and making them active agents of change. There are now a range of initiatives that encourage gender diversity in digital roles (and more broadly in science, technology, engineering and mathematics). Other types of workforce diversity also beginning to be addressed through initiatives such as *the CARE Principles for Indigenous Data Governance* (Research Data Alliance International Indigenous Data Sovereignty Interest Group, n.d.[32]).

There is an important role for the private sector, including the not-for profit sector, both as an employer of digitally skilled scientists and as a partner with various public sector actors in supporting the development of this workforce. This is particularly the case in 'hot' fields such as Artificial Intelligence, where on-the-job training opportunities such as mentorships and internships are being adopted. There are commercial initiatives for training in digital research skills, such as DataCamp and the Khan Academy, that are an important part of the overall ecosystem. Multinationals such as Google, Microsoft and Amazon also engage in various ways, including provision of financial support and placements for ECRs, or involvement of students in activities, such as the Google Summer of Code.

Box 2. Addressing training needs for digital workforce capacity

Examples from the case studies include:

- TU Delft have partnered with the University of Edinburgh, Digital Curation Centre (DCC) and Research Data Netherlands to develop a MOOC on delivering research data management. The course teaches development and delivery of effective data management services to improve research in an organisation.
- The Carpentries uses a train-the-trainers model to train researchers with digital skills in educational pedagogy and inclusive teaching practices through their instructor training program. This prepares trainers to be able to effectively organise and teach Carpentries workmeshops. There are currently more than 2000 trained volunteer instructors who ran approximately 600 workshops around the world in 2019. Learners who go through Carpentries workshops value the training and approach and some go on to become instructors themselves. ELIXIR has a similar train-the-trainer programme that focuses on bioinformatics (Morgan et al., 2017[33]).
- The CODATA-RDA School of Research Data Science has a focus on the needs of low to middle-income countries, and aims to develop into an international network which makes it easy for partner organisations and institutions to run the schools in a variety of locations. An annual event is organised in Trieste, and other schools have been organised with local partner institutions in Brazil, Rwanda, Ethiopia, USA and Australia.
- The Carpentries also have a substantive focus on diversity, as part of democratising access to digital tools. Their Nine Core Values include inclusive of all, access for all, and strength through diversity. Actions to support these are contained in *the Equity, Inclusion and Accessibility Roadmap* (The Carpentries, 2019[34]).
- Promotion of gender equity and diversity in the digitally skilled workforce is one of CIFAR's overall aims. To this end, it supports specific training programmes for schoolgirls, to raise awareness of AI, and for young women.

5.3. Community building

Professional communities are an important part of the development of digital workforce capability, enabling diverse groups of people with common interests to come together to achieve change at various levels. The focus of these communities varies, ranging from standardisation of curricula, to sharing of best practice and mutual support, or influencing policy change around career paths. This section examines issues relevant to community building, including the need to support communities of trainers, learners, and leaders.

A number of professional communities have developed 'bottom up' and are increasingly gaining recognition for certain skill groups or roles, including RSEs, data stewards and librarians. From a policy perspective, the UK's Technician Commitment, is a good example of how communities can be supported elsewhere in science. It is a sector-wide initiative led by the UK Science Council to help address key challenges facing technical staff working in research and was created as a response to the demand for technicians increasing by 5% annually (Science Council, n.d.[35]). Universities and research institutions from across the UK have backed the Technician Commitment, which is ensuring greater

visibility, recognition, career development and sustainability for technicians across all disciplines.

Communities of trainers are valuable for a number of reasons, including knowledge transfer and mutual learning, enabling collaborative curriculum development and evaluation methods, and encouraging reuse of training materials. Trainers are usually volunteers and face challenges in receiving recognition and credit for their training work. Training and accreditation for trainers can also be a challenge. There are a number of initiatives that create and support communities of trainers as part of their train-the-trainer programs, and this is a core part of the Carpentries approach. In another example, researchers and research support professionals associated with the Social Sciences and Humanities Open Cloud (SSHOC) are now developing a SSHOC Train the Trainer Toolkit and establishing a SSHOC Training Network.

The existence, but lack of recognition, of this volunteer work force is a key problem in scaling up training efforts and is not often acknowledged. Scroggins and Pasquetto have undertaken analysis that highlights the invisible labour that underpins data science, noting that "behind data-intensive science's technological facade lies a bewildering array of human labour, some performed in the spotlight by star scientists, but most performed behind the scenes by the precariously employed in conjunction with computational machines." They identify invisible labour in five main areas: authoring, administering, maintaining, archiving and collaborating, arguing that "a full and nuanced understanding of data-intensive science can only be obtained by starting with the in situ work and labour of scientific practice in all its manifold forms" (Scroggins and Pasquetto, 2020[36]).

Communities of digital science leaders are also emerging, particularly through programs that emphasise development of leadership skills and the establishment of formal and informal networks for decision makers. It is important to bring together key influencers to share best practice and collaborate on international solutions, and these communities need to continue to grow. Open, Data-driven Science for Decision-Makers is a community-driven initiative to target decision makers in biological and biomedical sciences, who need to be able to understand the wider policy implications of open science and new developments in data intensive science. Training is focused on funding agencies/program managers, grant reviewers, investigators, and educators who review research proposals or training curricula that make use of data integration, software development, and large-scale computation.

Box 3. Community building for digital workforce capacity

Examples from the case studies include:

- ADSA supports academic data science leadership to share and promote the institutional changes needed to integrate data science into the full spectrum of university research and education. ADSA is a community building and networking organisation for academic data science practitioners that enables better sharing of ideas and solutions to common challenges, including facilitating and supporting academic science leaders to make institutional changes. ADSA program components include sharing of learnings through white papers and a data science community summit/conference; an annual meeting for leads of data science institutes on university campuses; and support and facilitation of Moore-Sloan Data Science Environments alumni to spread the development of data science practices across universities.
- SSI builds communities through activities such as the SSI Fellowships Program, **which provides funding for researchers who want to improve how research software is used in their domains and/or area of work**. While Fellows are funded for a specific term, they can continue to use this title and are linked to the continuously growing set of fellows, building a community that become more senior over time. Some SSI Fellows have now moved into leadership positions.
- An example of both successful community development and professionalisation of a new role is the work of the RSE Association, with chapters of this new professional society now found in Europe, USA, and Australia. These groups campaign for the recognition and adoption of the RSE role within academia along with the need for appropriate reward and career opportunities for RSEs.
- The CODATA-RDA schools are often attended by individuals who are working in isolated environments and the schools are sometimes the first opportunity they have to work with many others with similar skills and experiences. The students have thus established their own community network, enabling them to develop further. Although not a planned part of the schools the value of this community is recognised and supported by students and organisers alike.

Community-led initiatives can provide mutual support and learning and help to strengthen the essential role that individuals can play in the development of a digitally skilled workforce. Individual action is emphasised also in the Open Science Policy Platform established by the European Commission to encourage practical commitments for implementation of open science by different stakeholders. Both institutions and individuals are encouraged by the platform to consider what they can do to make open science happen, and to make commitments within their jurisdiction (Méndez, 2019[37]).

5.4. Career paths and reward structures

Building and maintaining the digitally skilled workforce that is needed for data intensive science requires attention to careers and reward structures. New roles are emerging for digitally skilled professionals in research, who may come from diverse educational

backgrounds, and appropriate human resources structures and policies are needed to support this.

There is a need for digitally skilled researchers as well as a new cadre of professional support staff, most notably data stewards and RSEs (see earlier section 5.1). In addition, research support professions that have historically played a critical role in scientific information management, such as librarians, archivists, and curators, are adopting the 'digital' prefix as they take on new roles and acquire new skills in coordinating and managing digital assets. Individual scientists may have some of the digital skills that might be expected of these support professions, and in different contexts they may assume these support functions as part of research teams, with varying degrees of recognition.

Pathways into the newly emerging digitally skilled roles vary. For example, some RSEs start off as researchers who spend time developing software to advance their research. Because they enjoy this part of their work and have invested in developing specialist skills, they continue to focus on software and its use in research. Others start off from a more conventional software-development background and are drawn to research by the challenge of using software to further research (Software Sustainability Institute, n.d.[38]). Recognising the value of each skill set and providing progression opportunities is critical. However, the long-term career pathways needed for these new professional research support roles are only emerging very slowly.

There is a need for incentive mechanisms both for researchers and research support professionals to encourage and reward acquisition and application of digital skills. There are a number of published reports that emphasise the lack of incentives and career paths for digitally skilled personnel in academia (Berente et al., 2018[39]; Working Towards Sustainable Software for Science: Practice and Experience, n.d.[40]).

As mentioned earlier in this report, there is considerable competition between the academic and other sectors for digitally skilled personnel, particularly in 'hot areas' such as artificial intelligence. With regards to training provision, different sectors are to some extent working together and, as illustrated by some of the cases included in this report, there are also good examples of exchange schemes and joint appointments between academia and industry. Whilst academia may find it difficult to compete in terms of salaries, the academic research environment can be attractive for digitally skilled personnel at different stages of their careers. However, it is currently difficult to move between industry and academia, or between vocational and academic career tracks. Again, one of the main obstacles is the expectations for an 'academic cv' – no matter what an individual may have achieved in a commercial research setting, without publications in science journals he or she will struggle to be accepted in academia. When it comes to consideration of digital skills, this 'academic cv' needs to be re-considered.

Addressing these issues requires considerable adaptation of traditional academic promotion criteria. In hiring, promotion, and tenure-review, data and software need to be recognised alongside publications as valuable outputs and assets for science. This is not a new issue: the Declaration on Research Assessment (DORA, 2012[41]), which is widely acclaimed, recommends that research assessment consider the value and impact of all research outputs, including datasets and software, and this is beginning to be implemented in some countries. However, the predominant approach in academia is still to value and reward individual academic research on the basis of publication outputs. Recognition of the value of digital outputs – data, software, algorithms and code – remains limited.

Box 4. Improving career paths and reward structures

Some examples from the case studies in this area include:

- TeSS is the training registry for the ELIXIR community that enables training providers to register training events and materials, facilitating the attribution of credit and ownership to their authors. In April 2019 the Training Platform contained links to 1 208 training materials, 289 upcoming events, and more than 9 200 previous events (Europe's distributed infrastructure for life-science data, n.d.[42]).
- TU Delft have a significant focus on developing data stewardship within the institution. Since 2016, every TU Delft faculty has had a dedicated data steward to assist researchers with research data management. The focus of these roles is based on high-level understanding of the areas where improvement by researchers in research data management would make the most impact, based on a survey in 2017-18 of around 700 staff. The data stewards begin by providing a core of generic training and then provide discipline-specific training that builds on this. The establishment of data steward posts embedded throughout faculties is part of a process of recognising and valuing the contribution to research of those with this set of skills, and ultimately to providing onward career paths for them.
- Close collaboration between academia and knowledge users is a focus for CIFAR. Many of the 80 national AI Research Chairs in Canada have cross-appointments in industry. The AI institutes, supported by the Pan-Canadian AI Strategy, are designed to foster interaction, the exchange of ideas, and side-by-side collaboration between academia, industry and innovators.
- MAS identify their legal status as an autonomous centre that is embedded in the higher education system in Chile but formally independent of universities, as having enabled valuable flexibility in human resources management. Salaries are an area where MAS has some flexibility, relative to Universities, which enables it to recruit and retain digitally skilled staff in the face of competition from industry.

5.5. Broader enablers of digital workforce capacity

Whilst the main focus of this project is on the science enterprise, it is situated in broader research, societal, and economic contexts. Development of a digitally skilled research workforce is intertwined with the adoption of broader enabling policies, such as support for open science and guidelines on research integrity. At the same time, having strong digital workforce capacity is necessary to deliver the objectives of these broader policies. This section explores some of the broader policy actions and contextual issues that can have positive influences on the development of the digital workforce.

Understanding of the broader socioeconomic environment is important as this provides the enabling context in which to situate scientific research. This broader environment can be broken down into a number of key parts, in which policy development can directly or indirectly have a substantial influence on the development of digital research workforce:

- Economic environment, e.g., open science and open data funding.
- Demographic environment, e.g., digital literacy, labour market.
- Technological environment, e.g., digital connectivity.
- Legal and political environment, e.g., open science policies, database legislation.
- Sociocultural environment, e.g., values around education may affect the success of grassroots training initiatives if top-down, formal training is the main approach to education.
- Global/international environment, e.g., international relations between countries may affect the degree to which some countries will cooperate on digital capacity building initiatives

A major enabler is an education system that incorporates digital skills training at all levels, creating a pipeline of appropriately skilled students going into research careers. In this context, one can cite Hungary's Digital Education Strategy, which aims to increase digital literacy in harmony with sectoral strategies and professional objectives at all levels of the education system (EC/OECD, 2020[43]). Similarly, Portugal's National Initiative for Digital Skills e.2030 is a joint initiative of several governmental sectors that aims to stimulate and ensure the development of digital skills as a tool for paving the way for a future-oriented society (EC/OECD, 2020[43]).

Broader national and international trends in science policy are helping to make the case for digital skills acquisition, and were widely acknowledged by the thirteen case studies as important enablers for their capacity building efforts. The increasing emphasis on ensuring transparency and openness to enable reproducibility is an important driver as it has a clear requirement for open data and a workforce that has the digital skills to generate reproducible results that stand up to open scrutiny. Systems and processes that ensure that open research practices are supported, encouraged, and rewarded are increasingly being adopted and all of these depend on having a workforce with the necessary digital skills.

Other parts of the wider research ecosystem can also have a positive influence on the development of digital workforce capacity, with several case studies noting that requirements by organisations such as journal publishers that data and code are accessible have a major influence on the behaviour of the research community and increase the demand for relevant digital skills. Similarly, promotion of Open Source Software raises awareness of the need for digital skills.

Box 5. Enabling factors and digital research workforce capacity building

Example of actions, identified in case studies, that are not directly targeted at digital capacity building but that enable it include:

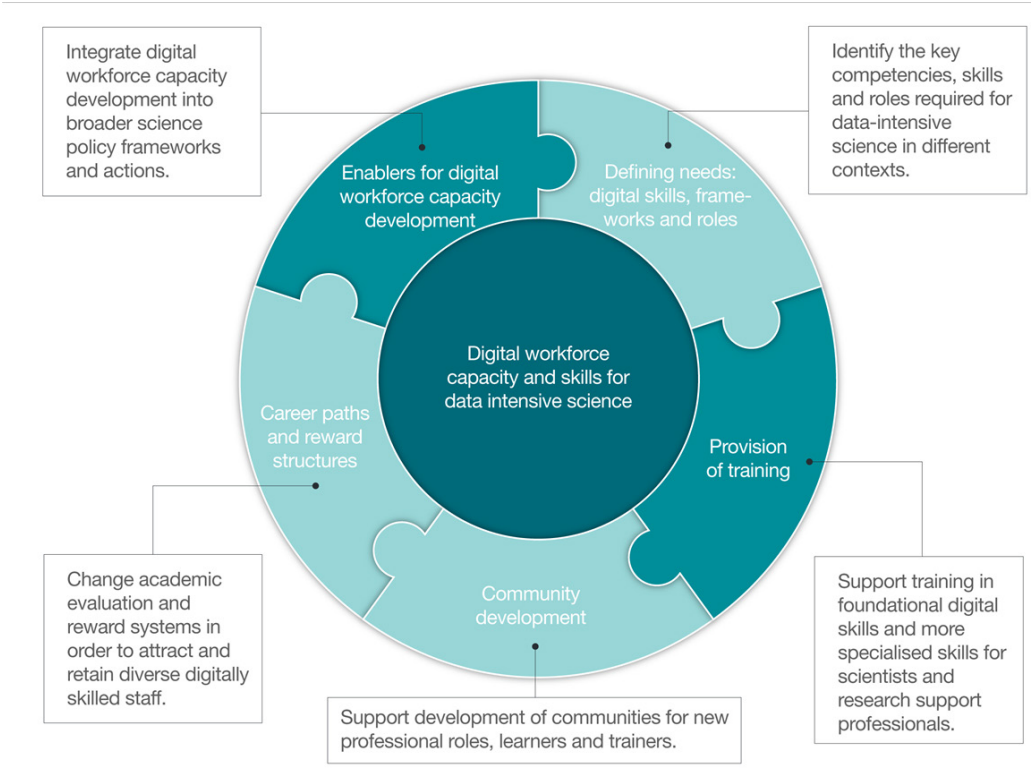
- The Japan Agency for Medical Research and Development (AMED) now requires that information on data management plans in funding applications include details of the person who will be responsible for implementing them. It has also funded training in data management and reuse for its staff and for researchers that it funds. This supports the work of JPCOAR in their ongoing development of research data management training tools for librarians. This training helps librarians acquire the skills needed to meet data management plan requirements.
- TU Delft note the positive effect of funder requirements for sharing of data and code and data management plans in driving researchers to acquire new data skills. The Dutch Research Council (NWO) and the European Commission, have such requirements and are enforcing compliance. Journal requirements for data access are also driving change.
- GESIS has benefitted from the work of the German Council for Scientific Information Infrastructures in mapping out future educational needs at both vocational and scientific research levels. Whilst in the past there were no vocational training programs in their field, the first graduates are now emerging from a vocational studies course that includes development of skills in recording and documenting data throughout the data lifecycle, to support survey work in the social sciences. In addition to the science education in digital skills, this assists GESIS in achieving their vision of enlarging services around new data sources as a complement to more traditional survey data.
- ARDC's work is enhanced by Australian government initiatives including the National Collaborative Research Infrastructure Strategy requirement that all of its funded projects embrace the FAIR data principles. The ARDC plays a key role in shifting national research culture to making FAIR data a priority, and provides an extensive repository of online training guides.

5.6. Linking action areas to goals

Building on the five key action areas described above, where advancements are needed, it is possible to identify five corresponding goals that need to be achieved to advance digital workforce capacity. Creating and maintaining a productive science workforce is a shared responsibility between multiple stakeholders and work towards these goals needs to be undertaken by diverse actors.

The five goals are illustrated in tandem with the five key action areas that they relate to, in Figure 4.

Figure 4. Five key action areas and goals for digital research workforce capacity development



Source: authors' design.

The next two chapters of this report examine how various actors can focus their energies towards achieving these goals.

6. Recommendations for various actors

Having defined a number of main action areas and goals, a return to the earlier discussion of the science ecosystem (Section 4) helps to identify who the key actors are and what actions they can take to deliver these goals. From a policy perspective, five key actors can be identified: national and regional governments, research agencies, professional science associations, research institutes and infrastructures, and universities. Recommendations for the first four groups are laid out in this chapter and illustrated by examples of relevant efforts that are already underway. This finishes with a short section on international cooperation and potential actions for all actors in this area. The role(s) of universities are discussed in a dedicated chapter that follows this one.

An overview is presented in Table 2, which indicates where actors might have the greatest opportunities to effect change across the five main action areas.

Table 2. Opportunities for actors to effect change across the five main action areas

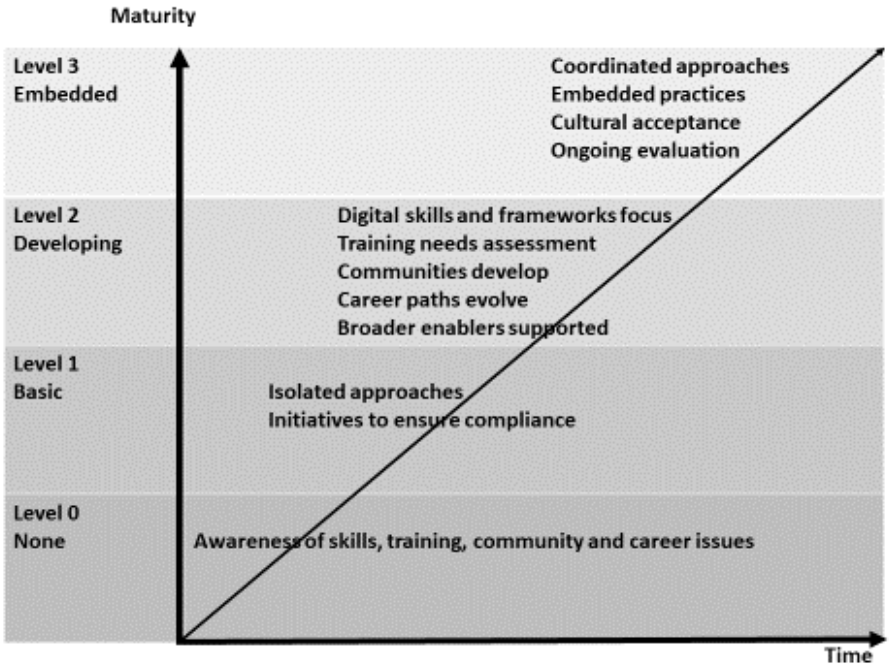
	Defining needs	Provision of training	Community building	Career paths rewards	Broader enablers
National/regional governments	✔	✓	✓	✓	✔
Research agencies	✔	✓	✓	✔	✔
Professional science associations	✓	✓	✔	✓	✓
Research institutes and infrastructures	✓	✔	✔	✔	✓
Universities	✔	✔	✔	✔	✓

Key: Large ticks denote areas where actors can exert significant change; small ticks indicate where actors can have a smaller influence. NOTE: This table shows a generalised view of where different actors can have the most significant impact; an individual organisation's actual opportunity areas may differ in practice.

Table 2 shows a major focus on the role of universities, which is perhaps not surprising as these are the main centres of education, training, and public research in most countries. However, from a policy perspective, it is recognised that these are largely autonomous bodies that need to be encouraged and steered rather than directed by top-down national policies. At the same time, having a digitally skilled academic workforce is critical for these institutions and it is in their self-interest to ensure that they cooperate closely with other actors to design and implement appropriate digital capacity building steps and create an enabling environment for data intensive science. Detailed actions that universities can implement are described in more detail in the next chapter.

A general recommendation for any organisation or community is to evaluate and improve the maturity of their digital workforce capacity strategy. Maturity models are commonly used to help organisations assess effectiveness in a given area and to support understanding of what is needed to improve performance. They are most effective when linked with strategic leadership. Figure 5 provides a digital workforce capacity maturity model that can be used at a variety of levels – by national governments to evaluate their national strategy, or by any type of organisation or community to improve their internal strategy

Figure 5. Digital workforce capacity maturity model



Source: adapted from (Cox et al., 2017[44]).

6.1. National or regional governments

There are a number of actions that governments can take to facilitate the development of a digitally skilled research workforce. First and foremost, it is important for the responsible authorities to understand the urgency of this issue, and the potential negative effects on research competitiveness of not taking action. Governments are recommended to:

- Recognise at the policy level the need for a digitally skilled workforce in research, and the importance of strategic planning that integrates the five key areas that must be addressed in parallel to build and maintain this workforce: defining needs; provision of training; community building; career paths and rewards; and broader enablers of digital workforce capacity.
- Analyse national digital workforce capacity needs and the status of the research ecosystem to meet these needs in order to inform strategic planning and investment. It is important to take account of, international and disciplinary initiatives as well as local developments, and to consider how government actions can most effectively leverage and support these.
- Facilitate coordination of the effort needed to build workforce capacity at the speed and scale necessary to optimise the benefits of data intensive science, including the implementation of regular monitoring and assessment processes that can keep pace with the evolving landscape.

There are also a range of broader actions that governments can take that help to provide an enabling environment for data intensive science:

- Provide policy support for open science, and efforts to ensure the reproducibility of research and research integrity.

- Implement research assessment systems that reward data and software outputs as well as publications.
- Encourage proactive and flexible initiatives to address digitally skilled workforce needs in a constantly evolving landscape, including, training provision, community building, and career paths and reward structures.
- Develop digital skills at all educational levels, with clear delineation of which ministries have responsibility for digital skills for science at different education and training levels.

Box 6. What are governments doing?

Examples of national and regional government initiatives include the following:

- The Finnish Ministry of Education and Culture established a data management and scientific computing development programme in 2017, as a key part of the national initiative on Open Science and Research. This focuses on the development of data management and computing research infrastructures, services and expertise to support research and education at Universities.
- Germany's Council for Scientific Information Infrastructures has a focus on the digitally skilled workforce; the key areas to be supported through targeted initiatives in the area of research data in the years ahead include policies and regulations, digital skills and the federation of infra-structures (German Council for Scientific Information Infrastructures, 2019[45]).
- Japan's Ministry of Education, Culture, Sports, Science and Technology supports the Doctoral program for Data-Related InnoVation Expert (D-DRIVE). D-DRIVE enables universities and companies to collaborate on development and implementation of training programs to acquire skills including those related to data science for doctoral students and PhD holders. The programs foster the acquisition and development of practical skills through problem-solving learning and internships that utilise company data (Ministry of Education, Culture, Sports, Science and Technology, n.d.[46]).
- South Africa's Department of Science and Innovation supports the multi-institutional National e-Science Postgraduate Teaching and Training Platform. The Platform is developing a suitable qualification, curricula and pedagogic interventions to advance the training of postgraduate students in the rapidly developing cross-disciplinary fields involved in e-Science (National e-Science Postgraduate Teaching and Training Platform, n.d.[47]).
- The European Commission is supporting a number of trans-national initiatives (policies, programs and/or projects) dealing with digital skills in relation to the European Open Science Cloud (EOSC). An EOSC Skills and Training Working Group has been established to identify how skills development and training can be embedded in different levels of EOSC (EOSCsecretariat.eu, n.d.[48]).

6.2. Research funding agencies

The division of responsibilities between governments and research agencies for strategic planning and funding varies across countries. The actions identified above for national or regional governments with regards to strategic planning are often shared with research

agencies; what is important is that these actions are effectively implemented. Likewise, when it comes to research funding, the mechanisms differ across countries and the following actions are targeted at the relevant responsible authority(ies):

- Ensure that funding schemes include elements that support digital workforce capacity development, such as funding opportunities to build and develop training for researchers and research support professionals, exchange programmes between academia and industry, and flexibility on staff recruitment profiles and salaries.
- Align planning, policies and investment for physical research infrastructure and for digital workforce capacity, recognising that both physical and human resources are needed to support data intensive science.

Funding agencies also share responsibility with governments and other actors for creating a broader enabling environment for data intensive science. They play a major role in establishing the mandates and incentives that shape how research is conducted. In this context, the following actions are particularly relevant for research agencies:

- Ensure that data and software management plans are a requirement in funding schemes, and that there are processes in place to monitor and enforce compliance.
- Recognise data and software as outputs that should be properly valued and taken account of in peer review and research assessment processes.
- Include digital workforce issues in guidelines and processes that relate to responsible conduct of research, research integrity and ethics.

Box 7. What are research funding agencies doing?

Aside from providing funding for capacity building and training, funders are taking a number of actions to promote digital workforce development. Examples include:

- The Scientific and Technological Research Council of Turkey, TUBITAK, has launched a Research Data Training Portal as part of the TUBITAK Open Science Policy (TUBİTAK, n.d.[49]). The portal provides a resource where sample data management plans are shared to assist the researchers in data management through sharing of sample data management plans, as part of opening data sharing to conduct scientific studies in a much more efficient and productive way (TUBİTAK, n.d.[50]).
- In the Netherlands, NWO is changing the rewards system for researchers and piloting a narrative curriculum vitae format in the Veni scheme, its major funding instrument for early career researchers. This new format has two categories, academic profile and key output, where key outputs can include diverse types of output, including data and software (DORA, 2019[51]).
- The UK's Arts and Humanities Research Council (AHRC) requires that PhD students include digital skills in their training. Based on the Vitae framework, students are encouraged to acquire knowledge and understanding of existing and new methodologies, such as numerical, data management, and statistical techniques or software, web and social media communication tools (Arts and Humanities Research Council, 2014[52]).
- Several countries have deployed systems that manage information on science, technology, and innovation related to individual researches and institutions, such as the Science Experts Network Curriculum Vitae (SciENcvUSA), Lattes (Brazil), National Academic Research and Collaboration Information System (NARCIS, Netherlands) and Researchfish (UK). These systems recognise a wide variety of research outputs, such as curated data, software and workflows. As a result, higher value is placed on them and the skills required to produce them, driving behaviour change in researchers and the organisations that they work for.

6.3. Professional science associations and academies

Academies and professional societies can play a leading role in ensuring a strong focus on digital workforce capacity in their communities through initiatives that:

- Define the community-specific needs for digital workforce capability and advocate for resources and mechanisms to address these needs.
- Promote discussions and networking in conferences and events on: digital skills, frameworks and roles; provision of training; community building; and, career paths and reward structures.
- Develop and openly disseminate digital training materials and contribute to curricula development.

- Contribute to the development of workforce transformation plans with, and on behalf of, their communities, with regard to data intensive science.
- Produce reports on data intensive science and skills that are influential at the research community and science policy levels and promote exchange of good practices across countries.

Box 8. What are professional science associations doing?

Examples of initiatives that science associations are undertaking in this area include:

- The American Astronomical Society is supporting the Carpentries to develop curriculum for a Data Carpentry workshop for astronomy. This workshop will be offered at the annual American Astronomical Society meeting. Society support provides a strong connection with the community for both development and delivery of the curriculum.
- The Australian Academy for the Humanities initiated a three-year **Future Humanities Workforce** project in 2018, to provide a comprehensive account of Australia's humanities research workforce and plan for its future knowledge and skills requirements. This aims to identify skills and knowledge priorities for future research environments, with a focus on data and digital literacy (Australian Academy of the Humanities, n.d.[53]).
- The UK Academy of Medical Sciences report on how to fully realise the benefits of AI in health identified five key themes that included data and computing technology, working across sectors and disciplines, and training and capacity building. The report notes: "The lack of a clear career pathway and lower salaries in academia were cited as two of the main reasons for leaving academic research, though this is often at the expense of scholarly and intellectual freedom" (The Academy of Medical Sciences, 2019[54]).

6.4. Research institutes and infrastructures

Research institutes and research infrastructures produce and analyse large amounts of data and are an important focus for data intensive research. A number of organisations that were case studies in this project, including ARDC, ELIXIR, GESIS and SSI, can be categorised as research infrastructures that provide a service to the broader community. MAS and Turing are autonomous research institutes or centres of excellence. There are similarities between these organisations and the universities in their roles of producing research and employing research staff and research support professionals; consequently research institutes and infrastructures can also play a role in enacting many of the recommendations that will be detailed for universities and libraries in the next section. As research service providers and centres of excellence and expertise, they can play a particularly important role in training. To perform this role effectively, capacity building should ideally be part of their overall mission, and they have to be well connected with universities, have strong links into national and international networks, and have the necessary resources.

Box 9. What are research infrastructures doing?

Some case studies in this area include

- The Digital Research Infrastructure for the Arts and Humanities (DARIAH-EU, 2019[55]) is a European research infrastructure supported jointly by a consortium of 18 member countries, which includes a focus on guiding and training researchers in terms of data management practices. DARIAH's mission to empower research communities in the arts and humanities with digital methods to create, connect and share knowledge about culture and society has resulted in training approaches including the DARIAH ERIC Sustainability Refined (DESIR) Winter School**.** This aims to strengthen the skills of the arts and humanities communities in research data management, curation, sharing, preservation and reuse
- The Research Software Program of the Canadian Network for the Advancement of Research, Industry and Education (CANARIE, 2020[56]) is a research infrastructure that has funded software teams at Canadian institutions to work directly with researchers since 2018. CANARIE designs and delivers digital infrastructure as a non-profit corporation, with the majority of its funding provided by the Canadian Government. CANARIE's 2020 funding round aims to enhance the availability of software teams to researchers, regardless of discipline, and will provide guidance, training, expertise, and software development specific to advancing research.
- The Consortium of European Social Science Data Archives (CESSDA, n.d.[57]) Training Working Group creates a place where CESSDA service provider staff, archivists, data producers and researchers working in the wide area of social science and humanities can find training, advice and educational resources. These cover a range of topics including: research data management, data discovery and use, digital preservation and data archiving.

6.5. International cooperation

As described in Section 3, several of the case studies that were considered in this project are inherently international, i.e., they exist to serve an international community. There was a strong interest across all the case studies in understanding other digital workforce programs, with almost all case studies citing other organisations or initiatives locally or internationally they either engaged with or looked to as an exemplar. This engagement took different forms, including partnerships between organisations (such as CODATA and RDA), and sharing of training materials and/or trainers. The international initiatives were often strongly linked with national and institutional training activities, although that they were not always recognised as formal partners. Whilst some competition between institutes and countries can help drive forward the training agenda, there is much to be gained from cooperation and open sharing of experience and materials.

Digital research capacity needs and the extent to which they are being met, *i.e.* overall digital preparedness levels (Figure 5) vary across countries (Figure 2) Whilst much of the responsibility for building sustainable capacity resides at the national and sub-national level, there is also a shared global responsibility to ensure that scientists, no matter where

they reside, can carry out research that contributes both to science and to socioeconomic development. Scientific research is inherently international and public research data can be considered as a global public good that should contribute to addressing the sustainable development goals. As discussed in this report, access to data is only useful when there is the capacity to exploit it, and the global community has a shared responsibility to help build digitally skilled research capacity wherever it is required.

All actors should take advantage of opportunities to:

- Engage in international collaboration wherever relevant and possible
- Share training materials, good practices and experiences across countries and communities
- Support digital research capacity development efforts in countries or research communities that can benefit from data intensive science but currently have extremely limited capacity

Box 10. What international collaboration is occurring?

Some of the examples in this area include:

- Since 2018, different international data common initiatives have been meeting regularly to advance collaboration. These include: ARDC, EOSC, the African Open Science Platform and the National Institutes of Health (NIH) Data Commons. There is interest in convergence on a framework to facilitate alignment, efficiency and interoperability.[5]
- The many collaborators working with ELIXIR's training platform reflect the importance of partnerships in building a global life sciences data community. ELIXIR engages with many organisations and initiatives including the Carpentries, CODATA, RDA, H3ABionet (a Pan African Bioinformatics Network for the Human Heredity and Health in Africa consortium), Partnership for Advanced Computing in Europe (PRACE), USA's NIH's Big Data To Knowledge (BD2K, now Data Commons), and the Global Organisation for Bioinformatics Learning, Education and Training (GOBLET). Formal training collaboration agreements are in place with several of these organisations. ELIXIR is also an important contributor to the planning for the European Open Science Cloud in terms of training and capacity-building.
- The Research Data Alliance (RDA) is a global grassroots organisation that brings together different communities with an interest in data intensive science in order to build technical and social bridges. It has a number of open interest groups and working groups that are focused on developing practical solutions. For example, RDA Plenary 14 in 2019 included a session to explore what is needed and what challenges need to be addressed to professionalise data stewardship (Research Data Alliance (RDA), n.d.[58]).
- The World Data System and a number of other international networks of data repositories have been established and many of these play either a formal or informal role in training and community building (OECD, 2017[6]).

7. Recommendations for universities

In most countries, universities are the central actors with regards to building digital workforce capacity and skills for data intensive science. They have responsibility both for supporting and conducting academic research (internal) and for the provision of education, including science education, to serve society more generally (external). Whilst governments and research agencies have a major role in supporting and incentivising universities, universities have both autonomy and capacity to set their own policies.

All the case studies identified the pivotal role of universities in producing and employing digitally skilled researchers and research support professionals. There is a pressing need for universities to respond to the need for a digitally skilled scientific workforce. This must be done at scale and will require leadership and investment. There is a perception in some communities that universities are not only unable to meet this need at present but will continue to be unable to do so in the future. At the same time, there are many good practices in individual institutions that could be more widely promoted. There are an increasing number of cross-institutional initiatives that align best practice undertaken by individual universities.

Academic libraries and librarians are a natural focus for digital skills support and capacity building within universities. Librarians both use digital assets and tools in their own work and become advocates for and train others in data and software practices, particularly in relation to foundational skills and data stewardship. "Because training is a well-established service-category for academic librarians (in areas such as information literacy) and is accepted as a key component of existing staff liaison roles, libraries have the potential to integrate digital skills training with existing training activities relatively easily" (Cox et al., 2017[44]). Consequently, libraries can be an important resource for universities to increase their digital workforce capacities, provided that the necessary investment is made.

Computing science or research IT departments are also natural places for seeding digital skills within a University, particularly skills relating to coding and software development. Institutions that host facilities, such as high-performance computing or specialist research data storage facilities, can also use these as a training resource. In some cases, enterprise IT organisations[6] have been established to provide digital research support services and training both for university staff and industry. In all of these cases, as for libraries, training needs to be properly supported and valued for it to be effective and sustainable. In practice, most large Universities have a mixture of different digital skills training capabilities and the challenge is to make sure that these work together coherently to meet evolving needs across all scientific domains.

Key recommendations for universities, including libraries, are listed in Table 3. These recommendations are separated into the five main action areas described earlier in this report: defining needs: digital skills, frameworks and roles; provision of training; community building; career paths and reward structures; and broader enablers of digital workforce capacity. They have also been divided into internal (academic research focussed), and external-facing activities.

Table 3. Key recommendations for universities and libraries

Note: Internal activities focus on strengthening (data-intensive) academic research. External activities focus on strengthening science education for society more broadly, for example through incorporation of digital skills in undergraduate science education.

Internally focussed: academic research	*Externally focussed: science and society*
1. Defining needs: digital skills, frameworks and roles	
• Conduct surveys of digital research needs and develop integrated strategies to address these. • Ensure that digital skills are integrated into frameworks to promote open science and research integrity	• Work with non-academic stakeholders, e.g. from industry, to define digital research skills and training needs
2. Provision of training	
• Provide wide access to both foundational and specialised digital skills training for existing researchers and research support professionals (with outsourcing of training an option). • Provide staff training and support to ensure that funding proposals have well-defined and well-resourced data management plans and software equivalents (e.g., software management plans). • Include mentorship in digital skills in the specified requirements for supervision of research.	• Provide generic undergraduate training for foundational digital skills for research across all science disciplines. • Integrate comprehensive, relevant, and up-to-date digital skills training from undergraduate level onwards into curricula, including for life-long learning. • Ensure flexibility in education curricula and training to keep pace with technological change. • Work with outside users, including from industry, to develop and support life-long digital research training programmes
3. Community building	
• Actively promote workforce diversity as a key element of digital capability. • Recognise of the importance of training provision, and community involvement and/or leadership in career evaluations. • Support the development of professional communities in emerging roles such as data stewards and RSEs, and for trainers and leaders of digital skills initiatives. • Establish digital skills support desks for data-intensive research support and assistance, and/or integrate staff with these skills into research teams.	

4. Career paths and reward structures	
• Develop and adopt frameworks to employ digital research support professionals in stable positions with opportunities for career advancement. • Implement systemic institutional change around careers paths and alternative metrics to recognise and reward a broader range of research contributions at all academic levels. • Recognise datasets and software as valued research outputs. • Incentivise collaborations between domain researchers and digital support staff with metrics that recognise the contributions of all members of research teams.	• Promote collaboration and exchange across industry and academia and recognise non-academic success for people wishing to re-enter academia. • Encourage and support cross-sectoral digital skills initiatives.
5. Broader enablers of digital workforce capacity	
• Adopt open science, reproducibility and research integrity principles. • Ensure that Institutional Review Boards have the support and capacity to assess ethical issues in data-intensive research.	• Support open science, reproducibility and research integrity at policy levels. • Engage with other sectors in defining and addressing digital research capacity needs

There is considerable potential for universities to work together at different geographic scales, learning from each other and combining expertise and resources where appropriate to address common digital workforce training and capacity challenges. This can be done via informal networks and exchanges or more formal partnerships, and it can be done at the level of the whole institution, at an appropriate sub-structural scale, or via individual networking. Libraries can be an important unit for engagement between universities. Some examples of inter-university cooperation are given in Box 11.

Box 11. How are universities working together?

Examples of cross-institutional university initiatives include:

- ADSA convened the 2019 Data Science Leadership Summit in the US to bring together the leaders of data science institutes, centres, and programs, and faculty interested in creating new initiatives on their campuses. The Summit aimed to form an academic community for data science; to share best practices where they face similar challenges and opportunities; and to take collective responsibility in preparing next-generation data scientists to contribute in the best interests of society (The 2019 Data Science Leadership Summit, 2019[59]).
- The UK Reproducibility Network (UKRN) is a consortium involving more than 20 universities and a range of other stakeholders that is investigating the factors that contribute to robust research, promoting training activities, disseminating best practice, and working with stakeholders to ensure coordination of efforts (University of Bristol, n.d.[60]). SSI are partnering with UKRN on a range of joint initiatives, including training in software development and data management skills (SSI, 2020[61]). The UK also has an emerging collaboration across eight

universities in the N8 Research Partnerships, examining similarities, challenges and opportunities for RSE groups (N8 CIR, 2020[62]).

- Leaders of eight university networks from multiple nations signed the Sorbonne declaration on research data rights in early 2020. The signatories commit to actions including: "Encouraging our universities in setting up training and skills development programs that create an environment to promote open research data management" (LERU, 2020[63]). The signatories are the Association of American Universities, African Research Universities Alliance, Coordination of French Research-Intensive Universities, German U15, League of European Research Universities, RU11 in Japan, Russell Group in the UK and the Group of Eight in Australia.

University libraries are combining their forces to build digital science capacity in a number of different ways:

- JPCOAR uses openly available training materials from other institutions to develop tailored on-line training courses for librarians in Japanese universities. These courses are delivered via a national MOOC platform and focus on foundational research data management skills.
- The Council of Australian Librarians (CAUL) has introduced a Digital Dexterity program with two objectives: that Australian universities recognise the importance of digital dexterity in accomplishing their missions and that they engage with CAUL on its delivery; and that Australian graduates have access to the digital skills to enable them both to thrive in a global work context and to become effective global citizens (Council of Australian University Librarians, 2019[64]).
- The Canadian Association of Research Libraries (CARL), has funded a multi-year activity aimed at supporting data services and skills development for its members and providing national services where appropriate, such as for data management planning. It has attracted funding from federal agencies to continue and enhance the work.
- Europe's research library network, Ligue des Bibliothèques Européennes de Recherche (LIBER) is developing research libraries as hubs for digital skills and services in both physical and virtual research environments. LIBER provides training and/or data services for members, and does advocacy with policy organisations to support data skills development.

8. Conclusion: the need for concerted policy action

Digital technologies are changing the practice of science, and a digitally skilled workforce is needed to maximise the potential scientific and socioeconomic benefits of data-intensive science. There is a critical need to accelerate the speed and scale at which the workforce issues associated with the digital transformation are being addressed. Whilst many organisations in different countries are involved in initiatives to improve digital workforce capacity for research, a more coordinated approach would be beneficial. There are many good initiatives underway and there is much that can be learned from what others are doing.

This project has explored the elements required to develop and sustain workforce capacity for data intensive science. There are five main action areas where focus is needed: defining

needs; provision of training, community building, career paths and rewards; and broader enablers of digital workforce capacity. Whilst there are many initiatives that are addressing these at different levels and in different domains, there are few comprehensive strategic approaches that integrate all five areas.

There are a range of actors throughout the science ecosystem who can implement actions to advance in these different areas. This report identifies the main areas where policymakers and other actors might concentrate their efforts, and includes specific recommendations on actions that can be taken to facilitate the development of a digitally skilled workforce.

Many individuals and institutions are working actively to promote digital skills for data intensive science and there is a critical role for policymakers to support and enable them so that their efforts can be effective at the scale necessary to achieve a common vision:

> *A world where effective policies, incentives, and investments sustain a future-focused, world-leading research workforce that can realise the full potential of data-intensive science and help provide solutions to complex societal challenges.*

Notes

[1] The terms science and research are used interchangeably throughout this report and include natural sciences, social sciences and humanities.
[2] See https://ec.europa.eu/jrc/en/digcomp
[3] This and similar program in a range of disciplines are described in the Canadian Vector Institute's list of AI programs (Vector Institute, n.d.[94]).
[4] Poll conducted at Australian eResearch Skilled Workforce Summit, 2019 (Australian Research Data Commons, n.d.[95]).
[5] Meetings have occurred at SciDataCon 2018 and RDA plenaries (under the Global Open Research Commons Interest Group).
[6] For example, the Edinburgh Parallel Computing Centre is an enterprise IT centre that provides specialised IT services and training for researchers at the University of Edinburgh and for local SMEs (EPCC, n.d.[96])

References

Apostel, L. et al. (eds.) (1972), *Towards Interdisciplinarity and Transdisciplinarity in Education and Innovation*, OECD Publications Center. [74]

Arnstein, S. (1969), "A Ladder Of Citizen Participation", *Journal of the American Institute of Planners*, Vol. 35/4, pp. 216-224, http://dx.doi.org/10.1080/01944366908977225. [75]

Arts and Humanities Research Council (2014), *AHRC Research Training Framework for Doctoral Students*, https://ahrc.ukri.org/documents/projects-programmes-and-initiatives/ahrc-research-training-framework-for-doctoral-students/ (accessed on 8 June 2020). [52]

Ashley, K. (2016), *Review: Developing skills for managing research data and software*, Wellcome Trust, https://doi.org/10.6084/m9.figshare.4133916.v1. [11]

Australian Academy of the Humanities (n.d.), *Future Humanities Workforce*, https://www.humanities.org.au/advice/projects/future-workforce/ (accessed on 8 June 2020). [53]

Australian Research Data Commons (n.d.), *The Australian eResearch Skilled Workforce Summit*, https://ardc.edu.au/events/the-australian-eresearch-skilled-workforce-summit/ (accessed on 5 June 2020). [95]

Babuska, I. and J. Oden (2004), "Verification and validation in computational engineering and science: basic concepts", *Computer Methods in Applied Mechanics and Engineering*, Vol. 193/36-38, pp. 4057-4066, http://dx.doi.org/10.1016/j.cma.2004.03.002. [67]

Bello, M. and F. Galindo-Rueda (2020), "Charting the digital transformation of science: Findings from the 2018 OECD International Survey of Scientific Authors (ISSA2)", *OECD Science, Technology and Industry Working Papers*, No. 2020/03, OECD Publishing, Paris, https://dx.doi.org/10.1787/1b06c47c-en. [8]

Berente, N. et al. (2018), *Organizing and the Cyberinfrastructure Workforce*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3260715. [39]

Buchhorn, M. (2019), *Surveying the scale of the research-IT support workforce*, https://ardc.edu.au/wp-content/uploads/2019/07/ARDC-National-Workforce-report-final-v3.pdf (accessed on 5 June 2020). [16]

CANARIE (2020), *Local Research Software Support – Call 1*, https://www.canarie.ca/software/funding/lrss-call1/ (accessed on 8 June 2020). [56]

CASRAI (n.d.), *Resources*, https://casrai.org/resources/ (accessed on 5 June 2020). [66]

CESSDA (n.d.), *Working Groups*, https://www.cessda.eu/About/Working-Groups (accessed on 8 June 2020). [57]

Council of Australian University Librarians (2019), *CAUL Digital Dexterity Position Statement*, https://www.caul.edu.au/caul-digital-dexterity-position-statement (accessed on 8 June 2020). [64]

Cox, A. et al. (2017), "Developments in research data management in academic libraries: Towards an understanding of research data service maturity", *Journal of the Association for Information Science and Technology*, Vol. 68/9, pp. 2182-2200, http://dx.doi.org/10.1002/asi.23781. [44]

Dai, Q., E. Shin and C. Smith (2018), "Open and inclusive collaboration in science: A framework", *OECD Science, Technology and Industry Working Papers*, No. 2018/07, OECD Publishing, Paris, https://dx.doi.org/10.1787/2dbff737-en. [5]

DARIAH-EU (2019), *DESIR Winter School: Shaping new approaches to data management in arts and humanities*, https://www.dariah.eu/2019/09/09/desir-winter-school-shaping-new-approaches-to-data-management-in-arts-and-humanities/ (accessed on 8 June 2020). [55]

Demchenko, Y., A. Belloum and T. Wiktorski (2017), *EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS) Release 2*, https://doi.org/10.5281/zenodo.1044346. [26]

DORA (2019), *Quality over quantity: How the Dutch Research Council is giving researchers the opportunity to showcase diverse types of talent*, https://sfdora.org/2019/11/14/quality-over-quantity-how-the-dutch-research-council-is-giving-researchers-the-opportunity-to-showcase-diverse-types-of-talent/ (accessed on 8 June 2020). [51]

DORA (2012), *San Francisco Declaration on Research Assessment*, https://sfdora.org/read/. [41]

Dutch Techcentre for Life Sciences (n.d.), *About Research Data Management*, https://www.dtls.nl/fair-data/research-data-management/research-data-management/ (accessed on 5 June 2020). [68]

EC/OECD (2020), *STIP Compass: International Database on Science, Technology and Innovation Policy (STIP)*, https://stip.oecd.org (accessed on 14 May 2020). [93]

EC/OECD (2020), *STIP Compass: International Database on Science, Technology and Innovation Policy (STIP)*, https://stip.oecd.org (accessed on 5 June 2020). [43]

Enengel, B. et al. (2012), "Co-production of knowledge in transdisciplinary doctoral theses on landscape development—An analysis of actor roles and knowledge types in different research phases", *Landscape and Urban Planning*, Vol. 105/1-2, pp. 106-117, http://dx.doi.org/10.1016/j.landurbplan.2011.12.004. [86]

EOSCsecretariat.eu (n.d.), *Skills & Training Working Group*, https://www.eoscsecretariat.eu/working-groups/skills-training-working-group (accessed on 8 June 2020). [48]

EPCC (n.d.), *EPCC*, https://www.epcc.ed.ac.uk/ (accessed on 8 June 2020). [96]

European Commission (2019), *Cost-benefit analysis for FAIR research data*, http://dx.doi.org/10.2777/02999. [9]

European Commission (2018), *Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data*, Publications Office of the European Union, http://dx.doi.org/doi: 10.2777/1524. [21]

European Commission (2017), *Evaluation of Research Careers fully acknowledging Open Science Practices; Rewards, incentives and/or recognition for researchers practicing Open Science*, Publications Office of the European Union, http://dx.doi.org/doi: 10.2777/75255. [23]

European Strategy Forum on Research Infrastructures (2018), *Big Data and e-Infrastructure Needs*, ESFRI, http://roadmap2018.esfri.eu/. [7]

Europe's distributed infrastructure for life-science data (n.d.), *Welcome to TeSS: ELIXIR's Training Portal*, https://tess.elixir-europe.org/ (accessed on 8 June 2020). [42]

FAIR4S (2019), *EOSC FAIR4S*, https://eosc-fair4s.github.io/ (accessed on 5 June 2020). [24]

German Council for Scientific Information Infrastructures (2019), *Empfehlungen zu Berufs- und Ausbildungsperspektiven für den Arbeitsmarkt Wissenschaft*, http://www.rfii.de/download/digitale-kompetenzen-dringend-gesucht. [45]

German Council for Scientific Information Infrastructures (RfII) (2019), *DIGITAL COMPETENCIES – URGENTLY NEEDED!*, http://www.rfii.de/?p=4015 (accessed on 5 June 2020). [20]

Grant, S. and K. Bouskill (2019), *Open Science and Institutional Review Boards: Aligning Transparency with Regulatory Protections for Human Research Subjects*, http://www.metascience2019.org/poster-session/sean-grant/. [29]

Gredig, D. (2011), "From research to practice: Research-based Intervention Development in social work: developing practice through cooperative knowledge production", *European Journal of Social Work*, Vol. 14/1, pp. 53-70, http://dx.doi.org/10.1080/13691457.2010.516624. [81]

Hadorn, G. et al. (eds.) (2008), *Handbook of Transdisciplinary Research*, Springer Netherlands, Dordrecht, http://dx.doi.org/10.1007/978-1-4020-6699-3. [73]

Hernandez Montoya, A. (ed.) (2017), "Interdisciplinary Collaboration between Natural and Social Sciences – Status and Trends Exemplified in Groundwater Research", *PLOS ONE*, Vol. 12/1, p. e0170754, http://dx.doi.org/10.1371/journal.pone.0170754. [76]

Houghton, J. and N. Gruen (2014), *Open Research Data report*, https://www.ands.org.au/working-with-data/articulating-the-value-of-open-data/open-research-data-report. [10]

Hulhoven, X. (n.d.), *Develop a sustainable future for Brussels through a co-creation project*, https://innoviris.brussels/co-creation (accessed on 27 May 2020). [89]

Jahn, T., M. Bergmann and F. Keil (2012), "Transdisciplinarity: Between mainstreaming and marginalization", *Ecological Economics*, Vol. 79, pp. 1-10, http://dx.doi.org/10.1016/j.ecolecon.2012.04.017. [82]

Kiser, G. and Y. Mantha (n.d.), *Global AI Talent Report 2019*, https://jfgagne.ai/talent-2019/ (accessed on 5 June 2020). [17]

Klein, J. (2008), "Evaluation of Interdisciplinary and Transdisciplinary Research", *American Journal of Preventive Medicine*, Vol. 35/2, pp. S116-S123, http://dx.doi.org/10.1016/j.amepre.2008.05.010. [78]

Kurata, K., M. Matsubayashi and M. Takeda (2017), "Research data management in Japanese universities and research institutions: Status report based on questionnaire survey", *Journal of Information Processing and Management*, Vol. 60/2, pp. 119-127, https://doi.org/10.1241/johokanri.60.119. [14]

Lang, D. et al. (2012), "Transdisciplinary research in sustainability science: practice, principles, and challenges", *Sustainability Science*, Vol. 7/S1, pp. 25-43, http://dx.doi.org/10.1007/s11625-011-0149-x. [83]

LERU (2020), *Data Summit in Paris*, https://www.leru.org/news/data-summit-in-paris (accessed on 8 June 2020). [63]

Lonie, A. and R. Francis (2017), *An Australian Bioscience Data Capability Project Report - Phase 2*, https://www.embl-abr.org.au/wp-content/uploads/2018/04/ABDC-Project-Report-Phase-2.pdf. [13]

Méndez, E. (2019), *Open Science?... Darling, we need to talk*, https://www.open-science-conference.eu/wp-content/uploads/2019/03/Eva-Mendez.pdf. [37]

Ministry of Education, Culture, Sports, Science and Technology (n.d.), *Doctoral program for Data-Related InnoVation Expert(D-DRIVE)*, https://www.mext.go.jp/a_menu/jinzai/data/index.htm (accessed on 8 June 2020). [46]

Molloy, L., A. Gow and L. Konstantelos (2014), "The DigCurV Curriculum Framework for Digital Curation in the Cultural Heritage Sector", *International Journal of Digital Curation*, Vol. 9/1, pp. 231-241, http://dx.doi.org/10.2218/ijdc.v9i1.314. [27]

Mons, B. (2020), "Invest 5% of research funds in ensuring data are reusable", *Nature*, Vol. 578/7796, pp. 491-491, http://dx.doi.org/10.1038/d41586-020-00505-7. [1]

Morgan, S. et al. (2017), "The ELIXIR-EXCELERATE Train-the-Trainer pilot programme: empower researchers to deliver high-quality training", *F1000Research*, Vol. 6, p. 1557, http://dx.doi.org/10.12688/f1000research.12332.1. [33]

Moser, S. (2016), "Can science on transformation transform science? Lessons from co-design", *Current Opinion in Environmental Sustainability*, Vol. 20, pp. 106-115, http://dx.doi.org/10.1016/j.cosust.2016.10.007. [69]

N8 CIR (2020), *N8 RSE Leaders and Aspiring Leaders Meeting*, https://n8cir.org.uk/events/rse-aspiring-leaders/ (accessed on 8 June 2020). [62]

National e-Science Postgraduate Teaching and Training Platform (n.d.), *National e-Science Postgraduate Teaching and Training Platform*, http://www.escience.ac.za/ (accessed on 8 June 2020). [47]

Noah, W. and M. Jean (1980), *Webster's New Twentieth Century Dictionary of the English Language, Unabridged*, W. Collins. [72]

OECD (2019), *Measuring the Digital Transformation: A Roadmap for the Future*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264311992-en. [19]

OECD (2019), *OECD Skills Outlook 2019 : Thriving in a Digital World*, OECD Publishing, Paris, https://dx.doi.org/10.1787/df80bc12-en. [22]

OECD (2018), *Bridging the Digital Gender Divide: Include, upskill, innovate*, http://www.oecd.org/internet/bridging-the-digital-gender-divide.pdf. [31]

OECD (2017), "Business models for sustainable research data repositories", *OECD Science, Technology and Industry Policy Papers*, No. 47, OECD Publishing, Paris, https://dx.doi.org/10.1787/302b12bb-en. [4]

OECD (2017), "Co-ordination and support of international research data networks", *OECD Science, Technology and Industry Policy Papers*, No. 51, OECD Publishing, Paris, https://dx.doi.org/10.1787/e92fa89e-en. [6]

OECD (2017), "Open research agenda setting", *OECD Science, Technology and Industry Policy Papers*, No. 50, OECD Publishing, Paris, https://dx.doi.org/10.1787/74edb6a8-en. [88]

OECD (2016), "Research Ethics and New Forms of Data for Social and Economic Research", *OECD Science, Technology and Industry Policy Papers*, No. 34, OECD Publishing, Paris, https://dx.doi.org/10.1787/5jln7vnpxs32-en. [30]

Ouellette, F. (ed.) (2017), "Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators", *PLOS Computational Biology*, Vol. 13/10, p. e1005755, http://dx.doi.org/10.1371/journal.pcbi.1005755. [12]

Piatetsky, G. (2018), *How many data scientists are there and is there a shortage?*, https://www.kdnuggets.com/2018/09/how-many-data-scientists-are-there.html (accessed on 5 June 2020). [2]

Pohl, C., P. Krütli and M. Stauffacher (2017), "Ten Reflective Steps for Rendering Research Societally Relevant", *GAIA - Ecological Perspectives for Science and Society*, Vol. 26/1, pp. 43-51, http://dx.doi.org/10.14512/gaia.26.1.10. [84]

Pohl, C. et al. (2011), *Questions to evaluate inter- and transdisciplinary research proposals*, td-net for Transdisciplinary Research. [90]

Research Data Alliance (2015), *Task Force on Defining data handling related competences and skills for different groups of professions - Working area*, https://www.rd-alliance.org/group/education-and-training-handling-research-data-ig/wiki/task-force-defining-data-handling (accessed on 5 June 2020). [25]

Research Data Alliance (RDA) (n.d.), *RDA 14th Plenary - Programme*, https://www.rd-alliance.org/rda-14th-plenary-programme (accessed on 22 June 2020). [58]

Research Data Alliance International Indigenous Data Sovereignty Interest Group (n.d.), *CARE Principles for Indigenous Data Governance*, https://www.gida-global.org/care (accessed on 5 June 2020). [32]

Roux, D. et al. (2010), "Framework for participative reflection on the accomplishment of transdisciplinary research programs", *Environmental Science & Policy*, Vol. 13/8, pp. 733-741, http://dx.doi.org/10.1016/j.envsci.2010.08.002. [77]

Schneider, F. et al. (2019), "Research funding programmes aiming for societal transformations: Ten key stages", *Science and Public Policy*, Vol. 46/3, pp. 463-478, http://dx.doi.org/10.1093/scipol/scy074. [87]

Scholtens, S. et al. (2019), *Life sciences data steward function matrix*, https://doi.org/10.5281/zenodo.2554973. [28]

Science Council (n.d.), *What is the Technician Commitment?*, https://sciencecouncil.org/employers/technician-commitment/ (accessed on 5 June 2020). [35]

Science Europe (2018), *Science Europe Data Glossary*, http://sedataglossary.shoutwiki.com/wiki/Main_Page (accessed on 5 June 2020). [65]

Scroggins, M. and I. Pasquetto (2020), "Labor Out of Place: On the Varieties and Valences of (In)visible Labor in Data-Intensive Science", *Engaging Science, Technology, and Society*, Vol. 6, p. 111, http://dx.doi.org/10.17351/ests2020.341. [36]

Software Sustainability Institute (n.d.), *Research Software Engineers*, https://www.software.ac.uk/research-software-engineers (accessed on 9 June 2020). [38]

Springer, R. (2019), *Counting Data Librarians*, https://sr.ithaka.org/blog/counting-data-librarians/ (accessed on 5 June 2020). [15]

SSI (2020), *Software Sustainability Institute joins forces with the UK Reproducibility Network*, https://www.software.ac.uk/news/software-sustainability-institute-joins-forces-uk-reproducibility-network (accessed on 8 June 2020). [61]

Stauffacher, M. et al. (2008), "Analytic and Dynamic Approach to Collaboration: A Transdisciplinary Case Study on Sustainable Landscape Development in a Swiss Prealpine Region", *Systemic Practice and Action Research*, Vol. 21/6, pp. 409-422, http://dx.doi.org/10.1007/s11213-008-9107-7. [85]

Swiss Academies of arts and sciences (n.d.), *td-net Network for Transdisciplinary Research*, http://www.transdisciplinarity.ch/en/td-net/Transdisziplinarit-t/Definitionen.html (accessed on 27 May 2020). [91]

Teal, T. et al. (2015), "Data Carpentry: Workshops to Increase Data Literacy for Researchers", *International Journal of Digital Curation*, Vol. 10/1, pp. 135-143, http://dx.doi.org/10.2218/ijdc.v10i1.351. [18]

The 2019 Data Science Leadership Summit (2019), *About the Summit*, https://sites.google.com/msdse.org/datascienceleadership2019/home/about-the-summit (accessed on 8 June 2020). [59]

The Academy of Medical Sciences (2019), *Artificial intelligence and health: Summary report of a roundtable held on 16 January 2019*, https://acmedsci.ac.uk/file-download/77652269. [54]

The Carpentries (2019), *The Carpentries Equity, Inclusion, and Accessibility Roadmap*, https://carpentries.org/files/assessment/equity_inclusion_accessibility_roadmap.pdf. [34]

The Royal Society (2019), *Dynamics of data science skills*, https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/ (accessed on 5 June 2020). [3]

The World Commission on Environment and Development (1987), *Report of the World Commission on Environment and Development - Our Common Future*. [79]

Tress, B., G. Tress and G. Fry (2006), "Defining concepts and the process of knowledge production in integrative research", in Tress, B. et al. (eds.), *From Landscape Research to Landscape Planning: Aspects of Integration, Education and Application*, Springer, Dordrecht. [71]

TUBİTAK (n.d.), *Research Data Management Training Portal*, https://acikveri.ulakbim.gov.tr/ (accessed on 8 June 2020). [49]

TUBİTAK (n.d.), *TUBITAK Open Science Policy Accepted*, https://ulakbim.tubitak.gov.tr/en/haber/tubitak-open-science-policy-accepted (accessed on 8 June 2020). [50]

University of Bristol (n.d.), *The UK Reproducibility Network*, http://www.bristol.ac.uk/psychology/research/ukrn/ (accessed on 8 June 2020). [60]

Van Noorden, R. (2015), "Interdisciplinary research by the numbers", *Nature*, Vol. 525/7569, pp. 306-307, http://dx.doi.org/10.1038/525306a. [70]

Vector Institute (n.d.), *List of Recognized AI-Related Programs*, https://vectorinstitute.ai/list-of-recognized-ai-related-programs/ (accessed on 5 June 2020). [94]

Working Towards Sustainable Software for Science: Practice and Experience (n.d.), *About WSSSPE*, http://wssspe.researchcomputing.org.uk/about-wssspe/ (accessed on 8 June 2020). [40]

Wright Morton, L., S. Eigenbrode and T. Martin (2015), "Architectures of adaptive integration in large collaborative projects", *Ecology and Society*, Vol. 20/4, http://dx.doi.org/10.5751/es-07788-200405. [80]

Zinsstag, J. et al. (2015), *One Health: The Theory and Practice of Integrated Health Approaches*, CABI. [92]

ANNEXES

Annex 1: Expert Group members

Country	Name	Affiliation
Australia	Michelle Barker (Chair of Expert Group)	Director, Skilled Workforce and Partnerships, ARDC
Belgium	Bart Dumolyn*	Department of Economy, Science and Innovation, Flemish Government, Brussels Area
Belgium	Inge van Nieuwerburgh*	Ghent University Library
Canada	David Castle	Vice-President Research, University of Victoria
Chile	Marcelo Arenas	Professor, Dept of Computer Science, Catholic University of Santiago
European Commission	Konstantinos Repanas	Unit for Open Science, Directorate-General Research and Innovation
European Commission	Carlos Casorran	Unit for Open Science, Directorate-General Research and Innovation
France	Nathalie Denos*	Higher Education General Directorate, Ministry for Higher Education and Research
France	Mehdi Gharsallah*	Higher Education General Directorate, Ministry for Higher Education and Research
Germany	Ingvill C. Mochmann	Head of EUROLAB, Knowledge Transfer, GESIS-Leibniz Institute for the Social Sciences, Vice President for Research and Knowledge Transfer, Professor of International Politics, Cologne Business School
Japan	Nobukazu Yoshioka	Associate Professor, National Institute of Informatics
Korea	Seo-Young Noh	Assistant Professor, Chungbuk National University, Korea Institute of Science and Technology Information (KISTI)
Netherlands	Karel Luyben	Rector Magnificus Emeritus, Delft University of Technology
Norway	Gard Thomassen	Assistant Director, University Centre for Information Technology, University of Oslo
United Kingdom	David McAllister	Associate Director, Research and Innovation Talent- UKRI-Biotechnology and Biological Sciences Research Council (BBSRC)
United Kingdom	Kevin Ashley	Director, DCC
United Kingdom	Lauren Clarke	International Policy Manager, UK Research and Innovation
United States	Daniel S. Katz	Assistant Director, National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign
United States	Todd K. Leen	Program Director, Computer Information Science and Engineering, Information Intelligent Systems, NSF
The Carpentries	Tracy Teal	Executive Director, The Carpentries
CODATA	Simon Hodson*	Executive Director, CODATA, International Science Council
SSI	Neil Chue Hong	Director, SSI, University of Edinburgh

* Participated in the 1st EG meeting only.

Annex 2: Case study interview questions

1. ***General information***
   a) Please give a brief description of this initiative.
   b) How is it funded/supported? What is its business model?
   c) When did it start and what is its expected duration?

2. ***Overview and focus***
   a) What types of skills does your initiative aim to help develop, for who and through what delivery method/s?
   b) How did you decide which skills to focus on (e.g., generic vs discipline-specific, technical skills versus softer skills like collaboration)?
   c) Can you map highlight the cells in the meta-framework (that is provided as an annex) that are the main focus of your initiative?
   d) Are the roles identified in the meta-framework (see annex) appropriate for your project/policy/initiative? Is the framework missing some key roles?
   e) Are the skills (horizontal axis) identified in the meta-framework appropriate for your project/policy/initiative? Is the framework missing some key skills?
   f) Are there other initiatives with a similar focus to this one and do you cooperate with these?

3. ***Broader context and drivers***
   a) What was the main driver for your program? Has your program been driven by an assessment of supply and/or demand (or some sort of market analysis), and if so, what?
   b) Does your program look at broader enablers to facilitate the development of a skilled workforce, such as career paths and recognition for different types of research outputs, inclusion of digital expertise in research teams, funder mandates to recognise the importance of digital professionals etc.? How does your initiative fit with this broader context?
   c) Can you identify broader enablers that support the success of your program, e.g. does your program implement more successfully when certain types of support from other organisations is present?

4. ***Policy requirements***
   a) What are the best examples of actions (e.g., mandates or incentives, including funding) which have led to improvements in digital skills development for data intensive research at scale?
   b) What actions are required to promote digital skills for research by the following actors:
      - National or regional governments
      - Research agencies
      - Professional science associations
      - Research institutes
      - Research community

5. ***Future perspectives***
   a) What are the most in-demand digital skills for science? What are the skills whose demand is growing fastest? Is there any skill that you think will not be so important in the future?
   b) Does your initiative take into account how research may be different in 10 years (influenced by digitalisation, e.g., AI)?
   c) Do our current skills programs/policies ready us for this?
   d) If not, what new skills or new programs will be needed? What enablers would be needed to support these?

Digital Skills Meta-Framework

This framework is designed to help define the scope and principle focus of initiatives that focus on digital skills for science, *i.e.* the case studies. Whilst some initiatives may cover all stages of the research cycle (horizontal axis) and all roles or skill groupings (vertical axis), it is likely that most initiatives concentrate on particular aspects. Interviewees are invited to highlight those cells that are the main focus of their case study.

Research cycle → / Roles ↓	Legal and ethical frameworks	Data generation/access & data retrieval	Data management, storage & sharing	Research methods & Data analysis	Data visualisation & interpretation	Publication, outreach & transfer
Early career scientist						
Experienced scientist						
Data analyst*						
Data Engineer*						
Data managers/data* steward/data curators						
Research software engineer*						
Research support staff						

Definition of terms:

Data analyst: This is someone who knows statistics. They may know programming, or they may be an Excel wizard. Either way, they can build models based on low-level data. Most importantly, they know which questions to ask of the data.

Data engineer: Operating at a low level close to the data, they are people who write the code that handles data and moves it around. They may have some machine learning background.

Data manager/steward: A data steward is a person responsible for the management of data objects including metadata. These people think about managing and preserving data. They are information specialists, archivists, librarians and compliance officers.

Research software engineer: A growing number of people in academia combine expertise in programming with an intricate understanding of research. These Research Software Engineers may start of as researchers who spend time developing software to progress their research or they may start of from a more conventional software-development background and be drawn to research by the challenge of using software to further research.

Annex 3: Workshop on Digital Skills for Data Intensive Science

The workshop was held on 28-29 October 2019 at the Leibniz Institute for Social Sciences (GESIS), Cologne, Germany.

The workshop was organised in four main sessions that focussed on: 1. The broad context and drivers for digital skills; 2. Defining skills needs; 3. Policy requirements; 4. Future perspectives.

Attendees

Country	Name	Affiliation
Australia	Ms. Michelle BARKER*	Director, Skilled Workforce and Partnerships, ARDC
	Ms. Rachel WEBSTER	Head of Astrophysics, University of Melbourne
Belgium	Mr. Alexander BOTZKI	Manager BITS Core Facility, Bioinformatics, Flemish life Sciences Institute VIB
Canada	Mr. David CASTLE*	Vice-President Research, University of Victoria
Chile	Mr. Andres JORDAN	Director, MAS
France	Mr. Cedric LOMBION	Open Data and Data Literacy Consultant, School of Data / Open Knowledge France
	Mr. François MICHONNEAU	Infrastructure Team Lead, Curriculum Development Lead, The Carpentries
	Ms. Laura MOLLOY	Committee on Data for Science and Technology, CODATA
Germany	Ms. Helene BRINKEN	State and University Library, Georg-August-University Göttingen
	Mr. Dietmar JANETZKO*	Professor of Information Systems and Business Process Management, Cologne Business School
Japan	Mr. Nobukazu YOSHIOKA*	Associate Professor, Information Systems Architecture Science Research Division, National Institute of Informatics
Netherlands	Mr. Alastair DUNNING	Head of Research Data Services at TU Delft and Head of 4TU, Centre for Research Data, Technische Universiteit Delft
	Ms. Celia VAN GELDER	Head of Training Platform, Dutch Techcentre for Life Sciences (DTL), ELIXIR
New Zealand	Mr. Nick JONES	Director, NeSI, University of Auckland
Norway	Mr. Gard THOMASSEN*	Assistant Director - Research Computing, University of Oslo
South Africa	Mr. Anwar VAHED	Director, Data Intensive Research Initiative of South Africa, Council for scientific and industrial Research (CSIR)
United Kingdom	Mr. Harriet BARNES	Head of Higher Education & Skills, British Academy
	Mr. Simon HETTRICK	Deputy Director, Software Sustainability Institute
	Ms. Corinne MARTIN	External Relations Officer, ELIXIR
	Mr. David MCALLISTER*	Associate Director – Research and Innovation Talent, BBSRC, UK Research and Innovation
	Mr. Ben MURTON	Head of Professional and Academic Development, the Alan Turing Institute
	Mr. Xavier POTAU	Principal Consultant, Technopolis
	Mr. Hugh SHANAHAN	Professor of Open Science, Department of Computer Science, Royal Holloway, University of London, Co-chair, CODATA-RDA School of Research Data Science
United States	Mr. Daniel S. KATZ	Assistant Director, National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign
	Ms. Tracy TEAL*	Executive Director, The Carpentries
European Union	Mr. Konstantinos REPANAS	Policy Officer, Open Science & EOSCDG Research
GESIS-Leibniz Institute for the Social Sciences	Mr. Fabian FLÖCK	Department Knowledge Transfer, Team EUROLAB, GESIS
	Ms. Ingvill Constanze MOCHMANN*	Department Knowledge Transfer, Team EUROLAB, GESIS
	Ms. Katrin WELLER	Team Leader "Social Analytics and Services, Computational Social Science, GESIS
Project consultant	Mr. Kevin ASHLEY*	Director, DCC, University of Edinburgh
OECD-STI/GSF	Mr. Carthage SMITH	Senior Policy Analyst
	Mr. Yoshiaki TAMURA	Policy Analyst

* Members of the OECD GSF Expert Group on Digital Skills for Data Intensive Science.

Glossary

One of the challenges in considering the digital workforce needs of science is reaching a common understanding across different communities, including scientists from different disciplines, and educators and policymakers from different backgrounds. This is further complicated by translation into different languages and cultures. Many detailed glossaries of digital terms in relation to science already exist. The key words that are used regularly in the current report – and which are likely to be useful to science policy makers - are defined below. These definitions have been adopted or adapted from other works in the field (Science Europe, 2018[65]; CASRAI, n.d.[66]; Babuska and Oden, 2004[67]) and are commonly used in the scientific research community.

Artificial Intelligence (AI)	AI is the theory and development of computer systems able to perform tasks normally requiring human intelligence
Big data	Big data is an evolving term that describes any voluminous amount of structured, semi-structured or unstructured data that has the potential to be mined for information. Big data is often dynamic and ensuring its reliability can be a challenge.
Data	Facts, measurements, recordings, records, or observations about the world, with a minimum of contextual interpretation. Data may be in any format or medium, including numbers, symbols, text, images, films, video, sound recordings, drawings, designs or other graphical representations.
Data curation	Data curation covers data selection, storage, preservation, annotation, provenance and other meta-data maintenance, and dissemination, and is needed to increase data interoperability. This includes the required hardware and software support for these tasks.
Data ethics	Data ethics is a new branch of ethics concerned with responsible use of data, algorithms and corresponding practices.
Data governance	The exercise of authority, control and shared decision making (planning, monitoring and enforcement) over the management of data assets.
Data-intensive science	Data-intensive science is considered to be the fourth paradigm of science after the three interrelated paradigms of empirical, theoretical, and computational science. It is seen as a data-driven, exploration-centred style of science, where IT infrastructures and software tools are heavily used to help scientists manage, analyse, and share data.
Data lifecycle	All the stages in the existence of digital information from creation to destruction. A lifecycle view is used to enable active management of the data objects and resource over time, thus maintaining accessibility and usability.
Data literacy	The ability to read, interpret, create and communicate data as information.
Data management	The activities of data policy development, data planning, data element standardisation, information management control, data synchronisation, data sharing, and database development, including practices that acquire, control, protect, deliver and enhance the value of data and information.
Data science	Data science encompasses the processes that deal with the extraction of meaning or knowledge from data.
Data stewardship	Data stewardship is the management and oversight of an organisation's data assets in order to provide professional users with high quality data that is easily accessible in a consistent manner.
Digitalisation	Digitalisation describes the way in which many domains of professional and social life are restructured around digital information and communication technologies.

Digital skills	Digital skills can be defined as a range of abilities to use digital devices, communication applications, and networks to access and manage information. In the context of science, these skills include an understanding of software, tools and data.
FAIR data	FAIR data is data which meet the principles of Findability, Accessibility, Interoperability, and Reusability.
Machine Learning	Machine Learning (ML) is a branch of Artificial Intelligence that is focused on developing systems that can learn from data. An ML algorithm is trained by learning from examples, which normally requires very large datasets.
Metadata	Literally, "data about data"; data that defines and describes the characteristics of other data, often using standardised formats.
Open data	Data that is accessible, freely shared. Open can be freely used, reused, built on and redistributed by anyone and may be subject to the requirement to attribute and share alike.
Open science	Open science promotes openness and early sharing of research ideas, papers, solutions, data and processes. Open science stresses the scientific, economic and societal benefits of increased and open scientific collaboration.
People-focussed skills	People-focussed skills differ from the system-focus of technical skills and include communication, teamwork collaboration, etc.
Research data	Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results.
Software	Software is a set of instructions, data or programs used to operate computers and execute specific tasks.
Roles	
Data analyst	This is someone who knows statistics. They may know programming, or they may be expert in spreadsheets. Either way, they can build models based on low-level data. Most importantly, they know which questions to ask of the data.
Data steward	A person responsible for planning and executing of all actions on digital data before, during and after a research project, with the aim of optimising the usability, reusability and reproducibility of the resulting data (Dutch Techcentre for Life Sciences, n.d.[68]).
Data scientist	A practitioner of data science. It is a generic term that encompasses many fields of specialised expertise. In the current report, data analysts, data stewards and research software engineers are considered as sub-groups of data scientists. In certain contexts, data scientist is also sometimes used in a more limited ways that make it equivalent to either the data analyst or software engineer roles.
Research Software Engineer (RSE)	A growing number of people in academia combine expertise in programming with an intricate understanding of research. These RSEs may start of as researchers who spend time developing software to progress their research or they may come from a more conventional software-development background and are drawn to research by the challenge of using software to further research.
Research support professionals	In the context of digitalisation, these are the people who support scientific researchers conducting data-intensive science. They are not necessarily part of a research team and might be considered as service providers. This is a broad category that can include data stewards, RSEs, data managers, librarians and archivists.