D7.1: Skills landscape analysis and competence model

Author(s)	Angus Whyte, Kevin Ashley (DCC-UEDIN)
Status	Final
Version	v1.1
Date	30/06/2017

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Abstract:

The report provides an analysis of EOSC stakeholder policy priorities for skills development, identifying these with data stewardship, a broad set of skills covering open science, data management, data science and engineering. Current professional development training is analysed by provider, with these including research infrastructures, e-infrastructures and a broad range of related initiatives, identifying the topical scope, nature of the events and materials involved. Drawing its scope from policy priorities the report also surveys relevant competence frameworks, and skills needs identified so far in EOSCpilot, including those from the first four active Science Demonstrator sub-projects. Informed by those needs, Section 6 provides a draft EOSCpilot competence framework, the first part of the skills framework. Conclusions for further engagement with stakeholders in skills development. These include the initial scope of training infrastructure, to catalogue existing provision and deliver events and materials to EOSCpilot users

The European Open Science Cloud for Research pilot project (EOSCpilot) is funded by the European Commission, DG Research & Innovation under contract no. 739563

Document identifier: EOSCpilot -WP0-D0.0	
Deliverable lead	**UEDIN**
Related work package	**WP7**
Author(s)	Angus Whyte, Kevin Ashley (DCC-UEDIN)
Contributor(s)	Femmy Admiraal (DANS), Themis Athanassiadou (EGI), Elly Dijk (DANS), Magdalena Getler (DCC-UEDIN), Marjan Grootveld (DANS), Eileen Kuehn (KIT), Jonathan Rans (DCC-UEDIN), Simone Sacchi (LIBER), Gergely Sipos (EGI), Jerry de Vries (DANS)
Due date	**30/06/2017**
Actual submission date	**30/06/2017**
Reviewed by	**Cath Brooksbank (EMBL-EBI),** **Simon Lambert (STFC)**
Approved by	**Brian Matthews (STFC)**
Start date of Project	**01/01/2017**
Duration	**24 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.1	24/05/2017	Angus Whyte <DCC-UEDIN>	Content outline
0.2	16/06/2017	Jerry de Vries (DANS), Marjan Grootveld (DANS), Gergely Sipos (EGI), Themis Athanassiadou (EGI), Elly Dijk (DANS), Jonathan Rans (DCC-UEDIN), Simone Sacchi (LIBER), Magdalena Getler (DCC-UEDIN), Femmy Admiraal (DANS), Angus Whyte <DCC-UEDIN>	Partially completed version for internal review
0.3	28/06/2017	As above	Revised following final internal review
1.0	**30/06/2017**	Angus Whyte, Kevin Ashley <DCC-UEDIN>	Final edits
1.1	**3/07/2017**	Angus Whyte, Kevin Ashley <DCC-UEDIN>	Added missing annexes

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The EOSCpilot work plan commits the project's Skills and Capacity workpackage to objectives in skills development, as follows: "…develop common standards and assessment frameworks to ensure organisations and individuals are motivated to develop the capabilities and competencies that the EOSC will rely on. It will establish the capabilities that organisations need to develop and reflect in the career development pathways of researchers and support service staff; and the skills needed by individuals to enhance their competencies in open data science and stewardship"

The *Skills landscape analysis and competence model* report (D7.1) is the first deliverable from that work. The contents shown in Figure 1 address specific workpackage objectives as follows:

1. Design an open data science skills framework that describes the individual competencies and organisational capabilities required to provide EOSC services of the required levels of quality

Drawing its scope from policy priorities reviewed in section 2, section 4 of the report surveys competence frameworks relevant to high priority skills gaps. Section 5 outlines skills needs identified so far in EOSCpilot, describing needs analysed from the first four active Science Demonstrator sub-projects. Informed by those needs, Section 6 provides a draft EOSCpilot competence framework, the first part of the skills framework.

2. Catalogue the current provision of education and training against this framework and identify gaps in delivery.

Section 3 analyses outcomes from the first phase of task 7.1 to map training needs being addressed by research infrastructures, e-infrastructures and related initiatives, and in task 7.2 to map existing provision in training and skills from a range of providers. The scope covers skills development activity in the areas of open science, data management, and data science, and identifies the topical scope, nature of the events and materials involved, and levels of competence addressed.

3. Develop an EOSC education and training strategy to address the gaps and set up a sustainable technical training infrastructure to ensure shared resources are openly accessible and reusable

Section 7 draws conclusions for further engagement with stakeholders in skills development. These include the initial scope of training infrastructure, to catalogue existing provision and deliver events and materials to EOSCpilot users. The next steps also include expansion of skills the framework, to allow organisations to identify their needs for individuals to develop competences, based on the capabilities the organisation itself will require for enabling effective use of EOSCpilot service. These will be identified in task 7.3, based on the capabilities these services deliver, informed by existing capability models and service certification frameworks.

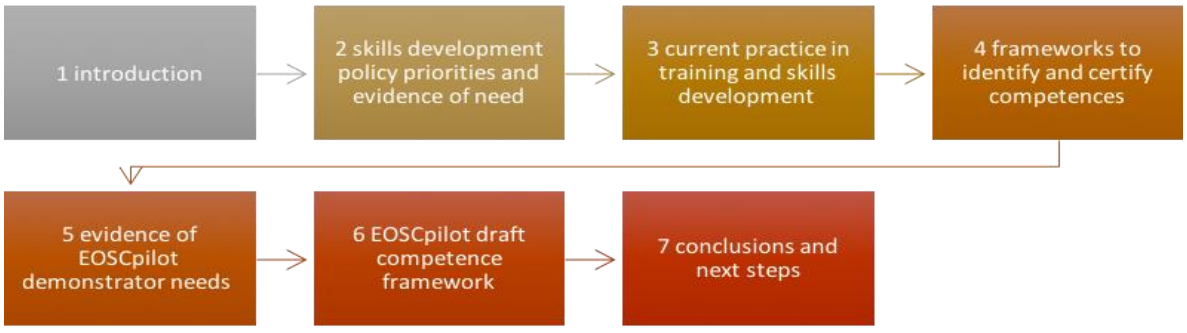

Figure A. EOSCpilot skills landscape analysis and competence model

1. SECTION 1 - INTRODUCTION

The report begins in its first section by establishing that policymakers and EOSC stakeholders have identified a priority to address skills gaps in data management and stewardship. The skills areas to be filled are among 'data experts', and among researchers themselves, and the report finds survey evidence to corroborate the need for training in these areas.

The D7.1 report is the first incarnation of open data science skills framework for EOSCpilot. This will shape the training activity within the pilot, and is intended to inform the broader EOSC vision and become a practical tool for organisations using EOSC services. The report also proposes an early draft representation, for further discussion, of the training infrastructure needed to deliver relevant events and materials. Along with recommendations on skills policy, the skills framework and training infrastructure will be the main elements the workpackage contributes to an EOSC strategy for education and training.

The Skills Framework, in this first version, comprises a set of competences in 'open data science stewardship'. These represent the union of two overlapping sets of competences; those to apply data science methods in research, and those to ensure that the resulting knowledge and evidence are FAIR (findable, accessible, interoperable and reusable). The report proposes that these competencies are needed to underpin the EOSC aims of ensuring wider access to scientific data and data-analysis services among European researchers. The competences should support the effective use of those services, and support the application of FAIRness principles in doing so. Figure 1.1 illustrates how the workpackage tasks and deliverables relate to each other.

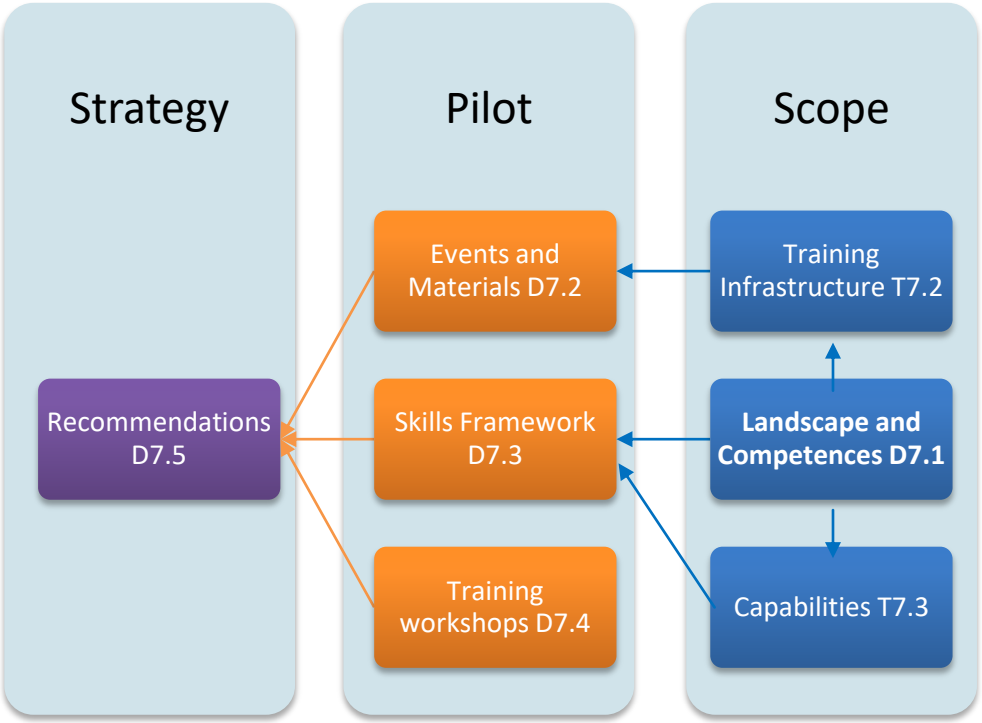

Figure 1.1 Skills and capacity workpackage tasks and deliverables

Section 3 of the report reviews the training and skills development landscape that EOSC is being built upon. While EOSC services will target researchers with services to support science and scholarship, the customers and stakeholders of those services include research organisations, institutions, and professional

organisations. As the report shows, each of these groups, plus a variety of community initiatives, are already providing training in open science, data management and data science. We highlight illustrative examples of a range of current or recent approaches, from formal training in data management to knowledge transfer 'on the job', with delivery approaches ranging from online to blended, and from summer schools to MOOCs.

Section 4 looks more closely at specific EOSCpilot needs, as far as these can be identified from the initial stages of the project's work, and from the first 4 of 15 Science Demonstrator sub-projects. These are developing applications to run as services for data science using the federated infrastructures the pilot provides. Demonstrator contacts indicated specific competence needs in the areas of service integration into the EOSCpilot, service management, and enabling effective use of the service. These competences extend across a number of skills areas including data science engineering as well as data management and data science research itself.

Competence frameworks are the focus of Section 5, which reviews current work directly relating to the scope of the workpackage. For training and skills development, as in other aspects of its work, EOSCpilot needs to impact on consensus-building and standardization if it is to underpin better integration of research and e-infrastructures.. Section 6 synthesizes competence frameworks with that aim in mind, and proposes use cases for applying the resulting draft framework in organisations. These will be used for further discussion with partners and related initiatives, on the scope of the open data science stewardship competences.

In the concluding section 7 we consider the next steps for the Skills Framework, and the scope of the training infrastructure in a skills development strategy for EOSC. Cloud-based data services require 'orchestration' of component services to respond to user needs on demand and at scale. Similarly, the EOSC strategy will need to devise infrastructure for a federation of training providers to orchestrate their delivery of skills and capacity as required, to support large scale collaborative R&D. We offer a high-level view of a stack of services for delivering 'Training as a Service', as a reference point to engage stakeholders in further defining the infrastructure.

Finally, we propose the next steps required to align our approach to skills development with the emerging outputs of EOSCpilot in the areas of governance, policy, demonstrators, services, integration and stakeholder engagement.

The main conclusions of the report are as follows.

1. EOSC training materials and events must be FAIR, i.e. materials and event descriptions must be provided with standard metadata to make them findable, they must be accessible from EOSC e-infrastructures, they must be in open standard formats so they are interoperable with each other and with the data services they are about, and they must be provided on license terms that are as non-restrictive as possible to encourage reuse.
2. The coordination and delivery of EOSC training materials and events should be managed on a similar basis to the services they describe i.e. provided in the form of 'Training-as-a-Service'.
3. EOSCpilot Wp7 should use opportunities provided through Wp2 (policy) and Wp8 (engagement) to highlight the relevance of the institutional role in enabling and rewarding data skills development.
4. EOSCpilot should consult stakeholders on a skills strategy for EOSC that, in addition to formal courses, includes skills development approaches embedded in data-intensive research environments, e.g. through Fellowships and staff exchanges.

5. EOSCpilot should focus its training provision on the *outcomes* from Science Demonstrators, e.g. illustrating how these help researchers apply FAIR principles, or provide lessons in service management, rather than attempt to deliver training within the Science Demonstrator projects.
6. EOSCpilot should refine its Skills Framework through engagement with stakeholders in the development of careers and expertise in data stewardship.
7. Further investigation of the ELIXIR Training e-Support System (TeSS) is needed to establish whether it may be recommended as a cross-domain solution for harvesting training events/ materials, and whether these may be tagged with competences from EOSCpilot or other frameworks.
8. Broader consultation is needed with the EOSCpilot governing bodies and WPs to establish how far certification will feature in the 'rules of engagement' for EOSC, and therefore how it should apply to education and training. Further consideration is needed on the balance between broad-based certification schemes applicable to services/management, and badging schemes applicable to specific events or materials.

2. POLICY PERSPECTIVES AND EVIDENCE OF A DATA SKILLS GAP

2.1. EOSCpilot in context

According to the first report of the High Level Expert Group on the European Open Science Cloud (EOSC),[1] the EOSC will be an accessible infrastructure for modern research and innovation implementing an internet of Findable Accessible Interoperable and Reusable data and services.[2] It should include the required human expertise, resources, standards, best practices and underpinning infrastructures. It will support Finding, Access, Interoperation and in particular the Re-use of open, as well as sensitive and properly secured data. It will also support the data related elements (software, standards, protocols, workflows) enabling re-use and data driven knowledge discovery and innovation. Professional data management and long-term data stewardship are therefore important to EOSC.

In Europe, domain-specific European Research Infrastructures and cross-domain ICT e-Infrastructures, as well as other disciplinary and cross disciplinary collaborations are already well established. These are the ground for EOSC. As noted in the introduction and described in more detail in section 3, these infrastructures are providers of training relevant to EOSCpilot. However, they are also one of the targets of that training. This is because realising the ambition of seamless access, reliable re-use of digital research objects, and greater collaboration across services and infrastructures, requires further enhancements to turn the "*ever increasing amounts of data […] into knowledge as renewable, sustainable fuel for innovation in turn to meet global challenges*". The EOSC is the instrument defined by the European Commission to foster such evolution towards the realisation of the so-called Open Science.

2.2. Open science, data science

Recent years have seen a proliferation of policy statements on Open Science and research data management by policy actors at international, national and institutional levels. In addition to the European Union these range from the OECD and G8 science ministers to learned societies, national research funding bodies and universities, through to research centres and departments. Such statements have played a key role in articulating researchers' responsibilities and stimulating service development to implement policy goals.

The policy rationale is articulated by the G8 science ministers, who state that 'to ensure successful adoption by scientific communities, open scientific research data principles will need to be underpinned by an appropriate policy environment.'[3] A common expectation is of cultural and behavioural change amongst researchers and research support environments. A link between such change and skills development is expressed explicitly in the UK Concordat on Open Research Data, which states that 'support for the development of appropriate data skills is recognised as a responsibility for all stakeholders'.[4]

The Joint position paper on the European Open Science Cloud issued by Germany and the Netherlands also explicitly calls for 'building up competences for research data management, including the training of data stewards capable of providing FAIR data services'[5] This echoes the EOSC High Level Expert Group report , which also highlights expectations that research data will be subject to appropriate data stewardship. The HLEG report recommended "a concerted effort to develop core data expertise in Europe" and to 'urgently develop adequate data stewardship capacity in European Member States". They refer to "a distinct and largely novel class of research professional... embedded data specialists that are able to support domain

[1] First report of High Level Expert Group on the EOSC: https://ec.europa.eu/digital-single-market/en/news/first-report-high-level-expert-group-european-open-science-cloud

[2] The FAIR Guiding Principles for scientific data management and stewardship: http://www.nature.com/articles/sdata201618

[3] G8 Science Ministers, https://www.gov.uk/government/publications/g8-science-ministers-statement-london-12-june-2013

[4] Concordat on Open Research Data. 2015. http://www.rcuk.ac.uk/media/news/160728/

[5] Joint position paper on the European Open Science Cloud, Germany and the Netherlands https://www.dtls.nl/wp-content/uploads/2017/05/DE-NL-Joint-Paper-FINAL.pdf

specific researcher throughout the entire knowledge discovery cycle." [1]

The EOSC HLEG also noted that lack of recognition for data skills by institutional promotional boards is a barrier to progress. The research career development implications of 'Science 2.0' were also highlighted in the report of a European Commission consultation on that topic (which can be considered a precursor of 'data science').[6] [7] Policy recommendations included that the Commission should consider revising the *Charter for Researchers & the Code of Conduct for the Recruitment of Researchers* (the 'Charter and Code'). Institutions taking part in the consultation highlighted better acknowledgement of Science 2.0 (open science, data science) activities in recruitment and career progression, as a key non-financial incentive.

Research skills and careers have been the focus for action by the European Commission on the European Research Area, with the Steering Group on Human Resources and Mobility (SGHRM) having a specific focus on this. The SGHRM has recently adopted a focus on the Digital Agenda, with a brief to inform the work of the more recently initiated Open Science Policy Platform, through 8 working groups considering a broad range of open practices, as illustrated in Figure 2.1 below.

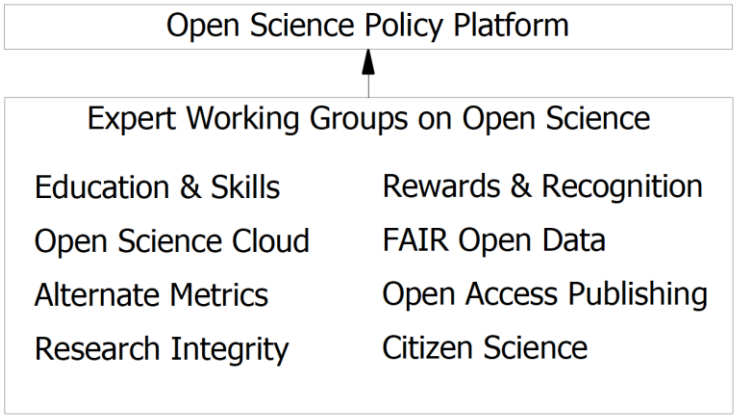

Figure 2.1 Scope of OSPP Expert Working Groups

Some national-level funders encourage institutions to address career progression issues, e.g. the Wellcome Trust.[8] [5] Increasingly, national level research data policies are explicitly identifying the development of relevant skills as a responsibility for research producing organisations. An analysis of research data policies by DCC for SPARC Europe identifies seven European national level policies doing this - Denmark, Estonia, Finland, the Netherlands, the UK, Norway and Switzerland.[9]

At an institutional level, the League of European Research Universities' (LERU) "roadmap for research data" acknowledges that, for many researchers, the necessary skills simply represent good practice for research conduct, whilst identifying that there is a need for discipline-specific doctoral courses to enable

[6] European Commission DG Research and Innovation (2016) 'Validation of the results of the public consultation on Science 2.0: Science in Transition' http://www.eesc.europa.eu/resources/docs/validation-of-the-results-of-the-public-consultation-on-science-20.pdf

[7] According to the above consultation report, "'Science 2.0' describes an on-going evolution in ways of doing and organising research. These changes are enabled by digital technologies, and they are driven by globalisation and growth of the scientific community as well as the need to address the grand challenges of our time. The changes impact the modus operandi of the entire research cycle, from the inception of research to its publication, as well as the way this cycle is organised." (p. 4).

[8] Ashley, Kevin (2016): Developing skills for managing research data and software. https://doi.org/10.6084/m9.figshare.4133916.v1.

[9] SPARC Europe (2017) An Analysis of Open Data and Open Science Policies in Europe http://sparceurope.org/what-we-do/open-data/sparc-europe-open-data-resources/

development of Data Scientists.[10] Policy support for research data skills development is also found among individual institutions. For example an analysis of UK university policies on research data management indicates that 45 of 57 policies examined define institutional research support roles.[11] These research support roles typically include provision of generic training in data management to research staff. The support roles themselves also require training in stewardship, whether that is carried out by an institution-level facility or by a domain-level repository, indicating a need to differentiate between the various roles, the competences relevant to each, and their respective levels. We return to this issue in sections 5 and 6.

2.3. Evidence of need from recent surveys and reports

There has been a number of studies published recently that highlight the need for a common vocabulary for the work of data scientists/experts, and for more training to build capacity in data expertise. There is also evidence of demand for reward structures that ensure data stewardship is evaluated on suitable criteria, beyond those used to evaluate research itself (e.g. publication of peer-reviewed papers). Although data sharing is of global interest, cultural and national factors continue to pose significant challenges. Data management requires significant effort, and training and resources are needed.

2.3.1. Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work

A report by O'Reilly/ STRATA (Vaisman et al, 2013)[12], describes a survey of data science practitioners self-perceptions of skills and careers. Seeking to clarify the skills associated with 'data scientist', 'analytics' or 'big data' in order to "efficiently match talent to projects", the survey focused on five areas: skills, experiences, education and self-identification. The report claims to offer more precise vocabulary to talk about the work the data scientists do based on how they describe themselves and their skills. The responses from the survey were clustered into categories and respondents fell into four self-identified groups of Data Developer, Data Researcher, Data Creative and Data Businessperson, and five Skills Groups of Business, Machine Learning/Big Data, Math/Operations Research, Programming and Statistics.

The authors assert that the most successful data scientists are those with substantial expertise in at least one aspect of data science (for example in statistics, big data or business communication). They relate their analysis to the notion, promoted by IBM among others, that information professions have a T-shaped skills profile (where the 'T' represents breadth of skills across the top, with depth in one area represented by the vertical bar). Data science, the authors argue, is an inherently creative and collaborative field, where the 'successful professional can work with database administrators, business people, and others with overlapping skill sets to get data projects completed in innovative ways'. The report emphasizes the importance of integrating data scientists with other roles, training provision, and career path development.

2.3.2. Belmont Forum Skill Gap Analysis

The Belmont Forum is a partnership of funding organizations, international science councils, and regional consortia committed to advancing interdisciplinary and transdisciplinary science. Its e-Infrastructures and Data Management group (e-IDM) commissioned in 2016 a skill gap analysis survey to support training and curricula in data-intensive environmental science for delivery to environmental, social and computer scientists. Survey responses (76%) were mostly from researchers and data managers from government or universities in Europe and North America. The gap analysis report [13] highlighted the that most respondents identified a need to address researcher reluctance to share data or models. Most also identified the most vital skill for global change research as data processing and analysis. Overall, the biggest data challenge

[10] LERU (2013) Roadmap for Research Data. http://www.leru.org/index.php/public/news/press-release-leru-roadmap-for-research-data/

[11] DCC (n.d.) UK Institutional Data Policies. http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies

[12] Full report available here: https://www.safaribooksonline.com/library/view/analyzing-the-analyzers/9781449368388/

[13] Full report available here: http://bfe-inf.org/document/skills-gap-analysis

was data complexity and lack of data exchange standards. The report highlights a need for more training on data management, and for capturing detailed metadata from the onset of a research project.

2.3.3. Leiden/Elsevier "Open Data: The Researcher Perspective"

The report is of a study co-conducted by Elsevier and the Centre for Science and Technology Studies (CWTS), Leiden University.[14] The study combines bibliometric analysis with a global survey of 1,200 researchers and three disciplinary case studies. The report investigates the extent of open data sharing practice, researcher perspectives of the barriers and drivers, and the effects of new data sharing practices and infrastructures on knowledge production processes and outcomes. The findings highlight that data-sharing practices are domain-dependent. In intensive data-sharing fields, data sharing is embedded into the research design and execution. Researchers acknowledge the benefits of open data, but sharing practices are still limited. Reasons mentioned include: research data management and privacy issues, proprietary aspects and ethics; insufficient training in data sharing, and lack of associated credit or reward.

2.4. Summary: skills policy and survey evidence

The vision of the EOSC High-Level Expert Group is of a commons of data, software, standards, expertise and a policy framework relevant to data-driven science. The EOSC will implement an internet of FAIR data and services, building on and further integrating the already established European Research Infrastructures and e-Infrastructures. EOSC must provide further enhancement to ensure seamless access and reliable reuse of digital objects.

International, national and local policies are placing increasing emphasis on building competencies to enable open science. Harmonisation across policies is required as is more granular detail and recommendations on delivery. While international and national-level data policy statements identify a need for research-producing organisations to promote data science, data sharing and good data management practices, there are relatively few policy statements or incentives for organisations and individuals to develop the necessary skills, or address the reported lack of credit or rewards for these practices.

14

3. SKILLS DEVELOPMENT INITIATIVES AND RESOURCES RESPONDING TO NEEDS

3.1. Skills initiatives from Research Infrastructures and e-Infrastructures

3.1.1. Introduction

The Research Infrastructures (RIs) and e-Infrastructures (e-Infras), the building blocks of EOSC, have a great deal of well-established skills development activity. According to the consultation report on *long-term sustainability of Research Infrastructures* a large majority of the 189 interviewed organisations indicated they already offer specific training for RI users.[15] According to the report, Research Infrastructures also provide skills development for their managers and, as well as formal training, this includes 'on-the-job' or embedded skills development, such as staff exchanges, and 'sharing of experience and best practices'.

Skills development activity is likely to expand, in keeping with policy expectations. For example, one of 12 guidelines in the *European Charter for Access to Research Infrastructures* encourages RIs to "offer education and training in the areas of their activities, and to collaborate with other institutions and organisations that benefit from using the Research Infrastructure for their education and training purposes."[16] And for e-Infrastructures, metrics relating to training activity have been included in the key performance indicators proposed by the e-Infrastructure Reflection Group.[17]

In view of the scale and scope of the skills development landscape, our analysis has highlighted illustrative examples, rather than attempt a comprehensive census or random sample. This report is a snapshot of ongoing activity to map existing provision of relevant events and materials, to inform EOSC training materials, to be collated in D7.2.

Our preliminary analysis provides a breakdown of examples by provider, focusing on RIs, e-infras, Horizon 2020 projects carrying out integrative activities across these ('cluster projects') and from related initiatives by other stakeholders including institutions and scientific or professional associations. The section describes the *range* of resource and event types (e.g. summer schools, workshops/ webinars, courses) *delivery mode*, e.g. face-to-face, online, blended) and *intensity* –(fulltime, part-time, occasional). We give separate attention at the end of the section to *embedded* approaches (e.g. fellowship, internship or exchange programmes) that offer longer duration opportunities for skills development.

3.1.2. Research infrastructures

AQUAEXCEL2020

The project *AQUAculture infrastructures for EXCELlence in European fish research towards 2020* (AQUAEXCEL2020) aims to advance aquaculture research and innovation in Europe by providing access to facilities, and services for researchers. Training courses provided by AQUAEXCEL2020 are multi-partner collaborations to create innovative modules promoting and enabling peer-to-peer networking. The project focuses on participative training design to ensure exchange and mutual learning between instructors and participants from academia and industry. Face-to-face and distance learning courses cover state-of-the-art knowledge and insights originating from research and services carried out and created by the project. Training includes experimental data management covering experiment protocols, data acquisition, processing and sharing. There is a focus on working on a national, cross-organisational scale. [18]

[15] European Commission DG Research and Innovation (2016) Report on the Consultation on Long Term Sustainability of Research Infrastructures https://ec.europa.eu/research/infrastructures/pdf/lts_report_062016_final.pdf

[16] European Commission DG Research and Innovation (2016) European Charter for Access to Research Infrastructures (p. 12) https://ec.europa.eu/research/infrastructures/pdf/2016_charterforaccessto-ris.pdf

[17] e-IRG Guidelines Document (2016) Evaluation of e-Infrastructures and the development of related Key Performance Indicators http://e-irg.eu/documents/10920/238968/Evaluation+of+e-Infrastructures+and+the+development+of+related+Key+Performance+Indicators.pdf

[18] ACQUAEXCEL Training http://www.aquaexcel2020.eu/training-courses/upcoming-training-courses-apply-now

CERN

The European Organization for Nuclear Research CERN provides instruments for physicists and engineers to study fundamental particles. The scale of data produced by experiments requires global effort to properly manage and analyse data. Reflecting the complex analysis and computing environment, the training focus is on providing a fundamental understanding of concepts, technologies and infrastructure across a broad range of CERN scientists from various domains as well as technical staff. CERN training combines the required understanding of data handling in globally interlaced efforts required by early career particle physicists, while teaching senior scientists about new technologies and concepts. There is an emphasis on exchange between junior and senior scientists to promote direct knowledge exchange.

CERN training also delivers in-depth information into current research in data preservation and long term analysis in high energy physics, fostering knowledge transfer by providing tools and best practices in data management, archival management and data ingest processes[19]. The CERN Open Data portal delivers educational material based on some of the implemented concepts. [20]

ELIXIR

ELIXIR integrates life science resources across Europe for managing and safeguarding the increasing volume of data being generated. The infrastructure integrates resources including databases, software tools, training materials, cloud storage and supercomputers, to enable users in academia and industry to find and share data, exchange expertise and agree on best practices.

ELIXIR develops and operates an online Training eSupport System (TeSS) registry/catalogue. This takes an innovative approach to aggregating and integrating life science training materials and courses from ELIXIR nodes, collaborating with participating sites to do this automatically, while enhancing their search visibility. The training platform integrates delivery of e-learning and face-to-face training to researchers, developers, and bio-informatics specialists. Scope includes effective use of data, tools, standards and compute infrastructure. A dedicated Train the Trainers programme is offered. ELIXIR collaborates closely with EMBL-EBI and with GOBLET, the Global Organisation for Bioinformatics Learning, Education & Training.[21]

SoBigData

The *Social Mining and Big Data Ecosystem Research Infrastructure* (SoBigData) aims to strengthen existing clusters of excellence in social data mining research, and create a pan-European, community of social data scientists through extensive training, networking and innovation activities.[22] These include summer schools for PhDs in science and engineering, training modules for stakeholders, and a series of hackathons. A mixture of practical hands-on approaches and accompanying reference material is used in face-to-face trainings. Open source training modules are offered through a training repository on big social data.

3.1.3. Cluster projects

PARTHENOS

Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (PARTHENOS) aims at strengthening the cohesion of research in the broad sector of Linguistic Studies, Humanities, Cultural Heritage, History, Archaeology. The PARTHENOS Training Plan is targeted at users of digital humanities research infrastructure. [23] Training modules are provided at beginner, intermediate, and advanced levels. Although organised to suit self-learners as well, the PARTHENOS Project training materials are primarily intended as 'train-the-trainers' support materials. A 'For trainers' page offers public access to all training resources: 'Research Infrastructures 101' videos, training slides, suggested course outlines (e.g.

[19] CERN Collaboration Agreement http://hep-project-dphep-portal.web.cern.ch/content/collaboration-agreement
[20] CERN Open Data http://opendata.cern.ch
[21] ELIXIR Training https://www.elixir-europe.org/platforms/training
[22] SoBigData Training http://project.sobigdata.eu/workpackages/wp4-training-and-best-practice-guidelines
[23] PARTHENOS Training http://training.parthenos-project.eu/wp-content/uploads/2016/10/D7.1-Initial-Training-Plan-FINAL.pdf

a weeklong summer school programme), brochures and other teaching resources.

SERISS

Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) aims to equip Europe's social science data infrastructures to play a major role in addressing the key societal challenges facing Europe today, and ensure that national and European policymaking is built on a solid base of the highest-quality socio-economic evidence. Training is provided online and face-to-face, and covers four main topics: how to use SERISS tools for cross-national research; Data management (collection, archiving, and dissemination); Statistical training for secondary data users, and Handling and harmonising survey data, the latter including train-the-trainer modules. Overall, SERISS training and dissemination aims to increase data literacy within the research community and to raise scientific standards.[24]

CORBEL

The *Coordinated Research Infrastructures Building Enduring Life-sciences Services* (CORBEL) is an initiative of eleven biological and medical research infrastructures. The project aims at creating unified user access to biological and medical technologies, biological samples and data services required for biomedical research. The project organises training regarding the competency requirements of newly proposed services, and uses these as the basis for a pilot training program.[25] The program consists of staff exchange to further develop operational expertise as well as webinars addressing challenges and best practices of biological and medical research infrastructures. The training focuses on data management and integration, physical access, ethics, and innovation and targets technical operators of research infrastructures.

ENVRIplus

ENVRIplus brings together Environmental and Earth System Research Infrastructures. ENVRIplus Theme 5 ensures the cross-fertilisation and knowledge transfer of new technologies, best practices, approaches and policies of the Research Infrastructures. The theme comprises two workpackages, covering training and service deployment/validation. In the training WP, topics emerging as high priority items from a user community survey included: creating data federations; Hosting data-intensive services in EGI; E-infrastructure services for the long tail of science; Security incident handling, methods and forensics. The WP on service deployment and validation aims to facilitate the exchange of (knowledge among) staff working in Research Infrastructures or related to the implementation of new RIs.

3.1.4. E-infrastructures

EGI

EGI is a collaboration of computing resource providers that delivers integrated computing services to European researchers. It provides access to computing (including closely coupled parallel computing normally associated with HPC), compute workload management services, data access and transfer, data catalogues, storage resource management, cloud computing. This involves 350 resource centres distributed across 56 countries in Europe, the Asia-Pacific region, Canada, and Latin America. EGI services aim to remove the need for research communities to develop and operate their own bespoke services. User communities gain access to EGI services by partnering with EGI, either directly through federating their own resource centres, or indirectly by accessing national or regional resource centres that already support their communities. [26] The funding support for access varies widely across the federation. EGI members run training events (both face-to-face and online), driven by the EGI training plan and with hands-on exercises using the cloud-based EGI training infrastructure [27],[28]. The EGI Foundation also offers training about FitSM,

[24] SERISS Training http://seriss.eu/about-seriss/work-packages/wp5-training-and-dissemination/

[25] CORBEL Training http://www.corbel-project.eu/work-packages/wp9-training.html

[26] EGI Services https://www.egi.eu/services/

[27] EGI Training http://go.egi.eu/trainingplan

a lightweight standard for IT service management. [29]

GÉANT

GÉANT delivers the pan-European network for scientific excellence, research, education and innovation by providing highly reliable, unconstrained access to computing, analysis, storage, applications and other resources. [30] The GÉANT network is the largest and most advanced Research and Education network in the world to support open innovation, collaboration and knowledge-sharing amongst the members, partners and the wider research and education networking community. The delivery of up-to-date training on GÉANT services is a core requirement of the project to enable the community to use the leading edge technologies provided by GÉANT to their best effect in supporting knowledge transfer across Europe. The GÉANT training, learning and development team provides self-paced, onlineand face-to-face training as well as a range of personal and skill development courses for participants in the GÉANT project and for the wider community. [31]

PRACE

The *Partnership for Advanced Computing in Europe* (PRACE) aims to create a pan-European supercomputing infrastructure, providing access to computing and data management resources and services for large-scale scientific and engineering applications at the highest performance level while it seeks to improve energy efficiency of computing systems and reduce their environmental impact. PRACE is in a transition towards a persistent and sustainable Research Infrastructure. With regard to the needs of researchers, PRACE provides a sustained, training and education service for the European HPC community through advanced training centres, seasonal schools targeting a broad HPC audience, workshops focusing on particular technologies and scientific and industrial seminars.[32]

EUDAT

EUDATis the largest pan-European data infrastructure initiative initiated under the European Commission's FP7 programme and is conceived as a network of cooperating centres. EUDAT's vision is data shared and preserved across borders and disciplines, and its mission is to enable data stewardship within and between European research communities through the EUDAT Collaborative Data Infrastructure.[33] The project offers a common model and service infrastructure, aim to address the full lifecycle of research data. [34] The EUDAT training programmeis a training and education service for European research communities, infrastructures and data centres on how EUDAT services support and facilitate research data management. [35] The training programme is delivered through e-training components for self-study (presentations as well as scripted hands-on modules), workshops for researchers and system administrators, webinars and a summer school for early-career researchers. EUDAT has developed a project-internal workflow in which the expertise of the service developers feeds the team that is responsible for user documentation, which in turn feeds into the training materials.

OpenAire

The *Open Access Infrastructure for Research in Europe* aims to establish an open and sustainable scholarly communication infrastructure responsible for the overall management, analysis, manipulation, provision, monitoring and cross-linking of all research outcomes (publications, datasets, software and services) across repositories. The project's objective is to promote the discoverability and reuse of data-driven research results across scientific disciplines and thematic domains. The current OpenAIRE2020 supports the

[28] EGI Training Infrastructure https://www.egi.eu/services/training-infrastructure/

[29] FitSM Training https://www.egi.eu/services/fitsm-training/

[30] GEANT https://www.geant.net/About/Pages/home.aspx

[31] GEANT Training http://cbt.geant.net

[32] PRACE Training http://www.training.prace-ri.eu

[33] EUDAT Collaborative Data Infrastructure https://www.eudat.eu/eudat-cdi

[34] EUDAT Services https://www.eudat.eu/services-support

[35] EUDAT Training https://eudat.eu/training

Horizon2020 vision of open access (OA) for scientific publications and research data. A current focus is delivery of a robust pan-European research infrastructure that monitors OA research outcomes from the EC, and other national funders, and towards enhancing the potential of international OA repositories collaboration to support global research and scholarly communication. In support of those goals, OpenAIRE offers training via a set of guidance materials (i.e. *OpenAIRE Guide for Researchers*, *Research Managers: What's in it for you?*, *OpenAIRE Guide for Journals*), workshops (the recent ones include Legal Issues in Open Research Data, Open Peer Review: Models, Benefits and Limitations) and webinars (OpenAIRE guidelines for data and literature repositories, FAIR Data in Trustworthy Data Repositories).

3.2. Open Science and data science training in the broader community

Alongside the work of the Research Infrastructures and e-infrastructures, parallel initiatives have addressed gaps in research community awareness of open science and data science. In this section we present indicative examples, categorised by the following stakeholder groups.

- *Collaborative projects* with fixed-term EU or US funding : EDISON, FOSTER/ Foster Plus, Data Information Literacy, Open Science Data Cloud
- *National-level programmes*: Research Data Netherlands, Open Science Finland
- *Institutional initiatives*: University of Cambridge, University of Edinburgh, Technical University of Delft, Karlsruhe Institute of Technology
- *Scientific or professional associations and communities of practice*: CODATA, Force 11, Open Knowledge Foundation, Research Data Alliance, Software and Data Carpentry

These initiatives vary in scale, but have each diversified their reach and participation beyond initial target groups, through a mix of seed funding and voluntary commitments.

3.3. 3.2.1 Collaborative projects

EDISON

The EU-funded EDISON Project is establishing mechanisms intended to increase the number of competent and qualified Data Science professionals across Europe and beyond. The core output of the project is the EDISON Data Science Framework, a collection of documents comprising a Data Science Competence Framework, Body of Knowledge, Model Curriculum and Professional Profiles.[36] The Data Science Competence Framework is the main source of the EOSCpilot draft competence framework in Section 5.

These documents elaborate skills and competences required of today's data science professionals, and can be used by a range of stakeholders to construct their own structured solutions for educating, training, certifying, recruiting, managing, and otherwise supporting data scientists and other data-dependent professionals.

Foster and FosterPlus

FOSTER was an EU-funded initiative that ran from February 2014 to July 2016. [37] FOSTER supported different stakeholders in the European Research Area (ERA), especially young researchers, to adopt open science practices and comply with the requirements of the Horizon 2020 programme. The training programme included different approaches and delivery options: eLearning, blended learning, self-learning, dissemination of training materials, helpdesk, face-to-face training, especial training-the-trainers, seminars, etc. Through the FOSTER Portal, more than 1800 training items were collected, categorized and made available for reuse, as standalone objects or organized into courses, and 11 different e-learning courses were created and offered, as self-learning or moderated courses, resulting in 25 e-learning initiatives. The FOSTER portal has been sustained to reference outputs from other EU-funded projects (e.g.

[36] Demchenko, Y. Belloum,A. and Witkorski, T, (2016.) EDISON Data Science Competence Framework v.0.7 EDISON Project, http://edison-project.eu/data-science-competence-framework-cf-ds

[37] FOSTER Project https://www.fosteropenscience.eu/

OpenMinTeD[38]), and is now used by the successor FosterPlus project. All the training materials on the FOSTER portal are classified according to the taxonomy shown in Figure 3.1.

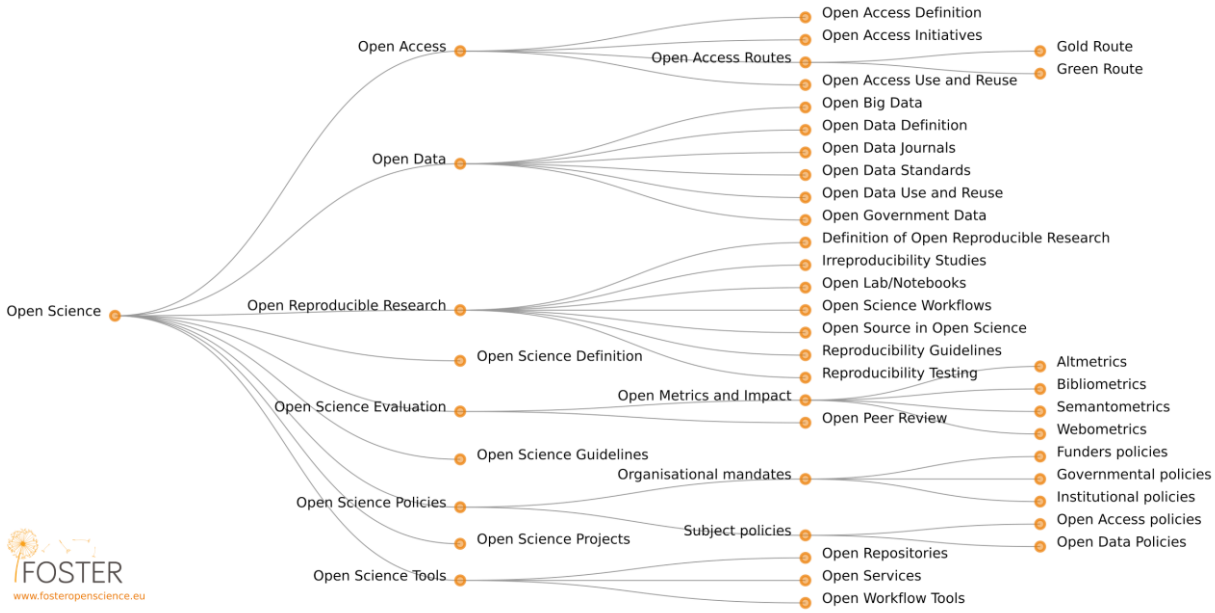

Figure 3.1 FOSTER Open Science Taxonomy

In May 2017 the successor EU-funded project FOSTER Plus started.[37] FOSTER Plus aims to contribute to a lasting shift in European researcher behaviour towards Open Science (OS) as the norm. While general awareness of OS approaches has improved among EU researchers, there is still a lack of practical guidance and training to help researchers learn how to open up their research within a particular domain or research environment. For this reason, FOSTER Plus places specific emphasis on creating discipline-specific guidance and is partnering with expert organisations representing the scientific areas of life science, social science and humanities. The project will enhance existing materials and co-produce new training content, focusing on practical and tangible outcomes directly applicable in researchers' daily practices. The training activities will be addressed to all relevant stakeholders in the European Research Area, with a focus on young scientists, academic staff and policy makers. A train-the-trainer approach and network of open science trainers will act as ambassadors to reach a wide audience.

Data Information Literacy

In 2007 The National Science Foundation (NSF) issued a report on technology-driven research ("e-science") in which it described a need to build public collections of digital data sets. It became apparent however that in most fields the researchers lacked the necessary skills and knowledge of data management and data curation to realize NSF's vision of 21st century scientific practice. As a result, the Purdue University Libraries partnered with the libraries of the University of Minnesota, the University of Oregon and Cornell University to address these issues through developing and implementing data information literacy (DIL)[39] instruction programs for graduate students in STEM disciplines. A set of 12 core competencies have been proposed that would comprise a data information literacy programme. As the authors explain, a key difference between information literacy and data literacy is the recognition of researchers as producers of data, as well as data consumers. Data literacy is a combination of data, statistical, information, and science data literacy, integrating them into a new kind of skill set.

The Twelve DIL Competencies are the basis for a framework reviewed further in Section 5. The Data Information Literacy (DIL) Toolkit (a set of tools, web pages and other resources) is also available to help

[38] OpenMinTeD Project http://openminted.eu/

[39] Data Information Literacy http://blogs.lib.purdue.edu/dil/what-is-data-information-literacy/

those interested in implementing the DIL Competencies at their institutions. The authors conclude that creating the model 'identified a hunger for support in data information literacy arena'.

Open Science Data Cloud

The *Open Science Data Cloud* (OSDC) is described as a "data science ecosystem in which researchers can house and share their own scientific data, access complementary public datasets, build and share customized virtual machines with whatever tools necessary to analyze their data, and perform the analysis to answer their research questions." [40] From 2010-16 it also hosted the NSF-sponsored Partnerships for International Research and Education (PIRE) programme. This used the OSDC "to train the next generation of scientists (graduate students and early career researchers) in data science and to sponsor their travel so that they can collaborate on data science research projects".

3.3.1. National-level programmes

Research Data Netherlands

Research Data Netherlands (RDNL) is a Dutch national coalition of data archives with a mission to promote long-term archiving and reuse of research data.[41] The current partners are 4TU.ResearchData, Data Archiving and Networked Services (DANS) and SURFsara. The RDNL coalition aims to fulfil a back-office function in the process of data curation by ensuring that the research data delivered to them is trustworthily archived and remains accessible. The front offices such as university libraries deal directly with researchers and can advise them in ensuring that research data is accommodated in one of the archives. Since 2011 Research Data Netherlands provides a blended training for front-office staff. This training *Essentials 4 Data Support* is an introductory course for those who provide support to researchers in storing, managing, archiving and sharing their research data: data support staff.[42] It is free and openly available online, but when support staff take the richer face-to-face version, make the weekly assignments, and give feedback on the assignments of their fellow students, they receive a Certificate of Attendance.

Open Science and Research Initiative (Finland)

The Open Science and Research Initiative (ATT) is a cooperation between Finnish ministries, universities, research institutions, and research funders, coordinated by the Ministry of Education and Culture of Finland. ATT aims at ensuring that the possibilities of open sciences will be widely used on a national level, thus making Finland one of the leading countries in openness of science and research. It offers a range of services for the Finnish research community and, through the National Library of Finland, coordinates high-level professional training activities in open science and research. This broadly targeted open science training programme includes an online introductory course on the practices of open science. [43] Face-to-face training sessions are also organised in collaboration with institutions of higher education, research institutes, and open science networks. Also available online are an Open Science and Research Handbook and a Data Management Guide.

3.3.2. Institutional initiatives

Gridka School

The international GridKa School is one of the leading summer schools for advanced computing techniques in Europe. [44] The range of topics includes state-of-the-art data processing, data analysis, cluster orchestration and modern programming. Over a period of five days GridKa School offers participants a

[40] OSDC PIRE Fellowship http://pire.opensciencedatacloud.org/pire-fellowship/

[41] Research Data Netherlands http://www.researchdata.nl/en/

[42] Verbakel, E., & Grootveld, M. J. (2016). 'Essentials 4 Data Support': Five years' experience with data management training. DOI: 10.1177/0340035216674027.

[43] Open Science Training http://openscience.fi/training

[44] GridKa School http://gridka-school.scc.kit.edu

comprehensive overview on cutting edge information technologies relevant to their research community, and practical skills on applications in different research domains. For each track both topical lectures and hands-on courses are provided by international researchers from academia and industry. The school provides a forum for scientists and technology leaders, experts, students and early stage researchers from different fields of science and industry to facilitate knowledge sharing and information exchange. Knowledge and information sharing as well as networking is furthermore incorporated by two social events. The organizers apply several workflows and techniques of modern learning models in a "form of blended learning combined with problem-based learning", [45] to enable a learner-centered education focusing on high interactivity and networking by extensive use of local and remote training platforms and real compute infrastructures.

University of Edinburgh MOOC

The University of Edinburgh has adopted the Massive Open Online Course (MOOC) approach to delivering open science training, offering a *Research Data Management and Sharing* module.[46] The training materials were developed in collaboration with University of North Carolina CRADLE project, and utilizes the Coursera platform. [47] The course provides an introduction into the topic of research data management and sharing, and aims to incorporate knowledge on the diversity of data and their management needs across the research data lifecycle. It builds on earlier material including the University of Edinburgh Data Library's MANTRA online course, which in turn grew from the early stages of development of the University's Research Data Management service for local support to researchers. The MOOC provides a self-paced learning experience by utilizing short, video-based lessons and assessments across a five-week period, supported by 'course mentors' and community discussion forums. After successful completion an electronic Course Certificate is available to learners for a small fee. The Course Certificate does not represent official credit from the partnering institutions that developed the course but can be used by participants for sharing for resume or LinkedIn profiles. If no certification is required, the course is free.

Technical University of Delft - Data Stewards Network

Technical University of Delft has eight faculties. The TU Delft Library Research Data Services (RDS)[48] team is responsible for the acquisition, management, archiving and sharing of scientific research data and manages the 4TU.Centre for Research Data. RDS recently started a Data Stewardship project, which aims to embed good research data management across the university. To this end RDS is in the process of hiring a dedicated data steward for each one of their eight faculties and has just appointed a coordinator. Apart from building a network of data stewards the coordinator will organise a platform for all data stewards to share knowledge on research data needs and requirements.

University of Cambridge - Data champions network

The Data Champions initiative at Cambridge was launched in September 2016 by the University's central Research Data Facility. The University is structured around more than 30 colleges and 100+ academic departments, and therefore challenging for a central data management facility to coordinate researcher engagement and particularly skills development in research data management. As is commonly the case, comprehensive discipline-specific training is beyond the capability of a small central RDM support service. This motivated the service to initiate a Data Champions Network in 2016, targeting individual researchers with specific expertise to encourage them to share it across the university. A Data Champions website offers individuals credit and recognition for the time and expertise that they contribute, as well as opportunities for networking, additional training and developing leadership skills. [49]

[45] M. Ernst et al., "GridKa school - Teaching information technologies since 2003," 2015 IEEE Global Engineering Education Conference (EDUCON), Tallinn, 2015, pp. 395-402. doi: 10.1109/EDUCON.2015.7096003

[46] Research Data Management MOOC https://www.coursera.org/learn/data-management

[47] CRADLE Project http://cradle.web.unc.edu/

[48] Technical University of Delft Research Data Service http://www.open.tudelft.nl/open-research/research-data-management/

[49] University of Cambridge Data Champions: http://www.data.cam.ac.uk/intro-data-champions

3.3.3. Scientific or professional associations and communities of practice

CODATA

The International Council for Science: Committee on Data for Science and Technology (CODATA[50]) aims at building data science capacity through promoting technical training for countries with emerging or developing economies. Through collaboration with the Research Data Alliance (RDA) it organises a Research Data Science 'summer school', and through collaboration with the Chinese Academy of Science an *International Training Workshop on Open Data for Better Science*.[51] The scope includes international development in Open Data and Big Data, Open Science and research management and curation, software development and the use of research applications, data platforms and infrastructures, techniques of large scale analysis, statistics, visualization and modelling techniques and case studies and best practices of research in the Big Data era. The delivery formats include theoretical lectures, advanced seminars, workshop and practical sessions as well as field visits. To foster knowledge sharing, promote interaction and develop opportunities for future exchange and collaboration the organizers involve elite scientists into a number of activities and explicitly promote interaction between experts and participants.

FORCE11

The FORCE11 Scholarly Communications Institute (FSCI[52]) is a week long intensive summer school incorporating intensive coursework, seminar participation, group activities, lectures and hands-on-training. The FSCI aims at providing knowledge on latest trends, technologies and opportunities that are transforming the way science and scholarship is done. The instructors are leading practitioners from research, libraries, publishing and research administration. The courses range from basic orientation through classes in the most advanced topics covering Scholarly Communication from a variety of disciplinary and regional and national perspectives[53]. The schedule includes courses for target audiences including scientists, social scientists and Humanities researcher, researchers in general and others involved in managing, organizing or publishing research.

Open Knowledge Foundation: Open Science Training Initiative

The Open Knowledge Foundation is a global non-profit organisation "…focused on realising open data's value to society by helping civil society groups access and use data to take action on social problems." The organisation has advocated for Open Science by articulating the *Panton Principles* and, in 2013, a programme of Fellowships to demonstrate their implementation.[54] Among its outcomes was the *Open Science Training Initiative (OSTI)*, a programme of micro-lectures and exercises on Open Science and research reproducibility, aimed at graduate students.[55] The pilot scheme for the initiative took place at the University of Oxford's Doctoral Training Centres for Systems Biology, Life Sciences and the Industrial Doctorate in January 2013. The OSTi programme is intended to fit around an existing subject-specific training course in an academic environment. All OSTI materials are released under a Creative Commons Attribution Licence, allowing to use, reuse, and/or remix the materials., including content, slides and advice sheets for the lectures and other training materials.

Research Data Alliance

The RDA is primarily a voluntary community-driven association launched in 2013 with the backing of by the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. In pursuing its goal of "building

[50] CODATA http://www.codata.org/about-codata

[51] CODATA International Workshop http://www.codata.org/2017-international-training-workshop-for-developing-countries-on-big-data-for-science

[52] FORCE11 https://www.force11.org/fsci

[53] FORCE11 Training https://www.force11.org/fsci/course-list

[54] Panton Fellowships. https://pantonprinciples.org/panton-fellowships/

[55] Open Science Training Initiative http://www.opensciencetraining.com/

the social and technical infrastructure to enable open sharing of data",[56] it has also become a provider of training. In late 2015 the RDA began a series of webinars, initially to explain and promote its first outputs and sets of recommendations. Currently the organisation also offers face-to-face workshops, hackathons/datathons partly organized as "summer schools" and special meetings on request. The RDA co-organises with CODATA a School of Research Data Science (see CODATA above). The RDA also hosts a number of Interest Groups and Working Groups with objectives directly related to those of EOSCpilot WP7, including the IG on Education and Training in Data Handling, and WG/IG on Data Science and data-related education and training certification and accreditation schemes.

Software and Data Carpentry

Software Carpentry is a volunteer non-profit organization dedicated to teaching basic computing skills to researchers in science, engineering, medicine and related disciplines[57]. The training is organized in the form of hands-on workshops teaching about program design, task automation, version control. The workshops aim to instill 'computational competence'. Independent assessments indicate that this training "increases participants' computational understanding [...], enhances their habits and routines, and leads them to adopt tools and techniques that are considered standard practice in software industry".[58] A typical workshop has up to 40 learners, two instructors, and two to three helpers to answer questions and provide guidance during practical sessions. Short tutorials alternate with hands-on practical exercises, and participants are encouraged both to help one another, and to apply what they are learning to their own research problem during, between, and after sessions.

Data Carpentry is a sibling organization to Software Carpentry, sharing its instructor training program and with overlapping steering committee membership, but focusing on data analysis skills rather than programming skills. Both organisations foster an active, inclusive instructor community that promotes and models reproducible research as a community norm.[59] Whereas Software Carpentry aims to help scientists learn to program better, Data Carpentry aims to teach people who are wrangling data manually how to automate their work and make it more reproducible. This is described as 'data competence' in core data skills for efficient, shareable, and reproducible research practices. Data Carpentry workshops are primarily designed for people with little computational experience and are domain-specific. Researchers work with data most relevant to their own work. They follow a narrative structure, working with one dataset through the whole data lifecycle, from data and project organization to data analysis and visualization. All provided lessons from Software Carpentry as well as Data Carpentry are freely reusable.

3.4. Embedded or immersive approaches

The training provision covered so far is, regardless of provider, mostly structured around conventional professional development courses typically lasting a number of days. There are exceptions to this however; programmes offering a more 'immersive' or embedded approach, i.e. whereby participants are offered structured experience of a role they do not ordinarily occupy. This includes skills exchanges, fellowship and internship programmes that we highlight here partly because they illustrate that some organisation favour these approaches, and partly for their potential suitability to address the need identified in chapter 2 to nurture 'embedded' data expertise in research teams.

3.4.1. Skills exchange e.g. RItrain

The Research Infrastructure Training Programme (RItrain) aims at improving and professionalizing the training of managerial and leadership staff in research infrastructures.[60] RItrain therefore defines required competencies in research infrastructures from preparatory to operational phases in the framework of the

[56] Research Data Aliiance https://www.rd-alliance.org/about-rda/who-rda.html

[57] Software Carpentry https://software-carpentry.org/about/

[58] Wilson G. Software Carpentry: lessons learned [version 2; referees: 3 approved] F1000Research 2016, 3:62 (doi: 10.12688/f1000research.3-62.v2)

[59] Data Carpentry http://www.datacarpentry.org/mission/

[60] http://ritrain.eu/home

RItrain organisational competency profile[61]. The core competencies are grouped into three broad areas including leading the organization, engagement within and beyond the organization and professional conduct. Building on this, training comprises an Executive Masters in Management of Research Infrastructures (add url), a series of webinars with experienced leaders in research infrastructures to gain insight into specific competencies, and staff exchanges to access managerial and leadership expertise directly from leading research infrastructures and foster cooperation. The staff exchange programme is an integral part of the Executive Masters in Management of Research Infrastructures, but can also be undertaken independently, helps in developing specific expertise through short knowledge-exchange visits to research infrastructures that are noted for their excellence in the required area. [62]

3.4.2. Internships e.g. CERN

CERN offers different programs for undergraduate students, graduates and professionals in the field of physics, computing and engineering to join day-to-day work of research teams participating in experiments at CERN. [63] Beyond the scientific value of internships, working in a multidisciplinary and multicultural environment provides an opportunity to discover science, technology, engineering and mathematics in the CERN context to strengthen the understanding of science and to develop skills in a data-intensive research environment. Internships for students offer participants a deep insight into particle physics by working on their own projects, through a variety of visits and through hands-on workshops at CERN's S'Cool Lab. Internships for doctoral students provide a unique experience to extend knowledge by getting involved in experiments of unprecedented scale and scope. [64]

3.4.3. Fellowships

Fellowship schemes are used by some organisations to facilitate knowledge exchange in research data and software management. Typically, fellowships offer financial support to augment early career researchers' opportunities to gain hands-on experience, engage in workshops or conferences, and contribute to advancement of best practices in their domain.

Several of the providers listed earlier in the section have offered fellowship schemes. CERN, for example, provides a Fellowship Programme offering scientists and engineers "challenging opportunities in particle physics research, and in related fields of physics and technology".[65] The *Open Science Data Cloud* (OSDC) PIRE project, a US National Science Foundation funded initiative, from 2010-16 offered a fellowship program providing international research and education experiences through training and study at universities and research institutes around the world. The program targeted graduate students, senior undergraduate, postdoc and early career faculty actively conducting research in Computer Science, Engineering or relevant data intensive science.[66]

The Software Sustainability Institute's Fellowship program provides a further example, intended in this case to support the development of careers in research software engineering by "gathering intelligence about research and software from all disciplines, encouraging Fellows to develop their interests in the area of software sustainability and aiding the Fellows as ambassadors of good software practice in their domains".[67] The program is for UK-based researchers from a wide range of areas, experience and career stages using software, writing software or sharing best practices in software.

[61] RItrain Competency Framework.
http://ritrain.eu/documents/210628/807721/Organisational+competency+framework+v30+April+2016.docx

[62] RItrain Staff Exchanges http://ritrain.eu/staff-exchanges

[63] CERN Internships https://cds.cern.ch/record/2153862/files/TakePart_Decisiontree.pdf

[64] CERN Doctoral Student Programme https://jobs.web.cern.ch/join-us/doctoral-student-programme

[65] CERN Fellowships https://jobs.web.cern.ch/join-us/fellowship-programme

[66] OSCD PIRE Fellowships http://pire.opensciencedatacloud.org/pire-fellowship/

[67] Software Sustainability Institute Fellowships https://www.software.ac.uk/fellowship-programme

3.5. Summary of the skills landscape

Research infrastructures (AQUAEXCEL2020, CERN, ELIXIR, and SoBigData) provide a mixture of face-to-face training, distance learning and amplification materials aimed at a variety of audiences including researchers, data scientists, and others (postgraduate, graduate, and high school students and teachers).

Analysis of four cluster projects (PARTHENOS, SERISS, CORBEL and ENVRIplus) demonstrates variation in target audiences for training, and in the focus of training topics. There is varying focus on using the RIs and e-infrastructure, as opposed to disciplinary research methods.

E-infrastructures (EGI, GEANT, PRACE) have a commitment to providing up-to-date training enabling users to exploit the full capabilities of the service. These are in some cases supplemented with more broadly focused courses aimed at raising skill levels in areas with general relevance across infrastructures (e.g. FitSM IT service management course)

Broader open science skills development is provided through related initiatives including the FOSTER and FOSTER plus programmes which offer a wide range of courses and materials, including discipline-specific materials. The 'train-the-trainer' approach is widely used to amplify the materials and develop a network of trainers.

Intensive workshop-based short course formats are widely promoted and popular. These include the Software Carpentry and Data Carpentry formats, and a number of summer schools. These commonly feature intensive hands-on training in the technical aspects of data science, stewardship and advanced computing. These courses tend to focus on problem-based learning, and on employing leading practitioners as instructors.

MOOCs, such as the Research Data Management MOOC produced by University of Edinburgh and University of North Carolina, offer an accessible and scalable method for learners to interact with leading practitioners.

Skills are also developed by embedding staff in environments that enable them to develop relevant skills. Embedded or 'on the job' approaches, coupled with self-paced learning materials, are prominent in some of the examples reviewed. They include placement, internship and fellowship programmes from the RItrain project, CERN, OSDC-PIRE, Open Knowledge Foundation and the Software Sustainability Institute.

4. IDENTIFYING EOSCPILOT SKILLS REQUIREMENTS

4.1. Skills implications of EOSC policy and governance frameworks

4.1.1. Policy frameworks

EOSCpilot Wp3 will develop a policy framework to address barriers to open sharing of research artefacts. It will set requirements for the policy framework, and review existing policies. Liaison with funders and other policy bodies will ensure the requirements also meet planned policies. Among the tasks that Wp3 is undertaking are activities to identify ethical and open science drivers and constraints. Wp7 liaises with Wp3 to monitor how the results are influenced by researchers' data knowledge and skills, or the lack thereof.

Furthermore, in May 2017 the Open Science Policy Platform adopted a "Report on the governance and financial schemes for the European Open Science Cloud".[68] This Report has eight recommendations, including that EOSC a) raise awareness of the EOSC benefits, b) develop Open Science and data skills, and c) align and develop ethical rules in data management, storage and analytics. The Report states: "It is recognised that the move to digital science, growing data volumes and growing influence of scientific data on policy have created skill gaps among researchers, emerging professions such as data science, and also among policy makers and advisers. To reap the benefits of the EOSC in support of open science, skill development in the area of Information Technology (IT) and data literacy should be supported at all levels, from the primary school up to policy makers." (ibid., p. 5) The Report recommends that advanced IT and data skills should be facilitated for researchers, librarians, IT-staff and citizen scientists, e.g. by organising appropriate advanced trainings and embedding expertise into their work and projects. This is in line with what various research infrastructures, cluster projects and e-infrastructures do (see Chapter 3 and in particular Section 3.4 on the embedded approach).

In a collaborative environment such as EOSC the ethical and legal rules are very important. The Report on governance and financial schemes mentions the processing and secure storage of personal data as well as research integrity (ibid., p.6). Researchers and data supporters should be (made) aware of rules and good practices and enabled to comply with them. This could well be part of the training topics for EOSCpilot.

4.1.2. Governance framework

In EOSCpilot WP2 a governance framework for the sustainability of the European Open Science Cloud will be developed from existing governance structures. The aim is to enable and encourage engagement from the key stakeholder communities such as European e-Infrastructures, Data/Research Initiatives, Cloud providers, Research funders, Cloud community, Research Communities and Institutions, Research Infrastructures, Policy makers. The governance framework must ensure the stability and sustainability of the EOSC. Sustainability, however, is a maturing process that has to be actively pursued. Therefore, the EOSCpilot governance framework will be based on a business model.[69]

Qualified and competent human resources are key to this business model, and Wp7 continues to work with Wp2 to deepen the project's understanding of how stakeholders' human resource processes can inform the EOSC business model. Workshops of the EOSC pilot Government Development Forum are a key input. For example the event in Helsinki, 9 May 2017 focused on the research infrastructure's need and expectations on EOSC. Regarding skills and training the participants suggested that EOSC could assist in providing training on how researchers can make the shift to cloud storage, e.g. by training on curation of data. Besides training also, user-friendly tutorials and help-desk support (outside the scope of WP7, however) for the various services were seen as necessary.

4.2. Requirements arising from Science Demonstrators and Services

EOSCpilot develops a number of demonstrators functioning as high-profile pilots that integrate services and

[68] https://drive.google.com/drive/folders/0B6gpAGKtp_HHM0otbXhvWmtNUkU

[69] EOSCpilot project proposal, p. 30.

infrastructures to show interoperability and its benefits in a number of scientific domains. Each of these science demonstrators provides scientific challenges in the context of Cloud infrastructures. The demonstrators aim to show the scientific excellence, relevance and usefulness of provided services and their enabling of data reuse to drive the further development of EOSC. To fully develop capabilities and competencies that the EOSC will rely on, skills requirements arising from science demonstrators were identified. EOSCpilot also includes service pilots, to test the integration of underlying services, however none had reached a stage in planning that provided opportunities to assess any additional training needs.

4.2.1. Liaison with science demonstrators on required skills

Each science demonstrator serves as an exemplary use case for EOSCpilot, including infrastructure integration and service management as well as usability for researchers and scientists. Demonstrators are selected to cover a broad range of distinct scientific areas, maximizing the coverage of scientific requirements. Each demonstrator exemplifies fundamental requirements that promote a broader impact beyond the demonstrator's native field of research. The science demonstrators selected in the first phase of EOSCpilot are each described on the project website and include ENVRI, Digital Preservation in High Energy Physics, Pan-Cancer Analyses, Photon-Neutron, and TEXTCROWD.[70] Of these, the first four were active during the first reporting period for the demonstrators (coordinated through Wp 4).

The approach to identifying skill gaps included a questionnaire, desk research on each of the demonstrator sub-project's planning documents, and (where possible) discussions with contacts in the demonstrator sub-projects. A questionnaire (Annex B) was used to elicit the demonstrators' self-assessment of the main skills required and any gaps in their availability to them and their user community. Three main categories were used: -

- Integration: skills needed by the sub-project to integrate the demonstrator application into the EOSC ecosystem,
- Service management: skills needed to turn this application into a service and manage that service
- Enabling effective use: skills needed by the demonstrator-service user community, to benefit from using the service and fulfil its data science use case(s)

Spreadsheets were used to collate responses (Annex C) and results of the needs analysis, an example of which is included in Annex D. The results are collated and condensed below for each of this first phase of four demonstrator sub-projects.

Data Preservation in High Energy Physics

This demonstrator aims for preservation of Open Data at a scale of 100+TB, including associated software for re-use, sharing and verification of results. For successful deployment in EOSC competences required include: -

- *evaluate, select, certify* and *integrate* reliable bit storage for archiving in trusted (certified) repository, preferably with integrated DOI generation
- *evaluate, select* and *integrate* storage solution for documentation with DOI with up to 64MB storage per element
- *set-up* and *provide* Cern virtual machines for storing software and associated configurations;
- *deploy* on on-demand compute resources by cloud providers
- *integrate* the different services
- *evaluate, select* and *integrate* authentication solution for uploading

Pan-Cancer Analyses

This aims to standardise genomic data processing and data redistribution to make the provided solutions interoperable and reusable for the EOSC. The project requires specific competences to: -

- *engineer* collaborative data sharing and analysis across countries and cancer types to maximize the statistical power for health-related discoveries

[70] https://eoscpilot.eu/science-demos/textcrowd

- *set-up* and *provide* interoperable IT frameworks to enable standardized sharing and large-scale processing of cancer genomes
- *evaluate, select* and *integrate* native Docker[71] support
- *improve* deployment experience towards a shingle shot deployment
- *evaluate* and *integrate* data staging mechanism
- *apply* operational monitoring and aggregation in federated clouds

Photon-Neutron

This will leverage on the photon-neutron community to improve computing facilities by creating a virtual cloud-based platform for all users by enabling sustainable, transparent access to distributed data. For deployment in EOSC the demonstrator requires capabilities to: -

- *deploy* ICAT data catalogue[72] as a data management solution;
- *evaluate, select* and *integrate* interoperability service supporting data type registry with machine digestible schemata
- *evaluate, select* and *integrate* AAI solutions to improve data access
- *evaluate, select* and *integrate* infrastructure interoperability services to support datasets ranging from few MB to hundreds of TB as well as varying I/O and compute requirements.

TEXTCROWD

This aims at structuring and integrating Humanities initiatives CLARIN, DARIAH and E-RIHS ERICs as well as Digital Humanities Organizations to offer advanced text-based services addressing common research needs in the Social Sciences and Humanities. These aims require capabilities for deployment to

- *set-up* Linux distribution and package dependencies required for the TEXTCROWD use case
- *evaluate, select* and *integrate* digital data repositories for publishing and storing of annotated data including proper authentication solution
- *set-up* and *integrate* text processing toolchain as well as natural language processing and machine learning tools in Docker container for cloud deployment;
- *engineer* and *deploy* REST-style web services for intercommunication between cloud-based storage and compute and
- *define* and *apply* terms of use for data providers, data consumers and data hosts

3.5.1. Skills gaps

Pan-Cancer and Photon-Neutron foresee general skills requirements regarding operation and utilization of large-scale cloud environments for data analysis, as common workflows need to be adapted with regard to cloud requirements. On the one hand, the science demonstrators for DPHEP as well as TEXTCROWD do not foresee any skills gaps and therefore do not foresee any specific training requirements at the present time. In the case of TEXTCROWD the experts predict instructions on how to use the provided tools to be sufficient. Experts of the DPHEP use case note the availability of documentation and resources in the scope of the science demonstrator itself.

Overall, the information provided in the questionnaires suggest that not all skill requirements for enabling effective use, including long-term data stewardship, can be identified yet. However, this is to be expected given the deliberately broad scope and impact of the selected science demonstrators. Given that the demonstrators were selected for their general applicability, each of the demonstrators also will help to identify generic skill requirements by relating each of the scientific use cases to application of the FAIR principles. It is that aim, rather than domain-specific needs, that underpins the selection of skills below.

Skills for integration

To integrate the different use cases of the four science demonstrators into the infrastructure of EOSC basic

[71] https://www.docker.com
[72] https://icatproject.org

needs related to system management and administration but also software development to adapt existing workflows and knowledge on specific tools are required:

- system management and administration including container technologies
- (long-term) data management, curation, preservation and provenance
- federated authorization and authentication including data security in the cloud
- domain knowledge and software development to adapt workflows to cloud workflows.

Skills for service management

To operate large-scale cloud-based analysis services a broad set of technical skills is required:

- system management and administration of virtual infrastructure
- federated authorization and authentication including data security in the cloud
- monitoring of federated data and compute services

Skills for effective use

Utilizing large-scale cloud-based infrastructures for domain-specific analyses requires a broad set of technical skills that are not commonly possessed by end-users:

- decision-making for when using the cloud is appropriate
- data security in the cloud
- data management, curation, including provenance to enable reusabilty and long-term preservation
- domain knowledge and domain-specific scientific methods

4.3. Conclusions on skills requirements for demonstrators and services

The needs reported so far are broader than just data management aspects of stewardship, also reaching into data processing, analysis and related topics. Many of the current science demonstrator requirements are focused on engineering and deploying services in the cloud. Given the recent start of most demonstrators, integration of workflows with the EOSC environment is a primary need shared across demonstrators and scientific fields. Similarly, skills requirements for end-users pivot on the awareness for implications of working in a multi-national, distributed infrastructure, though at a higher level of abstraction compared to service operators.

Many of the fundamental skills requirements are covered by existing resources which are publicly available, or at the least feasibly accessible. This is mainly guaranteed by a broad range of webinars, online tutorials and courses, including MOOCs as well as workshops and other training opportunities such as staff exchange opportunities or internships. Still, the scope of such resources is usually geared towards basic introductions, as opposed to the fundamental, conceptual and technical expertise required for scientific endeavours at the scope of EOSC. As such, contextualization of available resources as well as an expansion with expert training opportunities to cover the needs of EOSCpilot demonstrators and their communities may be useful. Additionally, the sustainability of training must transcend that of science demonstrators; the challenge is to ensure availability, applicability and up-to-dateness of resources well beyond the lifetime and scope of individual science demonstrators.

5. COMPETENCE-BASED SKILLS INITIATIVES

5.1. Introduction

This section informs the draft competence framework in section 6, by first reviewing relevant definitions and examples. The definition of competence is based on the European e-Competence Framework, as follows: "*demonstrated ability to apply knowledge, skills and attitudes for achieving observable results*".[73]

The draft competence framework represents the first step towards a broader Wp7 skills framework. In the current D7.1 report we identify the range of competences that need to be developed through professional development in the area of data stewardship. We focus on professional development in this area for several reasons. Data stewardship was highlighted by the High Level Expert Group as a key requirement for the establishment of EOSC. And while EDISON and other initiatives offer a concerted approach to development of university curricula for data science, similar coordination of short term professional development for researchers and support professionals in this area will be needed.

Fostering data stewardship means helping a range of people acquire new skills; researchers, including those in domains that are less data-intensive as well as data scientists themselves, engineers who build the e-infrastructures and services, data managers and other research support professionals who may occupy central advisory roles in institutions, or be embedded in research teams.

Competence *frameworks* are used to standardize domain knowledge and proficiency in a number of domains relevant to EOSC, including IT, service management, data management, and (through the EDISON project) data science. We offer an overview in this section, focusing on those that have been drawn on for the draft framework proposed in section 7.

The various frameworks are commonly hierarchical in nature, and described in one or more table. Frameworks relevant to the draft EOSCpilot competence framework are characterized below, using the following terms:

- Skills areas: the broad subject areas the competences relate to
- Competence groups: activities or practices the competences are typically used in
- Competences; topics indicating the scope of the knowledge to be applied, and/or statements defining progressive levels of application

5.2. Defining data stewardship

The term 'data stewardship' is associated with the care and handling of data. Commonly referenced definitions include the following: -

"*The formalization of accountability for the management of data resources."* Robert Seiner, TDAN.com http://tdan.com/the-data-stewardship-approach-to-data-governance-chapter-1/5037

"*The management and oversight of an organization's data assets to help provide business users with high-quality data that is easily accessible in a consistent manner."* Research Data Alliance IG Data Foundation and Terminology (citing publisher TechTarget) http://smw-rda.esc.rzg.mpg.de/index.php/Data_Stewardship

"…professional and careful treatment of data throughout all stages of your research project (i.e., the design, collection, processing, analysis, long-term preservation, and sharing of your research data)" Data4LifeSciences http://data4lifesciences.nl/hands/handbook-for-adequate-natural-data-stewardship/what-is-data-stewardship/

[73] European e-Competence Framework http://www.ecompetences.eu/methodology/

5.3. Characterising relevant frameworks

Frameworks reviewed in this section show a diversity of approach to defining the competences involved in data stewardship. Some identify it with a broad set of practices or activities that may be spread across a number of roles in the organisation. Other approaches associate stewardship with a specific role or function that can be distinguished from other roles, such as 'data manager', or 'data curator'.

The EOSCpilot draft competence framework represents a synthesis of elements from the following five frameworks in particular:

- Demchenko, Y. Belloum,A. and Witkorski, T, (2016.) EDISON Data Science Competence Framework v.0.7
- Hodson, S. et al (2017) Objectives, Scope and Activities of a Possible GO-TRAIN Implementation Network, Unpublished discussion document
- Lyon, L. & Brenner, A. (2015): Bridging the Data Talent Gap: Positioning the iSchool as an Agent For Change ; International Journal of Digital Curation 10(1) http://dx.doi.org/10.2218/ijdc.v10i1.349
- Molloy, L. Demchenko, Y, Jung, C. et al. (2016) Research Data Alliance Interest Group on Education & Training in Data Handling (ETDH-IG) Task Force on Defining data handling related competences and skills for different groups of professions
- Purdue University, & Sapp Nelson, M. (2017). A Pilot Competency Matrix for Data Management Skills: A Step toward the Development of Systematic Data Information Literacy Programs.

5.3.1. EDISON Data Science Competence Framework

EDISON, as outlined in section 3, has developed a competence framework as the basis for a 'Data Science Framework', also comprising a Body of Knowledge, Model Curriculum, Professional Profiles, and Data Science taxonomy and vocabulary. The competence framework references a variety of sources including the NIST Big Data Working Group, European e-Competence Framework, European ICT Professional Profiles, ACM Computer Science Classification, and ACM IT Competences Model. [74] The EDISON 'skills area' scope frames data science as a demand-led ICT professional development need, drawing on O'Reilly Strata Survey (2013), the NIST NBDIF Data Science and Data Scientist definitions, and a related literature review of ICT competence frameworks. 'Data Handling' is seen as a related competence area, but is not grounded in a review of research competencies or library/ information science frameworks.

Skills areas: *'Data steward'* is one of 4 role profiles associated with '*Data Management'*, along with "*Digital data curator*", *'Digital librarian' 'Data archivist";*

'*Data Management'* is in turn one of 6 'skills areas', others are: *Data Analytics, Data Science Engineering, Business Process Management, Scientific Methods, Domain Knowledge*

Competence groups: Competence groups are linked to a research lifecycle illustrated in Figure 6.1 below. The framework also maps these to Business process stages (Plan, Build, Run, Enable, Manage) in the European e-Competence Framework (e-CF)

Competences: approx. 6 per skills area, i.e. 24 total

[74] Demchenko, Y. Belloum,A. and Witkorski, T, (2016.) EDISON Data Science Competence Framework v.0.7 EDISON Project, http://edison-project.eu/data-science-competence-framework-cf-ds

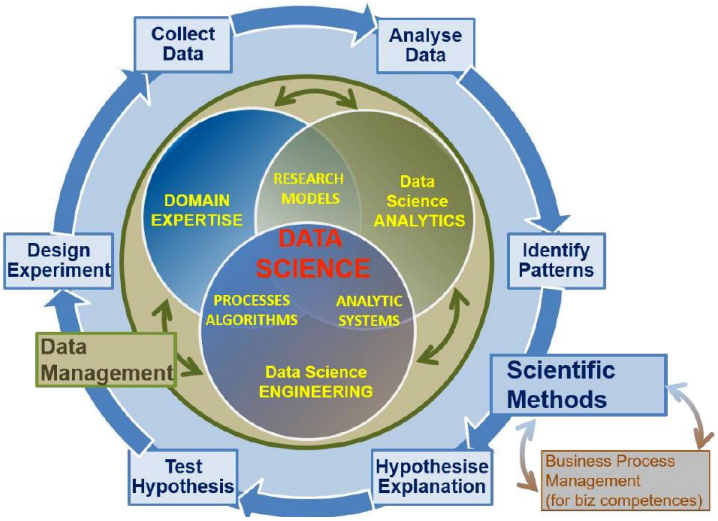

Figure 5.1 Data Science competence groups (source: Demchenko et al 2016)

5.3.2. Potential Go-TRAIN Implementation Network (discussion document)

Relevant skills areas were proposed in a discussion document resulting from a Feb 2017 workshop hosted by CODATA, which aimed to inform the wider GO-FAIR initiative. GO-FAIR is an early-mover-driven 'bottom up' initiative aiming to complement development of the EOSC and other global initiatives towards "an internet of FAIR data and services" (see https://www.dtls.nl/go-fair/).

The discussion document identifies skills areas that address current gaps in skills, and contrasts these with the question of skills required of data stewards as a distinct professional group or role.

Skills areas: *embedded research data specialist; institutional research data advisor and contact point; long term stewardship function; all researchers.*

Competence groups as above

Elements: a preliminary list of topics is identified with bullet points

5.3.3. Lyon & Brenner (University of Pittsburgh)

The paper identifies 'data steward/curator' alongside a number of other data science skills groups, within the context of an effort to examine "the role, functions and value of the 'iSchool' as an influential and effective agent of change in the data informatics and data curation arena." The six skills groups identified are described as an 'interpretation' to aid curriculum development, based on prior work with Microsoft on a capability model to inform requirements for academic libraries.[75]

Skills areas: *data analyst, data archivist, data engineer, data journalist, data librarian, data steward/ curator*

Competences: a brief list indicates 3 or 4 competences for each skills area and industry sector that equivalent roles are typically found in, as shown in Table 5.1

[75] Lyon, L. & Brenner, A. (2015): Bridging the Data Talent Gap: Positioning the iSchool as an Agent For Change ; International Journal of Digital Curation 10(1) http://dx.doi.org/10.2218/ijdc.v10i1.349

Table 5.1 'Family of data scientist roles' source: Lyon & Brenner, 2015

Role	*Focus*	*Typical Location*
Data analyst	Business/scientific analytics, mathematics, statistics, modelling	Corporate Sector
Data archivist	Long term preservation, repository management	National Archive
Data engineer	Software development, coding, programming, tools	IT Company
Data journalist	Telling stories and providing news using visualisations	Newspaper Publisher
Data librarian	Advocacy, research data management, training	University or Research Institute
Data steward /curator	Curation, cleansing, annotation, selection and appraisal	Data Center

5.3.4. Research Data Alliance IG Education & Training in Data Handling Wiki

The wiki includes listings of competencies produced through the RDA Interest Group. The role of this group is defined in its Case Statement (2014), which includes clarifying the career structure of relevant support professions; "data scientists, data librarians, data managers, data analysts, research administrators, infrastructure providers and developers, etc.". The statement also declares the intention to "make the case for creating taxonomies of the skills required by different group of data management specialists/professionals". [76]

Skills areas: *research librarians; research administrators; research infrastructure managers / operators, researchers*

Competence groups:

Research librarian competences are ungrouped, but shown alongside groups used in a US project on Data Information Literacy (see under Purdue University below)

RI manager competences are grouped according to the European e-Competence Framework, as used in EDISON, i.e. Plan/Build, Run, Enable and Manage

Research competences are ungrouped

Competences: Research Librarians x 45; RI managers x 35; researchers x 18

[76] Molloy, L. Demchenko, Y, Jung, C. et al. (2016) Research Data Alliance Interest Group on Education & Training in Data Handling (ETDH-IG) Task Force on Defining data handling related competences and skills - Working area. https://www.rd-alliance.org/group/education-and-training-handling-research-data-ig/wiki/task-force-defining-data-handling

5.3.5. Purdue U. Libraries; Pilot Competency Framework for Data Mgmt Skills

The approach is aimed at design of undergraduate curricula for data librarianship, and distinguishes between three organisational levels of competence; personal, team and research enterprise. The personal level is associated with basic "data information literacy', while data stewardship is associated with a mastery level, a set of skills that leaners encounter "… as they try to organize the data management of many individuals working on a common research endeavor (p. 3).

According to the author stewardship is defined by three goals: Can the learner make decisions about data knowledgably and efficiently? Can they teach members of the team a rationale and the routines resulting from those decisions? At this point, the learner has mastered most, if not all, data management skills to the point of being able to direct others to resources and find assistance with key tasks." (Purdue University, & Sapp Nelson, M. (2017). A Pilot Competency Matrix for Data Management Skills: A Step toward the Development of Systematic Data Information Literacy Programs. Journal of EScience Librarianship, e1096. https://doi.org/10.7191/jeslib.2017.1096)

Skills areas: *Data Management; data librarianship; data stewardship*

Competence groups 12 groups are used to identify competence statements for Personal Domain; Team Domain; Enterprise Domain. The groups are: Databases and Data Formats, Discovery and Acquisition of Data, Data Management and Organization, Data Conversion and Interoperability, Quality Assurance, Metadata, Data Curation and Re-use, Cultures of Practice, Data Preservation, Data Analysis, Data Visualization, Ethics, including citation of data

Competences: Personal x 36 Team x 36 Enterprise x 36; each applying three domains from Bloom's taxonomy: Knowledge (Cognitive) Skills (Psychomotor) Attitudes (Affective)

5.3.6. SFIA: Skills Framework for the Information Age

The SFIA is promoted as "...the globally accepted common language for the skills and competencies required in the digital world" and primarily focuses on skills required by IT professionals. Its development is overseen by the SFIA Foundation, a UK-based not-for-profit organisation, a collaboration between the Institution for Engineering & Technology, British Computer Society, IT Service Management Forum, the Tech Partnership, and the Institute for the Management of Information Systems. The development approach is described as "open consultation and input from people with real practical experience of skills management in corporate and educational environments." [77]

Skills areas: *(IT) Strategy and Architecture, Change and Transformation, Development and Implementation, Delivery and Operation Skills, and Quality Relationships and Engagement*

Competence groups: 7 levels of competence are defined, for four main aspects of responsibility (autonomy, influence, complexity, business skills).

Competences: Each of the above areas of responsibility is associated with each of the following seven competence levels: follow, assist, apply, enable, ensure, advise, initiate, influence, set strategy, inspire, and mobilise. This provides a 4 x 7 matrix of 28 competence statements.

5.3.7. Mozilla Web Literacy

Mozilla Learning, an initiative of the Mozilla not-for-profit internet development community, provides a web literacy model described as "a framework for entry-level web literacy & 21st Century skills". The scope of these is more generic and foundational than is needed for EOSCpilot, but the framework provides an exemplar of how to present a number of complex dimensions in an easily navigable format.

[77] SFIA Foundation: https://www.sfia-online.org/en/reference-guide

Skills areas: *read, write, and participate on the open web*

Competence groups: Four "21st century skills" groups are defined (problem-solving, communication, creativity, collaboration)).

Competences: The skills areas provide a top-level for describing a further 14 competences, each tagged with the relevant "21st Century" skill(s), and further described by 5 statements, providing 70 statements .

5.4. Linking competence assessment to certification

5.4.1. EDISON Approach to Certification

Nowadays, for almost every profession, and especially newly forming ones, it is important that individuals are able to demonstrate their ability through an internationally recognised certificate so that employers are confident to judge the level of skills and competences held by potential employees. At present, there are some certifications on the market for the Data Science profession that confirm the completion of classes, trainings and programs; however, they do not cover all knowledge areas identified by EDISON. (e.g. EMC2, aCAP – both data analytics focused). Guided by the EDSF document EDISON is providing a basis for the formal recognition of the Data Scientist as a new profession, including supporting the future establishment of formal certification for practicing "self-made" Data Scientists and related professionals.

EDISON analysed two types of certification schemes based on popular and recognised certifications (APM Group, PMI) and compared them with personal experiences or work in other European projects as well as types of certification structures (level-based; self-contained modules). This analysis was focused on the structure of the scheme and not the content per se.

Due to the fact that the Data Science profession is evolving over time and is based on a wide range of skills from different domains, it was decided to opt for self-contained certification as practiced in the field of project management, which has similar requirements. Thanks to that decision, all target groups (e.g. students, professionals and self-taught Data Scientists) can benefit and will be able to find an appropriate certification product best suited for them. It will also be easy to adjust the certification syllabus to the changing and not yet well-defined Data Science profession.

The EDISON project plans to define three certifications for: 1.) Learners wishing to demonstrate an understanding of the fundamental knowledge, terminology and activities of Data Science e.g. Data Scientist Associate, 2.) Experienced Data Scientists who would like to prove/improve their proficiency e.g. Data Science Specialist, and 3.) Experienced Data Scientists who would like to prove expertise in a given Data Science Profession, grouped as managers, professionals, technicians and associate professionals, and Clerical support workers. These schemes will be further elaborated by the end of the EDISON project, in September 2017.

5.4.2. Open Badges

An alternative, or complementary path to certification is to couple skills with digital badges. This can suit busy professionals that acquire knowledge in smaller chunks, and less formal settings. Schools, Universities, Employers and informal learning providers globally are using open badges to capture life-long learning which is currently unrecognised. Digital badges are validated indicators of skill, competences or accomplishments that can be earned in digital learning environments, and subsequently shared and displayed on the web. Many institutions that issue badges use open standards, such as the Open Badge Infrastructure. Open digital badging[78] makes it easy for anyone to issue, earn, and display badges across the web—through an infrastructure that uses shared and open technical standards.

A badge links to the underlying proof that the student has acquired the knowledge or skills or obtained the certificate in question. The badge also contains information about the issuer and possibly an expiry date.

[78] https://openbadges.org/

Employers and educational institutions can therefore check online who issued the badge and what a student had to do to obtain it. This increases the trustworthiness of a badge.

A prominent user, hence proponent of open badges is IBM, who is using digital badges as an acknowledgment of skills learned through their IBM Skills gateway[79]. As a pertinent example to our use case, they are offering badges for Big data foundations (skills: hadoop, big data). They classify badges as: Knowledge, Skill, Professional Certification, Proficiency and other.

Various academic institutions around the World already use or experiment with Open Badges. SURFnet recently published a white paper[80] to present the possible opportunities offered by open badges in Dutch education, e.g. as an aid to micro-credentialing.

5.4.3. Certification of Service Management: the FitSM approach

FitSM is a free and lightweight standards family aimed at facilitating service management in IT service provision, including federated scenarios. To support its work in building ITSM capacities in federates infrastructures, FitSM offers training and certification in lightweight service management, This training aims at providing those involved in operating federated infrastructures with the professional skills they need in order to effectively manage their services.

The training scheme defines a series of training courses, along with corresponding exams and achievable certifications for persons. Each course and certification is aimed at achieve the competencies required to fulfill specific roles in the context of an IT service management system. The three training levels and corresponding qualifications are: Foundation, Advanced and Expert. The design of this scheme is based on experience from training and certification systems for well-known ITSM standards like ISO/IEC 20000 and frameworks such as ITIL. The scheme is maintained and updated by ITEMO.

ITEMO (IT Education Management Organisation) is a non-profit organisation registered in Germany (a 'Verein'***[1]***) set up to take custody of a number of training schemes around IT management and development standards and approaches. It was established by a group of IT management professionals, including the FedSM project coordinator (the project that created the FitSM IT service management standard) and the chairman of an ITSM training organisation mITSM (the Munich Institute of Service Management).

A certification scheme needs to be backed by a certification authority, to provide validity to the certification, and increase the quality of the certification by ensuring standard practices are implemented through its development. The FitSM certifications are (were) backed by TÜV SÜD, an internationally recognised certification authority that is supporting a number of ISO standards as well as FitSM.

5.4.4. Conclusions

There are differences between certificates, and badging schemes in a number of respects, including their level of formalisation and granularity. Applying some of the existing certification schemas - particularly FitSM personal certification, or ISO20000 institutional certification - could help ensure that EOSC providers manage services professionally, with customer focus, for the benefit of EOSC users. Further engagement with the other EOSCpilot workpackages, in particular governance, is needed on the extent and nature of certification of training in EOSC.

[79] IBM Badges https://www-03.ibm.com/services/learning/ites.wss/zz-en?pageType=badges&id=35f7fa5d-d9b3-4a2f-afdc-444394d60d49

[80] SURFnet White Paper on Open Badges https://www.surf.nl/en/knowledge-base/2016/white-paper-on-open-badges-and-micro-credentials.html

6. DRAFT COMPETENCE FRAMEWORK FOR EOSCPILOT

6.1. Introduction

The WP Skills Framework aims to support development, through training and other means, of competencies relevant to the enhanced capabilities for data science that EOSC will provide. A key aspect of the framework is to align individual needs for skills development with organisational needs for capability development, since the competences should relate to the capabilities organisations need if they are to use EOSC services to best effect. This needs a staged approach;

- Identify the scope of the skills likely to be needed. This is the main aim of the current work to identify the elements and dimensions of a competence framework for EOSCpilot. We have approached it by synthesising current competence-based approaches reviewed in section 6, and describe in this section how we intend to make this usable by organisations to plan their skills development.
- Align the competences required of individuals with the capabilities required of organisations using the services and tools EOSC provides. This requires a capability model flexible enough to reflect evolving scope of the Science Demonstrators and underlying services emerging from EOSCpilot wp4 and wp5.

The aim of this section is to accomplish the first stage and lay the groundwork for later stages of this WP, in tasks 7.2 and 7.3.

The main targets of the framework are individuals and bodies who have responsibility for training or skills development, and need to identify the competences their organisation should develop. In practical terms the end result of applying the framework will be one or more *skills specification.*

To apply the framework a user will need some awareness of the following

1. What lifecycle stages and related competences need to be covered
2. Who needs the skills, i.e. which roles (researcher, data manager, data scientist etc.)
3. What level of competence is needed, i.e. awareness (comprehension), application, or evaluation
4. What the culture is in the organisation, and the user community. In other words, how will skills development be take account of the division of roles and responsibility for stewardship between:
   a. individual researchers
   b. research teams i.e. one or more Principal Investigator with project-level responsibility, or an embedded data specialist with cross-project responsibility
   c. the organisation, at any appropriate level of delegated authority or outsourced responsibility

The framework invites the user to classify the level of organisation, competence and responsibility according to simple 3-5 level scales, basing their choices on their knowledge of their target audience and organisation(s). This should then equip them to pick from a list of indicative competences for each relevant data lifecycle stage.

6.2. Definitions and use of existing approaches

Adapting the definitions given in section 6, the draft EOSCpilot framework uses the following definitions.

Data stewardship - formalization of roles and responsibilities to ensure that data is managed for long-term reuse, and that roles are performed in accordance with FAIR data principles

Data stewardship competence - demonstrable ability to apply the knowledge, skills and attitudes needed to meet responsibilities for data stewardship

Current frameworks described in section 6 are mainly oriented to curriculum development, with the

exception of the Go-TRAIN initiative. The draft EOSCpilot framework borrows elements from each of them but takes the latter's approach of viewing data stewardship as a generic set of skills that need to be developed across a range of existing roles. This is informed by but differs from the EDISON approach, which views data stewardship as one of a number of distinct professional profiles. The main influences of the approaches are summarised in Table 6.1 below.

Table 6.1 Sources informing EOSCpilot draft competence framework

Approach	*EOSCpilot framework comparison*
EDISON	Adopts competences from the EDISON Data Management group and some of those for Data Science Engineering (DSENG) Data Science Analytics (DSDA) Scientific/ Research Methods (DSRM). EOSCpilot also aims to equip users with the elements to construct a role profile, but through guidance rather than a generic template.
GO-TRAIN draft	Identification of stewardship as a skill requirement for all researchers, as well as data specialists at research team and organisation levels. EOSCpilot combines the latter with long-term stewardship, while leaving it open to users to determine the location of the skills in the host institution or a third-party organisation
Lyon & Mattern	Linking of stewardship with data curation
RDA Educ. & Training IG	Competences listed for RI managers and research librarians map to higher level competence groups in EOSCpilot (fig 7.2), while those for researchers map to lower level (individual-level) groups, and those for research librarians map to team or organisation level
Purdue Uni. Libraries	Division of competences between individual, team and organisation. Competence statement examples are also adapted.
SFIA	Identifies competences according to levels of responsibility

6.3. Open data science stewardship skills

The framework synthesizes competences that other frameworks outlined in Section 6 have identified with the following skills areas, illustrated in Figure 6.1.

Data management skills to deploy and apply data services to improve understanding of research data management practices; ranging from those required to make data FAIR across domains, to those required to make data actionable for research in at least one domain

Data science engineering skills to build, deploy and maintain data services for research application e.g. requirements engineering, scripting or programming, software engineering, database management, security and authentication, storage management

Data science/ analytics skills to apply data services to research methods innovation, enabling researchers to deploy capabilities such as predictive modeling, machine learning, text/ data mining, data integration, or visualisation

Domain research skills to apply data science and analytic techniques to the research domain for innovative purposes e.g. to enhance research methods or their application to collaborative research

An earlier version of these skills areas was used in questionnaires which asked Science Demonstrator contacts about skills gaps. As mentioned earlier in Section 5, the term 'data steward' was used in that questionnaire to refer to the same competences identified above with 'data management'. We have changed to using stewardship as a more generic term, recognizing that the reported skills gaps were spread across other areas. The above definitions combine the skills areas of data science and data analytics, as the distinction between these is likely to be less significant in a professional development context than it is for curriculum development.

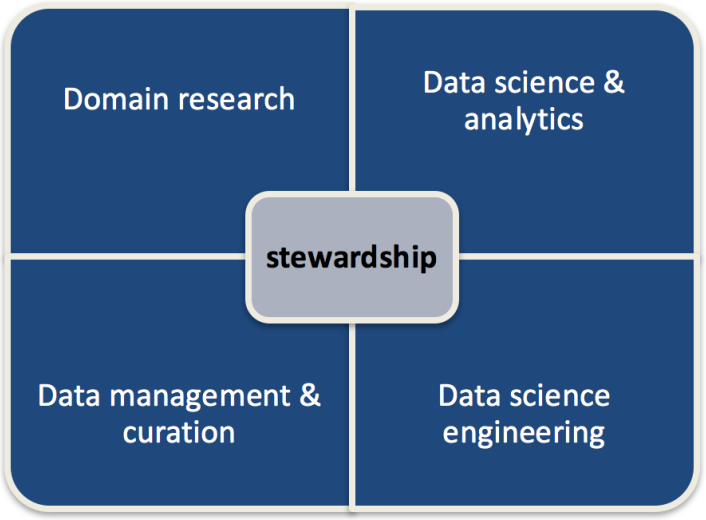

Figure 6.1 Skills areas

6.4. Principles: to govern the framework

The principles below are intended to make the framework flexible enough to use in varied professional development contexts, and alongside existing frameworks.

1. **Recognise FAIRness spans multiple data lifecycles**
    a. As the FAIR principles articulate, stewardship implies long-term and sustainable care across multiple lifecycles. This makes stewardship a collective endeavor, involving at least the individual researcher, colleagues in the study ('the team'), their host organisation (or and others providing services) and the research domain(s) or communities that care about the data.
2. **Recognise the policy and organizational context**
    a. The draft framework should enable users to construct role profiles, e.g. to define professional development goals, taking into account their organisation's professional development and data policy context.
    b. The framework should recognize that organisation and research cultures will influence the levels of responsibility that different roles have across the data lifecycle. There will also be variation in how funder and institutional policies (or legislation) define roles and responsibilities, e.g. by placing differing expectations on the research team, host institution and third-party organisation (such as data centres).

c. The framework should be flexible enough to allow competences to be identified with different levels of the organisation, reflecting the relationships in place between individual researchers and staff with stewardship responsibilities in their team, centralised research data management services, and other providers in their research community, including EOSC service providers. The extent of outsourcing or shared responsibility for stewardship will also depend on organizational characteristics e.g. research strategy and capability.

3. Recognise disciplinary norms in stewardship responsibility

a. The framework should help users deal with disciplinary differences across the data lifecycle in the level of responsibility for data stewardship that they can realistically assign to 'embedded' individuals – referring to those with a mandate to manage data at research team level (e.g. for one or more Principal Investigator).

In some domains it may be preferable to expect individual researchers to be primarily responsible for stewardship, supported by data librarians operating across higher level organizational groups (departments etc.). This variation is likely because research domains vary in their propensity to work in teams, availability of domain-level standards e.g. for formats and metadata, well- defined reuse cases, and availability of third-party repositories.

6.5. Applying the principles to the framework

6.5.1. Lifecycle–based skills groups

Several of the competence frameworks reviewed in section 6 use lifecycle models to represent areas of stewardship or curation skills, including the EDISON project model (Figure 6.1), and Purdue University's Data Information Literacy model. There are many such models, including the Digital Curation Centre's 'Curation Lifecycle' and 'RISE'. The model proposed for the EOSCpilot in Figure 6.2 is a synthesis of these and several others commonly used in data management training: from the University of Edinburgh MANTRA online course, UK Data Archive, and University of Bath Research 360 project.

The terms in the model have been adapted to emphasise that software and service/ infrastructure management is within their scope. For example 'Publish and Release' 'conveys more than 'publishing' alone does, that the competences include making software available, and versioning of dynamic datasets.

The inner ring of Figure 6.2 describes research project-level data stewardship activity, while the outer ring identifies cross-project support activity, whether at the individual researcher, research team or organisation level. These include the following, for example:

- Govern and assess: IPR management, legal and policy compliance, security and risk management, trend monitoring, strategy development,
- Scope and resource: business planning, costing of data management and preservation, service level management, project and portfolio management
- Advise and enable: personal development, education and training, consultancy,

Annex A further describes the indicative scope of each competence group, showing how these are derived from the three main source frameworks (EDISON, RDA-ETDH-IG and Purdue University)

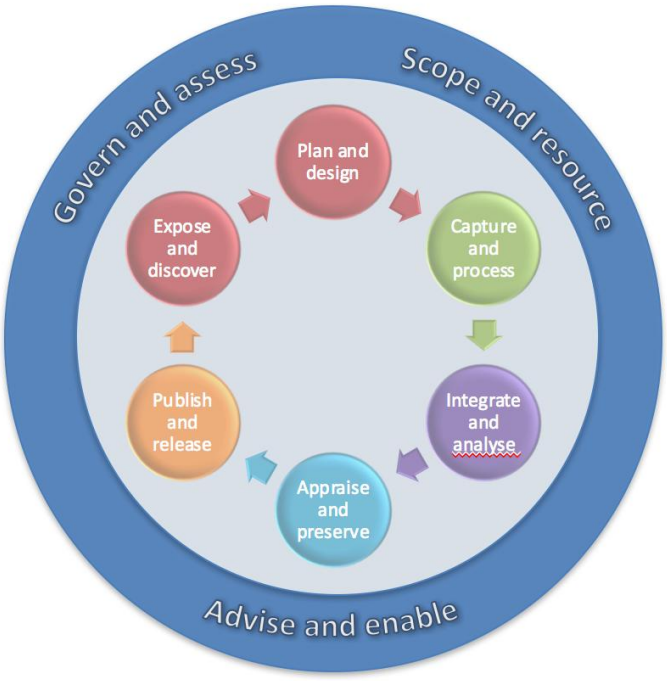

Figure 6.2. Open data science stewardship activities and skills groups

6.5.2. Competence levels

Both EDISON and the Purdue University frameworks categorise levels of competence by adopting aspects of Bloom's Taxonomy, a standard reference for curriculum development. The taxonomy categorises competence according to three domains, the cognitive, psycho-motor, and affective. The Purdue University – Sapp Nelson Framework identifies these domains with knowledge, skills and attitudes (respectively). The EDISON framework by contrast focuses on the cognitive domain and maps the competence levels in that domain to three "mastery/proficiency levels": Familiarity, Usage, and Assessment.

The EOSCpilot framework proposes similar but slightly different terms, on the basis they show a clearer progression of data stewardship skills. These are Comprehend, Apply, and Evaluate as shown in Figure 6.3

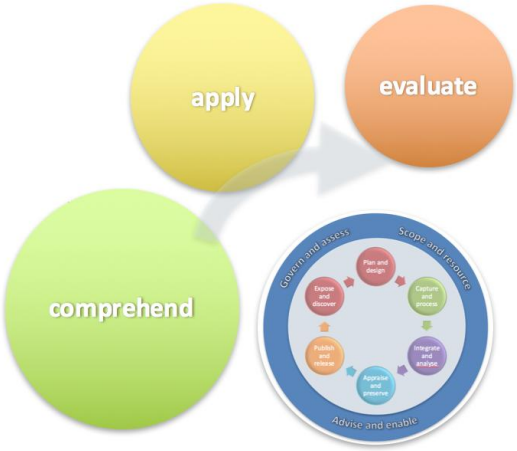

Figure 6.3 Competence levels

6.5.3. Organisational context and culture

The EOSCpilot framework adopts the Purdue University – Sapp Nelson (2017) approach of describing competences differently according to whether they are the remit of the individual researcher, research team or organisation. For our purposes we take these to refer to the following:

- Individual: individual researcher or data service user
- Research team: specialist supporting data stewardship across a research team, or acting with the delegated authority of its Principal Investigator(s)
- Organisation: professional service at any broader organisation level than the research team e.g. data library, research data management service, Research Infrastructure service

The framework offers competence statements that match higher level skills groups, for local adaptation to fit the lower-level topics that are locally relevant. The example below illustrates this.

Table 6.2. Competence statement examples for one skills group

Skills group: Plan and Design

	Comprehend	**Apply**	**Evaluate**
Organisation	Understand and describe funder and legal requirements, the service management context for plans, and their practical application to data stewardship	Develop guidance and support for local practice, balancing service management priorities with policy expectations, and complying with legal or regulatory requirements	Review guidance to reflect change in policy or legal environment, ensuring risks to quality and sustainability are managed, and planned service use is monitored
Research team	Understand and describe how local guidance for plans and protocols balance short term research quality needs and risks with long-term sustainability of outputs	Develop local guidance to support researchers in writing plans and protocols that balance short term research quality needs and risks with long-term sustainability of outputs	Review local guidance to ensure they reflect changing local or research community norms and to ensure risks are adequately addressed
Individual	Understand and describe how planning in data stewardship addresses FAIRness principles and research reproducibility norms in the context of local guidance and service user agreements, funder policy and legal requirements	Develop and implement plans and protocols to fit the needs of the research and follow local guidance, organising data/ software production and service use in accordance with these.	Review plans and protocols to ensure they reflect changes in local standards and service user agreements, funder policy and legal requirements

The example in table 6.2 for the 'Plan and Design' skills group shows text that can be adapted for use in a skills portfolio to fit the more specific scope of the local training and development need. Annex A matches each of the skills groups with these more detailed topics; e.g. in this case the text could be elaborated to refer to data and software management plans, metadata or database specifications or architecture designs as appropriate.

6.5.4. Responsibility levels

Viewing stewardship as a collective endeavor implies that individual researchers have responsibilities to each other as colleagues, to their research project, and to their organisation, as do each of these in return to the individual. These responsibilities would include the following for example.

- Codes of conduct for research integrity, and ethics processes. Legal frameworks such as copyright, data protection, or clinical regulations also often place obligations on individuals as well as organisations.
- The balance of responsibilities between the individual researcher, and specialist support roles embedded at the team or organizational level, for finding relevant data for potential reuse in the research, and making data produced in the research FAIR.
- Organisational responsibilities for research data beyond the lifetime of its originating project, to meet obligations defined by the legal and policy framework for the research.

As well as ensuring that competencies include these *kinds* of responsibility and their location in the organisation it will be important for skills management purposes to take into account the *level* of responsibility, as illustrated in Figure 6.4

Figure 6.4 Organisation and responsibility levels

The two main benefits of including levels of responsibility in the framework are:

- Take into account the different training needs of target audiences arising from their level of responsibility in the organisation
- To support its use in planning data stewardship career pathways

The EOSCpilot framework proposes 5 levels of responsibility, broadly informed by the levels 3 to 7 in the SFIA framework (reviewed in section 6). These are as shown in Table 6.2.

Table 6.3 Responsibility levels

level	description
5	Full accountability: leadership, trend monitoring, strategy and policy definition, resource allocation, and development of relationships with key partners, standards bodies and stakeholder groups
4	Objective setting: research/development proposals, standard operating procedures, guidelines, service level agreements, monitoring of overall compliance and performance against objectives
3	Substantial discretion: project management, stakeholder consultation, data/software management planning, preservation planning, negotiation of service levels, service portfolio management, service delivery, risk management and problem diagnosis
2	Discretion: monitoring data/ software management and service use against plans and agreements, research data and metadata management, documentation, and administrative records
1	Support: liaison with team members, managers, and research data or software/ service users to gather feedback and performance metrics to inform monitoring of outcomes and impacts against plans and user agreements

6.5.5. Service development level

The framework reflects the emphasis of EOSC on service integration and enabling use, rather than on earlier stages of service development. We have identified the following three levels as of most relevance to EOSCpilot skills development purposes.

- Application integration: Business plan development, Software versioning, Data transformation, Data versioning, Software component integration, Data service documentation
- Service operation: Service level management, Change management, Problem management, Security and risk management, Quality management,
- Enabling access and use: all topics in Annex A not mentioned above

The competences listed above are mapped in Annex A to those identified in the EDISON model, and in the RDA ETDH Interest Group wiki.

6.6. Using the framework to specify skills requirements

The framework is intended firstly as a reference point for discussion on the scope of open data science stewardship, and standardization of terminology around that, within EOSCpilot and with related skills initiatives. Beyond that, the main users and use cases for the competence framework are envisaged to be as follows.

Research centres and institutions (research-producing organisations), as EOSC service users

- For research data service managers; to identify the competences required to deploy and use EOSC

services, relate these to the people in their organisation with the appropriate stewardship roles, and to relevant training materials or events for them and the researchers they support.

- For HR professionals responsible for filling roles and developing career structures for open research and data science, to identify the stewardship competences that researchers and professional support need at different staff grades.

Research infrastructures as EOSC service providers

- To help training coordinators develop events and materials that will help the take-up of their services, by describing in a consistent manner the competences required to deploy and use those services, and the competences they should gain as a result.

Example: A university central Research Data Support service plans the skills development needed for Digital Humanities Coordinators to improve digital humanities researchers' awareness of software management planning, and enable their access and use of a software repository wizard.

1. skills groups: Identify from the lifecycle model the skills groups relevant to the user organisation's goals. (Note: the next version of the EOSCpilot framework will include a capability model to support this step, by showing how EOSCpilot services relate to the skills groups, their target groups and which support roles they are likely to impact on).

> *Example: the Research Data Management Service manager's goals include "integrating software management planning into data management planning".*

2. Based on the expected support needs and scope of the data science tools and services to be developed/ deployed/ used, and knowledge of current capabilities, Identify the gaps in the organisation's stewardship roles needed to deliver the improved data science capabilities.

> *Example: discussion with Principal Investigators has identified that researchers in the College of Arts and Humanities have low awareness of software management plans*

3. Identify where in the organisation you expect the relevant people in these roles to have the appropriate levels of responsibility to deliver the competences needed, taking into account the service target groups, and how their research tools and services are normally supported in your organisation, and in their domain.

> *Example, the service manager knows that the College of Arts and Humanities has digital humanities coordinators with senior data stewardship level roles, and has identified a central Library Systems Development team with a role in engineering the new software repository wizard.*

5. Matching the competence statements, and roles to relevant topics (see Annex A). create a skills plan with headings for each group of staff the development goals apply to, based on the roles and responsibility levels expected to be needed. Under each heading list the skills group and topics that need to be developed, and adapt the wording of the competence statements to describe what behaviors would be expected of people at this responsibility level and level of competence.

> *Example, see Table 6.4 below*

6. Use the EOSC training infrastructure to find relevant experts, resources or events available to help deliver the plan. i.e. sourced from Research Infrastructures and other networks (e.g. OpenAire NOADS, GO-TRAIN implementation Network).

Table 6.4 gives an example of Skills Plan for a Digital Humanities Coordinator embedded at local organizational level, to support increased focus on software management planning and the deployment of a (fictional) EOSC software repository wizard, by revising the guidance and support tools available to affected researchers.

Table 6.4 Skills plan example

Role and grade	Digital Humanities Coordinator, Grade 7-8, College of Arts & Humanities

Skills area	Key responsibilities
	Review DMP support template to reflect new funder expectations regarding software management planning, ensure digital humanities researchers are aware of relevant risks to software sustainability, and support roll-out across relevant departments of the EOSC software repository wizard.
Plan and design	*Comprehend:* Data management planning, Database design
	Apply: Service level management, Service planning
	Evaluate: Service level management, Software requirements
Integrate and analyse	*Comprehend:* Software component integration
	Evaluate: Software versioning
Appraise and preserve	*Comprehend:* Preservation planning
Publish and release	*Comprehend:* Software repositories
	Apply: Data service documentation
Govern and assess	*Comprehend:* Research reproducibility
Scope and resource	*Apply*: Project management
Advise and enable	*Comprehend:* Disciplinary practices, norms and values
	Apply: Training, advocacy, awareness
	Evaluate: User support

6.7. Conclusions on competences

This section has introduced a competence framework based on a synthesis of a number of relevant current frameworks, and outlined a process for applying it to fulfil three main use cases, shown in Figure 6.5 below.

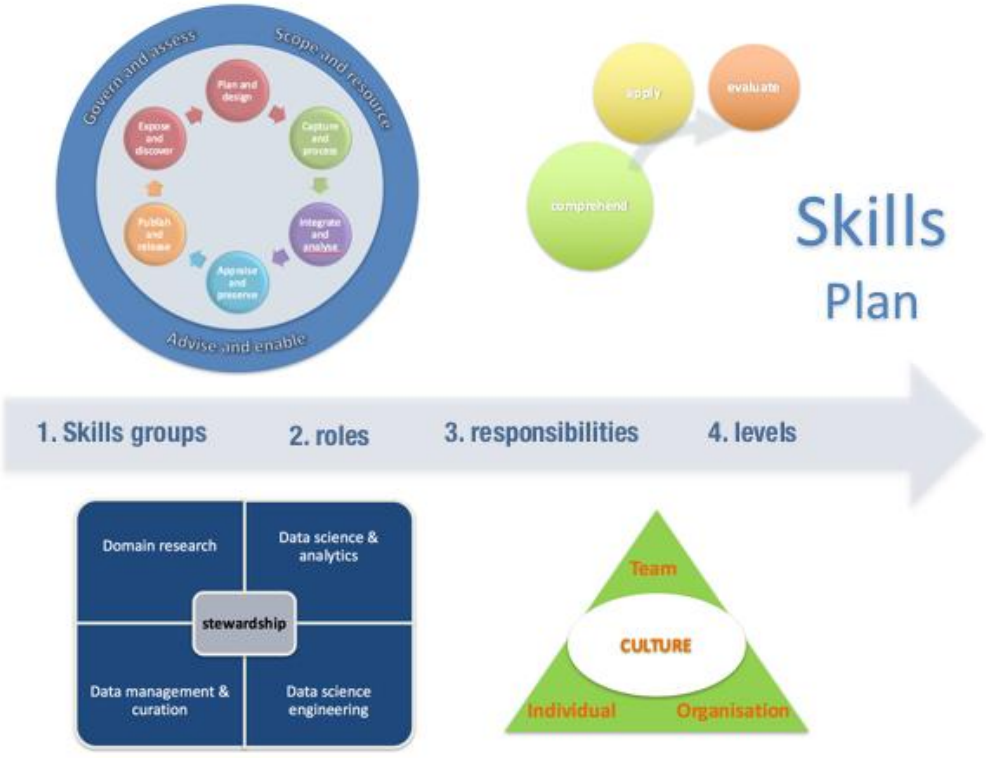

Figure 6.5 Applying the competence framework to produce a skills plan

This is a first draft of the Wp7 Skills Framework and needs further work to develop and validate the competence statements with EOSCpilot partners and stakeholders. This work, planned in Task 7.3, is needed to ensure the scope aligns well with that of the capabilities that EOSCpilot services will offer researchers and their organisations. There is also a need for further liaison with wp5 and wp6 to ensure a process is in place to link the service catalogue to information about the competences needed to integrate, run and enable the use of the listed services.

7. CONCLUSIONS: TOWARDS EOSCPILOT SKILLS DEVELOPMENT FRAMEWORK AND TRAINING INFRASTRUCTURE

7.1 Introduction

Before drawing overall conclusions from the previous sections, this section begins by outlining next steps in EOSCpilot Wp7. The relationships between the present report and forthcoming outputs of Wp7 were illustrated in section 1. As Figure 1.1 suggests, the draft competence framework in section 6 will be further developed into a Skills Framework, through capability modeling in Task 7.3. Also the various initiatives catalogued in section 3 will inform the set-up of training infrastructure in Task 7.2, which in turn will inform events and materials catalogued in D7.2. The next steps considered below begin with initial ideas on what that infrastructure would entail in a broader EOSC, and then return to the Skills Framework.

7.2 Training-as-a-service infrastructure layers from catalogue to storage

To maximize the outcome of training across EOSC, training infrastructure will need to draw on relevant services to provide access to events and materials. As the EOSCpilot architecture and first services become clearer it will be possible to further assess how extensive the requirement is for training infrastructure.

A central training catalogue, portal, delivery system and training content management will probably be required no matter how training is to be organized in EOSC, and what subject areas it will cover. However given the diversity of services and use cases that EOSC is expected to include and support, having a 'single' training infrastructure that encompasses all services may be unrealistic. For example, services may be relatively simple, requiring online user guides but no training events. And although complex services may need to be set-up separately for training purposes, in many cases this is not necessary as the production service can also be used for training (e.g. by supporting short-lived training accounts).

The layered service model shown in Figure 7.1 provides a draft for further discussion with training stakeholders, with the caveat that not all layers will be needed for every service to be used in EOSC. This section offers initial conclusions about the scope of each layer. For each, relevant exemplars and standards are proposed that would enable within-layer interoperability; and assumptions stated about the dependencies between layers implied in the diagram.

7.1.1. Registry/catalogue layer

Scope

To present information about research data skills training, and automatically harvest or mine that information from the training content portals or repositories of EOSC training providers.

Relevant exemplars and standards

ELIXIR Training e-Support System (TeSS), EOSCpilot Skills Framework; FOSTER thesaurus; generic research subject schema (e.g. DOAJ subjects); domain-specific subject schema (e.g. MeSH), training material and events schemas (e.g. bioschemas.org).

Dependencies

Websites and repositories participating in EOSC must be able to expose information on training events and materials in a harvestable metadata format, or enable that information to be automatically mined from their sites. These need to support the enabling metadata standards and web protocols (e.g. schema.org microdata, and/or Dublin Core and OAI-PMH).

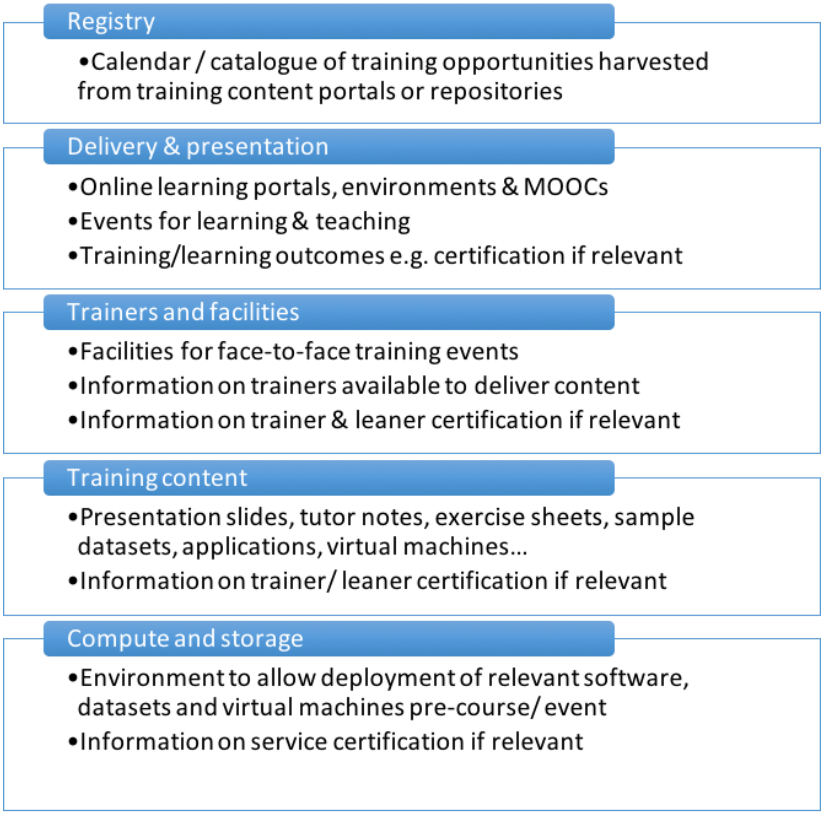

Figure 7.1 Infrastructure for training-as-a-service

Cross-linking between the catalogue and an EOSC service catalogue is a logical step, at least to include the catalogue as one of the services listed, and potentially also at the more detailed level of links between records about the pre-requisite competences to use (other) services and relevant records about training to provide those skills.

7.1.2 Delivery and presentation layer

Scope

Presentation of training events (online, face-to-face or blended), learning opportunities (e.g. Fellowships, internships, placements, staff exchanges), online learning materials, guidance materials and 'wizards'.

Relevant exemplars and standards

Project portals e.g. FOSTER, EDISON; national portals e.g. Research Data Netherlands, Open Science Finland; Infrastructure portals and websites of their individual nodes. MOOCS, their underlying content management systems and delivery platforms.

Dependencies

See layer above, and for the layer below; consensus on the content delivery and management technologies and distribution of effort required to maintain access to the training content and their metadata.

7.1.3 Trainers and facilities layer

Scope

EOSC will need shared information on the availability of specialist trainers to cover the variety of topics in the Skills Framework. Similarly, shared information on the availability of physical training facilities (venues, equipment, contacts) would be useful. A service offering CRM (Customer Relationship Management) would

facilitate this layer. The dynamic nature of this information and its dependence on local knowledge suggest this is more likely to be sustainable through a trainers' network or community of practice than through centralized administration.

Relevant exemplars and standards

This is already practiced to some extent by the EUDAT training team. An example of how to make specialisations explicit is the Speakers Directory provided by the FOSTER project (https://www.fosteropenscience.eu/speakers-directory). There are potential overlaps with the roles of 'data champion' networks, coordinated at the local level (see examples in section 3 for University of Cambridge).

Dependences

See layer above, and for the layer below; the ability of EOSC trainers to access and reuse content on relevant topics, at the appropriate time and place for training events, and to contribute revised or additional content for future use.

7.1.4 Training content

Scope

EOSC will need regularly updated high quality training content, covering the variety of topics in the Skills Framework. Section 3 headings cover the anticipated range of content providers.

Training materials have a pivotal role in the training process: not only for each individual training session, but also to be reused in new sessions and for the continued private study of learners after events. Training content can cover a wide range of types:

- Presentation content such as slides or notes to illustrate, lead and assist a trainer.
- Interactive content such as exercises and tutorials to challenge, involve and guide learners.
- Resource content such as virtual machines or sample datasets to demonstrate, contextualize and provide lessons.

Training events will commonly rely on multiple content types at the same time; for example, a presentation can provide the background for exercises conducted on virtual machine infrastructure. Whereas presentation content follows similar requirements regardless of scientific field, interactive and especially resource content must satisfy specific challenges.

Relevant exemplars and standards

See examples listed in section 3. To support interactive contents for a range of devices including mobile devices the Xerte Online Toolkits are utilized to embed explanations, descriptions, examples, exercises and video clips. The MOOC Research Data Management and Sharing on the other hand utilizes the proprietary technology stack provided by the Coursera platform. However, also open source solutions such as XBlock Courseware Components provided by edX.org are available to deliver interactive courseware.

Dependences

See layer above, and for the layer below; the ability to store, move and retrieve training content, and to run any software it must interoperate with to meet learning outcomes.

7.1.5 Compute and storage layer

Scope

Efficiently providing training on an international scale likewise requires efficient provisioning of infrastructure, namely both compute and storage resources. To satisfy a broad range of requirements for different training scenarios requires an agile, dynamic setup to prepare, deploy and share training environments. Ideally, such an agile infrastructure is suitable to provide training environments both on general topics, as well as the infrastructure itself.

Relevant exemplars and standards

The EGI Training Infrastructure provisions cloud-based computing and storage resources for training events, which themselves focus on skills required in the EGI environment. This training infrastructure is hosted as a dedicated resource pool on the EGI Federated Cloud infrastructure itself. The infrastructure provides resources and services for face-to-face events, online training courses (webinars, MOOCs) or self-paced learning modules, and can be extended with customised training environments on-demand. This breadth and flexibility is achieved by using the same high-quality computing and storage environment that EGI provides to researchers. In a similar vein some of the EUDAT services have a separate instance for training and self-paced learning.

The EGI Online Storage allows storing data in a reliable and high-quality environment and sharing it across distributed teams. This makes it ideal not only for the provisioning of research infrastructure, but also for the provisioning, organisation and distribution of training data and materials. In this regard, the various EGI storage services all offer the persistent availability of passive training resources independent of training volume. In contrast, EGI Cloud Compute provides on-demand deployment and scaling of active training resources to match current need and demand. It allows to create disposable training environments on virtual machines and sale training infrastructure as needed. Because of the cloud- and container-based operational model students can experience dedicated training environments, and organizers can benefit from the easy deployment, predictability and repeatability of courses.

7.1.6 Certification

Scope

The model shown in figure 7.1 includes certification at each level, and could be depicted as a vertical service. The rationale is that, while no certification scheme we are aware of would apply to every layer, certification ought to be treated in a consistent manner across the layer, whether it applies to training services, provider organisations, individual trainers, or the recipients of training.

Certification implies a certification authority, as discussed in Section 5; authorized to issue certificates with a level of trustworthiness among certificate users and stakeholders in the domain concerned, according to a transparent process and criteria. The scope and formality of this ranges from very broad and formal (e.g. SFIA) to the very narrow and informally (e.g. open badges).

Relevant exemplars and standards

EGI provides certification of FitSM training participants, while holding certificates for the organization and its IT service management itself. The training covers three levels: Foundation, Advanced and Experts and is accredited by TÜV SÜD, a global leader in standardisation and certification. The EGI Foundation itself has been awarded the two certifications ISO 9001:2015 and ISO/IEC 20000-1:2011. The two international standards certify appropriate Quality Management System of the organization and the continuous improvement as well as the excellence and proof of best practice of IT service management.

MOOC platforms such as Coursera also provide a Certificate of Attendance after successful completion of a course for a small fee. Education providers using such platforms need to consider whether and how any certification for MOOC participation relates to their formal qualification processes.

Formal certification of training and/or trainers may be desirable for training that focuses on deployment or application of EOSC services. We assume that EOSC will not seek to displace existing certification authorities for these, such as FitSM, but may potentially offer a mechanism to federate the administration of selected certification schemes, or at least information about these.

The pros and cons of extending formal certification (or accreditation) processes for training, further than service management, need further investigation. For example, they may act as an undesirable barrier to formation of networks of trainers knowledgeable on data stewardship topics relevant to EOSC. Informal more light-weight approaches to certification e.g. Open Badges, are worth further consideration as an

incentive for participation in training. Some evidence is available that this approach works as an incentive to apply data skills towards data sharing. E.g. https://researchintegrityjournal.biomedcentral.com/articles/10.1186/s41073-017-0028-9

7.3 Skills Framework next steps; mapping competences to capabilities

The EOSCpilot skills framework comprises both competences (which apply to individual behavior) and capabilities (which describe organizational behavior). Encompassing both elements will help make the Skills Framework useful to organisations planning skills development, by informing the scope of that work.

The draft competence framework described in section 6 describes competences needed in different dimensions. An important dimension is the "level of organisation" responsible for delivering the various competences described. The levels are: Individual researcher, research team and service provider organisation. These levels of organisation of the competence model need to be mapped to service capabilities which will define the EOSCpilot capability model.

This work will be informed by existing capability models, which offer a reference point for discussion with other workpackages on the EOSCpilot –specific capabilities that may be missing. Competences described so far in the draft framework which cannot be mapped to the existing capability models may also need to be represented in the EOSCpilot skills framework. An iterative approach is needed to evaluate the competences and map them to capabilities.

7.3.1 Data Seal/ Core Capabilities as a basis for EOSCpilot

The basis for an EOSCpilot capability model could be provided by existing models e.g. the Data Seal of Approval (DSA). The guidelines of the DSA first give data producers the assurance that their data and associated materials will be preserved in a reliable manner and can be reused. Second, the DSA supports data repositories in the efficient archiving and distribution of data. Recently the DSA implemented the "Core Trustworthy Data Repository Requirements" which has been developed in collaboration with ICSU/WDS.

With these aspects the DSA can be a good basis to meet some of the requirements of the EOSCpilot, by helping to determine the organizational and service capabilities. The DSA certification process is based on a self-assessment of how its guidelines are met by a repository, and the assessment is then peer-reviewed by an external reviewer. Using this model would involve extending its scope from data repositories to data services or e-infrastructures.

Extending the scope is the first goal. Experience in EUDAT indicates it is challenging to formulate DSA-like guidelines for data services or e-infrastructures other than repositories, as the DSA is limited by the scope of its reference model, the OAIS (Open Archival Information Systems) standard.[81] The recent implementation of the Core Trustworthy Data Repository Requirements will help in this respect, as the focus of the DSA is extended with infrastructures and security. When the first goal is achieved the missing service capabilities for e-infrastructures as EOSCpilot can be detected and the model can be extended.

7.4 Overall conclusions

Finally, we offer conclusions on the Skills Framework and Training infrastructure to be piloted and tested in EOSCpilot wp7. The EOSC vision is one of an integrated environment providing data and services that are easily discoverable, accessible to all, interoperable across their underlying infrastructures, and reusable for innovative purposes. EOSC will have to consider and develop skills for both service provision and service use. The range of expertise that we refer to as 'open data science stewardship' in this report covers roles

[81] EUDAT (2017) D2.3 Strategic Certification Plan (to appear) http://www.euda.eu

and topics that may need different approaches, and different levels of support as a service matures.

If EOSC should in general terms 'implement the FAIR principles', there are implicit requirements to ensure researchers have the skills to do so when they use EOSC services, their organisations enable and reward such skills, and that research support staff are also enabled and rewarded for sustaining the FAIRness of research data or code produced using EOSC services.

The report has synthesized evidence of training and other skills development in two broad areas; data management and open science, focusing on the main groups of stakeholders in EOSC as the providers of that training. We draw conclusions in this section about the needs identified so far within the EOSCpilot, and about the competence framework that we have drawn from other initiatives with similar scope. We then consider the infrastructure needed to provide training, both within the pilot and in the broader EOSC and ask; should the FAIR principles apply to training and skills development itself? Our tentative response is 'yes, and we should identify what that means for EOSC and how to do it'.

As section 2 shows, there is an explicit policy-driven need to delivering skills improvements that will enable data intensive research. The section identified the policy drivers and demand to expand the availability of data skills to a much larger scale, a demand that also underpins the cross-domain aspects of EOSC. The section also reviewed survey evidence that researchers are unable to find, access and/or apply suitable training in data skills. So, with reference to 'FAIRness', a first conclusion is as follows:

Conclusion 1. EOSC training materials and events must be FAIR, i.e. materials and event descriptions must be provided with standard metadata to make them findable, they must be accessible from EOSC e-infrastructures, they must be in open standard formats so they are interoperable with each other and with the data services they are about, and they must be provided on license terms that are as non-restrictive as possible to encourage reuse.

Conclusion 2. The coordination and delivery of EOSC training materials and events should be managed on a similar basis to the services they describe i.e. provided in the form of 'Training-as-a-Service'.

Training-as-a-Service is not a particularly novel idea; the concept has been promoted by EUDAT since 2014 ([82]) and as Section 3 shows, data and e-infrastructures participating in the EOSC pilot have for several years offered training programmes consistent with the Training-as-a-Service idea, using advanced facilities that include online learning and the cloud-based compute/store services to underpin it. It is reasonable therefore to aim for training and other forms of skills acquisition to be part of the federated, distributed approach envisaged for EOSC.

Section 3 also indicates that Research Infrastructures are potentially key players in delivering Training-as-a-Service. Reports identify that most offer training in the use of their services, i.e. as part of the user support and outreach activity for other services. However, Research Infrastructures are rarely offering training as a service in its own right; as of June 2017, the MERIL portal for the Research Infrastructures landscape listed 1931 services offered by the RIs, of which only 4 mention 'training'. A notable exception (not currently listed in MERIL) is the Training e-Support Service operated by the ELIXIR Research Infrastructure.

A number of the Horizon 2020 cross-infrastructure 'cluster projects' are providing training content that we see as relevant to EOSC, in that the topics addressed are about enhancing the capabilities of Research Infrastructures to collaboratively enable data science or manage the data/ software objects produced as a result. The workpackage is cataloguing the current provision of education and training in the project. This task (7.2) entails gathering further information about individual Research Infrastructure training practices and content, to offer a thematic analysis using the competence framework drafted in section 6, and will assess the relevance of both to EOSC stakeholder needs. It also involves selecting and setting up technology to deliver the catalogue. Our next deliverable (D7.2) will report on this work.

At this stage we have sought a range of examples to illustrate the range of approaches being taken across

[82] Gentzsch, W., Lecarpentier, D., & Wittenburg, P. (2014, April). Big data in science and the EUDAT project. In *Global Conference (SRII), 2014 Annual SRII* (pp. 191-194). IEEE.

infrastructures, and other categories of training provider, as described in Section 3. The diversity of approach across providers is not therefore a surprise, but it is notable that each category of provider offers a similarly diverse range of formal course provision, and online or hybrid self-paced learning coupled with reference material.

Research institutions are shown in Section 3 to play a role in the ongoing skills development of researchers and the stewardship roles of professional support staff. One MOOC example was highlighted, and this approach is likely to have a strong potential to deliver training at the broad scale EOSC requires. However, while it is clear that institutions *can* play a role as training providers to EOSC it is not clear that national-level funders believe they should. Section 2 indicates that few European national-level public research funders address skills provision in their policy statements on open science or management of research data.

Conclusion 3. EOSCpilot Wp7 should use opportunities provided through Wp2 (policy) and Wp8 (engagement) to highlight the relevance of the institutional role in enabling and rewarding data skills development.

Prompted by the reported need for data expertise to be 'embedded' in research teams we also considered approaches that involve skills acquisition through being embedded in research as part of the learners' work, outside their normal (or previous) working environment. This highlighted a significant role for fellowship programmes and internships, whereby early career researchers are offered relatively small funding awards that enable learning-by-doing and knowledge exchange.

Conclusion 4. EOSCpilot should consult stakeholders on a skills strategy for EOSC that, in addition to formal courses, includes skills development approaches embedded in data-intensive research environments, e.g. through Fellowships and staff exchanges.

An important aspect of the EOSCpilot wp7 is setting up training infrastructure for the pilot, to identify lessons that inform EOSC skills strategy. Skills development approaches that involve offering people experience of data-intensive research environments will need support infrastructure that is different from that needed to find and access (or deliver) specific training events and materials. To avoid over-complication this should be considered separately from the training infrastructure.

Section 3 shows that the providers and approaches commonly used for training in this area are more diverse than can be accommodated within the scope of EOSCpilot. If the training infrastructure is to be relevant to the broader community and successor projects it will need to be described in more abstract terms. Section 7.3 below proposes a layered model for providing Training-as-a-Service, drawing on examples described in earlier sections.

Section 4 describes the requirements identified so far in the project, beginning with those emerging from the EOSCpilot governance and policy processes. The section also describes the training requirements of the initial four Science Demonstrators operational in the first six months of the project. These are the first of fifteen, each Science Demonstrator running for one year. The needs reported were broader than data management aspects of stewardship, overlapping with the competences identified in the EDISON project with 'data science engineering' and analytics, for example about deploying services in the cloud. However, it was not clear from the Science Demonstrator reports whether they had identified the skills required for their services to have an impact with users, or to address long-term data stewardship.

The main issue is how to structure training delivery (or skills acquisition) to align the limited opportunities for training events in wp7 with the relatively short lifecycle of Science Demonstrators. Designing training to fit the specific needs of individual Science Demonstrators personnel would inevitably mean skewing resources towards those running earlier in the project, and disadvantage those running later in the project. Also, as any training content that is Demonstrator-specific will take time to develop, it is unlikely to be feasible to schedule delivery to align with the running period of each respective Demonstrator.

Conclusion 5: EOSCpilot should focus its training provision on the *outcomes* from Science Demonstrators, e.g. illustrating how these help researchers apply FAIR principles, or provide lessons in service management, rather than attempt to deliver training within the Science Demonstrator projects.

Training should differentiate between a) what users / researchers need to know, e.g. about applying FAIR principles using the Science Demonstrators, and b) what the relevant service providers and intermediaries would need to do, to be able to operate the Science Demonstrator applications as services in EOSC.

Section 5 describes competence frameworks relevant to EOSCpilot, and its focus on stewardship roles for open data and data science ('open data science stewardship'). The reviewed frameworks are based in initiatives with very similar concerns to the workpackage. They include the EDISON project, Research Data Alliance Interest Group on Education and Training in Data Handling, GO-Train element of the GO-FAIR initiative, and a US project on 'data information literacy' described in work from Purdue University Libraries. These were analysed across a number of dimensions.

The current frameworks share a view that stewardship entails giving all researchers at least awareness level competences, and that these overlap between the data science, research and engineering skills areas. Examples of the topical scope include moving data or code to the cloud, managing software repositories, and research reproducibility.

Section 6 sets out a draft competence model for EOSCpilot, first articulating principles for synthesising elements of the frameworks reviewed. These include, for example, recognizing cultural differences between research domains, and between organisations, in the levels of responsibility expected of roles in different parts of the organisation. Some initial ideas on how organisations could use the framework to develop a skills plan are presented. More work will be needed on this however, through engagement with EOSC stakeholders involved in the development of research careers, and of data expertise.

Competences are the core of the Skills Framework, but these need to be relevant to the capabilities that EOSCpilot services will enhance. So while the first draft in Section 6 represents generic aspects of stewardship, the next stage (carried out in wp7 task 7.3) is to refine and elaborate the framework by mapping the competence groups to a capability model for EOSCpilot, and further describing the competences. This also requires further liaison with other workpackages (primarily wp4 – 6) on the capabilities represented in the EOSCpilot service portfolio.

Conclusion 6. EOSCpilot should refine its Skills Framework through engagement with stakeholders in the development of careers and expertise in data stewardship.

The WP7 partners are already represented in relevant stakeholders including EDISON, FOSTER Plus, GO-FAIR, Research Data Alliance, and Belmont Forum, and will seek further engagement on the Skills Framework with these and other relevant groups. These include the Open Science Policy Platform working groups on Skills, and on Rewards, ESFRI, PLAN-E (Platform of National eScience Centers in Europe) and GEDE-RDA (Group of European Data Experts in RDA)

8.3 Conclusions on Training Infrastructure.

The workpackage will continue to develop the Training-as-a-Service model to contribute to EOSCpilot education and training strategy. In relation to the project work plan this related to task 7.2, which includes the set-up of training infrastructure, and results in a Training Materials Catalogue (D7.2). This corresponds to the top layer in Figure 7.1, the 'Registry/Catalogue' layer. While there is very limited scope for development in this WP, some further investigation of the potential to apply the TeSS system for this purpose is warranted. Other forms of delivery will be considered.

Conclusion 7. Further investigation of the ELIXIR Training e-Support System (TeSS) is needed to establish whether it may be recommended as a cross-domain solution for harvesting training events/ materials, and whether these may be tagged with competences from EOSCpilot or other frameworks.

Certification is the other main area of the workpackage planned for the next six months, and the subject of

the final conclusion from this stage of the project.

Conclusion 8. Broader consultation is needed with the EOSCpilot governing bodies and WPs to establish how far certification will feature in the 'rules of engagement' for EOSC, and therefore how it should apply to education and training. Further consideration is needed on the balance between broad-based certification schemes applicable to services/management, and badging schemes applicable to specific events or materials.

In the latter half of the first year of EOSCpilot the scope of Wp7 will be further defined, to incorporate service capabilities into the Skills Framework, set up the Training Infrastructure, and further catalogue relevant materials from sources including the Research Infrastructures, e-Infrastructures, and related initiatives. We look forward to consulting the skills development community interested in EOSC, to ensure this work responds to their needs and perspectives.

ANNEX A. ANNEXES

Annex A. Indicative scope of data stewardship competences and their mapping to other frameworks

Topic lists below are adapted from the identified sources referenced in section 6, and identified with the following EOSCpilot skills areas:

DSA Data Science/Analytics **DM** Data management **DE** Data science engineering [83] **DR** Domain Research

EOSCPILOT COMPETENCE GROUP	EDISON DATA SCIENCE COMPETENCE FRAMEWORK [84]		RDA INTEREST GROUP EDUC. & TRAINING IN DATA HANDLING (WIKI)		PURDUE UNIVERSITY/ SAPP NELSON - DATA INFORMATION LITERACY [85]	
PLAN AND DESIGN	Data management planning	DM	Data management planning	DM / DR	Data management planning	DM / DR
	Data model development	DM	Data repository requirements	DM	Database specification and design	DM / DR
	Metadata specification	DM	Service level management	DE	File format selection	DM / DR
	Software requirements	DE	Service planning	DE	Metadata specification	DM / DR
	Application design	DE	Application design	DE		
	Database design	DE	Architecture design	DE		
	Research design	DR				
CAPTURE AND PROCESS	Software prototyping	DE	Data documentation practices	DM / DR	Data documentation	DM / DR
	Data collection and normalization	DM	Data backup and security	DM	Controlled vocabularies and ontologies	DM / DR
			File naming and organisation	DM		
	Database management	DE	Data wrangling/ cleaning	DR	Workflow management tools	DM / DR
	Investigation & experimentation	DR				

[83] Includes most competences associated with research Infrastructure managers/operators in the RDA ETDH-IG wiki https://www.rd-alliance.org/group/education-and-training-handling-research-data-ig/wiki/task-force-defining-data-handling

[84] Demchenko, Y., Belloum, A. Wiktorski, T. (2016) EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS) Release 1 http://edison-project.eu/data-science-competence-framework-cf-ds

[85] Nelson, M. S. (2016). **Scaffolding for data management skills: From undergraduate education through postgraduate training and beyond**. Purdue University Research Repository. doi:10.4231/R7QJ7F9R

EOSCPILOT COMPETENCE GROUP	EDISON DATA SCIENCE COMPETENCE FRAMEWORK		RDA INTEREST GROUP EDUC. & TRAINING IN DATA HANDLING (WIKI)		PURDUE UNIVERSITY/ SAPP NELSON - DATA INFORMATION LITERACY	
INTEGRATE AND ANALYSE	Analytics platforms	DSA	Data preparation	DM/ DR	Data transformation	DM / DR
	Predictive analytics	DSA	Data mining	DM	Data processing and statistical analysis tools	DM / DR
	Statistical techniques	DSA	Data versioning	DM		
	Decision analysis	DSA	Software component integration	DE	Analysis workflows	DM / DR
	Data integration	DSA	Maths and statistics knowledge			
	Creative problem solving	DR	Statistical programming languages	DR		
	Software versioning	DE		DR		
			Database query languages			
			Machine learning algorithms	DR		
			Big data frameworks (e.g. MapReduce)	DR		
				DR		
APPRAISE AND PRESERVE	Data provenance	DM	Data review and appraisal	DM	Data review and appraisal	DM / DR
	Data quality	DM	Preservation planning	DM	File format migration	DM / DR
			Data preservation	DM/ DR	Preservation planning	DM / DR
PUBLISH AND RELEASE	Data access and publication	DM	Data citation	DM/ DR	Workflow documentation	DM / DR
	Visualisation	DSA	Visualisation	DM/ DR	Visualisation tools	DM / DR
	Security management	DE	Data publication	DM/ DR		
			Data repository platforms	DM		
			Data marketing	DM		
			Data service documentation	DE		

EOSCPILOT COMPETENCE GROUP	EDISON DATA SCIENCE COMPETENCE FRAMEWORK		RDA INTEREST GROUP EDUC. & TRAINING IN DATA HANDLING (WIKI)		PURDUE UNIVERSITY/ SAPP NELSON - DATA INFORMATION LITERACY	
EXPOSE AND DISCOVER	Repository management	DM	Data standards	DM	Database searching	DM / DR
			Metadata standards	DM	Domain repository evaluation	DM / DR
GOVERN AND ASSESS	IPR management	DM	Research reproducibility	DM	Research reproducibility	DM / DR
	Ethical compliance	DM	IPR management	DM/ DR	Data use agreements	DM / DR
	Research strategy	DR	Data policy and funder requirements	DM	Data quality management	DM / DR
			Ethical and legal compliance	DM/ DR	Ethical and legal compliance	DM / DR
			Research strategy	DM	Storage security management	DM / DR
			Information security and risk management	DM		
			Data governance	DM / DR		
			Quality management	DE		
SCOPE AND RESOURCE	Project management	DR	Business case	DM	Costing data mgmt. and preservation	DM / DR
			Business plan development	DE		
			Requirements management	DE		
			Service level management	DE		
			Change management	DE		

EOSCPILOT COMPETENCE GROUP	EDISON DATA SCIENCE COMPETENCE FRAMEWORK	RDA INTEREST GROUP EDUC. & TRAINING IN DATA HANDLING (WIKI)		PURDUE UNIVERSITY/ SAPP NELSON - DATA INFORMATION LITERACY	
ADVISE AND ENABLE		Data rescue	DM	Standards body participation	DM / DR
		Building collaborations	DM	Disciplinary practices, norms and values	DM / DR
		Social interaction & negotiation	DM		
		Training, advocacy, awareness	DM		
		User support	DE		
		Personnel development	DE		

Annex B Skills Questionnaire for Science Demonstrator Reporting

EOSCpilot WP7 aims to pilot training activities for selected topics, and develop a skills framework and strategy. Your input will help ensure these are also relevant to your aims.

1. Considering the skills needed to integrate the demonstrator application into the EOSC ecosystem, and to turn this application into a service that others can use, do you foresee any skills gaps and, if so, what are they?

2. What skills and competencies will the demonstrator-service user community need, to benefit from using the service and achieve its planned impact?

   Please indicate which of the four competence areas (below) are relevant, what level of expertise in these areas is needed, and available

Data science research methods to apply knowledge of the demonstrator capabilities and related analytic techniques to new research investigations, to create new knowledge and achieve research goals, e.g. cross-disciplinary study design, statistical analysis, interpretation and theory-building.

(please insert **x** in the box to indicate the degree of relevance, or 'don't know')

How relevant are skills in this area to achieving the planned impact for the demonstrator?

Low				High	Don't know	

What **levels of expertise** in these areas does the demonstrator user community need?

Low				High	Don't know	

What level of **availability** of this expertise is there for the user community?

Low				High	Don't know	

Data science analytics to effectively use the demonstrator to support research investigations, by applying its capabilities alongside tools and techniques for predictive modeling, machine learning, text and data mining, data integration, visualisation

(please insert **x** in the box to indicate the degree of relevance, or 'don't know')

How relevant are skills in this area to achieving the planned impact for the demonstrator?

Low				High	Don't know	

What **levels of expertise** in these areas does the demonstrator user community need?

Low | | | High | Don't know |

What level of **availability** of this expertise is there for the user community?

Low | | | High | Don't know |

Data science engineering technology or systems development to enable application of demonstrator capabilities, e.g. requirements engineering, scripting or programming, software engineering, database management, security and authentication, storage management

(please insert **x** in the box to indicate the degree of relevance, or 'don't know')

How relevant are skills in this area to achieving the planned impact for the demonstrator?

Low | | | High | Don't know |

What **levels of expertise** in these areas does the demonstrator user community need?

Low | | | High | Don't know |

What level of **availability** of this expertise is there for the user community?

Low | | | High | Don't know |

Data management and stewardship to support researchers' sustained and effective use of the demonstrator over a period of years, while addressing disciplinary differences in data practices and reproducibility norms e.g. by enabling data policy compliance, acquisition, quality assurance, metadata management, conversion and interoperability, preservation, and secure handling

(please insert **x** in the box to indicate the degree of relevance, or 'don't know')

How relevant are skills in this area to achieving the planned impact for the demonstrator?

Low | | | High | Don't know |

What **levels of expertise** in these areas does the demonstrator user community need?

Low | | | High | Don't know |

What level of **availability** of this expertise is there for the user community?

Low | | | High | Don't know |

3. What existing training/skills learning activities (e.g. courses, workshops, seminars, webinars, MOOCs, internship or mentoring schemes, self-paced learning resources) are you aware of that you would recommend to users of the demonstrator, to help them use it effectively?

   Please identify, and give a link pointing to any further information available.

4. EOSCpilot training workshops will be held in M10, M14 and M19 (approx.). What aspects of the skills challenges for EOSC would you most like to see addressed in these workshops?

Annex C Responses to Skills Questionnaire from Science Demonstrators

Demonstrator/ Area of Expertise	**Pan-Cancer**	**HEP**	**Photon-Neutron**	**Textcrowd**
Data Science research methods				
• Relevance	high	high	med	low
• Level needed	**high**	low	low	**med**
• Availability	low	low	low	low
Analytics				
• Relevance	low	low	low	low
• Level needed	low	low	low	med
• Availability	low	low	low	?
Engineering				
• Relevance	high	low	high	med
• Level needed	**high**	low	med	**high**
• Availability	low	low	low	?
Data Management				
• Relevance	high	high	med	med
• Level needed	**high**	low	low	**med**
• Availability	low	low	low	?

Bold- highlights areas where level needed exceeds availability

Annex D Science Demonstrator skills requirements catalogue

(Example entry for Textcrowd)