D7.2: Interim report and catalogue of EOSC skills training and educational materials

Author(s)	Eileen Kühn, Achim Streit (KIT)
Status	Final
Version	v1.0
Date	22/12/2017

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:

This report provides an overview of the current work of the EOSCpilot Skills and Capacity workpackage (WP7). This includes an update of the EOSCpilot Skills Framework, which is used to categorise, assess and catalogue training resources. Three topics are addressed as a result: First, the design and implications for the EOSCpilot Skills Framework with regard to requirements of service providers and different target groups. Second, the current availability of training resources and their coverage with respect to these requirements. Third, the requirements and basic architecture of a dedicated EOSC training service to apply, reuse and share training resources at the scale of the EOSC. Our efforts include explicit thoughts on the establishment of FAIR training provision, but are limited to findability at the current state.

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 –WP7-D7.2	
Deliverable lead	KIT
Related work package	WP7
Author(s)	Eileen Kühn, Achim Streit (KIT)
Contributor(s)	Kevin Ashley (DCC-UEDIN), Valentino Cavalli (LIBER), Elly Dijk (DANS), Magdalena Getler (DCC-UEDIN), Jonathan Rans (DCC-UEDIN), Gergely Sipos (EGI), Rahul Thorat (DANS), Angus Whyte (DCC-UEDIN)
Due date	31/12/2017
Actual submission date	22/12/2017
Reviewed by	Birgit Schmidt (SUB Göttingen) Simon Lambert (STFC)
Approved by	
Start date of Project	01/01/2017
Duration	24 months

Versioning and contribution history

Version	Date	Authors	Notes
0.1	10/10/2017	Eileen Kühn (KIT)	Content outline
0.2	05/12/2017	Valentino Cavalli (LIBER), Elly Dijk (DANS), Magdalena Getler (DCC-UEDIN), Eileen Kühn (KIT), Jonathan Rans (DCC-UEDIN), Gergely Sipos (EGI), Rahul Thorat (DANS), Angus Whyte (DCC-UEDIN)	Internal review draft
0.3	18/12/2017	Kevin Ashley (DCC-UEDIN), Eileen Kühn (KIT), Rahul Thorat (DANS), Angus Whyte (DCC-UEDIN)	Response to internal review
0.4	18/12/2017	Kevin Ashley (DCC-UEDIN)	Extract from google docs
0.5	20/12/2017	Eileen Kühn (KIT)	QA copy for STFC
1.0	22/12/2017	Eileen Kühn (KIT), Angus Whyte (DCC-UEDIN)	Final version

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The EOSCpilot Skills and Capacity workpackage (WP7) aims at developing standards and assessment frameworks in order to ensure that organisations and individuals are motivated and able to develop the capabilities and competencies that the EOSC will rely on. On the one hand, it will establish the capabilities that organisations need to develop and reflect in the career development strategies, both for researchers and support service staff. On the other hand, it will formalise and reveal the skills needed by individuals to enhance their competencies in open data science and stewardship.

The report *Skills landscape analysis and competence model* (D7.1) was the first deliverable from WP7. The objective of this second deliverable – *Interim report and catalogue of EOSC skills training and educational materials* (D7.2) – is to describe the outcome of cataloguing relevant materials and the progress on the three WP7 objectives:

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.
2. Catalogue the currently available education and training with respect to the skills framework and identify gaps in covering requirements by users and organisations.
3. Develop an EOSC education and training strategy to address the gaps and set up a sustainable technical infrastructure to ensure shared resources are openly accessible and reusable.

This report addresses the first of these aims by progressing on the EOSCpilot Skills Framework that aims to help organisations answer the question

"What skills and competences are required from people in order to develop, provide, or use the services in the EOSC, and meet the goals for open science and data science?"

With the help of this framework, the question can be answered in two ways: firstly, by helping managers plan skills development of their staff and secondly by helping individuals identify relevant resources for their own professional development needs. To achieve this, the EOSCpilot Competence Framework described in D7.1 is the first part of the EOSCpilot Skills Framework.

The EOSCpilot Skills Framework provides the base for cataloguing of relevant training materials and events and the identification of gaps with regard to current user requirements. This directly addresses the second aim. Based on a sample of 90 training materials and 50 events, we validated the process of cataloguing. The most important findings from this process involve the availability and quality of metadata about training materials and events: hardly any training provider exposes structured metadata. Even when metadata is provided, the quality varies greatly, making an automated cataloguing a complex and difficult task that currently requires manual classification of training.

This report further provides findings on current gaps in training materials with regard to skills requirements that we derived from the needs of Science Demonstrators, service architecture of EOSC-hub, and consultation with stakeholder on expected competence gaps. The skills requirements suggest major skills gaps for cross-project skills groups *govern and assess*, *scope and resource*, and *advise and enable* at organisational level. The preliminary analysis of gaps in training materials confirms this gap.

This report also addresses the third objective in terms of proposing a Training as a Service infrastructure that builds on the concept of orchestration of services to support user requirements by researchers, organisations, trainers, curators, and services. In doing so, we have attached particular importance to the FAIR provision of training materials and events in terms of findability and the minimization of operational costs.

1. INTRODUCTION

Given the scale and scope of EOSC, both in terms of infrastructure and scientific domains, it is an unprecedented environment for scientists to conduct research. Preliminary work as part of EOSCpilot and WP7 in specific revealed notable knowledge gaps between e-infrastructure providers and their scientific users. In specific, the first EOSC High Level Expert Group highlighted data stewardship as one of the most significant gaps. Yet, data handling is a key competence required by users in order to efficiently and effectively employ EOSC in their scientific work. As such, EOSCpilot WP7 takes the first steps towards identifying how the EOSC can provide the skills and capacity for open science and data science.

Section 2 summarises the EOSCpilot Competence Framework introduced in D7.1 and provides results of consultation with stakeholder on the consolidation of competences. This section further presents the progress on the EOSCpilot Skills Framework that builds on the EOSCpilot Competence Framework. The mapping of service capabilities to competences that is presented enables a systematic analysis for gaps in training. As a basis for this, this section is concluded with an update to the EOSC skills requirement based on an analysis of workshop outcomes, Science Demonstrator reports, and EOSC service architecture.

Section 3 contributes to the analysis of training gaps by supplying a preliminary comparison of available training materials with skills requirements of the previous section based on the EOSCpilot Competence Framework. To enable this analysis, the report first provides the progress and experience of defining the extent and structure of a training materials and events catalogue for EOSCpilot to help users acquiring the skills to apply EOSC services and enable new data science in accordance with FAIR principles. The section specifically focuses on FAIRness of training materials and events to enable findability as a key driver to support the use cases from the different EOSC users. The section concludes with a preliminary skills gap analysis based on terms from the EOSCpilot Skills Framework.

Section 4 builds on the FAIRness of training resources and defines our solution to an EOSCpilot Training as a Service (TaaS) infrastructure to expose FAIR training resources and make them discoverable. The infrastructure model consists of four distinct layers including presentation and delivery, catalogue/registry, training resource and infrastructure as well as cross-layer functionality. Our analysis in D7.1 as well as an in-depth analysis of existing training infrastructures provided in this section show excellent solutions and tools for the considered layers. The section therefore concludes with an infrastructure that builds on the orchestration of services to delegate responsibilities for the different layers as well as recommendations for services and tools to consider in the future.

The concluding Section 5 summarises the concluding points of each distinct section covering our Progress towards a skills framework for EOSC, the Catalogue of training materials and events, and our solution to provide Training as a Service in EOSCpilot. The section further provides our main conclusions to be considered in the second year of EOSCpilot.

The main conclusions of the report are as follows.

1. The EOSCpilot Skills Framework requires further development towards D7.3, including validation of its conceptual model against the EOSCpilot and EOSC-hub service architecture, and consultation with the target users of the framework in Research Infrastructures and institutions.
2. Current provision by partners and H2020 cluster projects has been analysed, with preliminary results indicating that some skills groups are under-represented in their current training offer. More specific skills gaps are evident from stakeholder consultation, and from Science Demonstrator outputs. These analyses will be expanded and refined in year 2 of EOSCpilot.
3. The EOSC training community should be consulted further on the gaps in skills provision, on whether applying FAIR principles to training resources could help improve provision and, if so, how FAIR principles should be applied in the training context.
4. A conceptual model for Training as a Service in EOSC has been developed and a need identified for consensus among stakeholders in EOSCpilot and EOSC-hub, particularly Research Infrastructures, on the preferred options. These include mechanisms for harvesting metadata, presenting content,

and interoperating with existing portals and catalogues.

5. To address identified gaps for training events and materials, gaps should be filled partly through an open call for participation in online training, and partly through face-to-face workshops by WP7 partners.
6. The EOSC Education and Training Strategy must consider steps necessary to ensure sufficient provision of relevant undergraduate and postgraduate education courses, and to meet the need for data literacy in the broader community.

2. PROGRESS TOWARDS A SKILLS FRAMEWORK FOR EOSC

In this section, we offer a brief summary of the Skills Landscape Analysis and EOSCpilot Competence Framework presented in D7.1. We describe the progress made in months 7-12 of the project, primarily on the EOSCpilot Competence Framework which identifies and formalises competences for data stewardship and other outputs of open data science. Furthermore, we provide an update to competences to reflect ongoing consultation with stakeholders to identify skills of highest priority for WP7.

We also provide a conceptual model for the EOSCpilot Skills Framework, relating stewardship competences of users to the capabilities enabled by EOSC services. Competences and capabilities are the two main elements of the EOSCpilot Skills Framework, to be presented in the forthcoming D7.3 (month 18 of EOSCpilot).

Furthermore, we update the analysis of skills requirements that is the basis for identifying training materials and events to support skills development for the EOSC aims based on a number of sources:

- Results of an EOSCpilot-OpenAIRE joint workshop in Berlin (24 October 2017), to consult stakeholders on stewardship competences to ensure FAIR outputs.
- Science demonstrator reports, highlighting the challenges faced by scientific domains targeted by EOSC.
- Service architectures from EOSCpilot and the EOSC-hub H2020 project, which focuses on the technical adoption, integration and development of existing infrastructure into EOSC.

2.1. Stewardship competences: consolidation and consultation

The EOSCpilot Skills Framework due to be delivered in M18 has progressed from the initial version of the EOSCpilot Competence Framework set out in D7.1. The EOSCpilot Skills Framework will link the competences needed by researchers and other professional groups for data stewardship in the EOSC context to the capabilities that EOSC offers to service developers, providers and users. In the following, we offer a brief recap of the motivations for the framework, its aims, and scope and sources.

2.1.1. Motivations for focusing on stewardship

The EOSC Declaration [1] includes three paragraphs that set out in general terms why this framework is needed, as follows:

1. On skills, the Declaration calls for “the necessary skills and education in research data management, data stewardship and data sciences” to be provided throughout the EU “as part of higher education, the training system and on-the-job best practice in the industry.”
2. Regarding data stewardship, the Declaration states that researchers need the support of adequately trained data stewards, and calls for investment in their education “...via career programmes delivered by universities, research institutions and other trans-European agents.”
3. On rewards and incentives, the Declaration notes these are essential for researchers who make research data open and FAIR for reuse, and/or actively reuse and reproduce data. As well as in project and career evaluation, these rewards should be reflected in “…other career policies in universities and research institutions (appointments, promotions etc.).”

2.1.2. Aims, scope and sources for the framework

The EOSCpilot Skills Framework aims to offer a reference model for planning professional development in stewardship. The framework allows individuals to identify their needs for specific skills development resources, and helps organisations to formulate requirements of skills to reflect new capabilities they wish to develop.

We use *stewardship* to refer to research output management in the context of open science and data science. In D7.1 we defined stewardship more precisely as the “roles and responsibilities to ensure that data is managed for long-term reuse [...] in accordance with FAIR data principles” [2, p. 39]. In light of stakeholder comments, we revise this definition to apply not just to data but research objects in general.

This reflects the pivotal role of code, workflows, and other outputs in addition to data.

The initial work on the framework in D7.1 proposed a set of competences for stewardship by synthesising elements of other competence frameworks: the EDISON project (data science), RDA Education & Training in Data Handling Interest Group (data management) and a data information literacy framework (data librarianship). Accordingly, the revised framework inherits these initial sources.

2.1.3. Main elements of the first draft

The EOSCpilot Competence Framework links what is essentially a list of topics to a minimal set of parameters for planning training:

1. Skills groups that represent the main skills, competences, and capabilities of interest, i.e. plan and design, capture and process, integrate and analyse, appraise and preserve, publish and release, expose and discover, govern and assess, scope and resource, or advise and enable,
2. Professional groups[1] that apply the competences, i.e. domain researchers, data scientists/analysts, data managers, and data science engineers,
3. Competence levels required by professional groups for a given skillset, i.e. comprehend, apply, or evaluate and synthesise,
4. Organisation levels at which the competences and capabilities are applied, i.e. individual researchers, or research managers at the level of a research team, or organisation-level service,
5. Responsibility levels required to deliver a capability, from support through to full accountability.

The framework also recognises that skills development planning should take account of variations according to the domains concerned, the local organisational culture, and the required capabilities.

The EOSCpilot data stewardship competences are organised into nine skills groups as illustrated in Figure 1 below. The skills groups are of two types: those applied across research projects and those applied differently in each research project. The skills groups applied across projects are shown in the outer circle of Figure 2.2, and include govern and assess, scope and resource, and advise and enable. The figure shows in its inner circle the 6 skills groups that follow a project-based lifecycle for managing research objects, i.e. data, software or other sources for research outputs. These skills groups are: plan and design, capture and process, integrate and analyse, appraise and preserve, publish and release, and expose and discover.

Figure 1 - Cross-project and project-specific groups of skills

[1] These were termed *roles* in D7.1 and are renamed here to avoid confusion with the use of this term in the EOSCpilot service architecture.

2.1.4. Consultation on the competence gaps

Annex A.1 of this report consolidates the mapping of skills and competences into a list of 78 topics, removing duplicates from those listed in [2, p. Annex A]. We invited feedback on the consolidated competence list in a breakout session at the EOSCpilot-OpenAIRE joint workshop in Berlin (24 October 2017) attended by representatives of academic libraries, institutions, and e-infrastructures. We asked the 21 participants in the session to identify their top 5 gaps in data stewardship from our skills group draft and at what scope the competences need to be provided, i.e. from an individual, a team, or a broader organisational responsibility.

Table 2.1 shows that the 105 posted gaps were seen to be mostly in the three cross-project skills groups and at an organizational level. Of the 78 competences in the draft list provided, 16 were revised by the participants, and 4 new ones added (software curation, FAIR and Open Access policy, preservation planning, and storage of large data volumes). The revised competence list is shown in Annex A.1.

Table 1 - Stewardship skills gaps – workshop responses

Skills group	Competences				Posts per organisational level			Total
	(a) proposed	(b) 1+ posts	(c) new	(d) revised	(e) individual	(f) team	(g) organisation	(h)
Plan and design	12	5	0	3	5	4	1	10
Capture and process	10	3	0	1	1	4	1	6
Integrate and analyse	12	7	0	1	3	3	2	8
Appraise and preserve	6	4	1	1	0	3	2	5
Publish and release	4	4	0	1	2	3	2	7
Expose and discover	7	3	0	0	5	0	1	6
Govern and assess	10	9	1	4	5	8	15	28
Scope and resource	10	8	2	2	2	3	10	15
Advise and enable	7	5	0	3	2	2	16	20
Total	78	48	4	16	25	30	50	105

The Table 1 shows in column (a) the number of competences proposed in the draft list provided to participants for reference, and in (b) the number of these that at least one person identified as a gap. Column (c) shows how many new suggestions were made, and (d) the number receiving suggestions of changes. Columns (e) to (g) show the number of times competences in each group were identified as gaps and at what organisational level these gaps were perceived to be at.

What is noticeable from these responses is that slightly more than half of the competences (48 of 82) were

identified as a priority by more than one person. Also, most of the posts were in the cross-project areas *govern and assess*, *scope and resource*, and *advise and enable*. The responses are given in the Annex A.2. There were 9 competences that 4 or more people indicated as gaps, accounting for almost half of those perceived as gaps. These are shown in Table 2.

Table 2 - Frequently identified gaps in available competences

Skills group	Competence	Posts
Govern and assess	Data policy, legal and funder requirements	9
	Research strategy/ open research potential	7
	FAIR and Open Access policy	5
	Research reproducibility	4
Advise and enable	Tools and domain standards awareness	7
	Personnel and skills development	5
	User support and training	4
Plan and design	Metadata, persistent id. specification	4
Scope and resource	Costing of data management and preservation	4
Total		49

Further consultation will continue until M18, drawing on the parallel work on D5.1 (Service Architecture) and D5.2 (Service Portfolio). This will enable the anticipated service capabilities to be mapped to the competence list.

2.2. Skills framework model update

The capability-competence mappings in the EOSCpilot Skills Framework will be expressed in two forms, to meet the two main use cases:

- Skills development planning: For a research team or organisation needing to plan skills development at a broad level, e.g. for a project or program to deploy services, the relevant capabilities are expressed for each organizational level. The wording can then be adapted to suit the level of responsibility appropriate to the context (as proposed in [2, p. Table 6.3]).
- Training planning: For an individual professional group, specific skills needs may be set out in a more granular user story format that also identifies the competence level needed for the relevant service role. According to the EOSCpilot Service Architecture D5.1, these roles are service developer, service provider, and service user.

EOSCpilot WP7 focuses on the skills requirements of various professional groups that share responsibility for stewardship. These are derived from the EDISON competence framework: domain research, data science/analytics, data management, and data science engineering. In EOSCpilot, data stewardship is conceived as the application of competences from these groups, towards managing research outputs for long-term reuse and in accordance with the FAIR principles.

In Annex A.1, we propose which professional group needs which competences. To emphasise the shared responsibility, we identify each of the competences associated with more than one professional group. The framework is also intended to be flexible enough to accommodate variation in the organisational and disciplinary context. Stewardship is likely to be partly embedded in what researchers do, as part of the

research. In some contexts, there may be a data manager or steward in the research team, but in most contexts, this will be the shared responsibility of research administrators and librarians, and IT support professionals.

2.2.1. Conceptual model and examples

The terms competence, skill, and capability need to be well-defined to support their application to EOSC services defined in EOSCpilot WP5. The working definitions are below, and shown in Figure 2.

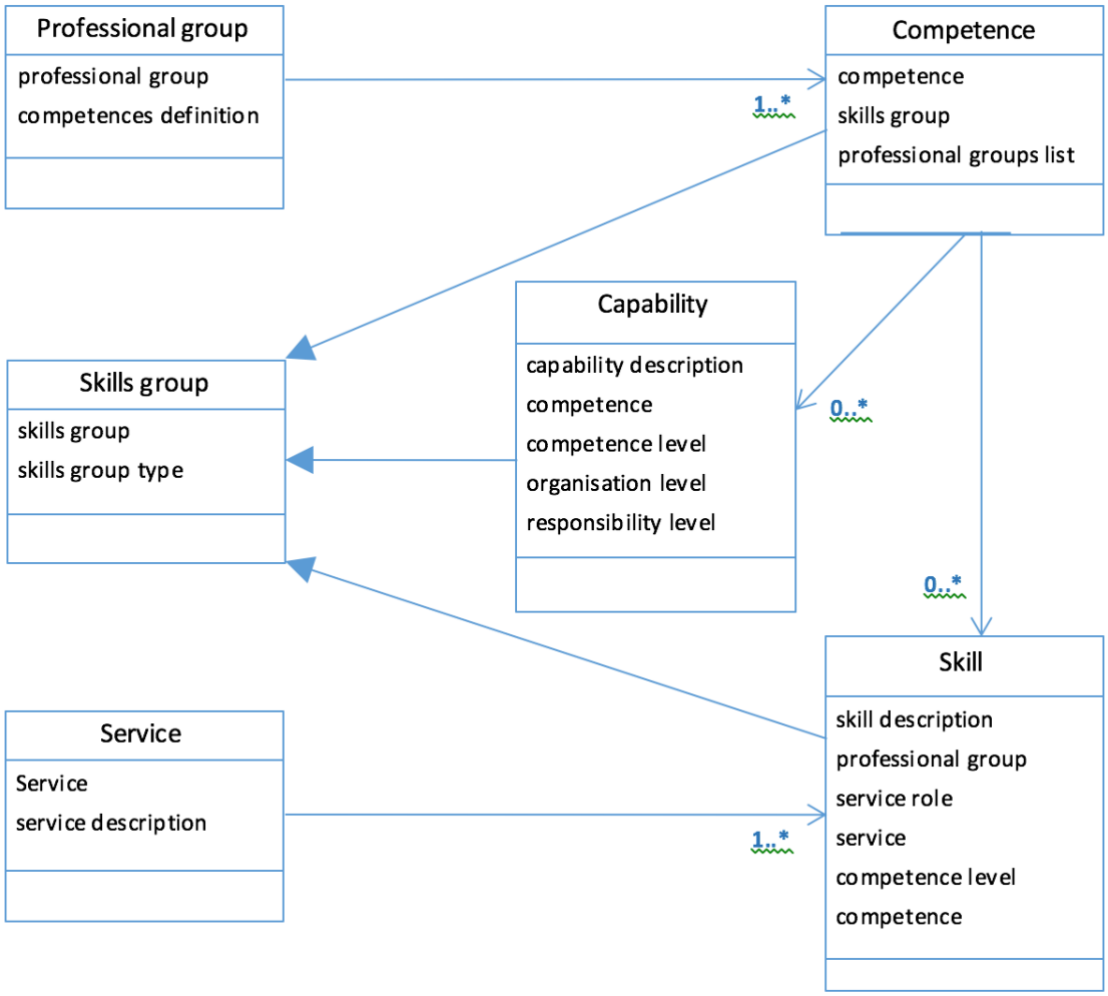

Figure 2 - Conceptual model for EOSCpilot Skills Framework

Competence: an element of stewardship theory or practice e.g. workflow set-up and management.

Competence levels: the capacity to comprehend, apply, or synthesise and evaluate the knowledge of a given competence.

Professional group: a person's domain of responsibility defined by a set of competencies, i.e. *domain research*, *data science/ analytics*, *data management*, and *data science engineering* (derived from [3]).

Skills group: a group of *skills*, *competencies*, or *capabilities*. Skills groups are of two types: those realised specific for individual (research) projects e.g. capture and process, and those applicable across (research) projects, e.g. govern and assess.

Skill: the level of *competence* that a *professional group* needs to perform a *service role* for a *service* e.g. "as a data manager, I need to apply repository evaluation and selection to use a repository service". As this example suggests, skills are expressed in the form of a user story for a training service.

Capability: the *competence level* and *responsibility level* needed to perform a *service role* at a particular

organizational level, in a form relevant to planning skills development for individuals, across a research team or organisation.

Responsibility levels: support, discretion, substantial discretion, objective setting, and full accountability.

Service: a service described in the EOSC service portfolio, or service architecture.

Service roles: service development, service provision, or service use envisaged by the EOSCpilot Service Architecture.

The nature of the services mentioned in the definitions above will be described in the EOSCpilot Service Portfolio, and the capabilities they offer will be inferred from service descriptions.

2.2.2. Mapping service capabilities to competences

Standards in two areas relevant to data repositories, and service management, have meanwhile offered a basis to illustrate the approach of mapping service capabilities to competences. Respectively the *Core Trust Seal* for repository capabilities, and the *FitSM* standard for service management enable relevant examples to be highlighted.

For example, AAI service management needs are identified in the FitSM standard as follows: "Access control, including provisioning of access rights, for information-processing systems and services shall be carried out in a consistent manner" [4]. This can be mapped to the EOSCpilot Skills Framework as follows, indicating competences needed by service-users, depending on whether the capability needs to be delivered by the individual, research team or elsewhere in the organisation:

- Individual: A user must be able to support the application of a single EOSC identity in a consistent manner, to access and use any provided e-infrastructure, service, tool and/or storage,
- Team: A user must have substantial discretion to authorise individual users to access and use any provided e-infrastructure, service, tool and/or storage in a consistent manner,
- Organization: The organization must take full accountability for ensuring that access control, including provisioning of access rights, for information-processing systems and services shall be carried out in a consistent manner.

User stories can be identified for each of the three competence levels. We will offer examples relevant to services in the EOSCpilot architecture, and indicate the competences and the level of these that are expected of relevant professional groups. For example:

- As a data science engineer, I need to be able to comprehend AAI technologies for access management, to use a recognised EOSC AAI service in a consistent manner.
- As a data manager, I need to be able to evaluate and synthesise policy to use AAI in my organisation.
- As a researcher, I need to be able to authorise individual users to access and use any provided e-infrastructure, service, tool and/or storage in a consistent manner.

The general form of the examples provided above follows the user story format proposed earlier (in the definition of *skill*), i.e.

"As a [professional group], I need to [competence level] [competence] to [use/develop/provide] [service]."

Using these statements or the competences themselves, individuals, research groups and organisations should be able to plan the skills development needed by finding trainings suited to the level of responsibility that is relevant to their context.

2.3. Skills requirements update

The first skills requirements reported in D7.1 were based on the experiences of Science Demonstrators after the first months within EOSCpilot. This section extends the analysis based on the currently available information from Science Demonstrators and most recent developments within EOSCpilot and EOSC-hub,

in particular the (service) architecture description.

2.3.1. Requirement analysis based on Science Demonstrators

As part of D7.1 we provided a first assessment on the skills requirements for *Science Demonstrators*, namely Digital Preservation in High Energy Physics (DPHEP), Pan-Cancer Analyses, Photon-Neutron, and TEXTCROWD. Each of these Science Demonstrators serves as an exemplary use case for integration of infrastructure, service management and usability for researchers and scientists. The assessment is based on a questionnaire, desk research on each of the demonstrators sub-project's planning documents, and, if applicable, discussions with contacts in the demonstrator sub-projects.

The assessment shows that the needs reported so far are broader than just data management aspects of stewardship, also reaching into data processing, analysis and related topics. Requirements mainly focus on engineering and deploying services in the cloud, especially the integration of workflows in the EOSC environment. These skills requirements are not only relevant to service developers and operators: end-users are faced with the side-effects of working in a multi-national, distributed cloud infrastructure [2, p. 31]. Although several of the current demonstrators expose low-level details about EOSC services to their potential end-users, this is expected to change as EOSC matures: The EOSC services will introduce better separation of user responsibilities and will expose different levels of abstractions to the different user groups. Developer users, such as the current demonstrator owners, interact with EOSC at a lower technical level than end-users.

To elaborate on the initial assessment of skills requirements in D7.1 and analyse training implications we focus in the following on the review of Science Demonstrators reporting documents. At the current stage, this analysis does not include user experience and feedback for the given Science Demonstrators. The Science Demonstrators are still in a progression from testing to a production use of EOSCpilot. Therefore, our analysis builds on the Science Demonstrator reports of WP4 that provide feedback at the level of service providers and engineers. Our analysis considers required skills derived from the challenges reported by the Science Demonstrators since D7.1.

The analysis not only covers the first phase of Science Demonstrators, but also includes a first analysis on the second phase of Science Demonstrators, including Virtual Earthquake and Computational Earth Science e-science environment in Europe (EPOS/VERCE), Linking distributed data and data analysis resources as workflows in Structural Biology with cryo Electron Microscopy: Interoperability and reuse (CryoEM workflows), Leveraging EOSC to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability (EGA datasets), LOFAR EOSC Pilot and PROMINENCE.

2.3.2. Liaison with Science Demonstrators on required skills

The review of challenges reported by Science Demonstrators (see Appendix B for details) from the first phase does not reveal immediate overlaps or commonalities. At the same time, it substantially confirms the assessment of skills gaps carried out in D7.1 on the basis of questionnaires. The summary in D7.1 concludes:

> "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." [2, p. 30]

Since the conclusion of D7.1, Pan-Cancer reported issues with user identity management when data is transferred to new storage location. The most extensive revision is with regard to DPHEP: At the time D7.1 was published, the demonstrator did not anticipate any skills gaps, whilst it has recently reported a number of issues both technical and policy related. From the available reports of DPHEP, nothing specific can be inferred about policy issues. On a technical level, analysis of those reports points to the need to re-scope

data used for demonstration. The goal is to provide a variety of data, presenting a more realistic use case than initially intended and focus on challenges of handling data itself, i.e. handling of data types and metadata. The contrast in both scope and implementation of EOSC compared to classical infrastructures suggests that many Science Demonstrators will encounter further skills gaps as their own workflow scope expands.

Besides the analysis of challenges faced by Science Demonstrators from the first phase we also provide results based on desk research on the Science Demonstrators of the second phase. The results are gathered and condensed below for each of the five demonstrators sub-projects.

CryoEM workflows

The Science Demonstrator CryoEM workflows aims at standardizing processing workflows in structural biology, to ensure provenance at the level of data and analysis tools, link workflow information with raw data and allow for reproducibility as well as interoperability with distributed data and analysis sources. The project requires specific competences to:

- *evaluate* and *select* a standard description language/format for workflow description,
- *engineer*, *integrate* and *deploy* adapted image processing workflow software for both local and cloud-based execution and re-execution,
- *evaluate* and *select* information to ensure reproducibility of workflows, and
- *engineer*, *integrate* and *deploy* a service to browse and analyse the workflow files.

EPOS/VERCE

The demonstrator EPOS/VERCE aims at expanding the pioneered Virtual Research Environment from computational Earth Science to the computing and visualisation of realistic scenarios while focusing on reusable workflows, deployment templates for data-reduction and provenance aspects. For extension to EOSC the project requires capabilities to:

- *evaluate* and *integrate* cloud infrastructure for simulations,
- *evaluate* and *integrate* data services by EUDAT for storing datasets,
- *engineer* and *integrate* authentication and authorisation services,
- *engineer* the provenance model and associated reproducibility and validation services,
- *integrate* and *deploy* virtual machines for simulation and processing, and
- *adapt* and *deploy* workflows for utilisation of cloud resources.

EGA datasets

The project aims at utilising EOSC for computing up-to-date versions of EGA datasets while increasing the homology and interoperability between data for crossed analysis. The project fosters reproducibility, reuse and FAIRness of data by introducing standardized workflows and container technologies in the biological sciences. The project requires specific competences for deployment:

- *evaluate*, *select* and *integrate* data storage that complies with security and privacy constraints,
- *evaluate*, *engineer* and *deploy* identity and management of sensible data in the cloud,
- *integrate* and *deploy* virtual machines for running automation and data management tasks,
- *evaluate*, *select* and *integrate* interoperability guidelines for FAIR (meta)data,
- *engineer* and *deploy* standardized workflow for data refreshing, and
- *analyze* reproducibility and portability of workflows.

LOFAR

The Science Demonstrator LOFAR aims at providing a service to locate, access, extract and process data from the distributed LOFAR data archive in the Physical Sciences to enable new scientific results based on the large-scale compute resources offered by EOSC. The project requires the following competencies:

- *evaluate*, *select*, *engineer* and *deploy* access to LOFAR data in accordance with FAIR principles,
- evaluate, engineer and deploy workflows based on Common Workflow Language,

- *integrate* and *deploy* container based processing of LOFAR data in the EOSC, and
- *evaluate*, *engineer* and *deploy* performance profiling.

PROMINENCE

The project aims at piloting an HPC infrastructure in the EOSC to make HPC machines and MPI/OpenMP jobs available as a cloud like service within the domains of Energy and Plasma Physics including Materials Engineering. These aims require capabilities for deployment to:

- *integrate* and *set-up* of OpenStack,
- *integrate* and *deploy* virtual machines and ensure access to OpenStack resources,
- *evaluate*, *engineer*, and *deploy* cloud scheduling across different cloud domains,
- *evaluate*, *select*, *engineer*, and *deploy* methods for support of AAI,
- *evaluate* methods for co-scheduling of batch system and cloud, and
- *evaluate* setup based on MPI workflows.

2.3.3. Requirement analysis based on EOSC architecture

Skills requirements will be derived from the EOSC Architecture as further details of this become available. As mentioned earlier in Section 2.2, the roles and activities defined in the Architecture, for EOSC service developers and service clients will be the raw material for WP7 to identify the capabilities that services offer to client organisations and professional groups, and the relevant skills that need to be developed.

Given the strong relationship between EOSCpilot and EOSC-hub, we can also consider skills requirements for EOSC service providers based on the architectural considerations made in the EOSC-hub H2020 project. EOSC-hub will start in January 2018, with EGI Foundation as coordinator.

The EOSC-hub mission is to contribute to the EOSC implementation by enabling seamless and open access to a system of research data and services provided across several nations and multiple disciplines. The project will offer these resources via the Hub – an integration and management system of the European Open Science Cloud, acting as a European-level entry point for all stakeholders. The Hub will deliver a catalogue of services, software and data from the EGI Federation, from EUDAT CDI, from INDIGO-DataCloud and from over 10 major Research Infrastructures. The Hub builds on mature processes, policies and tools from the leading European federated e-Infrastructures to cover the whole lifecycle of services, from planning to delivery.

The EOSC-hub project adopts a Service Integration and Management approach to managing suppliers and integrating them, to provide a business-facing EOSC Hub. The approach aims at integrating independent services from various internal and external service providers into end-to-end services. The project will federate four types of services (see Figure 3 below):

1. Common services covering baseline services on compute and storage as well as specialised services for compute, data, software management, and curation and preservation.
2. Thematic services covering research data, advanced data brokering and analysis capabilities for specific research communities and multidisciplinary research.
3. Collaborative services to provide tools for open science platforms for sharing of research digital objects such as scientific applications, pipelines and virtual appliances.
4. Federation services to integrate and manage access via federated identities, AAI, and to support IT Service Management processes. These services enable seamless operations and management of their services in the EOSC.

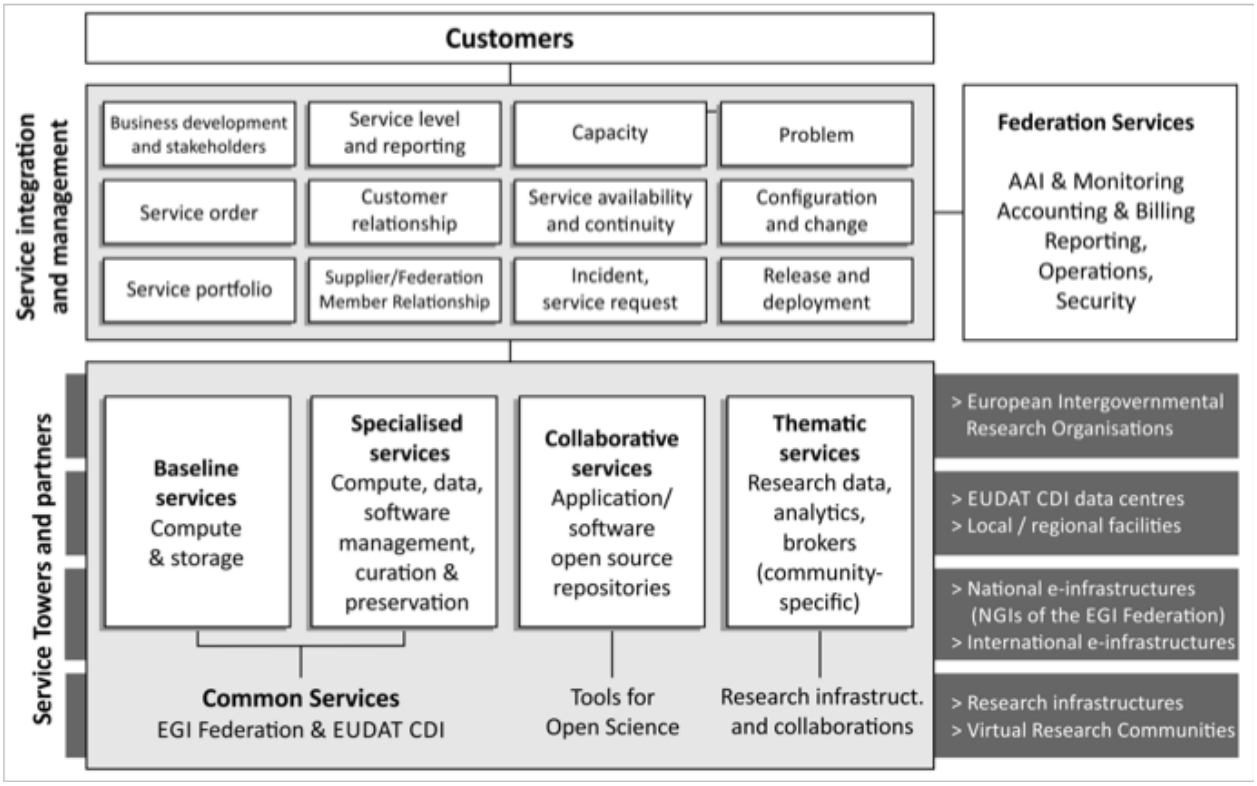

Figure 3 - Proposed service architecture for EOSC-hub [4]

We make an attempt to derive an initial set of skills requirements from this architecture. EOSC should offer skills development and training as follows

- For targeted service users: about the use of the *federated services*.
- For service providers: about service management practices, optionally aiming at reaching *certified level*.
- For service developers and providers: on *using baseline compute and storage services* for applications and servers that demand big-data and big-compute.
- For all groups: on *data management*, such as FAIR data principles, data curation, and preservation.
- For service developers: on *collaborative services*, primarily targeting software architects and developers within scientific communities.

The EOSC-hub project includes a work package on training with focus on the following topics:

- Data management planning (in collaboration with the OpenAIRE-Advance H2020 project),
- Federated Service Management Training and Certification,
- Common and federated services, and
- Training about thematic services.

Skill development activities are ideally developed and provided in collaboration of EOSCpilot and EOSC-hub. A key consideration should be on making this activity sustainable.

2.4. Summary and conclusions

In this section, we first summarise the main motivations and elements of the EOSCpilot Competence Framework introduced with D7.1 and provide information on our progress towards consolidation of data stewardship competences from different sources of competence frameworks into a list of 78 competences. We further present a revised competence list based on the feedback on competence gaps of participants from a joint EOSCpilot-OpenAIRE workshop in Berlin. Consultation about our current list of competences

will continue and we will report on the outcome in D7.3 at M18, drawing on the parallel work on D5.1 (service architecture) and D5.2 (service portfolio), which will enable the anticipated service capabilities to be mapped to the competence list.

After presenting the current progress about the EOSCpilot Competence Framework we offer an update on the EOSCpilot Skills Framework. For this, we provide a conceptual model that defines the terms competence, competence level, professional group, skills group, skill, capability, responsibility level, service and service role. We further establish the multiplicity and the connections between the individual terms. This model builds the framework to present the mapping of service capabilities to the competences. We illustrate this mapping using two relevant standards for the services of the EOSC service portfolio as an example.

The mapping of service capabilities to competences enables a systematic approach to skills gap analysis. As a basis for this, we conclude the section with an update to the EOSC skills requirements by taking examples from three main sources, namely workshop outcomes, science demonstrator reports, and EOSC service architecture.

Based on the analysis and evaluation on

- consultation with stakeholders on competence gaps at varying organizational levels based on the EOSCpilot Competence Framework,
- experience of skill requirements of Science Demonstrators from first and second phase, and
- analysis of service architecture and derived skills for EOSC-hub

we can conclude about a set of skill requirements for cataloguing of training materials.

The consultation with stakeholders suggests main skills gaps for cross-project skills groups including govern and assess, advise and enable, and scope and resource at an organisational level.

Experience of the Science Demonstrators calls for attention in

- integration, management and effective use of large-scale cloud infrastructure environments, in relation to data analysis, software algorithms as well as container technologies or virtual machines and
- handling of large data volumes and management of community requirements in terms of standardised workflow languages, data types and metadata.

This confirms skills gaps identified in D7.1.

Furthermore, the analysis of Science Demonstrators suggests that training needs to raise awareness for topics including reproducibility and provenance as well as identity management and security. We consider developing skills and awareness for FAIR data a key competence that is required in all the sources examined. EOSC should therefore focus competences required to apply FAIR principles to services, and to ensure that service-users can apply FAIR principles in doing research and reusing its outputs.

Finally, the analysis of service architecture for EOSC-hub confirms that EOSC should offer skills about collaborative services and data management including FAIR data principles, data curation, and preservation.

The analysis also suggests that it is important that training materials are reusable in order to benefit from existing expertise and materials. This enables creation of materials according to the target professional group, competence level, responsibility level and service.

In the next section, we contribute further to the analysis of gaps in training, supplying a preliminary comparison of available training materials based on the EOSCpilot Competence Framework.

3. CATALOGUE OF TRAINING MATERIALS AND EVENTS

The skills requirements derived in Section 2.3 outline the scope and extent of the catalogue of training materials and events for the EOSC. In the scope of D7.2, our cataloguing of training materials and events serves three distinct aims:

1. derive recommendations about the scope and structure of a training materials and events catalogue for EOSC,
2. identify training materials and events likely relevant to the EOSCpilot service and demonstrator users, and
3. derive a preliminary training materials gap analysis of available training materials and events with respect to the current skills requirements.

In the following, we approach these aims by identifying potential use cases for a catalogue and the extent to which available materials fulfil them. We further analyse general availability of training materials currently provided by partners and by the H2020 *cluster projects* of Research Infrastructures with respect to the EOSCpilot Competence Framework.

In addition, we provide a detailed analysis based on a sample of training materials, which we consider representative to identify gaps in the coverage of current training materials, using the EOSCpilot Competence Framework as a baseline. This sample is drawn from a number of recent Summer Schools offering intensive training in open science, data management and/or data science.

The analysis of gaps in currently available training materials and events is an ongoing task (T7.2), and will involve further updates to the catalogue to better address the scope of the EOSC Service Architecture and Portfolio. Further materials will be added to meet the needs identified from validation of the services and Science Demonstrators with their users, in WP4, WP5, and WP6.

3.1. Broadening the analysis of skills development provision

Our analysis of skills development provision uses the EOSCpilot Competence Framework to identify the extent to which selected organisations already provide training and other forms of skills development that are relevant to the WP7 focus, such as workshops and online guidance materials. The analysis is ongoing, and will inform the selection of training materials for a preliminary skills gap analysis in the context of this report as well as a survey of Research Infrastructures in the second year of the project.

A preliminary analysis of skills development provision in D7.1 offered a descriptive overview of the kinds of skills development provision offered by institutions, infrastructures and relevant projects. In this follow-up, we conduct further analysis of the 33 partners in EOSCpilot, plus the 8 Horizon 2020 cluster projects (see Annex C). These selected organisations may be considered the first *building blocks* of EOSC. Thus, the analysis aims at providing further insight into the extent of training material and skills coverage by EOSC related organisations.

We conducted the analysis through desk research. The criteria for selection of relevant training and training materials are:

- more than one professional group targeted, i.e. at least two of domain research, data science/ analytics, data management, or data science engineering,
- training events or materials readily identifiable – e.g. from news items identifying participation in external events, conferences or projects (e.g. conference tutorials are possible source of advanced training),
- materials explicitly about developing skills in data stewardship, open science and data science – whose content is within the EOSCpilot Skills Framework scope, i.e. topics from the main competence areas,
- materials either dated since 2015, or presented as currently relevant, and
- events provided within the last 12 months, or planned events.

The aspects of training/skills activity analysed included the research domains targeted, target audience professional groups, skill/competence areas from the EOSCpilot Competence Framework, the levels of competence covered, delivery formats and modes used, and the duration. The analysis also considers the availability of training materials.

Validation of the analysis is ongoing, but our preliminary results[2] show that:

1. The most covered research domains are Physical Sciences and Engineering, and Biological and Medical Sciences (32% and 29% respectively), and least covered are Material Sciences and Analytic Facilities, and Energy (9% and 15% respectively). Cross-domain topics are covered by 53% of organisations.
2. 91% of organisations target researchers, about 62% target research data/software engineers, 50% target research data scientists, and 44% target research data managers.
3. The topics least covered are those listed under the *expose and discover* category by 32% and the topics most covered are those listed under *capture and process* and *advise and enable* by 62% of organisations respectively.
4. Fewer organisations (50%) are offering advanced-level than basic (79%) or intermediate (76%) level of training.
5. Face-to-face courses and conferences are offered by a large majority of organisations of 91%, and around 50% are offering online provision.
6. The majority of trainings last at most 5 days (<1 day 53%, 1-2 days 71%, 3-5 days 47%) while fewer trainings are of longer formats (>1 week and <1 month 12%, >1 month 24%).

Results are available in Annex C and the underlying data are available [4].

These preliminary results offer an overview of the extent to which the EOSCpilot partners have information available about training that is relevant to the research domains they serve. Cross-domain topics are commonly available, enabling domain-agnostic skills training. All areas of the EOSCpilot Competence Framework appear to be covered, although the *expose and discover* skills groups are underrepresented relative to others.

We will further develop the analysis to inform the EOSC skills and education strategy (D7.5), by broadening the current scope to include the Research Infrastructures and e-Infrastructures represented in EOSC-hub. In addition, we will invite the partner organisations in both EOSCpilot and EOSC-hub to self-assess the availability of information about training, and materials. This will offer a basis for comparison with previous studies, e.g. by Knowledge Exchange and Belmont Forum.

The further analysis will include an assessment of capacity needs, based on available information about the numbers of researchers in each research domain, and the views of selected Research Infrastructures on the demand for skills development on the topics identified in this report as gaps.

3.2. Cataloguing training materials: towards FAIRness of training

The success of EOSC depends on the ability of its users to work with a large-scale, dynamic infrastructure that spans multiple scientific domains. For the users this presents a challenge for finding and using relevant information that will help them acquire the skills to effectively use EOSC services, enable data science and apply FAIR principles to its outputs.

Considering the potential benefits to EOSC skills development of having easily findable, accessible and reusable training materials, there is potential to adapt the FAIR data principles to derive guiding principles for the FAIRness of training resources, such as materials and events. The motivation is similar to that for

[2] We are currently validating the results with the organisations concerned. For 8% of the analysed organisations we have not been able to identify relevant information on skills development provision. These organisations are therefore excluded from the given analysis until the preliminary information have been validated with the respective organisations.

FAIRness of data: Training resources have potential value as a public good, contributing to knowledge exchange and innovation.

We already formulated this statement that EOSC training materials and events must themselves be FAIR in our first conclusion in the previous report D7.1:

- Materials and event descriptions must be provided with standard metadata to ensure *findability*,
- they must be directly *accessible* from EOSC,
- they conform to standard formats to be *interoperable* with each other and related services, and
- they must be licensed as non-restrictive as possible to encourage *reuse*. [2]

In this report, our aim is to define a Training as a Service infrastructure to expose FAIR training resources and make them discoverable. It is not our aim to define the FAIRness of training resources; a complex topic that is beyond the scope of this report and an important issue that requires further discussion within the EOSC training community, including the EOSC-hub project, and Go-FAIR initiative [5]. Within this report we are primarily concerned with the *findability* of training resources to offer a basis for further discussions. Further discussion on how to best make resources findable and accessible in EOSC are available in the third part of the report.

The needs for training resources to be *interoperable* and *reusable* are less clear. Training materials may become quickly outdated as the service and policy environment changes quickly. This indicates a need for further discussion of what kinds of material the EOSC skills development community would benefit most from having available in reusable form. For example, do case studies have a longer lifespan than presentation slides? What effort is saved by adapting materials across domains, and how may that be expressed in criteria for interoperability?

We propose to discuss these questions further with the community of training providers in Research Infrastructures and other EOSC stakeholders, especially academic institutions. FAIR guidelines for training resources may be seen as an adaptation to the EOSC context of prior work in the field of Open Educational Resources, and could potentially offer criteria for the governance of Training as a Service in EOSC.

3.2.1. Selection of materials for cataloguing

The selection of materials is based on prior decisions about the scope of the EOSCpilot Competence Framework, the skills requirements indicated in Section 2.3, and practical considerations of the resources available for cataloguing (see Section 3.1).

We define *training materials* pragmatically as any online guidance or training resource made available for independent use in a skills development context, at the highest level of granularity that material is presented in on the source site, such as a module or handbook. For the purposes of cataloguing we treat materials in a similar way to the ELIXIR TeSS system – “a link to a single online training material sourced by a content provider (such as a text on a Web page, presentation, video, etc.) along with description and other metadata (e.g. ontological categorization, keywords, etc.).

Given the enormous range of online resources *potentially* relevant to the EOSCpilot Competence Framework, we sampled a total of 90 materials from events and online resources whose scope and objectives were clearly relevant to data stewardship training in open science and data science context. In particular we focused on recent Summer Schools offering a comprehensive foundation, with some topics treated at advanced level. The main sources were:

- CERN School of Computing,
- CODATA-RDA Schools of Research Data Science,
- EUDAT Summer School,
- GridKa School,
- BSC PUMP Summer School,

- INFN International School on Architectures, Tools and Methodologies for developing efficient large scale scientific computing applications,
- PRACE Summer of HPC,
- UNC/UEDIN Research Data Management and Sharing MOOC,
- Karolinska Institute - postgraduate course in “Open Science and Reproducible Research”, spring 2017,
- Research Data Netherlands “Essentials 4 data” online course,
- RITrain project,
- Software Carpentry,
- Data Carpentry, and
- OpenAire webinar- “Open Research data in H2020”.

To exclude outdated material, we selected only recent contributions from the years 2016 and 2017. To avoid bias from overly specific trainings, all selected materials covered at least one of the skills groups in the EOSCpilot Competence Framework.

3.2.2. User requirements for FAIRness of training resources

The data stewardship skills development catalogue summarises training efforts of different formats by training institutes and infrastructures. The FAIRness of training materials and events affects various EOSC stakeholders. In the following, we consider and evaluate these stakeholders’ interests in FAIRness.

- *EOSC users* who seek to efficiently utilize the services provided by EOSC for their research and want to learn about and profit from relevant training resources,
- *Organizations* that seek to plan, organize and/or propose training materials and events for their employees which are EOSC users,
- *Trainers* who seek to provide relevant and up-to-date training materials and events for EOSC users,
- *Curators* of training catalogues who seek to provide a comprehensive collection of training materials and events, and
- *Services* that need to access provided training resources for a given purpose, such as checking availability of materials.

Each use case required by the four different target groups relies on various aspects of FAIRness. In the following, we derive the most important use cases for the given target groups based on interviews with researchers from different domains. A comprehensive list of use cases can be found in the Appendix (see Annex D).

The EOSC user seeks to:

- find relevant training materials based on its title;
- find relevant training materials based on free text search;
- find relevant training materials based on the EOSCpilot Skills Framework.

In contrast to an individual EOSC user, the organization or team seeks to:

- identify criteria for relevant training events based on the EOSCpilot Skills Framework;
- orchestrate relevant training materials to organize and deliver a training event;
- commission a training event by requesting specific training materials, trainers and/or venues.

The trainer seeks to:

- find relevant training materials for reuse with a specific license;
- deliver training materials based on demands by EOSC users, teams or organisations;
- be informed when referenced/used training materials are updated.

The training catalogue curator, including that of any EOSC catalogue, seeks to:

- identify outdated training materials;
- identify training materials that are unavailable or not relevant in the context of EOSC;
- integrate external trainings materials and events.

The service seeks to:

- collect and aggregate metadata on training materials and events;
- expose descriptions/annotations on training materials and events;
- get information about (external) changes of training materials and events.

In the course of this document we focus on the EOSC user as the most relevant target group. Users are our focus to derive guiding principles and recommendations for FAIRness of training materials, a FAIR catalogue of training materials and events and finally our proposal, enabling a FAIR provisioning of training within EOSC.

3.2.3. Minimum metadata requirements for training materials and events in EOSC

The FAIR data principles provide a template to enable FAIRness also for training resources. One of the principles of FAIR is the description of data objects with rich metadata. Each of the different types of training resources requires a set of specific metadata due to their scope. For example, a training event is constrained in its availability by time and venue availability, whereas a pre-recorded online course is not.

Many training providers do not provide structured information on training materials and events. However, many services such as integration services rely on structured training information to discover, filter and process available training materials. For example, the EOSCpilot Training as a Service infrastructure (TaaS) that we initially proposed in D7.1 relies on structured data to be populated. Several existing APIs of training providers do not follow a well-defined, open standard and are potentially subject to changes. This may may break integration workflows at any time without advance notice. There are annotation standards which enable the scraping of information from websites. Currently, such annotations are mainly based on bioschemas.org as a specific derivation of the schema proposed by schema.org. For example, TeSS both processes and exposes annotations following the bioschemas.org specification.

The adaptation to all principles mentioned in FAIR principles requires a conscious effort by training providers. As such, it cannot be expected that all FAIR data principles will be supported from the beginning. Instead, we recommend to initially focus on the findability of training resources, mainly via the provisioning of relevant metadata. To further minimise the effort to extend existing training portals and catalogues, our metadata schema focuses on a minimum set of metadata. With this set of metadata, we support the most important requirements by the target groups including EOSC users, organizations, trainers as well as curators (see Appendix D.3 for a mapping of metadata to specific user requirements) for our TaaS service.

The derivation of a minimal set of metadata for training materials and events follows the approach of Bioschemas. The user requirements outlined in Section 3.2.2 which are relevant for the findability of training materials and events guide the selection of fields for appropriate metadata for EOSC. The fields considered as part of the metadata are identified via a crosswalk of selected training portals and catalogues (see [8]) taken from skills landscape analysis. For both, training materials and events, we individually analysed existing fields from the crosswalk to determine the most relevant metadata while trying to keep the overall count of metadata at a minimum.

For the recommendation of metadata to expose, we distinguish three levels:

- a required or *minimal set* of metadata to ensure suitability for the most vital use cases for our target groups,
- a desired or *recommended set* of metadata which enables a convenient service and discovery for users, and
- a unrestricted or *optional set* of metadata, which has the potential to improve user experience for specific communities and use cases.

We aim at providing a cross-domain specification of metadata for training resources. Furthermore, we

intentionally consider the integration of specific fields required for searching in the context of the EOSC Competence Framework. This ensures interoperability between the different frameworks and services contributed by WP7.

Table 3 - Minimal set of properties for training materials

Proposed property	Type	Multiplicity	Description
Target professional group	Audience	many	Any appropriate EOSCpilot Competence Framework professional group: Data Science/Analytics (DSA), Data Management (DM), Data Science Engnineering (DE), Domain Research (DR)
Author	Person, Organisation	many	The author of the training material
Date modified	Datetime, Date	one	Date/time of most recent change of training material excluding metadata
Topic	Text, URL	many	A general subject/category covered by the training material
Title	Text	one	The title of the training material

Table 4 - Recommended set of properties for training materials

Proposed property	Type	Multiplicity	Description
Keywords	Text	many	Keywords describing the training material
License	CreativeWork, URL	one	The license of the training material
PID	Text, URL, PropertyValue	one	Persistent identifier of training material, e.g. DOI
Competence level	Text	one	The competence level of the EOSCpilot Competence Framework: comprehend, apply, analyse and synthesize
Skills group	Text	many	One of the skills groups specified by the EOSCpilot Competence Framework: Plan and design, Capture and process, Integrate and analyse, Appraise and preserve, Publish and release, Expose and discover, Govern and assess, Scope and resource, Advise and enable
Domain	Text	many	A field of science or expertise

Both the minimum and recommended set of metadata for training resources focuses on information independent from a specific domain. However, we explicitly support and call for extension of the metadata by the different communities with regard to domain-specific information. This ensures that training providers can specifically support their community but do not need to modify any pre-existing data models, implementations or annotations.

In Section 4 on providing training as a service, we discuss different possibilities to enable findability within

EOSC with regards to metadata.

3.3. Skills gap analysis

The collection of training materials and events forms a preliminary catalogue of relevant materials. This catalogue lists information on the minimal set of properties required for FAIRness of trainings materials, as identified in the introduction to this section. The current catalogue is available online [6]. Based on this material collection, we derive a preliminary analysis of skills gaps for the current requirements using the EOSCpilot Competence Framework.

By tagging the materials with the terms from the EOSCpilot Competence Framework we can derive an estimate of how far the selected materials address the skills needed. Based on this, we can evaluate the skills and competences both covered and missing.

The skills gap analysis is based upon two methods to analyse present training materials and events: a quantitative and a qualitative method. This reflects both concrete results from mapping available materials to the EOSC Competence Framework, as well as a broader view reflecting our experience from collaboration with the Science Demonstrators.

The first method directly applies the metadata properties derived from the EOSCpilot Competence Framework (see Section 2.1.3), namely *Target Professional Group, Competence Level* and *Skills Group*, and any possible combination of their values. The combinations of these properties represent the skills landscape, as defined by the EOSCpilot Competence Framework and, therefore, identifies gaps with respect to it.

Each of the Science Demonstrators represents a distinct scientific use case, and as such puts different weight and emphasis on different competences. Thus, analysing just the coverage without regard to actual skills requirements does not provide an accurate view of critical skills gaps. To reflect this, we also perform a second analysis that provides a weighting with current skills requirements derived from Science Demonstrators (see Section 2.3.2).

3.3.1. FAIRness of training materials

In general, there is a lack of metadata required to ensure findability by users. More than one fourth of training materials does not provide a description, making it difficult for users to choose appropriate materials without looking at the actual content. Furthermore, the keywords provided to search materials are not expressive in a broader context: on average, three fourths of the keywords match only a single material. This makes most keywords useless to search for general topics.

In contrast, metadata for accessibility and reusability is covered better than we anticipated. Almost two thirds of training materials provide clear licensing information. This is critical to ensure their reusability beyond the initial authors and venues. Also, more than one quarter of materials is accessible via a persistent identifier. Given that persistent identifiers for training materials are a recent trend, and that summer schools are usually focused on face-to-face formats, a persistent access to materials is a welcome trend.

Overall, our analysis reveals that there is considerable effort required to ensure training materials are findable. As even critical information is often missing, a dedicated culture change may be needed by training authors, providers and consumers.

3.3.2. Skills and competence coverage

The heat map shown in Figure 4 indicates the extent to which the sample training materials cover each of the 9 skills groups, for each of the 3 competence levels, and for the 4 targeted professional groups.

According to this analysis, there is a lack of advanced materials at the *evaluate and synthesise* level for all skills and target professional groups. More specific gaps are in the following areas:

- Appraise and preserve – for all target groups,

- Publish and release – for all target groups,
- Expose and discover – for all target groups,
- Govern and assess – for the domain research, data science, and engineering target groups,
- Scope and resource – for all target groups, and
- Advise and enable – for the domain research, data science, and engineering target groups.

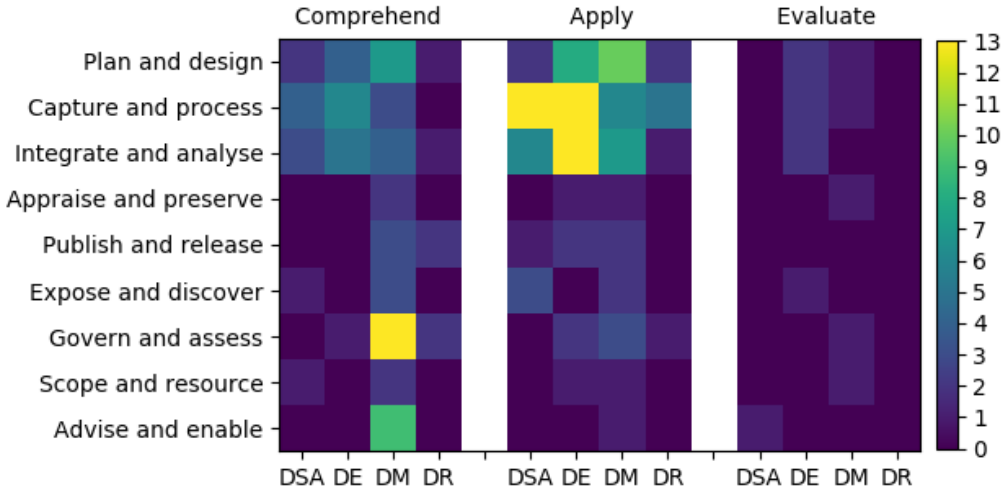

Figure 4 - Coverage of EOSCpilot Competence Framework by available materials

This analysis of a catalogue contents to show relative coverage is potentially useful to training providers, to help assess what they may offer in future. However, it has clear limitations for the purpose of the report:

- It is based on subjective analysis of available descriptions of courses.
- Analysis at the second level of the EOSCpilot Competence Framework would be more helpful to identifying specific gaps in current provision.

The first limitation could be addressed in future if training providers classified their own offerings, provided with a simple to use controlled vocabulary based on the EOSCpilot Competence Framework. For the more immediate needs of the gap analysis, a further analysis of materials was carried out by applying the second level of the EOSCpilot Competence Framework to training materials from the CODATA-RDA Research Data Science School.

3.3.3. Test case: CODATA-RDA Summer School

The CODATA-RDA Research Data Science School offers training in the form of residential workshops. The core school takes place for two weeks and intends to give early career researchers a grounding in data skills. The course offers seven modules and covers open science, research data management, author carpentry, software carpentry, machine learning, visualisation and computational infrastructure.

We selected the materials from this event as a test case for the applicability of the EOSCpilot Competence Framework for classifying materials. This test case follows the two main purposes:

- Identifying gaps in coverage of required skills with available training materials.
- Usefulness of gap analysis to guide the development of the specific course or curriculum.

The materials from the First CODATA-RDA Summer School that took place in Trieste in August 2016 were classified using selected terms from the EOSCpilot Competence Framework. The main difficulty in classifying the materials was a lack of description for the modules, making the classification subjective.

Each module was assessed regarding its competence level in the EOSCpilot Competence Framework. Three of the modules were assessed at the intermediate/application level: *Visualisation with R*, *Data Visualisation: Hands On*, and *Research Data Management*.

These three modules were then classified according to the Skills Group of the framework. In combination,

these modules were found to cover eight out of the nine groups:

- Plan and design: research design, metadata/persistent id specification, database design,
- Capture and process: data collection, data documentation, data processing, data cleaning,
- Integrate and analyse: theory building, data interpretation, data mining,
- Appraise and preserve: data review and appraisal, software curation and preservation,
- Publish and release: licensing, data marketing in EOSC, workflow documentation,
- Expose and discover: visualisation of research results, presentation of data collections,
- Govern and assess: research ethics, research reproducibility, FAIR/OA policy, and
- Scope and resource: storage provision and management, costing of data management and preservation.

The gaps in the course offering were found to be in the *advise and enable* Skills Group encompassing competences in:

- Building cross-sector collaborations,
- Communication and negotiation,
- Tools and standards awareness,
- User support and training,
- Personnel development,
- Standards body participation, and
- Data rescue.

The CODATA-RDA trainer's response to this analysis was that the EOSCpilot Competence Framework could be useful for training providers to tag their training in an EOSC catalogue. The gaps found in the analysis were considered to be valid, although the assessment of competence level was queried. So if the materials were tagged by a third party with the intention of offering a gap assessment, it would help to offer a matrix or table identifying what was rated, and the criteria used for assessing competence levels. This would also help others apply the criteria similarly. Analysis of this nature would also be more likely to be useful to providers in the early stages of curriculum development than to the CODATA-RDA Summer School, given that this has already coalesced around an agreed set of topics.

3.4. Summary and conclusions

In this section, we report on our progress and experience of defining the extent and structure of a training materials and events catalogue for the EOSCpilot to help users acquiring the skills to apply the EOSC services and enable new data science in accordance with FAIR principles. We specifically focus the FAIR data principles on different levels: First, to make users aware of the FAIR principles by ensuring coverage of appropriate training on the FAIRness of data itself and second, to enable the provision and exposure of FAIR training resources for the EOSC.

We therefore consider how the FAIR principles may be applied to training resources, particularly in the context of metadata to enable findability as one of the key drivers for successful training provision. We highlight drivers and objectives for making FAIR training resources available within EOSC. The drivers and objectives are aligned with the aims of EOSC to provide the trainings to its users at a large scale.

To provide a preliminary skills gap analysis based on available training materials and events, we first provide an analysis of the training initiatives currently offered by RIs and e-infrastructures. This analysis provides an overview on available training resources provided by EOSC-related organisations. This analysis further enables a selection of training materials for a detailed analysis of gaps in competence provided by current training resources.

We present use cases for different target professional groups of users to analyse the implications of FAIRness for training. Based on the use cases, we propose minimum and recommended criteria for metadata of training materials. FAIR principles will be incubated by applying them to materials for forthcoming WP7 workshops. The results will inform further recommendations (in D7.5) about applying the

principles to training materials in EOSC.

We have illustrated an approach to skills gap analysis based on the tagging of a collection of 90 training materials and 50 training events with terms from the EOSCpilot Skills Framework. Although we utilised a quantitative approach to visualise skills gaps, the analysis is not intended to be representative for the entire landscape of training conducted by Research Infrastructures, institutions, or other potential training providers in EOSC.

The next step will be to extend the analysis with taking account of:

- Definition of the EOSCpilot Service Architecture,
- Development of the EOSCpilot Skills Framework to link competences to capability statements that reflect the Service Architecture and the application of broader FAIR principles to services and research outputs, and
- Further clarification of the FAIR principles applicability to training.

To enable the EOSCpilot Competence Framework to be used as a controlled vocabulary for tagging training materials and events it is essential to further refine the terms used, to ensure these can be reliably applied by domain experts to enable others to find the relevant materials by searching on the same terms. The approach used so far is limited to manual classification of the training materials according to target professional groups, competence levels, and skills groups. We are conscious of the subjectivity involved in this being carried out by work package members of different backgrounds and culture. We will give further consideration in year 2 of EOSCpilot to the possibility of automatically classifying training materials and events using the EOSCpilot Skills Framework and/or relevant vocabularies, with the aim of providing information to training providers on gaps between their training materials and those already on offer in the EOSC environment.

4. A SOLUTION TO PROVIDE TRAINING AS A SERVICE IN EOSCPILOT

Short innovation and advancement cycles for hardware and software put continuous pressure to sustain the value of technology and services. This also applies in the context of EOSC, both to the infrastructure and services provided by EOSC. We, therefore, need a training approach that can keep up with the pace of advancements without disrupting day-to-day business to ensure the target audience is adequately equipped with relevant skills and competences. This not only requires a fine-tuned development of training materials to fill any skills gaps but also a timely performance of training events. We further aim to provide and design training materials that engage the intended audience without sacrificing scalability and efficiency of training.

To meet these demands, we propose a training infrastructure that integrates into the EOSC service portfolio as the EOSCpilot Training as a Service (TaaS). The proposed EOSCpilot TaaS aims to ensure

- quality of training experience by considering ease of access and convenience by providing a single TaaS that orchestrates external materials, components and services,
- dynamic composition of training materials from external resources to integrate easily into everyday business, including pausing and resuming the learning process, and
- user personalisation involves users in their own learning process by providing required information augmented with supplementary information,

while delivering scalability and efficiency with an agile but user-focused solution for the provisioning and creation of training materials. Scalability of the approach also includes efforts for developing, maintaining, operating and administrating the service as well as managing and curating the training resources. We, therefore, strive for a solution that minimises required efforts apart from the core competence to provide the required trainings and enable trainers to fill any gaps by benefiting from already existing training materials.

The preceding description of core competencies and related audience already suggests that our intended audience not only includes common trainees such as researchers within and outside academia using current or future services provided by the EOSC, such as EOSC-hub, but builds on the target groups and user requirements derived in Section 3.2.2.

Building on our results and analyses in the framework of the skills development initiatives and resources (see Section 3.1), we compare existing training infrastructures and set-ups in the following by focusing on a) evaluation of general concepts for training provisioning and b) technical evaluation that is driven by the FAIRness of training resources. We conclude with the description of our proposed solution for the EOSCpilot Training as a Service infrastructure as well as recommendations towards providing our solution within the framework of the EOSC service portfolio.

4.1. Existing training infrastructures and their set-ups

The various initiatives analysed in deliverable D7.1 and in Section 3.1 influence the design and set-up of the EOSCpilot training infrastructure. In the following, we present an overview on advantages and disadvantages of various approaches to expose and provision training resources in general, point out their relevance in the context of user requirements for the FAIRness of training materials and events within EOSCpilot and highlight appropriate examples where possible.

We complement the evaluation with a structured assessment of selected infrastructure set-ups for the different layers of the TaaS infrastructure proposed in D7.1. The assessment is based on evaluation criteria that focus on the technical implementation and exposure of training resources. The evaluation criteria are derived from the conclusions of D7.1 as well as FAIR data principles. We conclude the evaluation with a summary and discussion on existing training infrastructure set-ups and select appropriate solutions for the different layers of the TaaS where possible.

4.1.1. Exposing and provisioning of training resources

There is a variety of existing approaches to supply users with training materials. A number of training portals provide discovery and access to training materials, while several catalogues expose aggregated metadata of training materials provided elsewhere. However, all these different approaches and implementations ultimately attempt to solve similar issues.

Models to provision training resources

From a user perspective, a successful training event includes two distinct phases: the actual delivery of training and the provisioning of training materials and events to enable the user to find, select and book relevant trainings. From the perspective of a training provider the provisioning of training resources further includes the underlying data management and storage.

Examples for the provisioning of training materials and events range from full-fledged training portals, which provide access to all trainings, including course listings, course descriptions, training schedules, and online materials, down to lightweight training registries, which provide listing of and references to all trainings based on harvested metadata of external training resources. In the following, we differentiate three categories for the data management of training resources that we base on the terms of the Research Data Alliance where possible:

- *Training portals* provide actual training resources in conjunction with their metadata,
- *Training catalogues* [7] provide curated metadata of training resources, but not their content, and
- *Training registries* [8] provide the metadata of training resources, but not their content, by retrieving the metadata from external providers.

As a fluent transition can be identified between the different categories, it is also possible to establish a training service that combines the different advantages of two or more categories.

A training portal is centred around the concept of a repository for storing the actual training resources, including the content and metadata. It provides a flexible selection procedure for contents of training resources, technical solutions to store and deliver training and its metadata as well as internal data stewardship workflows. Managing these features requires a high amount of effort by the portal provider. The curation effort includes keeping available training resources up to date, responding to current user demands and the management and operation of required technologies. Also, adequate storage volume for fault-tolerant storing of all relevant information and materials for training is required.

A training catalogue implements a repository of training resources metadata. The training catalogue therefore only stores metadata of training materials, the content of which has to be provided by other services. It is comparable to the training portal with regards to provisioning and curation of training resources and technology, though without fault-tolerant storage or preservation requirements. Providers benefit from flexibility in selecting external training materials to provide high-quality trainings. On the other hand, additional effort is required to keep the catalogue up to date with external content providers. Compared to training portals, the engineering and training resource stewardship effort is reduced at the cost of operational effort compared to the training portal. The storage volume for storing all relevant information and materials for training is significantly smaller than for training portals.

A training registry is distinguished from training catalogues and training portals by merely storing how to acquire metadata from external training resource providers which in turn allow acquiring training resources. Thus, a training registry can be considered a catalogue of training catalogues. The catalogue of metadata is commonly built by harvesting information from external training providers including training portals or catalogues. This concept of integrating external training resources enables good scalability but requires additional engineering efforts for harvesting different providers. High-quality contents cannot be guaranteed by a training registry as the information depends entirely on external sources.

In the scope of EOSC, a TaaS infrastructure must be both scalable and considered high-quality in terms of available training resources. A high quality of training resources rules out training registries that harvest in a fully automated fashion: Storing only metadata of other training catalogues or portals ties the registry

directly to the quality of its sources. The level of quality is thus outside the control of the service. In contrast, scalability rules out classical training portals: Storing all training materials in a single service is not feasible at the scale of EOSC, especially if materials are already provided by other services. Only training catalogues offer both simplicity and control, required to handle both metadata and data at the scale of EOSC.

Protocols for depositing and harvesting training resources

Exemplary training catalogues and registries utilise various protocols for accessing training materials and harvesting training resource metadata. The utilised protocols include SWORD, OAI-PMH as well as ResourceSync. The adoption to any of these protocols by primary training providers enables the flexibility for selecting any protocol-compliant repository without potentially adverse effects on the other components of a training infrastructure.

Simple Web-service Offering Repository Deposit (SWORD) is a lightweight protocol funded by the Joint Information Systems Committee (JISC) for depositing content from one location to another [9]. Recently, the first draft of the SWORD v3 technical specification was released [10]. The most recent version aims at aligning current use cases around data publishing and complex objects as well as establishing community and maintenance mechanisms for the standard and supporting code libraries. The protocol can be utilised to interface to standard-compliant repositories [put reference to stash here].

The *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) is an application-independent interoperability framework for metadata repository harvesting. The protocol specification [11] describes how to expose and request structured metadata while distinguishing itself from the resource the metadata is about. The protocol describes a set of six verbs or services that are invoked within HTTP. Thus, the protocol can be considered a low-barrier mechanism for repository harvesting. The protocol is used by several training initiatives and repository services [12].

The *ResourceSync* specification [13] describes a synchronisation framework for evolving resources in the web. The protocol describes how a resource provider can advertise synchronisation capabilities, how a third-party system can inspect the synchronisation capabilities and finally how the third-party system stays synchronised. All the capabilities are implemented on the basis of the formats introduced by the Sitemap protocol. The modularity of the framework allows the combination of different capabilities to meet community-specific needs and enables a broad range of use cases including the provisioning of training resources and metadata. In contrast to OAI-PMH, ResourceSync enables the sharing of both metadata and contents.

Annotating and exposing training resources

As previously discussed, most existing training portals, catalogues and registries utilise a variety of annotation mechanisms, APIs or even custom, unspecified encodings to expose structured information if any. However, some pioneering solutions, such as the TeSS catalogue, expose structured, well-defined information and annotations, and must be considered as viable template and example for integration into the proposed TaaS solution.

TeSS supports an API as well as an annotation schema for events and materials that was developed by the Bioschemas Group at bioschemas.org. This schema is based on schema.org and fosters data interoperability in life science. The main outcome of Bioschemas is a collection of metadata and vocabularies built on top of existing technologies and standards. The Bioschemas community actively encourages people to use the defined markup in web pages and applications. The goal of the specification is to make it easier to discover, exchange and integrate life science training material information across the internet.

In contrast to harvesting protocols, annotations have the advantage that they can be easily adapted by content providers. We consider this to be important in the context of EOSC as we wish to make relevant, high-quality training materials from the largest possible number of domains available without the need for content providers to meet certain technical conditions we might pose.

Vocabulary and thesauri

The EOSC provides a multi-disciplinary, international environment that fosters innovation by reuse and cooperation between domains. This puts various challenges in terms of vocabulary and thesauri for the provision and exposure of training materials and events. Commonly, each domain utilises different vocabulary. The unification and harmonization of metadata therefore is a highly complex or even impossible task. For this, different standards and established schemas need to be considered including for example FOSTER thesaurus, generic research subject schema such as DOAJ subjects, but also domain-specific subject schema such as MeSH as well as available annotations (see Section on Annotating and exposing training resources). This is still an ongoing task and we, therefore, propose to report our recommendations in D7.5 after further discussion with relevant contributors in the field.

4.1.2. Evaluation of existing training infrastructure set-ups

The various initiatives analysed in Section 3.1 serve as inspiration for the set-up of the EOSCpilot training infrastructure. Each provides a viable example for a successful solution to one or several of the challenges for an EOSCpilot TaaS infrastructure. However, since none of the initiatives operate at the scale of EOSC, it is imperative to evaluate their respective approach in this context as well.

D7.1 concluded with a TaaS infrastructure consisting of five layers:

1. Presentation and delivery,
2. Registry,
3. Trainers and venues,
4. Training materials, and
5. Infrastructure for course deployment[3].

To evaluate existing training infrastructure set-ups, we concentrate on these five layers. To simplify the interpretation, we consider the layers *Trainers and venues* and *Training materials* as comparable from the point of a service and therefore evaluate whether Trainers and venues are supported in addition to Training materials. The evaluation aims to compare and identify existing solutions that can be utilised for the EOSCpilot TaaS.

The evaluation of existing training infrastructure set-ups is based on criteria that we derived based on the conclusion from D7.1 and the FAIR data principles. The criteria used in the evaluation are:

- *Metadata support* is a basic requirement to allow to search and find relevant trainings and is therefore considered important by D7.1 as part of Conclusion 1 and principle FAIR F2.
- Support for *open data formats* that are usable for the common target audience are key for accessibility and are demanded by Conclusion 1 of D7.1 and A1.1 of FAIR data principles.
- *License support* is a requirement for enabling reusability of training materials. This criterion is demanded as part of Conclusion 1 by D7.1 and principle FAIR R2.
- *Certification support* is considered an important criterion given Conclusion 8 from D7.1 to enable high-quality training for EOSC on a large scale.
- *Support for persistent identifiers* ensures findability of training resources and is formalized in principle FAIR F1.
- *Reference support* ensures proper credits for materials that are utilized by a training resource as well as interoperability. This principle is given by FAIR I3.
- *Standards and vocabulary* support interoperability and reusability are required for cross-domain and domain-specific knowledge representation. The principle is suggested by FAIR by principle I1.

We further need to deal with repeating events that may change over time, depending on demands by users or feedback. We therefore add two more criteria:

- *Support for version control* ensures accessibility of repeating events.

[3] For consistency, the naming of layers introduced in D7.1 are adapted with regards to most recent EOSC Training as a Service layer model presented within this document.

- *Feedback support* enables trainers and users to evaluate the value and relevance of training resources.

Another criterion we consider relevant is to provide the training infrastructure on a similar basis to the common EOSC services (cf. Conclusion 2 from [2]):

- *Scalability and elasticity* to ensure the support for international, large-scale deployment in cloud environments.

To benefit from existing experience and expertise in the context of European training initiatives and solutions we aim to build our proposed TaaS infrastructure upon existing tools. Another important criterion for the evaluation of existing tools and services is the availability of the underlying implementation and its license. Therefore, we also consider two more generic criteria related to the training infrastructure:

- Software availability and
- Software license.

We evaluate each layer in the scope of the criteria listed above. For this, we assign a score for the range from 0 for no support, 2 for full support and values in between meaning support to some extent. For each layer, we specify a score in the percentage range with regards to the maximum score by aggregating each criterion.

Within the context of the different layers, the criteria may refer to varying features, if any. For example, the criterion on open data formats includes evaluation of accessibility with regard to impaired people when considering the *presentation and delivery* layer while the other layers focus on services as consumers and therefore accessibility on a technical level via a particular API. A definition of criteria per layer can be found in Annex E.

The score itself does not necessarily evaluate the complete coverage of a value range for a given criterion, but instead the general support. For example, the score for metadata does not evaluate the support of a specific metadata property defined in Section 3.2.3 but the overall support for metadata.

The evaluation is applied to a selected number of training infrastructures and related tools and solutions: Digital Humanities Course Registry [14] and ELIXIR TeSS [15] as examples for training catalogues, EMBL-EBI Train Online [16], FOSTER Portal [17] and Up2Universe [18] as examples for full-stack training portals, DataONE Dash [19] as an example for a combined approach of a catalogue with a portal interface as well as Zenodo [20] as an example for a repository.

We expect training in the context of EOSC and EOSC-hub to require on-demand deployment infrastructures to deliver outcome-oriented training and enable users to apply learned theory in a real-world environment. We, therefore, also include EGI Training Infrastructure [21], GÉANT Testbed Service [22] and SWAN [23] as examples for Infrastructure as a Service to deploy training materials.

The summary of the evaluation of the selected training infrastructures and solutions is given in Table 5. For a detailed evaluation please refer to [26]. The evaluation is based on desk research using information available online. The evaluation does not claim to be exhaustive, but to give a general overview.

Table 5 - Summary of evaluation of existing training infrastructure set-ups

Solution/Tool	Score in %					
	Presentation and delivery	Trainers and venues	Catalogue/ registry	Resources	Deployment	Extensibility Adaptability
DataONE Dash	-	0	75	75	-	100
Digital Humanities Course Registry	-	0	50	56	-	0
GÉANT Testbed Service	-	-	-	-	88	25
EGI Training Infrastructure	-	-	-	-	88	50
ELIXIR TeSS	-	0	65	100	-	50
EMBL-EBI Train ONline	38	50	60	61	-	0
FOSTER Portal	50	0	60	61	-	50
SWAN	-	-	-	-	75	25
UP2Universe	75	50	80	50	63	50
Zenodo	-	-	70	78	-	100

While we already opted for training catalogues as the preferable foundation to build a Training as a Service infrastructure (compare Section on Models to provision training resources) the evaluation of existing training infrastructure set-ups also confirms this decision. Almost all solutions that build on the concept of catalogues show a better score given our evaluation criteria for the catalogue/registry layer.

From the three exemplary training infrastructures that are based on the principle of catalogues, the Digital Course Registry shows the worst score. This is due to the fact that it neither builds on existing standards for harvesting nor does it provide a solution for version control of training resources.

Furthermore, neither the software nor the source code are currently available with a proper license. The two solutions DataONE Dash and ELIXIR TeSS on the other hand show very promising results.

DataONE Dash

The solution provided by DataONE builds on the Open Source project called Dash [11]. Dash focuses on an integration of services to provide a combined approach for the provisioning of resources. The software Dash does not just focus on the provisioning of training resources but on the provisioning of resources in general. Dash allows to integrate metadata from standards-compliant repositories that support either of the protocols OAI-PMH or ResourceSync for harvesting.

The catalogue provided by Dash conforms to each of the evaluation criteria except for certification of resources. A simplified architecture of Dash is visualised in Figure 5. Persistent identifiers are supported for digital as well as human resources.

The approach taken is of specific interest for the design of the EOSCpilot Training as a Service infrastructure

as it provides a great flexibility and scalability in terms of usability and efficiency. The approach supports both major use cases we need to support with the training infrastructure: the integration of existing materials from external providers as well as the possibility to create new materials without posing a huge overhead for long-term management of data as this can be delegated to external services.

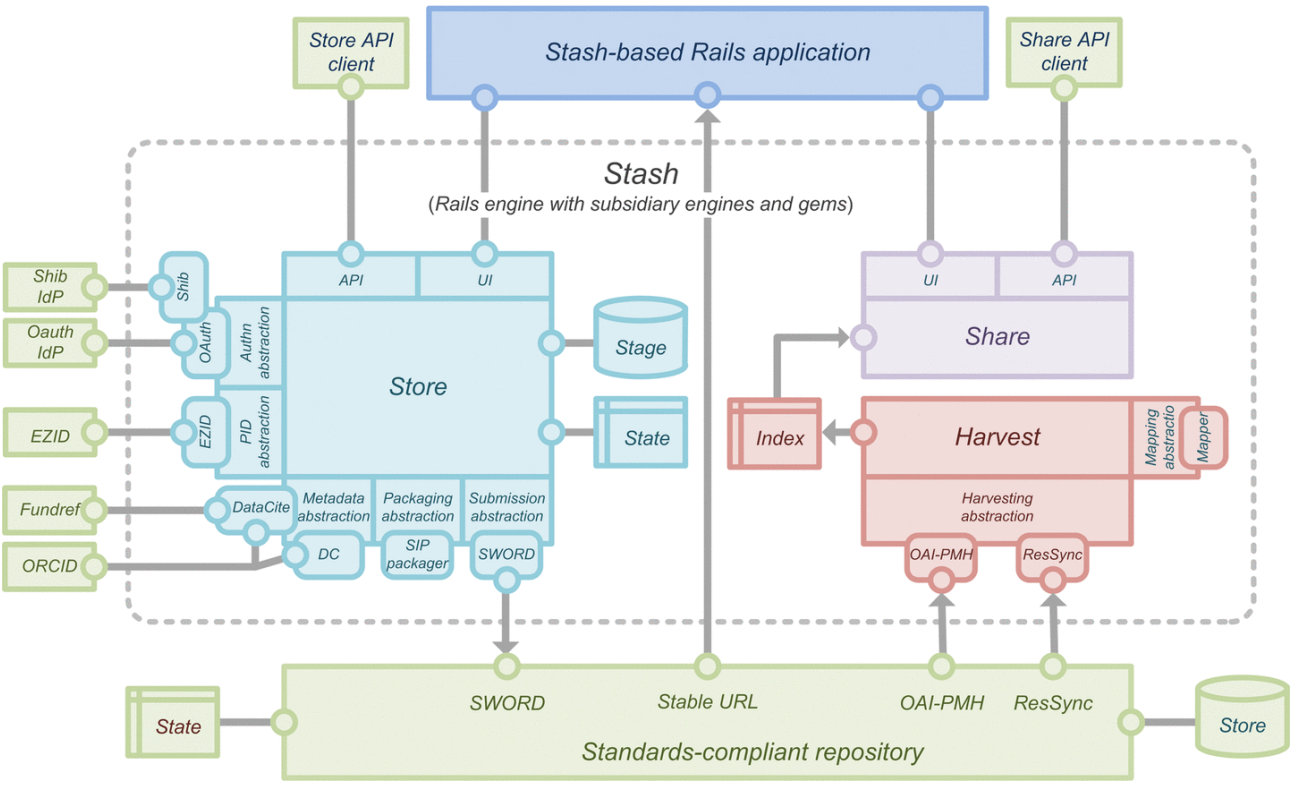

Figure 5 - Dash architecture as utilised by DataONE

ELIXIR TeSS

While Dash provides a solution for provisioning of resources in general, ELIXIR's Training e-Support System TeSS provides a specialised solution for training resources. TeSS focuses on disseminating, discovering and packaging training resources, primarily by aggregating information from various external training resource providers. The system is a registry and catalogue of training materials and events and not a repository of training materials and events.

The aggregation of metadata about training resources from websites is either automatic or manual, as shown in Figure 6. Automatic aggregation scrapes training information from various content providers using specialised scraper scripts which add links to training resources and their metadata to TeSS. This can also be done manually, by adding links to training resources manually by users. The automatically extracted information are kept up to date by regularly executing the scraper scripts. For each content provider to be aggregated, the developers provide a new script or re-use one of the existing scripts. Whenever the structure of information from a given provider changes, the appropriate script needs to be adapted accordingly. However, the responsibility for adaptation can also be taken by the community, as the scripts as well as the TeSS portal source code are Open Source, licensed under the BSD 3-clause license and under version control at GitHub.

It is not possible to upload training materials and events directly to TeSS. Instead, the user can refer to third-party repositories such as FOSTER Portal. Still, registered users of TeSS can provide links to training materials manually from the training materials page and attribute the content to a specific provider.

Although the implementation currently has a limited scalability due to manual adaption of scripts and their repeated execution even when not necessary, the approach is a pioneering solution in the context of training infrastructures. TeSS does not just use annotations to gather materials. TeSS also exposes

aggregated training resources via an API and annotated HTML for querying by other training infrastructures. This enables the deployment of a catalogue-of-catalogues, aggregating selected training materials from multiple TeSS instances to provide high-quality training resources from different domains. This degree of abstraction as well as the open approach chosen by TeSS makes this solution highly relevant to establish an EOSC Training as a Service infrastructure.

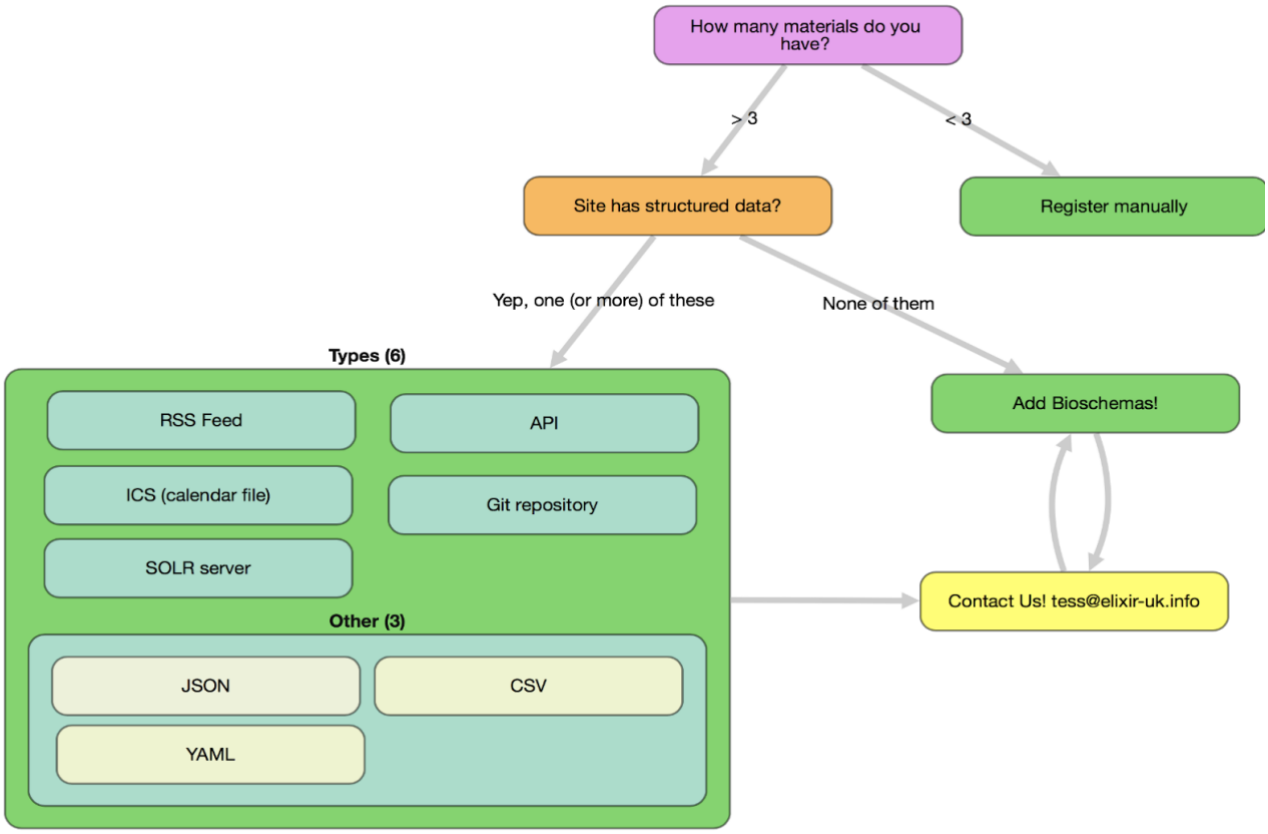

Figure 6 - Workflow to register training materials at TeSS

Up to University

The solution Up2Universe by Up to University follows a similar approach to TeSS or Dash for managing training resources, but on a different level of abstraction: Up2Universe utilises the European level OER metadata aggregation service (eduOER) to harvest data based on the OAI-PMH protocol. The eduOER service is designed to search, find and re-use trusted, quality, educational and research multimedia content. The functionality further includes a shared metadata repository, training resources, multi-language support and creative commons licenses.

In general, the project coordinated by GÉANT aims to provide an innovative learning platform building on cloud-based tools and services. The integration of relevant services allows high scores with respect to our evaluation criteria for all given layers of the TaaS and, therefore, makes this approach a relevant candidate model. The Up2University design is built on composing several independent technologies and services, as shown in Figure 7.

Up2University supports the integration of formal and informal learning scenarios by bringing together a number of tools for the provisioning of learning content, document sharing, recording and publishing, real-time interaction and social interaction. The platform is based on a MOODLE Learning Management System that provides a modern, easy to use interface, a custom dashboard, collaborative tools, calendar, file management, text editor, notifications and track functions. The platform further includes functionality for learning analytics.

For delivery and presentation, modern tools based on HTML5 such as H5P are used. Further, the Open Badges services is utilised to reward the user and record his training progress. Logging of a user's progress and making it reusable is further improved by integrating the experience API.

At the deployment layer, Up2Universe supports federated file sync and sharing of resources based on OwnCloud. The project started in January 2017 and is still still working on the integration of other tools such as the Jupyter Notebook, an Open Source web application allowing users to create and share documents containing live source code, equations, visualizations and narrative text. The platform does not provide direct access to compute, storage or network resources but can be used in combination with the GÉANT Testbed Service, which provides an integrated environment for the on-demand provisioning of segregated infrastructure resources to be used in the lab for testing or training purposes.

In contrast to Up2Universe, the other examples providing a full-stack training portal, namely EMBL-EBI Train Online and FOSTER+, do not utilise a harvesting approach but internally manage training resources. Both solutions are also comparable by their scores.

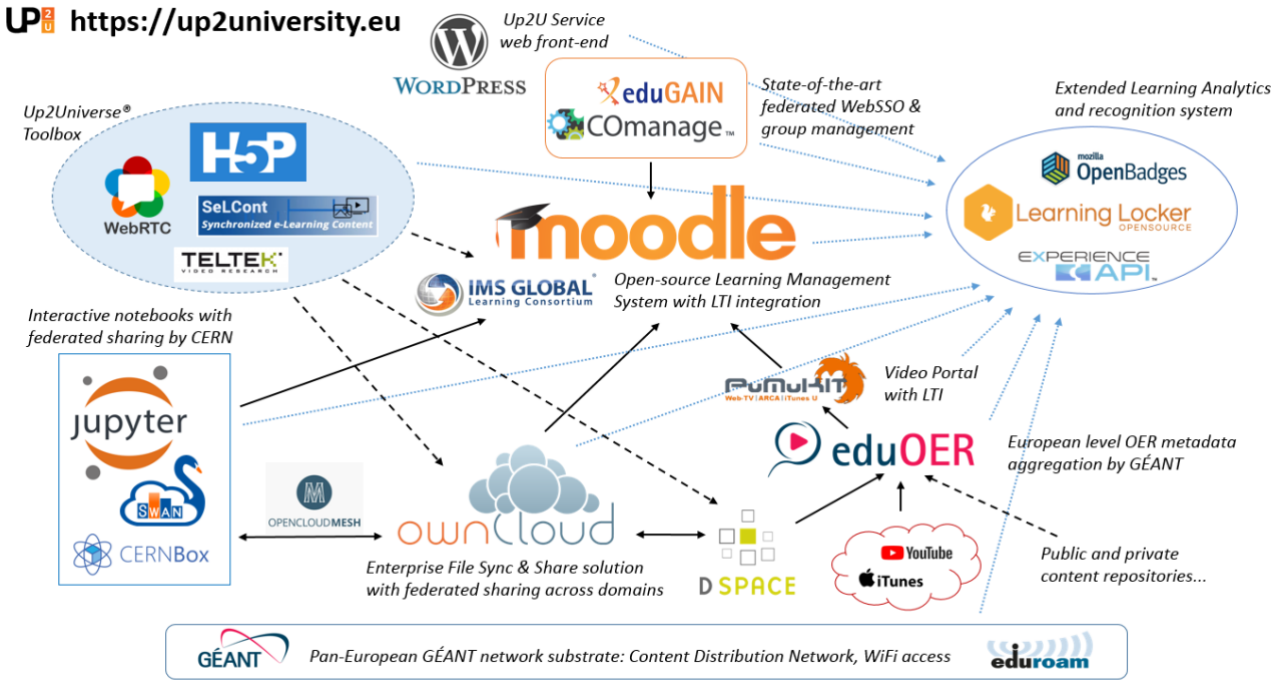

Figure 7 - Set-up and services of Up2Universe

EMBL-EBI Train Online

The EMBL-EBI Train Online platform offers an intuitive and user-friendly way of searching and browsing through training resources classified into discipline-specific subjects, learning levels, duration, type and project. The focus is on delivery and presentation of training materials to the users. Therefore, the developers explicitly focus on responsive design as well as accessibility. All materials are made available with the CC-BY-SA license.

Further details on the technical set-up of the given training platform can neither be found on the website nor related sites.

FOSTER Portal

The *Fostering the practical implementation of Open Science in Horizon 2020 and beyond* (FOSTER) Plus addresses the current skills and content gaps of European researchers, both at community/discipline and institutional level, on the practical implementation of Open Science. The FOSTER portal is an e-learning platform that provides access to various resources, including a growing collection of training resources. The technical setup of the FOSTER service is shown in Figure 8.

FOSTER Plus is pursuing its objectives through the combination of the following main activities:

- Delivery of face-to-face training events, blended and elearning courses focusing on the provision of practical, outcome-oriented lessons,
- Creation of high-quality, advanced-level training resources including a multi-module Open Science toolkit and an Open Science training handbook, and
- Consolidation of an Open Science trainers network involving the disciplinary communities of humanities, social sciences and life sciences.

The FOSTER portal gives access to a great amount of training resources that were collected, categorised and made available for reuse as standalone objects or organised into courses. Furthermore, FOSTER created and offers different elearning courses as self-learning or moderated courses. The main strategy for this portal is to gather a critical mass of contents on the various topics of Open Science as well as provisioning of high-quality content. The portal was meant to evolve from an initial digital-library-like software facilitating the collection and management of training material to an Open Course environment providing the necessary functionality to support elearning activities. This approach is further meant to enable the selection and reuse of relevant training content as part of a course.

To maximise comprehensiveness, barriers for uploading material on the portal are set very low, with a lightweight peer-review mechanism that involves the users of the portal. With this approach, FOSTER has sought to augment European researchers' understanding of open access, open data and open science requirements among the participants of the European Research Area. The portal supports their compliance with the open access policies and rules of participation set out for Horizon 2020.

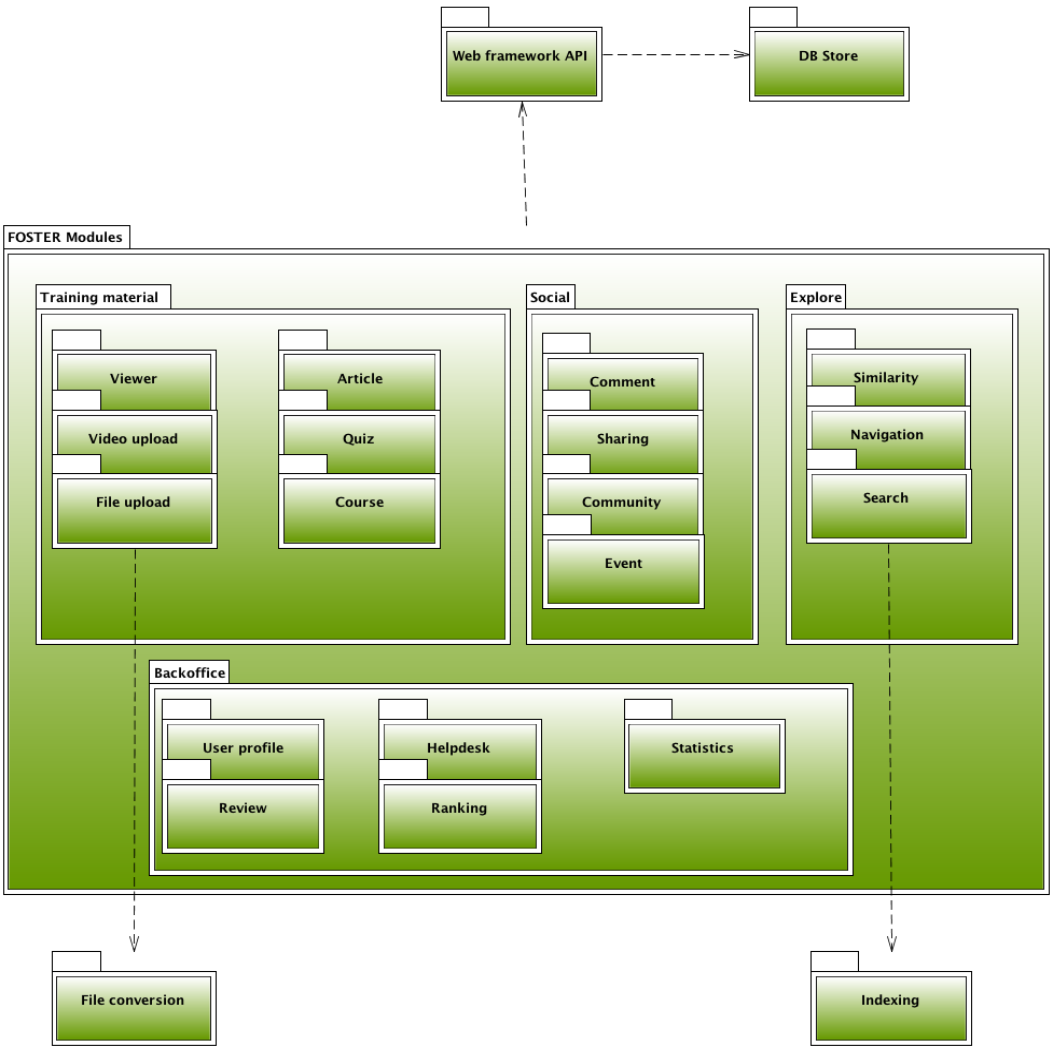

Figure 8 - Set-up as utilised by FOSTER

The most recent update to the FOSTER portal improves responsiveness of the portal and integrates certification of participation based on Open Badges. However, detailed information on the set-up of the portal and internal management are not available. Liaison with the project will be highly desirable in the second year of EOSCpilot to identify the scope of potential collaboration.

Zenodo

Zenodo is a service offered by CERN and serves as an example as a repository provider. The service offered by Zenodo is utilised by several solutions presented within this evaluation. In specific, Zenodo enables the sharing, curation and publication of data and software for researchers across different communities, disciplines and nations. To be an effective solution for all kinds of data, zenodo tries to eliminate barriers to adopting data sharing by not imposing any requirements on format, size, access restriction or license.

Zenodo provides a simple web interface that is supplemented by a rich API which allows third-party tools and services to use Zenodo as a backend in the workflows. For example, FOSTER+ utilises the service provided by Zenodo to publish some of the materials and create a persistent identifier, a DOI, for the respective resources.

Zenodo strives to comply with the FAIR data principles and offers the required services to enable FAIRness for data that are managed within Zenodo [29]. This makes Zenodo as a service a highly relevant solution to the establishment of a TaaS for EOSC. The support of managing and creating data via Zenodo allows to support FAIR training materials. It further supports version control for resources and therefore also

complies with regard to FAIR data principles.

SWAN

The project Up2Universe uses the Service for Web based Analysis (SWAN) provided by CERN to enable interactive data analysis, documentation and preservation of training and sharing of results. SWAN combines CERN IT services with modern technologies of distributed computing. The service is not yet officially provided by CERN but already enables users to analyse data without the need to install any software. Users can write and run their data analyses with only a web browser or even a shell, leveraging on the widely-adopted Jupyter notebook interface.

However, the service is currently centred around the concept of Jupyter notebooks and therefore does not enable the processing and use of a variety of data formats, programming languages and processing paradigms. Although access is currently limited to CERN members the approach looks promising especially in the context of training (compare the current set-up of Up2Universe in Figure 7).

GÉANT Testbed Service

The GÉANT Testbed Service (GTS) provides an integrated virtual environment for on-demand provisioning of segregated infrastructure resources to be used by researchers in the lab for testing innovative technologies or arguably for training purposes. The GTS environments consists of computational servers, data transport circuits, and switching/forwarding elements.

Within GTS, researchers are able to describe the logical composition of their desired network. GTS allows them to specify desired attributes of each resource – such as the geographic location for a server or switch, or the desired bandwidth capacity of a transport link. GTS provides a generic application program interface (API) that allows researchers to maintain control of their respective testbed resources and which will allow the service administrative and operations staff to assert management control over the entire service facilities.

Testbed resources are dynamically allocated from real e-infrastructure that is distributed throughout the GÉANT core service area. However, GTS is logically isolated from the production GÉANT network to ensure the integrity of live applications. GTS testbeds are also insulated and isolated from one another; this allows GTS capable of supporting multiple projects simultaneously in a secure and protected way.

From information available on the GÉANT website, training appear amongst the GTS use cases. However, the isolation, scalability and flexibility features of the service, the capability of advance testbed scheduling makes it a suitable environment for the coordination and support of experiments and simulation for wider sets of activities including training. As the service aims at supporting the network research community it is therefore of high interest to enable hands-on exercises for trainings with a focus on architectures, clouds and workflows.

EGI training infrastructure

The EGI training infrastructure was established by the EGI community as a cloud-based e-infrastructure specifically for training. It is implemented as a resource pool of the EGI Federated Cloud infrastructure and provides physical resources and access services. The training infrastructure is integrated with the EGI AAI that allows trainers to generate short-living user accounts for training participants. Such accounts can identify students individually and for a limited duration – typically few hours or days, depending on the length of the training event. The accounts also allow a user to interact with the training infrastructure sites and services.

The training infrastructure is suitable for two operational modes of courses:

- Courses to teach students about IaaS and the management of Virtual Machines, storage blocks and other types of low-level resources. For this, no deployment of applications or services is required in advance.
- Courses to teach about a specific software. In this operational mode, the trainer deploys the domain-specific application/tool in advance on the training infrastructure. Students do not require knowledge on the set-up and deployment. Configuration in advance of training

enables the community to benefit from the easy deployment, predictability and repeatability of courses.

In both operational cases the deployment of applications/tools/services can happen in the form of Virtual Machines (VMs), and block storages – the latter basically behaving like a virtual USB drive that can be attached/detached to VMs to provide data and storage for applications.

VMs can be deployed on the infrastructure through the EGI Applications Marketplace [25]. A specific section within the catalogue is reserved for training-related application VMs [26]. The application Marketplace also includes a VM management dashboard [27], offering a web interface for both trainers and students to deploy and manage VMs from the Marketplace on the training infrastructure.

The system currently includes two cloud sites. Additional two sites are under integration. Between 2018-2020 the infrastructure will receive funding in the EOSC-hub project to serve the training needs of the project and its partners. The training infrastructure can be booked for a course through an online form [21].

4.1.3. Conclusions

Current training portals provide innovative interfaces for trainees as well as trainers in the presentation and delivery layer by utilising interactive tools such as H5P, considering responsive design and focusing usability. These interfaces nowadays support the access from a variety of devices including mobile devices as well as interfaces for the access by services. These portals therefore enable a broad range of use cases which we expect to adapt to evolving demands. Furthermore, they are subject to continuous improvement by a community of research engineers that are sensitive to a sustainable establishment of FAIR training in the context of Open Science. Building on the expertise and offerings of those providers therefore seems a reasonable approach.

The evaluation further shows the impact of the different models to provision training resources. However, by adopting a registry-based approach on top of repositories provided by training providers offering high-quality contents, a combination of both approaches becomes feasible. An appropriate set-up is capable to provide high-quality contents while keeping management and operational efforts at a minimum. The set-up taken by Dash further enables the creation of training materials by cross-fading a repository-based approach with a registry-based harvesting approach. Although the user gets the impression to locally manage materials, the responsibility can be delegated to an external repository provider.

The skill requirements for EOSC and EOSC-hub, for example the deployment of container technologies or the adaptation of scientific workflows to cloud environments (compare Section 2.3), require complex environment set-ups to enable users to directly test and apply theory from within a training. The Up2Universe platform utilises an approach to integrate interactive web-based analysis for their users to directly apply mediated contents. Still, this approach is limited due to the utilisation of Jupyter as the appropriate interface.

The flexibility and availability of the EGI training infrastructure compared to the previously evaluated deployment services makes the EGI training infrastructure highly relevant. This equips users with the necessary tools to apply contents in daily business. In particular, this approach enables on-demand utilisation of deployment by users to apply exercises for better traceability and competence-building. A service like this should also be utilised within the proposed Training as a Service infrastructure. At best, the federated infrastructure envisioned in EOSC can be used directly as a service from specific EOSC trainings for on-demand utilisation and application of training materials.

4.2. The EOSCpilot Training as a Service Infrastructure

In this section, we provide general information about the EOSCpilot TaaS infrastructure and outline the scope of the individual layers. For each layer, we discuss the integration of relevant standards and services to enable inter as well as intra layer interoperability.

4.2.1. Layers of the TaaS infrastructure

To ensure FAIRness of training materials and events, certain constraints and features are mandatory with regard to the TaaS infrastructure. To support the variety of user domains and communities represented in EOSC, the infrastructure focuses on a domain-agnostic approach and implementation.

Starting from the initial proposal for an EOSCpilot Training as a Service infrastructure model, we have further abstracted and refined the model. In the following, we introduce the adapted model that implements implications from FAIR data principles and adapts to the needs we identified during evaluation of existing infrastructure set-ups.

The main idea of the layered TaaS infrastructure fulfils two primary goals: On the one hand, presentation of and access to training resources, with focus on user needs. On the other hand, provisioning and curation of services and data with focus on implementation. These ideas are reflected by the layer structure as bottom-up (implementation) and top-down (user). Dependencies between the different layers are strictly top-down, to reflect consistency and integrity of materials but flexibility of presentation.

As noted in Section 4.1.2, our analysis of existing services shows that training materials, events, trainers and venues are equivalent from a service point of view. To avoid redundancy and ambiguity, we therefore consider them as specific adaptations of the same generic concept. Compared to D7.1, we combine the two layers *Trainers and facilities* and *Training content* to the unified *Training resource* layer representing both.

In the following, we differentiate four layers, including *Presentation and delivery*, *Catalgoue/Registry*, *Training resource* and *Infrastructure* for the EOSCpilot Training as a Service infrastructure as visualised in Figure 9. The four layers serve as a basis for abstraction and allow the assignment of various services that are required to comply with the user requirements introduced in Section 3.2.2.

The layered infrastructure is designed for separate entry points depending on the use case at hand: Trainees access the *Presentation and delivery* layer, which allows access and retrieval of training resources. In order to find any relevant training materials and events, the user may skim or query the *Catalogue/Registry* layer on available contents. An organisation managing training, accesses the *Catalogue/Registry* layer to plan, organise and publish training events by picking trainer, venue and required training units. Curators and trainers may directly access the *Catalogue/Registry* layer to add, maintain and manage training materials over their lifecycle. By using stacked layers, each individual layer has to deal only with isolated set of user requirements of adjacent layers.

The fundamental FAIRness features are directly provided by individual layers: lower layers fulfil version control/provenance while the entire stack satisfies certification constraints.

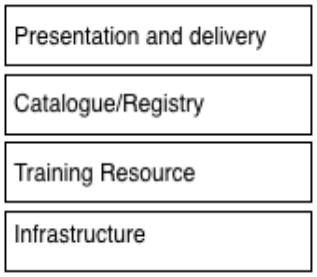

Figure 9 - Simplified EOSC Training as a Service 4 layer stack

In the following, we describe the scope of each layer and discuss relevant services that are required to meet the user requirements for the different target groups.

Presentation and Delivery

The presentation and delivery layer includes different formats of presentation of training materials. This may be in the scope of a training event by a trainer at a specific venue, or independent online, self-paced learning as well as various learning opportunities, such as fellowships, internships, placements, or staff exchange. The various available formats and possibilities of presentation and delivery have already been

shown in the framework of the skills landscape analysis (see Section 3.1 associated with the results from D7.1).

In addition, we evaluated the capabilities of various existing training portals with regard to the presentation and delivery (see Section 4.1.2). The available expertise and existing tools and solutions as well as current developments for delivery and presentation of training allow to benefit from intensive efforts. Therefore, we propose to delegate the actual delivery and presentation of training material to existing training portals specialising in presentation and delivery. To enable the tracking of a user's progress, we further propose to conform to the experience API as well as Open Badges. The delegation of the interaction between trainee and trainer via the EOSC TaaS and external providers is shown in Figure 10.

To enable proper presentation and delivery, information on the various training resources, such as trainers and facilities but also information about the training materials and possible events need to be available and accessible. In addition, the training contents and materials need to be in a supported format that can be visualised/accessed following the FAIR data principles.

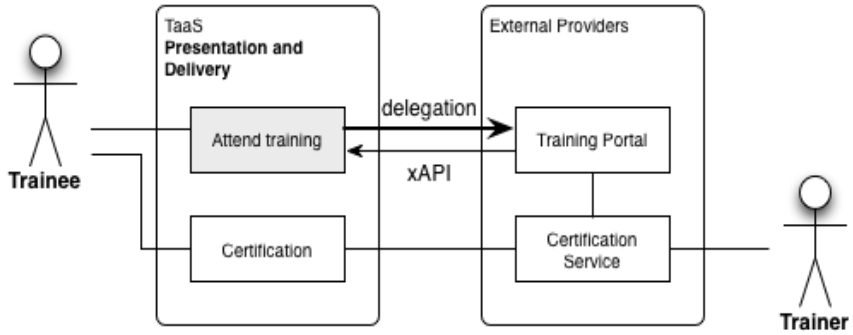

Figure 10 - Services and protocols for presentation and delivery of training

Catalogue/Repository

The Catalogue/Registry layer plays an essential role in the proposed EOSCpilot Training as a Service infrastructure. It provides two components to enable presentation and delivery: The first targets the findability and accessibility of training resources for different target audiences while the second targets the management and provisioning of metadata including the harvesting of resources from an operational point of view. Both components seek to support the FAIRness of training resources for different target audiences. In the remainder of this section, we refer to the first component as front end, while referring to the second component as back end. Figure 11 provides a high-level view on the various responsibilities and workflows in this layer.

The front end of the TaaS Catalogue/Registry layer mainly focuses on findability for EOSC users, organisations, trainers and curators: The EOSCpilot TaaS provides search functionality over the metadata of harvested training resources provided by the back end. Furthermore, accessibility is supported in terms of usability of the user interface. Accessibility hereby includes features for people with vision loss and other disabilities as well as responsive design to enable an engaging experience for all, independent from the devices in use.

The front end also provides functionality to use the TaaS as a publishing interface. However, the front end only acts as the interface for the user while the actual functionality of publishing training contents is maintained by an external service. The back end is tasked with the delegation of tasks to the external service provider. It is the purpose of the infrastructure layer to model the various external repositories of external providers.

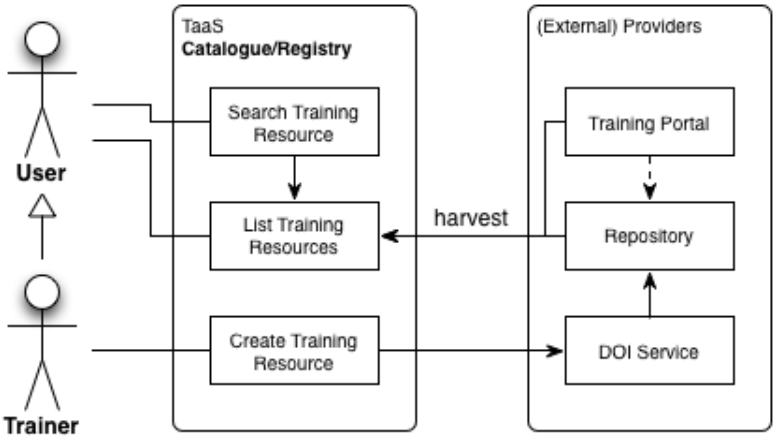

Figure 11 - Services and use cases for Catalogue/Registry

In general, the functionality provided by the back end aims at realising the balance between the provisioning of high-quality, up to date and relevant training resources while keeping the efforts for stewardship of these resources at a minimum. For this, the back end builds on harvesting of training resource from external training portals and repositories (see Section on Models to provision training resources for a comparison of the available approaches). The harvesting process and its components are shown in Figure 12.

To enable the harvesting from a variety of training resource providers, our proposed solution builds on existing standards for harvesting from repositories as well as scraping scripts and services to also include providers using annotations as well as REST APIs. This enables providers to benefit from the TaaS service with minimal adaptations to our proposed metadata schema (see Section 3.2.3).

Further, the back end aims to provide a FAIR service by exposing the EOSC training metadata specification as well as the harmonised and enriched EOSC training metadata. This enables external training content providers to adapt their existing models for improved integration into the EOSC training service and the consumption of provisioned training resources by external training providers.

To enable a valuable EOSCpilot TaaS the service relies on the availability and exposure of information on training events and materials by websites and repositories participating in EOSC. This is modelled by the infrastructure layer. The information need to be in a harvestable metadata format, preferable the EOSC training metadata schema provided in Section 3.2.3, or it should be possible to automatically mine the relevant information from the respective websites. The content providers need to support the enabling metadata standards and thesauri (see Section 4.1.1).

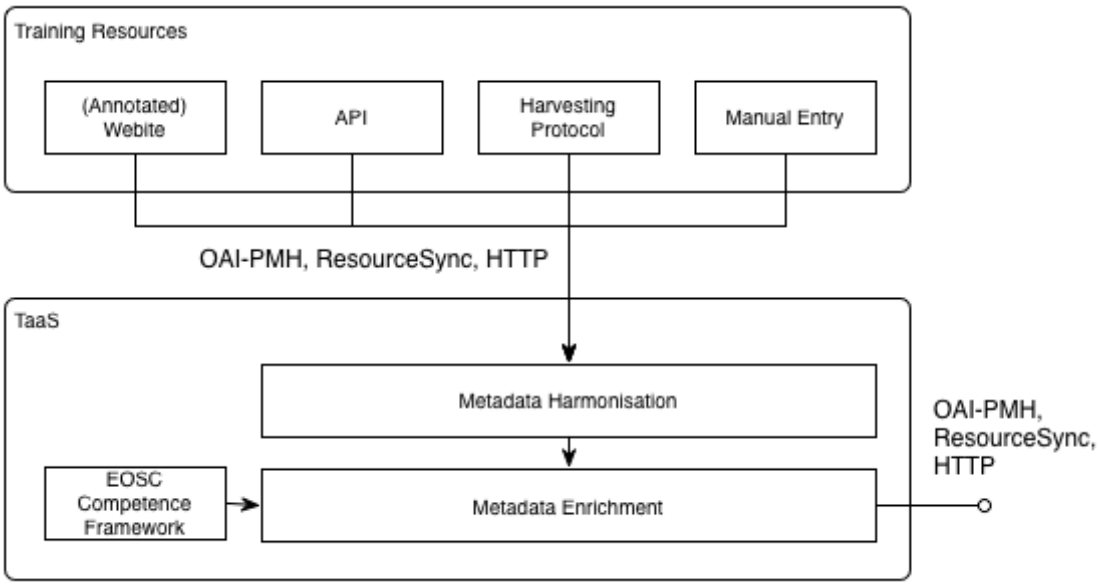

Figure 12 – Overview of the harvesting process in TaaS

Training resource

The training resource layer is an abstraction of all relevant entities that are required to provide trainings for the EOSC and to fill the skills gaps. This includes training materials, training events but also trainers and facilities for the delivery of trainings. Evaluation on the current available training platforms (see Section 4.1.2) shows only a basic support for managing information and capabilities on trainers and facilities in terms of sustainability, scalability and efficiency of management.

To enable sustainable delivery of high-quality and certified training, information on the availability of specialist trainers and adequate training venues is needed. The dynamic nature of information on trainers and venues and its dependency on locality and time suggests the application of a trainers network or community rather than a centralised administration. We therefore propose to foster further discussions within the community of trainers through EOSCpilot WP7 workshops, and collaboration with EOSC-hub to establish a scalable and efficient approach that enables utilisation by multiple training providers.

To provide a unified view onto the relevant training resources we focus on providing metadata of training resources that are enriched with properties given by the EOSCpilot Competence Framework. For this, the TaaS relies on the availability of information from external training providers. The layer relies on the ability to access stored metadata from the infrastructure layer to serve contents to higher layers.

It is important to note that responsibility for assessing the relevance and quality of the resources is undefined in this model. Where that responsibility lies, e.g. with training providers or with EOSC, is a governance issue that we make further comments on at the end of this section.

Infrastructure

The infrastructure layer handles the dynamic provisioning of compute, storage and networking capabilities. The layer abstracts from two different methods to provision and utilise infrastructure: First is the requirement to abstract from internal and external repositories. Second is the provisioning of on-demand infrastructure for deployment of training materials required for hands-on training and exercises.

The proposed TaaS tightly integrates with internal and external repositories as well as content providers. For harvesting of training resources in the context of the catalogue/registry layer standard protocols, APIs and annotation schemes are utilised to build a repository of metadata for training resources. To keep the storage and operational efforts and dependencies at a minimum, the repository of metadata itself is based on available repository services, such as services provided within EOSC. However, the requirement of high-availability and fault-tolerant storage needs to be fulfilled to provide a valuable training service.

Furthermore, the TaaS integrates with Infrastructure as a Service providers to enable deployment of training resources. Whenever a training requires availability of compute, storage and/or network an appropriate environment is requested. This on-demand infrastructure complements the information that are conveyed via the presentation and delivery layer and serve the purpose to enable exercises and hands-on trainings.

Again, this layer aims at delegating responsibility by orchestrating the different services and providers but ensures to meet the requirements for a given task. However, this layer needs to ensure the availability and responsibility of the registry of training resources for overall availability of the TaaS.

Cross-layer services and solutions

Some of the services discussed above only apply to one layer or functionality. However, to enable a highly professional TaaS service within EOSC clear values need to be proposed. One of the main functionalities that need to be supported to enable pan-European relevance is certification activities. We therefore aim to provide certification if applicable on all layers. This includes certification of infrastructure, e.g. certified storage and quality of service, and most importantly certification in terms of different training criteria. This includes certification of training materials, trainers and venues but also the delivery and certification of users after successful participation.

Some of these aspects can be solved on a technical level and can be made a requirement for external providers, e.g. quality of service and service level agreements. The thematic certification on the other hand is more complex and requires future activities to enable an appropriate service. Our current concept however is mainly driven by the user requirements in terms of EOSC users to integrate Open Badges for enabling certification of participation as a first step. Our aim is to lay the lay the foundations to an encompassing certification for the EOSCpilot Training as a Service infrastructure.

4.2.2. FAIRness of the TaaS infrastructure

To enable the use cases defined by the different target groups including to find, access and use training resources and services provided by the EOSC Training as a Service infrastructure, the TaaS needs to comply with the FAIR data principles at all levels, i.e. all layers. Further we need to ensure a future-proof concept in a cloud environment to enable scalable and effective training. This enables to lay the foundation for outcome-oriented, sustainable training.

Our concept for the TaaS therefore focuses integration of external service and content providers. We therefore propose the utilisation of standard protocols to enable standard-compliant services and APIs. Further, we propose to adapt existing solutions whenever possible, e.g. the approach taken by TeSS for harvesting existing training resources. Whenever possible, we suggest the utilisation of services provided by the EOSC Service Portfolio.

The delegation of responsibilities however puts requirements to external content and resource providers that in turn influence the overall FAIRness of the proposed TaaS. An overview on the expected FAIRness if all requirements are met with regard to the FAIR data principles is given in Annex F.

4.3. Governance and policy aspects of the EOSCpilot Training as a Service infrastructure

One of the aims of the governance framework is to develop a number of pilots that integrate services and infrastructures for training. Training as a service infrastructure within EOSC services is a product of performance against needs, best practices and objective metrics [28]. The services are at the centre of EOSC governance model as facilitation and of strategic importance to commercial and non-commercial users for EOSC's wider application. Within the three layers of digital governance, there are three roles, of which the *primary* stakeholder role provides the services, data and other resources including training. In the governance model of Open Science Commons there are multiple stakeholders enabling national and European-wide sharing (e.g. training marketplace, open source software repository). These stakeholders are e-infrastructures, VREs, and other H2020 projects, service providers, enterprises and academic institutions and libraries.

The section on Data Culture and FAIR Data of the EOSC Declaration [1, p. 1] contains the intents about training the relevant skills in research data management, data stewardship and data science. This asks a policy of further development and dissemination of RDM training, and alignment of the training activities among the stakeholders. The stakeholders follow rules of engagement that specify the conditions for data management, storage and analytics. They are a minimal set of rigorously applied and enforced protocols for stakeholders to participate in EOSC training service infrastructures. Non-EOSC approved players are free to explore any role in open science ecosystem, without getting their services EOSC approved/certified. Research communities, institutes, infrastructures and e-infrastructures are responsible for rule of engagement while funders, commercial entities should be consulted. In recent discussion, it has been made clear that member states and EC should ensure long term funding of the services that are needed to enable the integration of and access to the possible federated resources in the EOSC [29].

The training infrastructure will be based on principle of Data Culture and FAIR Data and will be available in different countries and with different accredited stakeholders. As a part of Open Science Commons [30], the resources such as training materials, scientific publications and expertise of researchers will be shared. The stewardship part will be active maintenance of open science resources, such as technical development, certification of data repositories, and maintenance of training and education programmes. The trainings will have a certain standard for training and trainers, but of course not to set the bar too high, disqualifying enthusiastic researchers from taking the role of trainers. Institutes, infrastructures, funders, commercial entities and e-infrastructures are responsible for providing the skills trainings while research communities should be consulted. Similarly, service providers should be able to provide trainings for their services. Training infrastructure will facilitate sharing of relevant materials and event certification to support the recognition of data skills acquisition [31].

Not all layers from TaaS infrastructure are required to provide every service to be used in EOSC. Each level contains a certification, for example, certified core services, can be certified services and uncertified services. The certification should be treated in a consistent manner across the layers, whether it is applied to training services, provider organisations, individual trainers, or the recipient of training. The scope and formality of certification ranges from broad and formal such as the Skills Framework for the Information Age to the very narrow and informally such as Open Badges. A commonly-agreed European certification scheme would come for the services and there will be EOSC-certified providers to provide the service.

4.4. Summary and conclusions

The EOSCpilot TaaS infrastructure should be a single, efficient and scalable service in the future EOSC portfolio that supports

- the quality of training experience by orchestrating a number of external components and services and/or EOSC services,
- the composition of training materials at fine granularity, based on these external resources to integrate easily and in a flexible way into everyday business,
- user-driven approach in management of their own learning processes,
- agile user-focused solution to the provisioning and creation of training materials and events.

To maximise efficiency and scalability the EOSCpilot TaaS infrastructure should focus on the concept of a training registry. Instead of handling a repository of training resources, the TaaS should build upon a repository of metadata of resources that is harvested from external content providers.

To support training throughout the whole lifecycle the TaaS infrastructure should also support trainers to manage FAIR training materials. The TaaS should therefore extend interaction with repositories to comply with standards to depositing resources into repositories.

The TaaS should be setup in collaboration between EOSCpilot and EOSC-hub and eventually a component of the EOSC-hub service portfolio, to support the set of skills required by its architecture.

The evaluation of existing training platforms shows that many of them cover some but not all the layers of our proposed TaaS infrastructure. The recommendation is to build upon existing platforms but, as a result, it is not possible to recommend a single platform. In the context of the evaluation we have carried out we several existing platforms fulfil the requirements of the EOSCpilot TaaS infrastructure on the basis of the layers described in the previous sections.

5. CONCLUSIONS

The purpose of the EOSCpilot Skills and Capacity workpackage (WP7) is to promote a training environment to advance scientific research with the capacities and resources of EOSC for service users and providers. The work of WP7 consists in large parts of the ongoing survey, assessment, and proposal of existing, future and planned training environments. As an interim report, D7.2 has given a progress update on three of the main WP7 objectives:

1. Design of an open data science skills framework that describes the individual competencies and organisational capabilities required to provide and use EOSC services.
2. Catalogue the currently available education and training materials with respect to the skills framework and identify gaps in delivery.
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.

The conclusions in this final section also address the actions taken to in response to the results reported previously in D7.1. Further conclusions consider the next steps required to pursue the remaining WP7 objectives, namely:

4. Coordinate delivery of relevant education and training materials and events to address those gaps that stakeholders identify as of the highest priority.
5. Connect with established national and international training schools and institutions, and collaborate with *champion* universities in their implementation of open data science curricula.

Conclusion 1. The EOSCpilot Skills Framework requires further development towards D7.3, including validation of its conceptual model against the EOSCpilot and EOSC-hub service architectures, and consultation with the target users of the framework in Research Infrastructures and institutions.

The Skills Framework has now been modelled to enable clear relationships to be established between EOSC services, the capabilities they offer to researchers and research organisations, and the skills and competences they need to acquire. To align with WP5, the conceptual model recognizes that these skills may be different according to the relationship with each service, i.e. whether the organisation is its developer, provider or user.

The consultation has begun on the competences, aiming to focus on stakeholder priorities, and will be much more extensive in year 2 of the project. Where possible this will be done in conjunction with EOSC-hub. By identifying the capabilities that stakeholders want from the range of services envisaged by WP5-6, and aligning these with competence gaps, it will be possible to construct and validate skills *user stories* and capability statements constructed according to its conceptual model.

Conclusion 2. Current provision by partners and H2020 cluster projects has been analysed, with preliminary results indicating that some skills groups are under-represented in their current training offer. More specific skills gaps are evident from stakeholder consultation, and from Science Demonstrator outputs. These analyses will be expanded and refined in year 2 of EOSCpilot.

The analysis of training provision has so far highlighted the difficulty in obtaining a comprehensive and consistent picture of activity across infrastructures and institutions represented in EOSCpilot and by the H2020 cluster projects. The analysis of training provision remains to be completed and extended to Research Infrastructures participating in EOSC-hub.

Nevertheless, we identify a number of areas that we believe are gaps representing a demand for training across six of the main skills groups in the Skills Framework, as shown in Table 6.

The analysis of materials has so far used a very limited sample, based on outputs from a number of leading

Summer Schools. As with information on training provision, the materials analysis illustrates the lack of a comprehensive catalogue of resources available. There are however a number of excellent portals, including from FOSTER+ and EDISON, and catalogues on specific domains including life sciences, and cross-disciplinary areas such as marine science. The project will further liaise with their providers to establish their priority gaps.

Table 6 - Interim analysis of skills gaps

Skills group	Gaps
Appraise and preserve • All professional groups • All competence levels	Workflows for cloud resource utilization – description for reproducibility and portability
Publish and release • All professional groups • All competence levels	Research reproducibility
Expose and discover • All professional groups • All competence levels	FAIR metadata for interoperability
Govern and assess • Domain researchers, Data scientists/analysts, Data engineers • All competence levels	Data policy legal and funder requirements FAIR and OA policy
Scope and resource • Domain researchers, Data scientists/analysts, Data engineers • All competence levels	Research strategy and open research potential AAI integration/management Secure storage integration Virtual machines and containers setup and deployment
Advise and enable • Domain researchers, Data scientists/analysts, Data engineers • All competence levels	Tools and domain standards Personnel and skills development

Conclusion 3. The EOSC training community should be consulted further on the gaps in skills provision, on whether applying FAIR principles to training resources could help improve provision and, if so, how FAIR principles should be applied in the training context.

Conclusion 4. A conceptual model for Training as a Service in EOSC has been developed and a need identified for consensus among stakeholders in EOSCpilot and EOSC-hub, particularly Research Infrastructures, on the preferred options. These include mechanisms for harvesting metadata, and interoperation with existing portals and catalogues.

Planned workshop activity for early in year 2 of EOSCpilot is aimed at trainers in the infrastructures and research institutions. These workshops will elicit further views on stakeholder priorities and expectations of the *FAIRness* of training resources in EOSC, and their preferences with regard to training infrastructure. Training coordinators in Research Infrastructures participating in EOSC-hub will be

consulted directly, in conjunction with EOSC-hub. This will add to the feedback obtained from events at the end of year 1, and will be included in D7.4 Report on Training Workshops.

The qualitative feedback gained from these workshops will be used to structure a survey of infrastructures and institutions. This will obtain a more quantified picture of the capacity gaps, according to the main topics (skills groups) and research domains, and inform the strategy reported in D7.5.

Conclusion 5. To address identified gaps for training events and materials, gaps should be filled partly through an open call for participation in online training, and partly through face-to-face workshops by WP7 partners.

The training workshop activity planned by WP7 partners in year 2 aims to help the WP 'learn by doing'. This will include a collaborative workshop aimed at supporting interaction between data experts and service developers, and addressing a broad range of competences highlighted as priorities in our consultations. Nevertheless, there is advanced expertise in priority areas that would best be delivered by others. In the interest of quality and transparency an open call will be issued to conduct a number of webinars, with support from WP partners to provide training infrastructure and materials, publicise the events, and to share feedback from the participants.

Conclusion 6. The EOSC Education and Training Strategy must consider steps necessary to ensure sufficient provision of relevant undergraduate and postgraduate education courses, and to meet the need for data literacy in the broader community.

WP7 has chosen to focus on professional development, foreseeing that much of the skills gap (in the general sense) will be closed by re-skilling. However, it is important to recognize the wider WP objective to inform EOSC strategy on the provision of education to the new generation of researchers and other professional groups, and the need for data literacy among the broader public who are among the users of scientific research. As originally proposed, the WP intended to collaborate with the EDISON project and its network of 'champion' universities on these areas. That focus changed due to personnel changes, but the need and motivation remains to sustain the impact of EDISON, and particularly its network of universities engaged in data science curriculum development. We will liaise with EOSC-hub to address this need.

BIBLIOGRAPHY

[1] European Open Science Cloud, "EOSC Declaration," 26 10 2017. [Online]. Available: https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf.

[2] EOSCpilot WP7, "D7.1: Skills landscape analysis and competence model," 30 06 2017. [Online]. Available: https://eoscpilot.eu/sites/default/files/eoscpilot-d7.1.pdf. [Accessed 22 12 2017].

[3] A. B. T. W. Yuri Demchenko, "EDISON Data Science Framework: Part 1. Data Science Competence Framework," 03 07 2017. [Online]. Available: http://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_cf-ds-release2-v08_0.pdf. [Accessed 22 12 2017].

[4] FitSM, "FitSM-6: Assessment," 2017. [Online]. Available: http://fitsm.itemo.org/fitsm/6. [Accessed 22 12 2017].

[5] M. Brus, "Joint EOSC discussion session," 31 10 2017. [Online]. Available: https://envriplus.manageprojects.com/s/notebook/6nPzzuSGbmW2C/page/881. [Accessed 22 12 2017].

[6] EOSCpilot WP7, "EOSCpilot Wp7 LandscapeAnalysis (Responses)," 12 2017. [Online]. Available: https://docs.google.com/spreadsheets/d/1n-zoMk4EJ4ZMxU-oB89wlMb7QofxZyCQrI88a2ll47U/edit?usp=sharing. [Accessed 22 12 2017].

[7] "GO-FAIR Initiative," [Online]. Available: https://www.dtls.nl/fair-data/go-fair/. [Accessed 22 12 2017].

[8] EOSCpilot WP7, "EOSCpilot Wp7 Task 7.2.1.2 Metadata," 22 12 2017. [Online]. Available: https://docs.google.com/spreadsheets/d/1bmdsfiPl8lW613bWv6PKmyzq4NRvuAhbzFY_-Pflgzw/edit?usp=sharing. [Accessed 22 12 2017].

[9] E. WP7, "EOSCpilot Wp7 Task 7.2.2.2 Training Material Catalogue," 20 11 2017. [Online]. Available: https://docs.google.com/spreadsheets/d/1hYnZUdqH5PF-3NB3LuCFNXnn1VH-OmhZ7fz_Igm_h94/edit?usp=sharing. [Accessed 22 12 2017].

[10] Research Data Alliance, "Catalog," 16 06 2017. [Online]. Available: https://smw-rda.esc.rzg.mpg.de/index.php?title=Catalog. [Accessed 22 12 2017].

[11] Research Data Alliance, "Registry," 23 03 2015. [Online]. Available: https://smw-rda.esc.rzg.mpg.de/index.php?title=Registry. [Accessed 22 12 2017].

[12] "About SWORD," [Online]. Available: http://swordapp.org/about/. [Accessed 22 12 2017].

[13] N. Jefferies, "SWORD V3 First Drafts," 03 11 2017. [Online]. Available: http://swordapp.org/2017/11/sword-v3-first-drafts/. [Accessed 22 12 2017].

[14] H. V. d. S. M. N. S. W. Carl Lagoze, "The Open Archives Initiative Protocol for Metadata Harvesting," 08 01 2015. [Online]. Available: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm. [Accessed 22 12 2017].

[15] zenodo, "General Policies," [Online]. Available: http://about.zenodo.org/policies/. [Accessed 22 12 2017].

[16] NISO, "ResourceSync Framework Specification (ANSI/NISO Z39.99-2017)," 02 02 2017. [Online]. Available: https://www.openarchives.org/rs/1.1/resourcesync. [Accessed 22 12 2017].

[17] H. Schmeer, “Digital Humanities Registry - Courses,” 2017. [Online]. Available: https://registries.clarin-dariah.eu/courses/. [Accessed 22 12 2017].

[18] elixir, “TeSS (Training eSupport System),” 2017. [Online]. Available: https://tess.elixir-europe.org. [Accessed 22 12 2017].

[19] EMBL-EBI, “Train online,” 2017. [Online]. Available: https://www.ebi.ac.uk/training/online/. [Accessed 22 12 2017].

[20] FOSTER consortium, “FOSTER,” 2017. [Online]. Available: https://www.fosteropenscience.eu/. [Accessed 22 12 2017].

[21] Up to University, “Up to University,” 2017. [Online]. Available: https://up2university.eu/up2universe/. [Accessed 22 12 2017].

[22] DataONE, “DataONE,” 2017. [Online]. Available: https://www.dataone.org. [Accessed 22 12 2017].

[23] zenodo, “Zenodo - Research. Shared.,” 2017. [Online]. Available: https://zenodo.org. [Accessed 22 12 2017].

[24] EGI, “Training Infrastructure,” 2017. [Online]. Available: https://www.egi.eu/services/training-infrastructure/. [Accessed 22 12 2017].

[25] G. project, “GÉANT Testbed Service,” 2017. [Online]. Available: https://gts4.geant.net. [Accessed 22 12 2017].

[26] CERN, “The SWAN Service | SWAN,” 2017. [Online]. Available: https://swan.web.cern.ch. [Accessed 22 12 2017].

[27] EOSCpilot WP7, “Training Infrastructure Evaluation,” 23 11 2017. [Online]. Available: https://drive.google.com/file/d/14z4fy6IUJwSLlow8EkjRRJmcYhJYnr_q/view?usp=sharing. [Accessed 22 12 2017].

[28] Dash, “Dash - Data sharing made easy,” [Online]. Available: http://cdluc3.github.io/dash/. [Accessed 22 12 2017].

[29] Zenodo, “Principles,” 2017. [Online]. Available: http://about.zenodo.org/principles/. [Accessed 22 12 2017].

[30] EGI.eu, IASA, “Cloud Marketplace,” 2017. [Online]. Available: https://appdb.egi.eu/browse/cloud. [Accessed 22 12 2017].

[31] I. EGI.eu, “VO,” 2017. [Online]. Available: https://appdb.egi.eu/store/vo/training.egi.eu. [Accessed 22 12 2017].

[32] EGI.eu, “Dashboard,” 2017. [Online]. Available: https://dashboard.appdb.egi.eu/vmops. [Accessed 22 12 2017].

[33] EOSCpilot WP2, “D2.2: Draft Governance Framework For the European Open Science Cloud,” 23 11 2017. [Online]. Available: http://eoscpilot.eu/sites/default/files/eoscpilot-d2.2.pdf. [Accessed 22 12 2017].

[34] European Open Science Cloud Working Group, “Report on the governance and financial schemes for the European Open Science Cloud,” 04 05 2017. [Online]. Available: https://ec.europa.eu/research/openscience/pdf/ospp_euro_open_science_cloud_report-.pdf. [Accessed 22 12 2017].

[35] Open Science Commons, “Open Science Commons – Enabling collaboration between e-Infrastructures and research communities,” 2017. [Online]. Available: https://www.opensciencecommons.org. [Accessed 22 12 2017].

[36] A. Whyte, “EOSCpilot Data Skills: what it is about and how it will help research, academia and industry?,” 29 05 2017. [Online]. Available: https://eoscpilot.eu/eoscpilot-data-skills-what-it-about-and-how-it-will-help-research-academia-and-industry. [Accessed 22 12 2017].

ANNEX A. EOSCPILOT-OPENAIRE JOINT WORKSHOP RESOURCES

WP7 organised a joint workshop at the EOSCpilot-OpenAIRE event in Berlin on 24 October 2017 to consult stakeholders on stewardship competences most needed to ensure FAIR outputs. Following is the draft list of competences, showing the responses by participants identifying gaps and revisions.

A.1. Consolidated list of competences with revisions

Revisions are indicated below in green text			Competences seen as biggest gaps (number of sticky notes)		
	Level 1	Level 2	Individual	Team	Organisation
	Govern and assess	Research strategy/ open research potential	2		2
		Data policy, legal and funder requirements		1	5
		Research reproducibility	2	2	
		FAIR & Open Access policy	1	2	2
		Research ethics and integrity		1	
		IPR management, licensing			2
		Information security and risk management			1
		Data governance, handling third-party data			2
		Storage security management			
		Data quality management			1
	Scope and resource	Business case & plan			
		Personnel and skills development		1	3
		Project management			
		Requirements management	1	1	
		Service level management			1
		Change management			1
		Repository set-up and management			1
		Workflow set-up and management	1	1	1
		Storage of large data volumes			1
		Preservation planning		1	1
		Costing of data mgmt and preservation			4

	Advise and enable	Building collaboration (cross-sector)	1	1	1
		Social interaction & negotiation			1
		Tools& domain standards awareness	1		6
		User support and training			4
		Personnel development			1
		Standards body participation			
		Data rescue			

	Plan and design	Research design			
		Data management planning	2	1	
		Open data model development		1	
		Metadata, persistent id. specification	2	1	1
		Research software requirements			
		Application design			
		Architecture design			
		Database specification			
		Database design			
		Service requirements identification		1	
		File format specification			
		Repository options and selection	1		
	Capture and process	Study set-up			
		Workflow set up and documentation		1	
		Database management			
		Software prototyping			
		Data collection			
		Data/software documentation for reuse	1	2	
		Data cleaning/ wrangling		1	1
		Data processing workflow application			
		File naming and organisation			
		Data and software versioning			

	Integrate and analyse	Maths and statistical knowledge	1	1	
		Critical thinking and theory building		1	
		Creative problem solving			
		Software component integration			1
		Data preparation and write-up			
		Data transformation and integration			1
		Analytic workflow application			
		Data mining	1		
		Database querying			
		Predictive modelling and analytics		1	
		Machine learning application	1		
		Data interpretation			

	Appraise and preserve	Data quality assessment/ assurance		1	1
		Data provenance		1	
		Data review and appraisal			1
		Format and media migration			
		Data transfer			
		Software curation		1	
	Publish and release	Workflow documentation		1	
		Access control and management		1	
		Data and software licensing		1	1
		Data and software marketing in the EOSC *	2		1
	Expose and discover	Vocabulary/ ontology application	1		1
		Metadata and persistent id. exposure			
		Presentation of data collections			
		Visualisation of research results	1		
		Repository/ database evaluation			
		Repository/ database searching			
		Data and software citation	3		

* looking at data from a user perspective

A.2. Summary of responses

Skills Group	No. of competences				No. posts at each organisational level			Total
	proposed	1+ posts	new	revised	individual	team	organisation	
Plan and design	12	5	0	3	5	4	1	10
Capture and process	10	3	0	1	1	4	1	6
Integrate and analyse	12	7	0	1	3	3	2	8
Appraise and preserve	6	4	1	1	0	3	2	5
Publish and release	4	4	0	1	2	3	2	7
Expose and discover	7	3	0	0	5	0	1	6
Govern and assess	10	9	1	4	5	8	15	28
Scope and resource	10	8	1	2	2	3	10	15
Advise and enable	7	5	0	3	2	2	16	20
Total	78	48	3	16	25	30	50	105

ANNEX B. ISSUES REPORTED BY FIRST PHASE SCIENCE DEMONSTRATORS

DPHEP

- re-scoping of data volumes to focus on provisioning a range of data to present a more realistic use case than initially proposed in the workplan and focus on the challenging parts including handling of data itself, i.e. handling of data types and metadata[4]
- evaluation by the Science Demonstrator has shown that Trustworthy Digital Repositories have specific needs with regard to a community in terms of data types, volume, conversion etc. that need to be considered for a production solution[5]
    - time consuming and tough task to handle long term data preservation in a generic way with domain-specific data formats and metadata
    - not only technical issues but also policies
    - EOSC engineers need to cope with these issues to make this demonstrator a medium or long-term solution in a production environment

Photon-Neutron

- adaptation of existing software and algorithms to meet the requirements of cloud environments[6]
- improved docker usage in science community via simulation pipelines might require relevant training materials/events[7,8]

TEXTCROWD

- Current status: In general software development within TEXTCROWD is on schedule. A working demo will be made available for the cloud. A beta version of the demonstrator is expected to be available in late October/early November.
- In the previous period people from TEXTCROWD came together to refine the terminological tool. A virtual meeting was held with D4Science team of CNR for the deployment of TEXTCROWD planned for next period.
- TEXTCROWD will create corpora of manually annotated Italian archaeology reports to fill the gap of no available corpora.
- TEXTCROWD requires no plan for management and handling of sensitive data because it works with open and publicly available textual documents.
- Possible use of TEXTCROWD: researchers in conservation and restoration, the Italian Ministry of Culture (MIBACT) to manage and index text reports, other potential research teams within EOSCpilot framework (e.g. “Visual media” pilot). TEXTCROWD will be considered as a service in the forthcoming cloud-based version of the ARIADNE portal.
- TEXTCROWD is an innovative tool: neither training resources nor tutorials are available to date for working with it. Instructions on how to use the tool will provided together with the tool itself once deployed in the cloud.

Pan-Cancer

- Issues with utilisation of UIDs when moving data[9]

[4] https://repository.eoscpilot.eu/index.php/f/6588
[5] https://repository.eoscpilot.eu/index.php/f/7614
[6] https://repository.eoscpilot.eu/index.php/f/6590
[7] https://repository.eoscpilot.eu/index.php/f/6590
[8] https://repository.eoscpilot.eu/index.php/f/7417
[9] https://repository.eoscpilot.eu/index.php/f/6589

ANNEX C. EDUCATION & TRAINING PROVISION (PARTNERS & RI CLUSTERS)

NB. Entries have not yet been validated with the organisations concerned

Organisation	Uedin (DCC)	ATHENA	BSC	CNRS
outline	DCC provides training in research data management, including service development and data management planning. Collaborated with data services in UoE on MOOCs and online training. UoE provides a range of data science undergraduate and postgraduate courses.	The English eebsite does not provide specific information about training, so it is not clear what type of formats are adopted. The Service section lists a Technology Transfer Office and an education section which provides a link to an external masterclass offered by the National Technical University of Athens.	BSC offers different types of education activities aimed at HPC. Main categories are PATC courses (PRACE), RES (Red ES) and an Education Programme on Parallel Programming, based on CUDA a cluster-aware programming environment.	Provides doctoral programmes and postdoctoral fellowships, through a number of Institutes in data and computing domains including INS2I.
Research domains targeted	Cross-domain	Cross-domain	Cross-domain, e-Infrastructure, Environmental Sciences, Physical Sciences and Engineering	Biological and Medical, Energy, Environmental Sciences, Physical Sciences and Engineering, Social Sciences and Humanities
Target audiences (professional groups)	Researcher, Research data manager	Researcher	Researcher, Research data/ software engineer	Researcher
Skills/ competence areas covered	Plan and design, Capture and process, Appraise and preserve, Publish and release, Govern and assess, Scope and resource, Advise and enable		Plan and design, Capture and process, Integrate and analyse	Plan and design, Capture and process, Integrate and analyse, Publish and release, Expose and discover

Organisation	Uedin (DCC)	ATHENA	BSC	CNRS
Competence levels covered	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)		Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)	Evaluation, synthesis (advanced)
Delivery formats used	Conference/Seminar, Course, Materials, Webinar, Workshop	Conference/Seminar, Internship/exchange / fellowship	Course, Summer School, Workshop	Post-doctoral Corses and fellowship
Duration of events	< 1 day, 1-2 days	Not applicable - no events offered	1-2 days, 1 month +	1 month +
Delivery modes	Face-to-face, Online		Face-to-face	Face-to-face

Organisation	CEA	DESY	ECRIN	EUROPEAN XFEL
outline	No information found on http://www.cea.fr/english	DESY offers opportunities for further education in the form of open events targeting public audiences and special internal courses (these are in a protected part of the website and no specific information can be found http://fortbildung.desy.de/) as well as specific events for pupils. Open seminars are listed at http://www.desy.de/news/events/index_eng.html	No information found, training provided by French National partner F-CRIN http://www.fcrin.org/en/home-formation	Offers workshops and seminars targeting scientists in the scientific disciplines covered by the organisation https://www.xfel.eu/news_and_events/events/index_eng.html
Research domains targeted		Physical Sciences and Engineering		Biological and Medical, Cross-domain, Energy, Physical Sciences and Engineering

Organisation	CEA	DESY	ECRIN	EUROPEAN XFEL
Target audiences (professional groups)		Researcher		Researcher
Skills/ competence areas covered				Advise and enable
Competence levels covered		Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)		Comprehension (beginner/awareness level), Application (intermediate)
Delivery formats used		Conference/Seminar, Course		Conference/Seminar, Workshop
Duration of events		< 1 day, 1-2 days		< 1 day, 1-2 days
Delivery modes		Face-to-face		Face-to-face

Organisation	GEANT	ICOS - ERIC	INAF	INFN
outline	The geant.org website provides information on the TRANSITS course for security incident response teams, these come in two flavours, a basic and an advanced course. A new website was recently launched https://learning.geant.org/ offering a wider number of learning opportunities. Announced are a course on Security and Identity Management with OAUTH2 – OIDC and one on Software Code Management. GEANT provides also courses and webinars on infrastructure and services for internal project participants, but these are not listed on the public websites.	ICOS does not mention training or skills on the website but capacity building is mentioned on slide 12 of a presentation https://www.icos-ri.eu/materials which gives access to a number of keynotes presented at ICOS conferences.	INAF provides an educational website (in Italian) http://edu.inaf.it/ offering access to courses, educational resources and information. The courses section provides bullets on f2f and MOODLE. No courses are announced at the time of this review.	The INFN website provide access to a large number of conferences, seminars and events (including international schools) organised by affiliated institutions in Italy.
Research domains targeted	Cross-domain, e-Infrastructure	Environmental Sciences	Physical Sciences and Engineering	Energy, Physical Sciences and Engineering
Target audiences (professional groups)	Researcher, Research data/ software engineer	Researcher		Researcher, Research data/ software engineer
Skills/ competence areas covered	Plan and design, Scope and resource, Advise and enable	Advise and enable	Advise and enable	Advise and enable
Competence levels covered	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level)	Comprehension (beginner/awareness level)	Comprehension (beginner/awareness level), Application (intermediate)

Organisation	GEANT	ICOS - ERIC	INAF	INFN
Delivery formats used	Course, Webinar, Workshop	Keynote presentations	Course, MOOC	Conference/Seminar, Workshop
Duration of events	< 1 day, 1-2 days, 3-5 days	< 1 day	1-2 days	< 1 day, 1-2 days, 3-5 days
Delivery modes	Face-to-face, Online	Face-to-face, Online	Face-to-face, Online	Face-to-face

Organisation	INGV	LIBER	MPG	PRACE
outline	The INGV site http://www.ingv.it/it/ mentions that the institute organises frequently conferences and seminars at various locations in Italy, but there don't seem to be recent events. The education section (Formazione) focuses on dissemination and awareness raising at school and general public level.	LIBER supports building capacity, development of skills and adoption of best practice within the community of European research libraries. Through activities coordinated by working groups or in the context of EU funded projects, LIBER supports the organisation of webinars and workshops on themes that are relevant to its priorities in relation to Scientific Information Infrastructures, RDM, TDM, copyright etc.	The Max Planck Society provides training for junior scientists and doctoral students, in close cooperation with the universities through which the doctorates are officially awarded. The International Max Planck Research Schools (IMPRS) supports Ph.D. students in such areas as biology and medicine, chemistry, physics, humanities and social sciences.	PRACE provides well structured information about training activity on its website. The PRACE training portal provides information about corses, tutorials and materials. In addition to a list of upcoming events the organisation maintains access to a Summer of HPC section https://summerofhpc.prace-ri.eu/
Research domains targeted	Environmental Sciences	Cross-domain, Social Sciences and Humanities	Biological and Medical, Energy, Environmental Sciences, Material Sciences and Analytic Facilities, Physical Sciences and Engineering, Social Sciences and Humanities	e-Infrastructure

Organisation	INGV	LIBER	MPG	PRACE
Target audiences (professional groups)		Researcher, Research data manager	Researcher	Research data scientist, Research data/ software engineer
Skills/ competence areas covered		Appraise and preserve, Publish and release, Advise and enable	All	
Competence levels covered	Comprehension (beginner/awareness level)	Comprehension (beginner/awareness level), Application (intermediate)	Evaluation, synthesis (advanced)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)
Delivery formats used	Conference/Seminar	Conference/Seminar, Webinar, Workshop	Summer School, Bachelor, Doctoral, Post-doc courses	Conference/Seminar, Course, Materials e.g. self-paced learning guide, MOOC, Summer School, Workshop
Duration of events		< 1 day, 1-2 days	1 month +	< 1 day, 1-2 days, 3-5 days, > 1 week < 1 month
Delivery modes		Face-to-face, Online	Face-to-face	Face-to-face, Online

Organisation	STFC	TRUST-IT	UGOE	UNIMAN
outline	STFC provides a significant set of skills, experience and expertise. This involves STFC in-house scientists and engineers, technical support specialists, science communications and international relations experts, and also IT, finance and management teams. A section of the STFC website is devoted to skills http://www.stfc.ac.uk/skills/.	No information found	The website of the University gives access to information about faculties, courses and Phd. The University libary website www.sub.uni-goettingen.de provides access to courses as well as self-study materials. Actual information about skill courses and learning material is in German and does not seem to be very detailed.	The website of the University of Manchester has a useful search tool that provides access to courses. Specific search for data courses starting in 2018 leads to a postgraduate research programme Data Engineering, and Masters on ACS: Data and Knowledge Management, Data Science and Health Data Science
Research domains targeted	Cross-domain, Energy, Physical Sciences and Engineering			Cross-domain
Target audiences (professional groups)	Researcher, Research data/ software engineer		Researcher	Researcher, Research data scientist, Research data manager, Research data/ software engineer
Skills/ competence areas covered	Plan and design, Capture and process, Integrate and analyse, Appraise and preserve, Govern and assess, Scope and resource, Advise and enable			Plan and design, Capture and process, Integrate and analyse, Appraise and preserve, Publish and release, Expose and discover, Govern and assess, Scope and resource, Advise and enable

Organisation	STFC	TRUST-IT	UGOE	UNIMAN
Competence levels covered	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)			Application (intermediate), Evaluation, synthesis (advanced)
Delivery formats used	Course, Internship/exchange/ fellowship, Post-doc, Phd		Course, Internship/exchange / fellowship, self-study material	Course
Duration of events	> 1 week < 1 month, 1 month +			1 month +
Delivery modes	Face-to-face		Face-to-face, Online	Face-to-face

Organisation	INGV	LIBER	MPG	PRACE
outline	The INGV site http://www.ingv.it/it/ mentions that the institute organises frequently conferences and seminars at various locations in Italy, but there don't seem to be recent events. The education section (Formazione) focuses on dissemination and awareness raising at school and general public level.	LIBER supports building capacity, development of skills and adoption of best practice within the community of European research libraries. Through activities coordinated by working groups or in the context of EU funded projects, LIBER offers webinars and workshops on themes that are relevant to its priorities in relation to Scientific Information Infrastructures, RDM, TDM, copyright etc.	The Max Planck Society provides training for junior scientists and doctoral students, in close cooperation with the universities through which the doctorates are officially awarded. The International Max Planck Research Schools (IMPRS) supports Ph.D. students in such areas as biology and medicine, chemistry, physics, humanities and social sciences.	PRACE provides well structured information about training activity on its website. The PRACE training portal provides information about corses, tutorials and materials. In addition to a list of upcoming events the organisation maintains access to a Summer of HPC section https://summerofhpc.prace-ri.eu/

Organisation	INGV	LIBER	MPG	PRACE
Research domains targeted	Environmental Sciences	Cross-domain, Social Sciences and Humanities	Biological and Medical, Energy, Environmental Sciences, Material Sciences and Analytic Facilities, Physical Sciences and Engineering, Social Sciences and Humanities	e-Infrastructure
Target audiences (professional groups)		Researcher, Research data manager	Researcher	Research data scientist, Research data/ software engineer
Skills/ competence areas covered		Appraise and preserve, Publish and release, Advise and enable	All	
Competence levels covered	Comprehension (beginner/awareness level)	Comprehension (beginner/awareness level), Application (intermediate)	Evaluation, synthesis (advanced)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)
Delivery formats used	Conference/Seminar	Conference/Seminar, Webinar, Workshop	Summer School, Bachelor, Doctoral, Post-doc courses	Conference/Seminar, Course, Materials e.g. self-paced learning guide, MOOC, Summer School, Workshop
Duration of events		< 1 day, 1-2 days	1 month +	< 1 day, 1-2 days, 3-5 days, > 1 week < 1 month
Delivery modes		Face-to-face, Online	Face-to-face	Face-to-face, Online

Organisation	CORBEL	ENVRIplus	ASTERICS	EMBRIC
outline	Training addresses competency requirements of staff in BMS RIs for newly identified services and use them as the basis for a pilot training programme. Main target audience is technical operators of RIs in BMS RI hubs and nodes. Activity focuses on the four Cluster areas: data management and integration, physical access, ethics and innovation.The CORBEL website links to webinars but no one is scheduled at the moment. CORBEL partner Instruct-Eric offers internship opportunity starting autumn 2017.	A report on planned training highlights areas for development mainly relating to use of the EGI infrastructure including: Containers based applications in the EGI Federated Cloud infrastructure with Docker; Creating data federations with the EGI Open Data Platform: Security incident handling, methods and forensics. 6. One of the six themes of the ENVRIplus cluster is Data for Science, lead by UvA, but no specific information about skills is provided at http://www.envriplus.eu/themes/t2/	Three of the main strands of activity are Dissemination, Engagement and Citizen Science, includes aim of attracting young people via citizen science initiatives and through open educational resources. (OBservatory E-environments LInked by common ChallengeS - holds workshops to train researchers and data scientists. Data Access, Discovery and Interoperability) includes training and support about the Virtual Observatory framework including hands-on exercises, which will allow participants to become fully familiar with the VO capabilities on their own laptops.	The cluster provides a training catalogue offering "a non exhaustive overview of existing trainings " The catalogue can be filtered by 'field of education' (ISCED classification), Free text search on 'data' produces 200+ results, 'data management' 100+ of which 7 in 2017. These include workshops and courses of various levels including Masters and doctoral, and information on research placements / internships. The latter include contributions to a biodiversity data portal. A catalogue of video resources is provided, some of which have 'data' in the title, although none appear to focus on data or software management specifically.
Research domains targeted	Biological and Medical	Environmental Sciences	e-Infrastructure, Physical Sciences and Engineering	Biological and Medical, Cross-domain, Environmental Sciences
Target audiences (professional groups)	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data/ software engineer	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data scientist, Research data manager, Research data/ software engineer

Organisation	CORBEL	ENVRIplus	ASTERICS	EMBRIC
Skills/ competence areas covered		Advise and enable	Plan and design, Capture and process, Integrate and analyse, Publish and release, Expose and discover, Scope and resource, Advise and enable	Plan and design, Capture and process, Integrate and analyse, Appraise and preserve, Publish and release, Govern and assess
Competence levels covered	Application (intermediate), Evaluation, synthesis (advanced)	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)	Application (intermediate), Evaluation, synthesis (advanced)
Delivery formats used	Course, Internship/exchange/ fellowship, Webinar	Workshop	Conference/Seminar, Materials, Summer School, Workshop	Conference/Seminar, Course, Internship/exchange/ fellowship, Materials, Summer School, Workshop
Duration of events	< 1 day, 1-2 days, 3-5 days, 1 month +	1-2 days	< 1 day, 1-2 days, 3-5 days	< 1 day, 1-2 days, 3-5 days, > 1 week < 1 month, 1 month +
Delivery modes		Face-to-face	Face-to-face, Online	Face-to-face, Online, Embedded

Organisation	SINE2020	DANS - KNAW	BBMRI – ERIC	CINECA
outline	Education and Training Activities are on two fronts: 1) e-learning : freely accessible, e-learning modules with online experiment simulation sessions; 2) Schools : financially supporting neutron scattering schools, organised by European facilities and research institutions, which generally include hands-on sessions . The e-learning is offered through a dedicated platform; e-neutrons.org ;. Schools are both introductory and advanced. Data management challenges for neutron facilities user communities are reported in a blog	Data Archiving and Networked Services (DANS) supports skills development with respect to EOSCpilot with community workshops, webinars, and summer schools; and spanning across EUDAT, OpenAIRE, CESSDA, CLARIAH, DARIAH, EHRI, and PARTHENOS. DANS has defined users for its programmes, i.e., Essentials 4 data support:.).	Training is not included in the BBMRI-ERIC service portfolio (http://www.bbmri-eric.eu/bbmri-eric-services/), however BBMRI-ERIC is the main organiser of the annual Global Biobank Week events and these often include workshops that are skill development / training events. BBMRI-ERIC is coordinator of the H2020 RItrain project (http://ritrain.eu/) which recently launched the 'Executive Masters in Management of Research Infrastructures',	CINECA is s one of the main Data Management infrastructure in Italy and Europe, providing data services to several research communities,. CINECA does not have training or skills development explicitly in its service portfolio, however a training portal (http://www.hpc.cineca.it/content/training) which serves as a repository of training materials, events and online courses. The resources here all relate to HPC, The resources from the CINECA SCAI training portal are federated into the PRACE training portal (http://www.prace-ri.eu/trainings/),
Research domains targeted	Cross-domain, Material Sciences and Analytic Facilities	Biological and Medical, Cross-domain, e-Infrastructure, Environmental Sciences, Social Sciences and Humanities	Biological and Medical	Cross-domain
Target audiences (professional groups)	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data scientist, Research data/ software engineer	Researcher, Research data scientist, Research data/ software engineer

Organisation	SINE2020	DANS - KNAW	BBMRI – ERIC	CINECA
Skills/ competence areas covered	Capture and process, Integrate and analyse	Plan and design, Capture and process, Appraise and preserve, Expose and discover, Govern and assess, Scope and resource, Advise and enable	Plan and design, Capture and process, Integrate and analyse, Govern and assess, Advise and enable	Plan and design, Capture and process, Integrate and analyse, Expose and discover, Govern and assess, Scope and resource
Competence levels covered	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)
Delivery formats used	Course, Internship/exchange/ fellowship, Materials, Workshop	Conference/Seminar, Course, Webinar, Workshop	Conference/Seminar, Webinar	Course, School
Duration of events	< 1 day, 1-2 days, 3-5 days	< 1 day, 1-2 days, > 1 week < 1 month	1-2 days, 3-5 days	1-2 days, 3-5 days
Delivery modes	Face-to-face, Online, Embedded	Face-to-face, Online, Blended - online and f2f	Face-to-face	Face-to-face

Organisation	CNR	KIT	SURFsara	JISC
outline	Supports postgraduate studies and research training. The training office promotes, revises and enhances staff professional development by continuous training. In addition, the Cnr S&T Digital Library (ensures permanent, certified and effective access to information resources, scientific and technical data, expertise, research activities and programmes. CNR also provides events and workshops in Italian language.	KIT, in particular SCC as the contributing project in EOSCpilot, offers training via a dedicated Summer School, the GridKa School. The school provides a forum for scientists and technology leaders, experts and novices to facilitate knowledge sharing and information exchange. The target audience are different groups like graduate and PhD students, advanced users as well as IT administrators. The school is composed of topical presentations and hands-on workshops. The materials are accessible by the public.	SURFsara provides training and support including * Collaboratorium: supporting collaboration with presentation and visualisation aids * Consultancy: expert advisors ensure perfect ICT solution for research projects * SURFacademy: intensive knowledge and skills-transfer opportunities on specific current ICT issues * Tendering: support for tendering issues * Training courses: Hands-on systems training courses for researchers and research supporters	Jisc provides learning and research resources on a broad range of topics for further education, digital collections and e-learning. Further guides are provided on on e.g. research data management, education sector data and analytics, digital literacy, open access. Sharing of best practice in research data management is promoted through a Research Data Network focused on the UK sector. Training events cover a number of specific topics e.g. discoverability of digital collection, connectivity, cloud, cyber security as well as on trust and identity for higher and further education.
Research domains targeted	Biological and Medical, Environmental Sciences, Material Sciences and Analytic Facilities, Physical Sciences and Engineering, Social Sciences and Humanities	Cross-domain	Cross-domain	Cross-domain

Organisation	CNR	KIT	SURFsara	JISC
Target audiences (professional groups)	Researcher	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data scientist, Research data manager, Research data/ software engineer
Skills/ competence areas covered	Plan and design, Capture and process, Integrate and analyse	All	All	All
Competence levels covered	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)
Delivery formats used	Workshop, Internal training of staff	School	Conference/Seminar, Course, School, Webinar, Workshop	Conference/Seminar, Course, Materials e.g. self-paced learning guide, Webinar
Duration of events	1-2 days, 3-5 days	3-5 days	< 1 day, 1-2 days, 3-5 days	< 1 day, 1-2 days, 3-5 days
Delivery modes	Face-to-face	Face-to-face	Face-to-face, Online	Face-to-face, Online

Organisation	CSC - IT	EMBL	ESS
outline	CSC offers versatile and high-quality training in scientific computing, data networking and data management. Some courses target on bioinformatics as the research domain. Most of the courses are applicable cross-domain. Courses offered by CSC are offered as public intensive courses, workshops and other events. CSC supports university education by sharing training material; produces training content that can be incorporated into larger courses; maintains well-equipped training facilities, and participates in European training collaborations and networks.	Scientific training activities at EMBL are coordinated by the International Centre for Advanced training (EICAT) EMBL further provides a Scientific Visitor Programme to enable collaboration with EMBL staff or pursuing own scientific studies. Further collaboration is provided through a Collaborative Training Programme for PhDs Training for schools for example to enable professional development of secondary school biology teachers. EMBL provides a broad spectrum of training materials and opportunities in the context of bioinformatics. Courses range from beginner to intermediate and cover face-to-face events as well as online courses.	ESS interaction with future users and potential employees is provided through education and collaboration with NNSP in the EU-funded BrightnESS project. The first event in this direction was the first Swedish-Nordic-Baltic Summer School on Neutron Scattering that took place over two weeks in September 2017. The school focuses on introducing of PhD students to domain-specific content and applications in life science, energy, quantum materials; as well as in engineering and industrial R&D.
Research domains targeted	Biological and Medical, Cross-domain	Biological and Medical, Cross-domain	Physical Sciences and Engineering
Target audiences (professional groups)	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data scientist, Research data manager, Research data/ software engineer	Researcher, Research data scientist, Research data/ software engineer
Skills/ competence areas covered	All	All	Plan and design, Capture and process, Integrate and analyse
Competence levels covered	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level)

Organisation	CSC - IT	EMBL	ESS
Delivery formats used	Conference/Seminar, Course, Materials e.g. self-paced learning guide, Webinar, Workshop	Conference/Seminar, Course, Workshop	School
Duration of events	< 1 day, 1-2 days, 3-5 days	< 1 day, 1-2 days, 3-5 days, 1 month +	3-5 days
Delivery modes	Face-to-face, Online	Face-to-face, Online	Face-to-face

Organisation	SERISS	PARTHENOS
outline	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.	The PARTHENOS Training Plan is targeted at users of digital humanities research infrastructure. 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. a weeklong summer school programme), brochures and other teaching resources.
Research domains targeted	e-Infrastructure, Social Sciences and Humanities	Cross-domain, e-Infrastructure, Social Sciences and Humanities
Target audiences (professional groups)	Researcher, Research data scientist, Research data manager	Researcher, Research data scientist, Research data manager, Research data/ software engineer
Skills/ competence areas covered	Plan and design, Capture and process, Integrate and analyse, Appraise and preserve, Publish and release, Govern and assess, Advise and enable	Capture and process, Integrate and analyse, Appraise and preserve, Publish and release, Govern and assess, Scope and resource, Advise and enable
Competence levels covered	Comprehension (beginner/awareness level), Application (intermediate)	Comprehension (beginner/awareness level), Application (intermediate), Evaluation, synthesis (advanced)
Delivery formats used	Course, Materials e.g. self-paced learning guide	Conference/Seminar, Course, Materials e.g. self-paced learning guide, School, Webinar, Workshop

Organisation	SERISS	PARTHENOS
Duration of events	1-2 days	< 1 day, 1-2 days
Delivery modes	Face-to-face, Online	Face-to-face, Online

ANNEX D. USER REQUIREMENTS

Term	Explanation
EOSC user	• find relevant training materials and events based on its title • find relevant training materials and events based on free text search • find relevant training materials and events based on the EOSCppilot Skills Framework • filter relevant training materials and events related to a given research domain • find relevant training events based on a specific date or location • find relevant training materials and events based on topicality • access relevant training materials and events based on searches
Organisation or team	• identify criteria for relevant training events based on the EOSCpilot Skills Framework • orchestrate relevant training materials to organize and deliver a training event • find relevant training materials and events for a team of EOSC users • commission a training event by requesting specific training materials, trainers and/or venues
Trainer	• find relevant training materials for reuse with a specific license • provide a proper attribution to referenced/used training materials • deliver training materials based on demands by EOSC users, teams or organisations • be informed when referenced/used training materials are updated • adapt training materials based on feedback by users
Training curator	• identify frequently and infrequently used training materials • identify outdated training materials • identify training materials that are unavailable or not relevant in the context of EOSC • integrate external training materials and events • identify training materials that need to be adapted based on feedback by users
Service	• access training materials and events by a public application programming interface (API) • collect and aggregate metadata on training materials and events • ingest training materials and events from external sources • expose descriptions/annotations on training materials and events • get information about (external) changes of training materials and events

ANNEX E. EVALUATION CRITERIA FOR TRAINING INFRASTRUCTURES

Category	Layer	Criterion	Definition
Delivery	Presentation and Delivery	Open data formats	Accessibility (impaired people) and possibility for execution
		Certification support	Badges, certificate for participation
		Feedback support	Evaluation of training and trainers
		Scalability and elasticity	On-demand provisioning of required infrastructure and software
Visualisation and interface	Trainers and venues	See training catalogue/registry	See training catalogue/registry
	Training catalogue/ registry	Metadata support	Support to search and access metadata
		Open data formats	Support to search and access data
		License support	Support to search and access licenses
		Certification support	Support to filter for certified resources
		Persistent identifiers	Support to search and access via persistent identifiers
		Version control	Support to access different versions
		Reference support	Support to search and access referenced resources
		Standards and vocabulary	Support to search and access utilised vocabulary
		Feedback support	Support to filter and access feedback
		Scalability and elasticity	Support of standard protocols for service extension
Resources	Training resources inclusive trainers and venues	Metadata support	Support to store metadata per resource
		Open data formats	Support for composition from open data formats
		License support	Support for storing license information
		Certification support	Support for certified resources
		Persistent identifiers	Support for creation of persistent identifiers
		Version control	Support to track versions of resources

		Reference support	Support to assign references to other resources
		Standards and vocabulary	Utilisation of thesauri and standard vocabulary
		Scalability and elasticity	Curation and management efforts
Infrastructure	Storage, compute, network	Metadata support	Support for operational metadata including creation, editing, last access date
		Certification support	Support for certified storage, QoS (availability, network, …)
		Version control	Support for version control of data
		Reference support	Support for reference to external data
		Scalability and elasticity	On-demand provisioning of infrastructure

ANNEX F. TOWARDS FAIRNESS OF TRAINING PROVISION

- Findability
  - F1: (meta)data are assigned a globally unique and persistent identifier
    - The creation of each training material or event is delegated to a service that ensures issuing of DOIs, e.g. Zenodo
  - F2: data are described with rich metadata (defined by R1 below)
    - We recommend the implementation of the minimum set of metadata given in Section 3.2.3 for training resource providers to enable proper harvesting. Based on this we provide a metadata harvesting service to aggregate training resources from different content providers, ensure harmonisation as well as further enrichments of metadata based on the EOSCpilot Competence Framework.
  - F3: metadata clearly and explicitly include the identifier of the data it describes
    - When training resources are published via the TaaS we ensure the DOI to be included in the metadata. Furthermore, our minimal set of metadata ensures the persistent identifier to be a recommended field (compare Section 3.2.3).
  - F4: (meta)data are registered or indexed in a searchable resource
    - Metadata of each training resource is indexed and searchable directly in the TaaS infrastructure immediately after publishing and harvesting
- Accessibility
  - A1: (meta)data are retrievable by their identifier using a standardized communications protocol
    - Metadata for individual training resources as well as resource collections are harvestable using the OAI-PMH protocol by the record identifier and the collection name
    - Metadata is retrievable through annotated contents
    - Metadata is also retrievable through a public REST API
  - A1.1: the protocol is open, free, and universally implementable
    - See point A1. OAI-PMH and REST are open, free and universal protocols for information retrieval
  - A1.2: the protocol allows for an authentication and authorization procedure, where necessary
    - Metadata are publicly accessible and licensed under public domain. No authorisation is necessary to retrieve it.
  - A2: metadata are accessible, even when the data are no longer available
    - The infrastructure utilises a repository for publishing that ensures long-term availability of data and metadata for harvesting.
    - Harvested metadata are stored in high-availability database servers which are separate from the data itself.
- Interoperability
  - I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
    - Metadata are stored utilising the JSON Schema as internal representation and offers export to other popular formats such as Dublin Core.
  - I2: (meta)data use vocabularies that follow FAIR principles
    - For certain terms we refer to open, external vocabularies (see Section 3.2.3). However, further consideration is needed to ensure applicability in a multi-disciplinary environment.
  - I3: (meta)data include qualified references to other (meta)data
    - Each referenced resource or collection is qualified by a resolvable URL
- Reusability
  - R1: (meta)data are richly described with a plurality of accurate and relevant attributes

        - Each record contains a minimum of our proposed mandatory metadata with optionally additional recommended terms and further enrichments
    - R1.1: (meta)data are released with a clear and accessible data usage license
        - License is one of the recommended terms in our proposed metadata
        - The external training resource providers ensures that data downloaded by the user is subject to the license specified in the metadata.
    - R1.2: (meta)data are associated with detailed provenance
        - All data published via the TaaS is traceable to a registered user
        - Metadata contain mandatory information about the original author of the creative work
    - R1.3: (meta)data meet domain-relevant community standards
        - EOSC TaaS does not implement a domain-specific solution for provisioning and delivery of training and targets cross-domain applicability.
        - Metadata contain recommended information on the domain of training resources and further enable the supplementation with additional domain-specific terms.

ANNEX G. GLOSSARY

Term	Explanation
AAI	See *Authorisation and Authentication Infrastructure.*
API	See *Application programming interface.*
Application programming interface	Subroutine implementation or protocol specification to communicate and interact with a service, application, or library.
Authorisation and Authentication Infrastructure	Authorisation and authentication infrastructures are middleware systems consisting of a set of protocols allowing for delegation of authentication and authorization issues to different instances, commonly the user's organisation.
Harvesting	Harvesting refers to gathering metadata from a number of distributed content providers or repositories into a combined data store.
Javascript Object Notation	Human readable data format for structured text, numbers and boolean data. Commonly used to transmit data to remote services.
JSON	See Javascript Object Notation.
OAI-PMH	See *Open Archives Initiative Protocol for Metadata Harvesting.*
OER	See *Open education resources.*
Open Archives Initiative Protocol for Metadata Harvesting	The Open Archives Initiative Protocol for Metadata Harvesting provides an application-independent interoperability framework based on metadata harvesting for data providers and service providers.
Open education resources	Types of open educational resources include: full courses, course materials, modules, learning objects, open textbooks, openly licensed (often streamed) videos, tests, software, and other tools, materials or techniques used to support access to knowledge. OER may be freely and openly available static resources, dynamic resources which change over time in the course of having knowledge seekiers interacting with and updating them, or a course or module with a combination of resources.[10]
Repository	A repository stores data/digital objects and provides an interface for accessing and searching. Commonly we assume a repository to provide functionality also for managing, maintaining and curating the data.
Representational state transfer	Abstract class of protocols which do not rely on state being preserved between communication attempts. Enables fault tolerance and increased scalability of services.
REST	See *Representational state transfer.*
Simple Web-service Offering Repository Deposit	The Simple Web-service Offering Repository Deposit is a lightweight protocol for depositing any contents from one location to another.

[10] https://en.wikipedia.org/wiki/Open_educational_resources

SWORD	See *Simple Web-service Offering Repository Deposit*.
TaaS	See *Training as a Service.*
Training as a Service	We define Training as a Service as an agile approach to training provision with a focus on the user. This allows the adaptation to a user's needs with regard to amount and combination of training materials.
Training catalogue	Training service which provides curated metadata of training resources, but not their content.
Training portal	Training service which provides actual training resources in conjunction with their metadata.
Training registry	Training service which provides the metadata of training resources, but not their content, by retrieving the metadata from external providers.
Training resoures	Training resources include all entities related to training including training materials, training events and also trainers as well as training venues.