

Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782

Start Date of Project: 01/01/2019

Duration: 40 months

Deliverable 6.7 Inventory of existing learning materials

Dissemination Level	PU
Due Date of Deliverable	31/12/2019, M12
Actual Submission Date	18/12/2019
Work Package	WP6 Fostering Communities, Empowering Users, & Building Expertise
Task	Task 6.3 Empowering Users: Training Materials and Online Learning Paths
Type	Report
Approval Status	
Version	V1.0
Number of Pages	p.1 – p.32

Abstract: This report gives an overview of the landscape of training material in the SSH disciplines cluster and the broader context of EOSC based on inspecting main existing sources of training materials and their content. The collected information serves as input for further activities in WP6 as well as an important source of information for WP7.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.



History

Version	Date	Reason	Revised by
0.0	10/10/2019	Started implementing the application for the curation	Matej Durco Klaus Illmayer
0.1	22/10/2019	Started curation sprint	all curators
0.2	14/11/2019	Initial outline	Matej Durco
0.3	06/12/2019	Refined outline, introduction of main topics to be present in the final report	Matej Durco
0.4	10/12/2019	Analysis section - statistics from the curation application	Klaus Illmayer
0.5	13/12/2019	Formulation of the main parts of the document (Introduction, State of the art, etc.)	Klaus Illmayer Matej Durco
0.6	17/12/2019	Conclusion, Reformulation, proof reading of individual sections	Klaus Illmayer, Matej Durco
0.7	17/12/2019	Final edits	Vicky Garnett; Vasso Kalaitzi; Martina Torma; Erzsébet Tóth-Czifra
1.0	17/12/2019	Final	Matej Durco

Author List

Organisation	Name	Contact Information
DARIAH/OEAW	Matej Ďurčo	matej.durco@oeaw.ac.at
DARIAH/OEAW	Klaus Illmayer	klaus.illmayer@oeaw.ac.at
DARIAH ERIC	Laure Barbot	laure.barbot@dariah.eu

Contributor List

Compiling list of sources and curators of sources in the inventory curation application:

Organisation	Name	Contact Information
CESSDA/UKDA	Veerle Van den Eynden	veerle@essex.ac.uk
CLARIN ERIC	Jakob Lenardič	jakob.lenardic@ff.uni-lj.si
CNR	Francesco Gabellone	francesco.gabellone@cnr.it
CNR	Alessandra Chirivi	alessandra.chirivi@cnr.it

DARIAH ERIC	Laure Barbot	laure.barbot@dariah.eu
DARIAH/OEAW	Matej Durco	matej.durco@oeaw.ac.at
DARIAH/OEAW	Klaus Illmayer	klaus.illmayer@oeaw.ac.at
KNAW	Ellen Leenarts	ellen.leenarts@dans.knaw.nl
KNAW	Elly Dijk	elly.dijk@dans.knaw.nl
UCL	Alejandra Albuerne	a.albuerne@ucl.ac.uk

Executive Summary

For planning further activities in *WP6 - Fostering Communities, Empowering Users, & Building Expertise* as well as input for the development of the SSH Open Marketplace in *WP7- Creating the SSH Open Marketplace* it is indispensable to survey the current landscape of existing training materials in the SSH cluster as well as in the broader context of EOSC. Numerous similar activities have been performed in various recent EOSC- and FAIR-related projects and these have served as partial input while compiling this inventory.

Around 30 sources of training material of various kinds and scopes have been gathered initially. These were analysed along a number of dimensions, like intended audience, relevant discipline, topics covered, extent, status etc. Additionally, for most of the sources a few individual training material items were picked and described in further detail. All this information has been collected through a dedicated Drupal-based application. The collected information is not exhausting, but it should provide a sufficient probe into the training materials landscape to inform further actions.

Particularly it reveals following findings:

- There seems to be a host of material on Research Data Management and application of FAIR principles, as well as on basic scripting skills for Data Science, i.e. working with the shell, programming with Python/R, for purposes of automatically processing data.
- We can distinguish four types of training materials: e-learning courses, carpentry-style, blog-style, static (slides, videos, PDFs)
- There are numerous sources that seem to be abandoned or in danger of becoming obsolete.
- There is little to none harmonisation with respect to metadata about the training material. This information is seldom available in structured and automatically harvestable form.

The findings imply following recommendations for further work in WP6 and WP7:

- Perform a gap analysis on the topic coverage combined with the set of topics which could be provided by the individual WPs of the project
- Reserve sufficient capacities for the manual curation of existing sources, defining metadata and communicating with the providers. Start as soon as possible.

Abbreviations and Acronyms

ACDH-OEAW	Austrian Centre for Digital Humanities - Austrian Academy of Sciences
API	Application Programming Interface
CESSDA	Consortium of European Social Science Data Archives
CMS	Content Management System
DARIAH	Digital Research Infrastructure for the Arts and Humanities
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
PARTHENOS	Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
SERISS	Synergies for Europe's Research Infrastructures in the Social Sciences
SSH	Social Sciences and Humanities
SSHOC	Social Sciences and Humanities Open Cloud
SSK	Standardization Survival Kit
TEI	Text Encoding Initiative
TeSS	Training eSupport System
WP	Work Package
XML	Extensible Markup Language

Table of Contents

1	<i>Introduction</i>	7
2	<i>State of the art / Previous work</i>	7
3	<i>Methodology</i>	9
3.1	Inventory curation application	10
3.2	Controlled vocabularies / Analysis dimensions	13
4	<i>Analysis of the training material landscape</i>	15
4.1	Statistics	15
4.1.1	By discipline	15
4.1.2	By audience	16
4.1.3	By format	17
4.2	Topics & Learning paths	18
4.3	Harvestability	20
5	<i>Use cases</i>	21
5.1	e-Learning platforms: #dariahTeach	21
5.2	Carpentry: SSK	22
5.3	Blogs: CESSDA Training, DARIAH-Campus, PARTHENOS Training	24
5.4	Slides, PDF, videos, webinars: SERISS, CLARIN Videolectures and DARIAH-Campus	25
6	<i>Input for Marketplace</i>	26
7	<i>Conclusion</i>	27
8	<i>Annex – Overview of sources</i>	28

1 Introduction

This document provides an overview of the current landscape of training materials in the social sciences and humanities disciplines, as well as in the broader context of EOOSC.

It shall serve as input for the task 6.3 - *Empowering Users: Training Materials and Online Learning Paths* aiming to provide a coherent set of training materials for the SSH community. Given the vast amount of resources already existing and the limited capacities to generate new material within the project, it is important to have a good coverage/overview of what already exists so that T6.3 can concentrate on filling the gaps. The intended approach here, as with the WP6 as a whole is to liaise with other work packages to identify topics being tackled in these and work collaboratively on developing material where still needed, also in alignment with corresponding training events.

This document will furthermore inform the curation work on the SSH Open Marketplace in WP7. By the end of the project the Marketplace is intended as the primary user-facing discovery application to find training materials and other information regarding research practices.

The task of inventorising the landscape operates in a very dynamic and heterogeneous environment with a multitude of sources of training material and numerous initiatives and projects pursuing similar goal to aggregate the information about training materials and make it available and discoverable to the potential learners.

2 State of the art / Previous work

The landscape of training materials in the broader context of EOOSC, shift towards Open Science and Data Science has been extraordinarily dynamic and heterogeneous. This has prompted a number of mapping activities, similar to the one described in this report. Following section touches upon some of the recent outcomes that have served as partial input for compiling this report.

As already stated in the EOOSCPilot deliverable *D7.5: Strategy for Sustainable Development of Skills and Capabilities*¹: “The emerging EOOSC ecosystem provides the context for focusing on developing data stewardship skills of researchers and professional groups supporting research, including data stewards themselves”. Joint effort from various actors contribute to creating the conditions for quality training and skills, and the *D6.2 Building Expertise Strategy*² deliverable of the SSHOC project provides an overview of the different training initiatives shaping the SSH and the EOOSC landscapes.

As a non-domain specific initiative gathering a diversity of actors, but also a wide range of practices and materials used for training, EOOSC has to address the way training materials will be integrated in its global architecture. So far, a dedicated “Training and Support” category is available in the services & resources section

¹ <https://eoscipilot.eu/content/d75-strategy-sustainable-development-skills-and-capabilities>

² To be published in December 2019

of the EOSC portal. The EOSCpilot projects³, with the above-mentioned deliverable provided a first attempt to determine a minimal set of properties for training materials. Important factor in this endeavour is the harmonisation and standardisation of metadata. In this respect, in the various initiatives following metadata schemas have been devised: the EOSC Datasets Minimum Information (EDMI) guideline⁴, a dedicated module for training materials in Bioschemas⁵ (a set of schemas for describing resources in the life sciences based on the schema.org standard) and Learning Resource Metadata Initiative (LRMI)⁶ as an extension of Dublin Core Metadata Initiative (DCMI)⁷.

Next to the upcoming EOSC Working Group Training & Skills, the most relevant current initiatives and activities to closely observe and coordinate with are found in the scope of OpenAire: OpenAireAdvances⁸, a project fostering application of Open Access/Open Data, and OpenAire Communities of Practices Training Network⁹, an informal network to share training experiences among people coordinating training programmes of research and e-infrastructure. It is to be noted that although OpenAire produces a range of training materials (mainly webinars and workshops, the former usually with a very generic scope, the latter usually embedded in specific national contexts), best practices for the curation and for ensuring the long-term availability of the materials are not yet fully established and topic of discussion in the CoP.

Furthermore, OpenAire's huge central catalogue¹⁰ of all kinds of research resources intended to become (part of) the central catalogue of EOSC¹¹ has to be taken into consideration. Even though it has general (not discipline-specific) scope and currently no special focus on training material, given its sheer volume and its central prospective position in the EOSC infrastructure, it is paramount to align with its developments and assess the feasibility of using it as a source of information as well as ways to contribute to it. In this respect, a relatively recent development is OpenAire allowing for and encouraging discipline or community-of-practice specific views, or dashboards, on the huge OpenAire Research Graph¹².

A whole separate area are initiatives and projects dedicated to the spreading of FAIR principles: FAIRsFAIR¹³, FAIRsharing¹⁴, FAIRassist¹⁵ as well as the dedicated EOSC working group on FAIR¹⁶ have a natural strong focus

³ <https://www.eoscpilot.eu/>

⁴ cf. EOSCpilot D6.3: 1st Report on Data Interoperability: Findability and Interoperability:
<https://www.eoscpilot.eu/sites/default/files/eoscpilot-d6.3.pdf>

⁵ <https://bioschemas.org/specifications/TrainingMaterial/>

⁶ https://www.dublincore.org/specifications/lrmi/lrmi_terms/

⁷ <https://dublincore.org/>

⁸ <https://www.openaire.eu/advance/>

⁹ <https://www.openaire.eu/cop-training>

¹⁰ <https://explore.openaire.eu/>

¹¹ <https://catalogue.eosc-portal.eu/>

¹² <https://doi.org/10.5281/zenodo.2600275>

¹³ <https://www.fairsfair.eu/>

¹⁴ <https://fairsharing.org/>

¹⁵ <https://fairassist.org/>

¹⁶ <https://www.eoscsecretariat.eu/working-groups/fair-working-group>

on training to foster the application of FAIR principles. The initiative [terms4FAIRskills](#)¹⁷ aims to build “a terminology for the skills necessary to make data FAIR and to keep it FAIR”. This terminology needs to be taken into account in further activities of WP6 as input for shaping the terminology to describe the training materials.

Finally, also to highlight is the platform Training eSupport System (TeSS)¹⁸ developed in the life sciences cluster project [eliXir](#)¹⁹. While it is not directly relevant regarding the topics covered, it represents a well-established solution for collecting and presenting for discovery both training materials and events. Confronted with a scattered heterogeneous landscape, the creators of TeSS invested substantial resources into mechanisms for collecting metadata about the resources, primarily by scraping the content of the websites of individual providers, while encouraging providers to offer the metadata as semantic tags adhering to the Bioschemas standard, directly in the HTML landing pages of the individual items. These scrapers are available for inspection and reuse on [github](#)²⁰. Though specific to the sources, they provide a valuable blueprint for similar endeavours in other domains. The approach taken by the TeSS is described in greater detail in EOSCpilot D7.5.

3 Methodology

Starting point was a hand-crafted list of sources on training material. This list was composed of a base of publicly well-known websites in the different research communities of SSHOC and extended by interviewing members of WP6. It contained very basic information on the sources and was compiled and shared internally as a spreadsheet.

This initial list served as starting point for a short intensive curation campaign. For this purpose, a dedicated curation application has been set up, and the list of 35 sources has been imported. An overview of the sources can be found in the [annex of this deliverable](#).

The inventory application is based on the content management system [Drupal 8](#)²¹, which allows for quick implementation of custom data models, including customized forms for entering the information. It is also possible to import data from spreadsheets directly into Drupal 8.

The application is currently available under <https://training-inventory.acdh-dev.oeaw.ac.at/>. While the application is only intended as a working tool for specific project task and there are currently no plans to sustain it beyond the end of the project, it has proven itself very useful for the curation task and we plan to keep it running and use it for continuous collection of information during the project, until it will become superseded by the corresponding curatorial mechanisms in the *SSH Open Marketplace*.

Six curators from the partners participating in T6.3 selected five sources each for deeper inspection and used the application to collect the required information about the sources, including a small sample of actual training material items from these sources. The selection of the sample items was at curator’s discretion, based

¹⁷ <https://terms4fairskills.github.io/>

¹⁸ <https://tess.elixir-europe.org/>

¹⁹ <https://elixir-europe.org/>

²⁰ <https://bioschemas.org/specifications/TrainingMaterial/>

²¹ <https://drupal.org>

on the principle of maximising the diversity and heterogeneity with respect to the size, the age, the type and the topics covered. There is one exception, that is the “Standardization Survival Kit” (SSK), where we performed automatic harvesting and import (that is the reason, why for the SSK 27 curated items show up). This had the two-fold motivation of collecting more data and experimenting with automatic collection of data. SSK was chosen because it is well known to the authors, has data available in a structured form and contains highly relevant material. The SSK use case is described in more detail in [section on use cases](#).

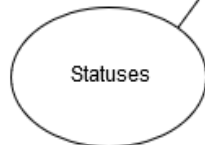
3.1 Inventory curation application

For the curation task a dedicated application has been built as a simple website based on Drupal 8 CMS, with custom data model (see [figure 1](#)). Developing a dedicated application for such a limited task, instead of using e.g. a shared google-spreadsheet may seem as unjustified effort and overengineering. However, the invested work paid off

- in significantly more efficient curation (editing forms supporting the curators with autocompletes etc.),
- resulting in better quality of the collected data (values controlled by well-defined vocabularies, possibility to assign multiple values to an item, notorious problems in spreadsheets resulting in messy data)
- through simple means to generate analyses and statistics
- as valuable fast prototype for the task of collaborative curation, which will be critical to the development and maintenance of the Marketplace in WP7.

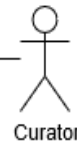
Vocabularies

used for sources only

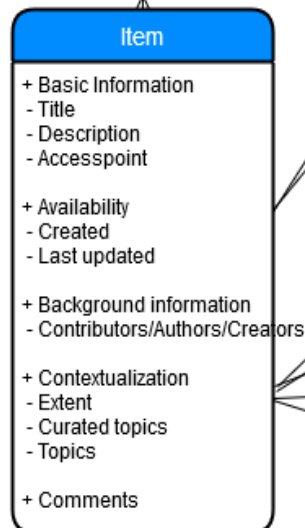


Entities

Source, Item,
Curator



hasitem



Vocabularies

used for sources and
items

Licenses

Languages

Disciplines

Audiences

Formats

Topics

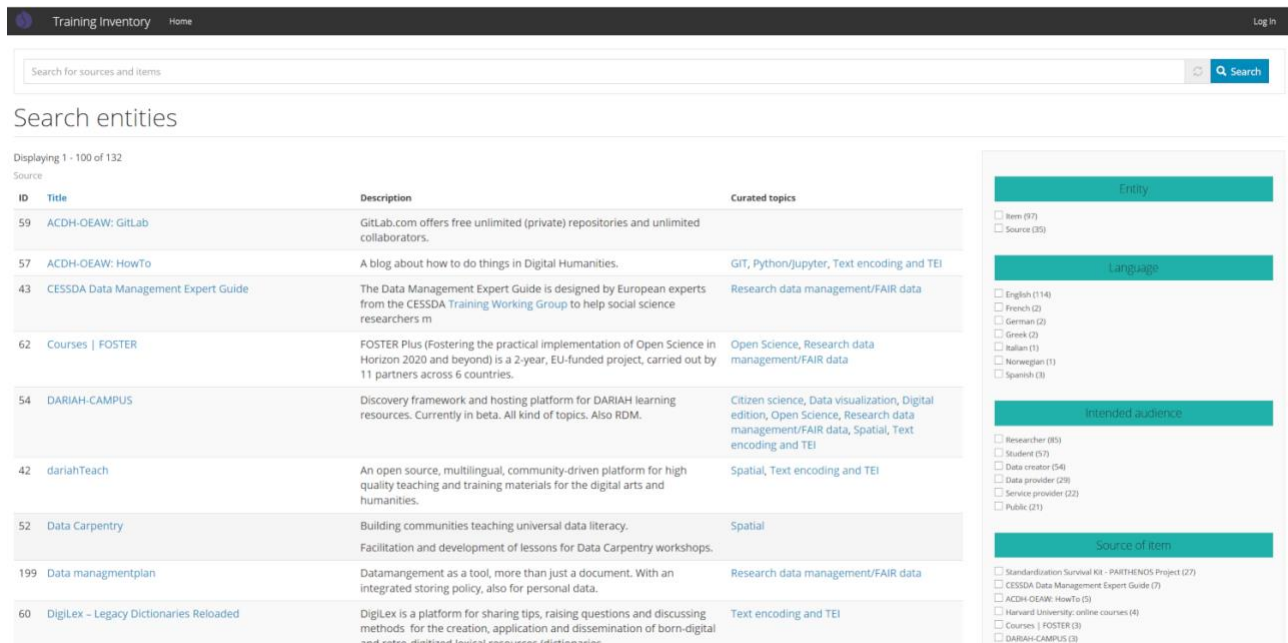
Curated
topics

Figure 1: Data model for sources and items in the curation application

In the following paragraph, we describe the implemented functionality of the application:

Dedicated users were created for the curators, that allow them to add and edit sources and items. A minimal design was applied, as the main task for the curation application was to collect items for selected sources. To

get an overview of the on-going process, the start page of the curation application shows all available content. As primary means for exploring the data, a faceted search view was implemented.

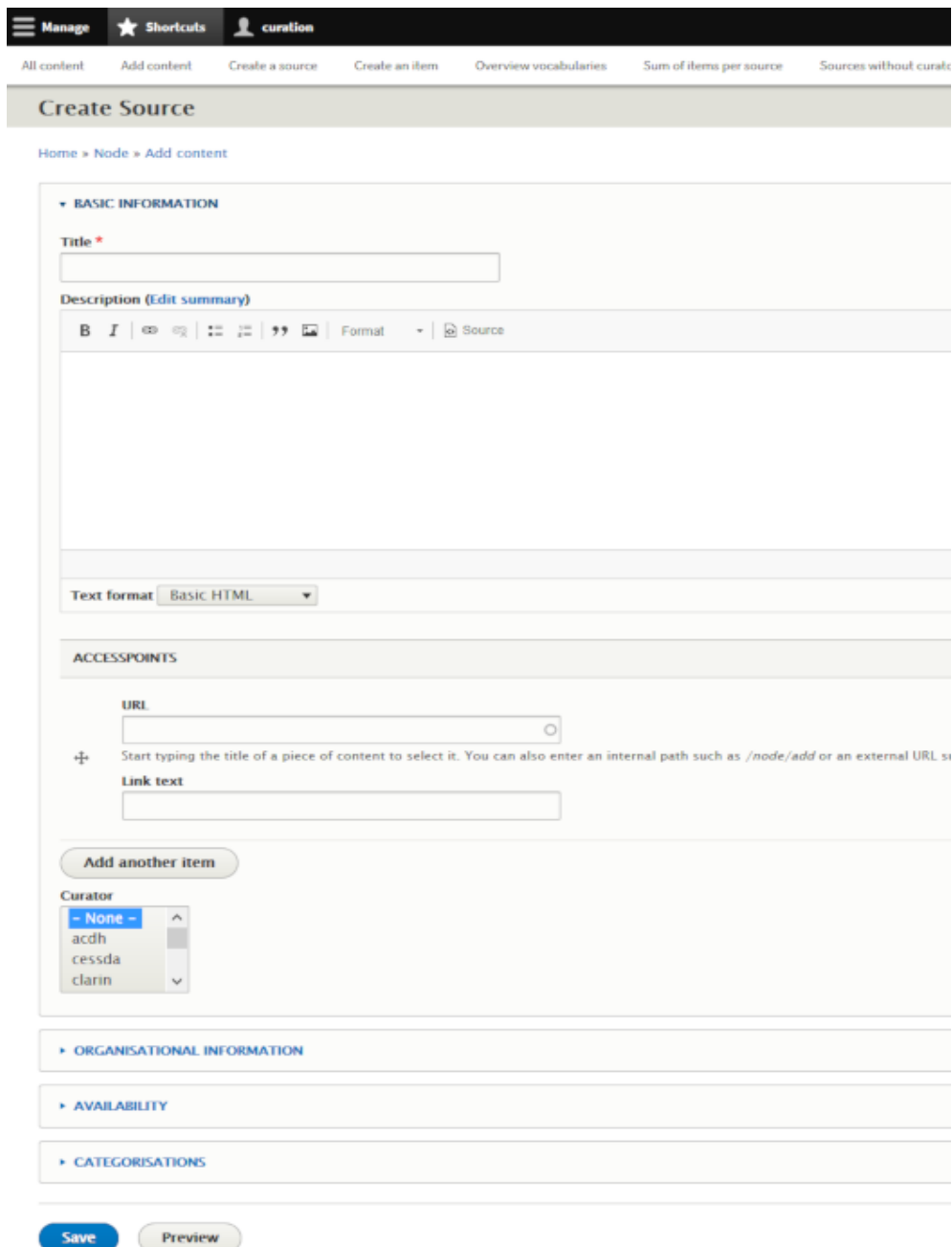


The screenshot shows the 'Training Inventory' application interface. At the top, there is a search bar with the text 'Search for sources and items' and a 'Search' button. Below the search bar, the page is titled 'Search entities' and indicates 'Displaying 1 - 100 of 132' items. A table lists various sources with columns for ID, Title, Description, and Curated topics. On the right side, there are four faceted search filters: 'Entity', 'Language', 'Intended audience', and 'Source of item', each with a list of checkboxes and counts.

ID	Title	Description	Curated topics
59	ACDH-OEAW: GitLab	GitLab.com offers free unlimited (private) repositories and unlimited collaborators.	
57	ACDH-OEAW: HowTo	A blog about how to do things in Digital Humanities.	Git, Python/jupyter, Text encoding and TEI
43	CESSDA Data Management Expert Guide	The Data Management Expert Guide is designed by European experts from the CESSDA Training Working Group to help social science researchers	Research data management/FAIR data
62	Courses FOSTER	FOSTER Plus (Fostering the practical implementation of Open Science in Horizon 2020 and beyond) is a 2-year, EU-funded project, carried out by 11 partners across 6 countries.	Open Science, Research data management/FAIR data
54	DARIAH-CAMPUS	Discovery framework and hosting platform for DARIAH learning resources. Currently in beta. All kind of topics. Also RDM.	Citizen science, Data visualization, Digital edition, Open Science, Research data management/FAIR data, Spatial, Text encoding and TEI
42	dariahTeach	An open source, multilingual, community-driven platform for high quality teaching and training materials for the digital arts and humanities.	Spatial, Text encoding and TEI
52	Data Carpentry	Building communities teaching universal data literacy. Facilitation and development of lessons for Data Carpentry workshops.	Spatial
199	Data managementplan	Datamangement as a tool, more than just a document. With an integrated storing policy, also for personal data.	Research data management/FAIR data
60	DigiLex – Legacy Dictionaries Reloaded	DigiLex is a platform for sharing tips, raising questions and discussing methods for the creation, application and dissemination of born-digital text. Historical linguistic resources. Edition science	Text encoding and TEI

Figure 2: Screenshot of the start page of the curation application

Drupal allows implementing complex data models and the maintaining of vocabularies (called “taxonomies” in Drupal). Next to the two main content types *source* and *item*, nine vocabularies were defined and populated. (See schema of the data model in [figure 1](#)) Moreover, Drupal features a powerful mechanism for creating custom dynamic presentations, overviews and statistics of the content, so called “Views”. A template was created for adding and editing a source (see figure 3), based on the data model. Additionally, a manual for the curators was created, that explained the use of Drupal and gave basic agreements on the form of the input data. Items are created the same way, only having some differences in the fields to add.



Manage Shortcuts curation

All content Add content Create a source Create an item Overview vocabularies Sum of items per source Sources without curate

Create Source

Home > Node > Add content

BASIC INFORMATION

Title *

Description (Edit summary)

Text format Basic HTML

ACCESSPOINTS

URL

Link text

Add another item

Curator

- None -
- acdh
- cessda
- clarin

ORGANISATIONAL INFORMATION

AVAILABILITY

CATEGORISATIONS

Save Preview

Figure 3: Screenshot of the form for creating a new content type source

3.2 Controlled vocabularies / Analysis dimensions

Numerous fields are connected to controlled vocabularies. These also represent the main analysis dimensions, as they allow to categorize the data along well-defined distinctions: We are interested in the formats, that are offered by the training material sources. We further investigate the topics that are outlined and the disciplines that are involved. Further analysis includes the extent and the harvesting possibility.

We distinguish between closed vocabularies, where the set of possible values is fixed, and curators choose from this pre-defined set.

Closed vocabulary	Description
Curated topics	A set of terms selected from the full list of topics, keywords deemed especially relevant for the purposes of WP6, WP7, or in the context of SSH in general.
Formats	Types of training materials, like webinar, e-learning module, blog-entry, data notebook, etc.
Audiences	The kind of learners the material is intended for, with respect to their position in the research process (student, vs. researcher, vs. data provider).
Statuses	Tentative indication of the stage in the lifecycle of a source, i.e. is it a prototype still in development, or rather a legacy resource that is in danger to disappear.

Table 1: Overview on the controlled vocabularies

The open vocabularies are a kind of folksonomy. Curators are allowed to add new terms but are encouraged to use already existing terms.

Open vocabulary	Description
Disciplines	scholarly disciplines for which given material primarily applies, mostly taken as indicated by the sources
Languages	language of the content of the material
Licenses	license formulating the conditions of (re-)use of the material, e.g. Creative Commons Attribution. Ideally this should be an unambiguous reference (URL) to a license definition
Organisations	Legal entities responsible for providing (maintaining and curating) a given source
Topics	keywords describing what a training material is about. As opposed to “curated topics”, this vocabulary is meant to collect keywords as they are applied in the individual sources.

Table 2: Overview on the open vocabularies

The identification and definition of this vocabularies is based on recurring information in the training materials themselves and on relevance for WP6. Vocabularies offer a common coherent view over the diverse sources and items, allowing a systematic analysis of the coverage along multiple dimensions. I.e. for which topics, or for which audiences there is enough material available, and where there are still potential gaps.

4 Analysis of the training material landscape

In total 35 sources were collected, of which 30 went through a curation process. For these selected sources it was tried to add information about at least three representative sample items (concrete training materials like slides, e-learning modules, webinars etc.) that also fits the scope SSHOC research domain of social sciences and humanities. 97 items were created, so that on average nearly 4 items were collected ($97 / 26 = 3,7$ items per source). As we did a complete collection of 27 items for one source (the Standardization Survival Kit - SSK), this leaves us with 70 items for 25 sources, having nearly 3 items per source ($70 / 25 = 2,8$ items per source). Thus, even though it was not feasible to achieve a comprehensive coverage of the landscape, by means of such sampling we believe to have collected a dataset sufficiently representative of the inspected domain.

4.1 Statistics

In the following we provide various summaries on the collected data along the main dimensions. It is to be noted that these numbers can only be seen as rough indications of the trends, as they come with some uncertainty. For one they represent only a fraction of the overall set of existing training material. Additionally, the categorisation performed by the curator team is an approximation and could be disputed in many cases. Nevertheless, the data reveal certain trends, on which solid conclusions can be based.

4.1.1 By discipline

The relevance of the chosen sources for SSHOC is validated by the discipline tag that is assigned to every source and item (selecting multiple disciplines was allowed). The statistics shows that eleven sources cover the discipline "Digital Humanities", eight sources cover "Social Sciences" and eight sources are multidisciplinary oriented ("All sciences"). Having a look on items, we see a slight bias towards training material that is humanities oriented (showing only the ones with at least five items in overall). All in all, we can conclude that we chose sources relevant to the disciplinary scope of SSHOC.

Discipline	Number of sources	Number of items
Digital humanities	11	19
History	1	17
Social Sciences	8	14
Cultural heritage and museology	0	8
Art and art history	0	7
All sciences	8	6
Library and information sciences	1	5

Humanities	2	5
Archaeology and Prehistory	0	5
Linguistics	3	4

Table 3: Discipline tags of sources and items (only the ones with at least four items)

The [table with the overview on sources](#) in the annex of this document shows an estimation of the extent of a source, where it was possible to collect this information. The range of items differ very strongly between the sources. There are some that have collected a lot of training material like *Humanities Commons* (around 9000 items) or *TeSS* (1285 items and 343 events), others offer only a small set like *Library Carpentry* (4 items) or *dariahTeach* (10 items). Many sources have around 100 items, like *DARIAH-CAMPUS* (70 items) or *EOSC Hub training material* (150 items). Therefore, the analysis of selected items sums up a source either very detailed or very broad. As it was not the goal to have a precise statistic on all of the training material but instead a first overview to evaluate the available sources in regard to the work in WP6, these numbers are still useful for the approach of this deliverable.

4.1.2 By audience

Another key factor is the intended audience of the training material. It shows that the chosen training material sources are mostly oriented towards researchers and students. This makes sense, as for data and service providers tailored training material is available either in their education or is part of their application range.

Intended audience	Number of sources	Number of items
Researcher	21	64
Student	17	40
Data creator	12	42
Data provider	8	21
Public	7	14
Service provider	6	16

Table 4: Sources and items by intended audience

However, these numbers are especially uncertain, as researcher naturally create new resources in the course of their scholarly activities and thus could be considered in the role of data creator or data provider. Having a look at the intended audience of the items, it shows equal distributed material for students and data creators, having researchers still in the lead.

It is worth noting that only a small portion of the training materials is oriented to the public. This may be due to the fact that in the field of data science and digital humanities at least some basic occupation with research

data is a prerequisite for starting a training in the field of digital methods and tools. Another possible explanation is that dealing with *research* data, is naturally oriented primarily towards researchers and not the general public.

4.1.3 By format

Training materials point to different types. In general, we can distinguish these basic types based on the way how they were created and in which form and through which system they are presented:

- **e-learning platforms** - complex dedicated online applications, which allow to enrol students, structure the material into smaller units (lessons etc.), employ interactive resources like quizzes or assignments, etc. Typical example is moodle or blackboard.
- **carpentry-style** - static online resources on a specific topic typically in the form of one long page with internal structure generated from a markup document (usually markdown), mostly hosted in a public versioning repository like GitHub. Usually they feature some form of structured metadata, governed by a template.
- **blog-style** - static online resources, mostly in the form of one page, however authored through a dedicated content management system (like WordPress).
- **slides/pdf/video/webinars** - resources with rigid sequential display of the content, oftentimes with connection to a live event (a presentation or a webinar). This can be slides accompanying a presentation, or a video capture of a webinar given.

Having a look at the sources, the statistics shows that nine of them rely on e-learning modules. Videos, webinars, slides, and blogs are mentioned between five and four times. Three sources take data notebooks like Jupyter²² in account. Only two sources cover events.

This is mostly in accordance with the identified formats for the training material items. A majority of the chosen items are e-learning modules. A lot of them are videos whereas webinars occur more seldom. Blogs and slides are equally well distributed.

²² <https://jupyter.org/>

Format	Number of sources	Number of items
E-learning module	9	26
Webinar	5	4
Slides	5	7
Blog	4	8
Video	4	13
Data Notebook (jupyter)	3	4
Events	2	4
Course	1	0
Demo	0	3

Table 5: Format tags applied to sources and items

From the attached formats we can identify the basic types of e-learning platforms, blog-style, and slides/pdf/video/webinars. All of them are present. For the carpentry-style we didn't have an adequate format tag, therefore they do not show up directly.

Not surprisingly the e-learning platforms are in the majority. It is a well-accepted technology, widely used especially in the universities and therefore training resources are often offered in this way. It is also interesting to see that many items are videos. The difference in the number of sources relying on webinars and the smaller number of items being webinars, can be explained by the random sample of items. Most important is that every expected basic type is present in the sources as well as in the items. In the [use cases section](#) for every such basic type at least one source is described in more detail.

4.2 Topics & Learning paths

Probably the most relevant dimension is the topics, i.e. what given training material is about. At the same time, this is a very difficult dimension to capture, given that there is no clear closed set of categories to classify the material. To tackle this challenge, we distinguished between two types of topics. On the one hand, we collected keywords verbatim as we encountered them at the sources. Virtually each source has a way to categorize the content through some keywords or tags. Even though individual sources partially use the same keywords, a simple union of keywords from all the sources yields a rather incoherent vocabulary. Therefore, we introduced a second dimension called curated topics, where we were stricter about the introduction of terms and admitted only those that appear in multiple sources and seem relevant for the purposes of SSHOC and the work in WP6. Where possible we mapped the general topics to one of the curated topics, performing certain normalisation and even cautious semantic interpretation. The aspiration is to iteratively refine the collection of keywords based on the empirical data, to formulate/define a basic vocabulary of the most relevant topics. This can

gradually evolve as new material will be inspected, or specific requirements come from other stakeholders. Table 6 lists topics (including curated topics) with at least five mentions in items and sources:

Topic	Count sources	Count items
Research data management/FAIR data	18	26
Open Science	7	10
Text encoding and TEI	8	8
Python/Jupyter	7	6
Data	1	10
Programming with R	4	6
Quantitative analysis	6	4
Data visualization	3	6
Metadata	0	9
GIT	4	4
Scanning	0	8
Spatial	3	5
Copyright	2	4
Encoding	0	6
Artifacts	0	5
Citizen science	2	3
Command Line	4	1
Data sharing	2	3
Images (3D)	0	5
Linked open data	4	1
Mapping	2	3
Photography	0	5
research infrastructures	3	2

Technology Preservation	0	5
-------------------------	---	---

Table 6: Most frequent topics split in sources and items (having at least in overall 5 mentions)

One aspect to consider with respect to the topics is that they are a heterogeneous collection of keywords. In further iterations we envisage to refine the vocabulary and extract especially references to tools and technologies (including programming languages) on the one hand, and to scholarly practices or activity types on the other hand to separate vocabularies and match these against existing vocabularies for these dimensions. As an example, the current topics list contains terms like “Programming with R”, “Python/Jupyter”, or “GIT”, which for the sake of homogenisation would belong into a dedicated vocabulary on tools and technologies. “Scanning”, “Encoding” and “Data Visualisation” on the other hand should be mapped on a taxonomy of activity types, such as TaDiRAH.

Finally, topics lend themselves as good starting point for defining learning paths, i.e. the question how individual training materials would fit into overarching coherent sequences of acquiring specific skills. For this purpose, however, it would be required to identify more strictly which concepts are prerequisites for a given course, and which are the learning outcomes, i.e. what the learner is expected to know after taking a course. This information is not always available explicitly and would require extra manual curation work. An example of an implicit potential for fitting individual training materials into coherent connected learning paths is given in one of the use cases sections (5.3 - *Blogs: CESSDA Training, DARIAH-Campus, PARTHENOS Training*).

4.3 Harvestability

A key issue with respect to aggregation and discoverability, especially given the complex heterogeneous landscape is the option to harvest training material from sources automatically, as this is the only sustainable approach to make information from many sources available through one discovery application. The best-practice approach widely in use in data repositories and corresponding metadata catalogues is to employ the OAI-PMH protocol to expose the metadata on the provider side and to automatically regularly harvest it on the aggregator side. There are numerous other possibilities with varying degrees of the required case-by-case customisation.

The analysis of all training material sources shows that the situation is much less developed in the area of training materials. Only one source does have a dedicated API to harvest the metadata of the disposable training material, the TeSS: Training eSupport System that itself is an aggregator performing extensive harvesting to collect the data. TeSS however collects material for Life Sciences, thus the material itself is mostly not relevant in the SSHOC context. Nevertheless, in terms of exchange of experience on harvestability much can be learned by the creators of TeSS. (Faced with a similar situation of no structured metadata and no APIs readily available, TeSS relies largely on “scraping”, i.e. extracting the structured information directly from the HTML landing pages for the training materials. More details about their approach can be found in the eosc-pilot deliverable D7.5.)

All other sources do not have a dedicated API. But there are different possibilities to still retrieve descriptive information automatically. Nine sources store their training material data in Git repositories in some form of

structured format, which lend themselves well for automatic processing. Still, due to lack of semantic harmonisation, it will be necessary to analyse the data model in use (the descriptors) on a case by case basis. This holds especially for the eight sources that rely on the markup language Markdown²³. One source - the SSK - uses TEI²⁴ as native format. Given the well-defined structure and semantics of TEI documents, it was quite easy to harvest and extract structured information (see [use case: SSK](#)).

Three sources provide RSS feeds. This could be also a way to harvest the data. In such cases it depends on the mechanism how the RSS feed was created, as it could be limited to a certain number of items that show up.

One source has a dump of all the data available via Zenodo. The usage is limited, as this data is available only in a specific format bound to a dedicated software.

All other more than 20 sources do not offer an explicit way to harvest the data. It could be that there are ways to do harvesting, especially because many of these sources run on CMS like Drupal, Wordpress, and TYPO3, but it would be necessary to get in contact with the creators/maintainers.

All in all, the existing sources and platforms do not support harvesting, both due to the lack of a dedicated API, and due to missing agreed and implemented data model/ontology for most of the sources. Creating awareness about the need, investing in recommendations for standard-driven training material sources and especially also accompanying and supporting the technical implementation should therefore be taken up as activities for future work in the task *6.3 - Empowering Users: Training Materials and Online Learning Paths*. Given the need for this functionality for the *SSH Open Marketplace* being developed in WP7, these activities seem to lend themselves well for a tight collaboration with the tasks *7.3 Marketplace Interoperability* and *7.4 Governance: Population, Curation & Sustainability of the Marketplace*.

5 Use cases

In the following section, we present a few examples of platforms or training material sources, structured along the four main types we introduced earlier. The motivation for this categorisation is that the different types require distinct technologies and workflows both in the production and in the consumption or use. This implies different approaches when further processing them or aggregating information about them and is also an important information for the potential learners to clarify the intended mode of interaction and expected effort to digest these materials. At the same time we need to highlight that the introduced categorisation is only a rough one, and in practice we encounter hybrid combinations of the proposed types.

5.1 e-Learning platforms: #dariahTeach

#dariahTeach is a Moodle-based platform hosting “open-source, high-quality, multilingual teaching materials for the digital arts and humanities”²⁵. More precisely, users can find courses in English, French, Greek and

²³ <https://en.wikipedia.org/wiki/Markdown>

²⁴ <https://tei-c.org/>

²⁵ <https://teach.dariah.eu/local/staticpage/view.php?page=about>

Hungarian, on the following topics: Introduction to Digital Humanities, Text Encoding Initiative (TEI), Sound Studies, Digital Scholarly Editions, Conceptual Modelling, Multimodal Literacies, Digitizing Dictionaries and Spatial Image Analytics. #dariahTeach intended audience is students and academic researchers. A given number of European Credit Transfer and Accumulation System (ECTS) is indicated for each course available.

As a collaborative platform, supported by a dedicated DARIAH Working Group²⁶, #dariahTeach welcomes new materials or contributions to existing courses (see for example the “How to contribute” section in the Introduction to Digital Humanities course²⁷).

As it has been already described in an extensive deliverable of the DESIR project²⁸, *D7.2 DARIAH training materials assessment report*, to be published soon: “there is currently some variation how, how much, and where metadata is presented” in #dariahTeach. However, a new format is under development and should allow not only a more standardised way to present metadata, but also a “new Wordpress ‘home page’ for each course on the #dariahTeach blog, to ensure that it gets indexed and is therefore discoverable.”

5.2 Carpentry: SSK

The Standardization Survival Kit (SSK) was created in the EU Horizon 2020 funded project PARTHENOS. Its aim is to push researchers in using community-accepted standards when working with research data. As experience shows, many researchers are not aware of such standards. Therefore, the approach of the SSK was to collect best practice solutions - so called scenarios - that offer a low-barrier entry point. In describing a common practice and how to handle it (e.g. doing a specific research operation), researchers can compare their approach with the one described in the SSK. The description in the SSK gives clear references to standards, so that a researcher can verify, if her/his approach conforms.

The SSK can be therefore seen as a training material resource based on a carpentry approach. This means that it points researchers to different tools, methods, and practices, where they can then individually decide which parts to integrate into their workflow. It also gives a basic introduction to relevant skills, aiming for a train-the-trainer method²⁹. The carpentry approach is used by the considered training material sources “software carpentry”, “library carpentry”, and “data carpentry”. It is frequent, that such resources are listing often used tools or - in the case of SSK - standards. Library carpentry - as an example - does offer a list with “Top 10 FAIR Data & Software Things”³⁰. The carpentry type offers in general on the one hand a basic introduction to techniques, on the other hand an extensive collection of relevant material for further study. Therefore, it may miss detailed training sessions for specific research methods/communities but qualifies as a good starting

²⁶ <https://www.dariah.eu/activities/working-groups/dariahTeach/>

²⁷ <https://teach.dariah.eu/course/view.php?id=26>

²⁸ <https://www.dariah.eu/activities/projects-and-affiliations/desir/>

²⁹ The description of library carpentry gives an introduction to the concept of carpentry-based training: Baker, J., Moore, C., Priego, E., Alegre, R., Cope, J., Price, L., ... Wilson, G. (2016). Library Carpentry: software skills training for library professionals. *LIBER Quarterly*, 26(3), 141–162. DOI: <http://doi.org/10.18352/lq.10176>

³⁰ <https://librarycarpentry.org/Top-10-FAIR/>

point for getting an overview. From there follow-up questions can be derived, that are answered in detail from training material on other sources.

SSK was also the only source where we actually experimentally performed automatic harvesting of data. As members of the WP6 team were involved in the creation of the SSK in the PARTHENOS project, this was a good starting point for writing a script to machine-driven get the metadata of the SSK scenarios. The SSK is organized on base of XML data. Scenarios and steps are described in TEI format³¹. Using the TEI standard, the SSK follows itself good practice why standards are helpful for other projects. In combination with the documentation of the SSK, it was quite easy to create a mapping between the data model of the inventory application and the one of the SSK. The CMS Drupal provides a convenient ingestion of data. Based on this it was possible to integrate all scenarios of the SSK as items into the inventory application. Most of the fields could be filled automatically, the others (e.g. audiences, curated topics) were complemented manually.

From the statistics, it can be seen that the SSK is very humanities oriented. In terms of most frequent disciplines: Twelve items relate to “history”, eight to “cultural heritage and museology”, seven to “art and art history” whereas “sociology” is only mentioned in two items. For the topic section it is important to mention, that the SSK uses the TaDiRAH vocabulary³². Therefore, a lot of keywords are coming from there. In this respect, it can be seen that the SSK items are focused on data-related issues. The terms “metadata”, “data”, “encoding” are among the most frequent ones. Based on this general classification, it is possible to find similar sources, where training material could be connected. Proceeding manually, on the facet search the considered entities are limited to “source”, the discipline is limited to “digital humanities” and the covered topics are limited to such, that deal with data (“data curation”, “data manipulation”, “data modelling”). This leaves us with three potential related sources: “DARIAH-CAMPUS”, “Digital humanities Examples: Digital Humanities Resources for Project Building”, and “The Programming Historian”. Indeed, there are connection points to these sources in perspective of the SSK. As an example: SSK has the scenario “Extract textual content from images” that can be ideally combined with the lesson “Cleaning OCR’d text with Regular Expressions” from Programming Historian.

The complete harvesting of SSK shows the potential of having metadata on training material collected from different sources. At a first glance it is possible to identify the focus of a source and to look for related material from other sources. The example of SSK shows, that there is a good chance to establish such connections. For this, it is necessary that most of the metadata about training material from a source is harvested and that algorithms are developed, that establish the connections. A precondition is further detailed analysis of potential connection points and the adaption of well-balanced and harmonized vocabularies (it is termed here as “curated topics”) to apply on the sources and items. One important observation is about the type of training material source: the carpentry style of SSK makes it an ideal starting point, where further training material from sources more oriented to e-learning- or blog-style could be attached.

³¹ Text Encoding Initiative (TEI): <https://tei-c.org/>

³² TaDiRAH - Taxonomy of Digital Research Activities in the Humanities: <http://tadirah.dariah.eu/vocab/index.php>

5.3 Blogs: CESSDA Training, DARIAH-Campus, PARTHENOS Training

The blog-style format refers to relatively long and static content produced on training providers' website. We present three examples - CESSDA Training, DARIAH-Campus and PARTHENOS Training - to reflect the diversity of approaches covered by this format. These three sources present a diversity of content. The perimeter depends on the provider's area of expertise: CESSDA Training focuses on "finding, managing and preserving data"³³ for social sciences; DARIAH-Campus on "digitally-enabled arts and humanities"³⁴ when PARTHENOS Training chooses to focus on "digital humanities and research infrastructures"³⁵.

When it comes to CESSDA Training, slides and videos of training sessions provided by CESSDA and its service providers are freely accessible. CESSDA Training focuses on training for data producers, data users and data professionals or CESSDA Service Provider staff. Resources are classified into three categories: "Discovering and using data"; "Managing Research Data" and "Preserving Data". Furthermore, the CESSDA Data Management Expert Guide (DMEG)³⁶ covers every step of the data lifecycle, providing for each step recommendations, examples, as well as a Train-the-Trainers Package³⁷ gathering workshops outlines, exercises and presentations. A search function is also available in the DMEG, but no tags or metadata are attached to the different resources, including the DMEG.

In a different style, DARIAH-Campus (currently in beta version) is a "discovery framework and hosting platform for DARIAH learning resources"³⁸. DARIAH-Campus was developed as a pilot within the framework of the DESIR project³⁹ in order to consolidate and offer a coherent approach to DARIAH learning resources. DARIAH-Campus is hosted on GitHub and individual items are written in Markdown. This structure offers openness and flexibility for both contributors and reusers: the former can add or create training materials in any formats, length or granularity level while the latter can easily and openly reuse them under an open license (CC-BY 4.0 if not marked differently) and without registration. In the current phase of development, enabling the machine harvestability of the resources is also kept in mind and is under discussion. Four types of resources are presented in DARIAH-Campus: external learning resources (DARIAH-Campus as an aggregator of existing training materials); original hosted learning resources; Pathfinders (i.e. enriched, curated and contextualized summaries/collections of already existing training resources around a certain topic); and events. All resources are tagged by topics and a search mechanism is provided. Each resource, hosted or not in DARIAH-Campus, is presented by a short text. The most advanced feature when it comes to the blog-style aspect is the "DARIAH Pathfinders" that offers "Useful collections of external learning resources, curated and contextualized by the

³³ <https://www.cessda.eu/Training/Training-Resources>

³⁴ <https://campus.dariah.eu/about>

³⁵ <https://training.parthenos-project.eu/>

³⁶ <https://www.cessda.eu/DMGuide>

³⁷ <https://www.cessda.eu/Training/Training-Resources/Library/Training-Packages/Train-the-Trainers>

³⁸ <https://campus.dariah.eu/>

³⁹ <https://www.dariah.eu/activities/projects-and-affiliations/desir/>

DARIAH team⁴⁰. This resource type aims to provide a quick overview of and trusted navigation through available training landscape. The “Data Management Best Practices in the Humanities”⁴¹ entry presents six sections highlighting for each of them additional resources and a list of the most important questions. Providers are encouraged to contribute by adding their contents. A simple guide to reference or creates new learning resources is provided in DARIAH-Campus documentation⁴².

PARTHENOS Training has been developed under the PARTHENOS project⁴³ to address a gap of understanding related to infrastructures in humanities and digital humanities⁴⁴. The Training plan designed covers a wide range of modules, such as “Introduction to Research Infrastructures”; “Management Challenges in Research Infrastructures”; “Collaborations within Research Infrastructures”; “Manage, Improve and Open up your Research Data”; “Formal Ontologies: A Complete Novice’s Guide”; “e-Humanities and e-Heritage Webinar Series”; “Citizen Science in the (Digital) Arts and Humanities” and “Digital Humanities Research Questions and Methods”. The WordPress format was used because that was the platform used by the entire project, but the modules themselves cannot really be considered as blog posts. They are distinct purpose-written modules that take the format of an online course in that there is a linear progression that a user could take, while also allowing users the opportunity to navigate straight to the section they find most relevant if they don’t have the inclination or time to read the entire module.

These three examples - CESSDA Training, DARIAH-Campus and PARTHENOS Training - act on the one hand as aggregators combining existing training materials and highlighting, thanks to high-quality narratives and a curated set of type of materials (text, videos, slides, images, cartoons, exercises...), topics that were chosen by the editorial teams. On the other hand, these platforms have also created and host original learning resources.

5.4 Slides, PDF, videos, webinars:

SERISS, CLARIN Videolectures and DARIAH-Campus

Training materials are often presented with slides, pdf or videos formats. Very often, this is because the contents presented are the result of events, virtual like webinars, or real like workshops. Three examples, illustrating three different practices to manage these formats are given here.

Synergies for Europe’s Research Infrastructures in the Social Sciences (SERISS) project was financed by a H2020 grant between 2015 and 2019. One of the goals of this project was to “share learning and expertise between project infrastructures and with the wider social science community”⁴⁵. Under a dedicated work package, tools and related courses were developed. The website of the project still provides an overview of these different

⁴⁰ <https://campus.dariah.eu/source/dariah-pathfinders>

⁴¹ <https://campus.dariah.eu/resource/dariah-pathfinder-to-data-management-best-practices-in-the-humanities>

⁴² <https://campus.dariah.eu/docs/proposing-a-contribution>

⁴³ <http://www.parthenos-project.eu/>

⁴⁴ <http://training.parthenos-project.eu/about-parthenos-training/>

⁴⁵ <https://seriss.eu/about-seriss/work-packages/wp5-training-and-dissemination/>

training activities and points to videos hosted on a dedicated SERISS YouTube channel⁴⁶. Those videos captured webinars, lectures or Train the Trainer sessions that were given during the project lifetime.

Part of the Knowledge Sharing Infrastructure⁴⁷, CLARIN Videolectures⁴⁸ gathers all the video-contents produced by CLARIN or during CLARIN events. VideoLectures.net is a “free and open access educational video lectures repository”⁴⁹ that allows organisations to set up a portal, where they can host their videos. The platform provides additional features: it allows to link and synchronise slides and videos; some community features like ratings and different sharing options are available, as well as the possibility to organise contents (i.e. create an event for all the videos produced during a conference for example). Author name and affiliation, date of publication and of recording, as well as number of views and licences are the minimal set of information linked to every content. Several categories are used to describe and search for content: scientific categories (disciplines and topics), language, type of content (event, lecture, tutorial etc.) and date.

Finally, an interesting feature, when it comes to capturing events, is the workflow and the template provided by DARIAH-Campus⁵⁰. This workflow allows providers to gather “programmes, slides, videos, speaker biographies and much more” and the template “to compile and structure the content in a learner-friendly way”.

An example of this capture-event feature can be found here: <https://campus.dariah.eu/docs/types-of-resources#events>.

6 Input for Marketplace

Next to being an input for further work of WP6, the inventory will also inform the development of the SSH Open Marketplace in WP7. The Marketplace shall serve as a discovery application, a catalogue with focus on research practices. It should offer answers on the question of “How?”, i.e. what the methods and workflows are to achieve a certain goal in the research process. This can be a question on how to proceed to linguistically annotate a text, to retrodigitize a dictionary, doing network analysis on social network data or doing a statistical analysis of a dataset from a socioeconomic survey.

To achieve this, the marketplace is not meant to be yet another a directory of existing tools and services. While it will feature this kind of information, the focus will lie a) on contextualizing it with relevant training materials and publications, b) on collecting and presenting also “solutions”, recipes or workflows, i.e. specific instructions on how to perform a task, achieve a goal, or the individual action steps of such solutions. This will be further enriched with information on which tools are suited to perform these actions.

To achieve this contextualisation, this enrichment, the overview over the broad and rich landscape of training materials provided by this deliverable is paramount. The initial analysis with respect to the type of content, as

⁴⁶ <https://www.youtube.com/channel/UCVwDSeAGfZgwpCGAWXGZuhw>

⁴⁷ <https://www.clarin.eu/content/knowledge-sharing>

⁴⁸ <http://videolectures.net/clarin/>

⁴⁹ <http://videolectures.net/site/about/>

⁵⁰ <https://campus.dariah.eu/docs/types-of-resources#events>

well as the topics covered, the intended audiences, the covered disciplines, the activity types and tools tackled, will allow the integration of the training material metadata into the marketplace.

Another important aspect for which this inventory offers at least a preliminary analysis is the “harvestability” of the metadata. Given that the marketplace is intended as a live catalogue with continuous automatized population from identified sources, the availability of the metadata at the sources in a structured form through well-defined mechanisms (APIs) is a critical feature. The preliminary findings in this respect are somewhat sobering, prompting the recommendation for the interoperability task in WP7 (*T7.3 - Marketplace Interoperability*) to reserve sufficient capacities to handle this aspect.

7 Conclusion

In this report we provided an overview of the landscape of training material in the SSH disciplines cluster and the broader context of EOSC. The provided summaries and statistics are based on collaborative curation work, inspecting main existing sources of training materials and their content.

The collected information serves as input for further activities in WP6 as well as an important source of information for WP7. At the same time the curation process itself served also to formulate a conceptual model for training materials and iteratively refine it based on real-life examples, especially with respect to the relevant vocabularies to use for categorisation along dimensions like: audience, type of material, and especially topics, which are the natural primary dimension to allow exploring the materials.

Most relevant findings, outcomes for further work in WP6 and WP7 are:

- There seems to be a host of material on Research Data Management and application of FAIR principles, as well as on basic scripting skills for Data Science, i.e. working with the shell, programming with Python/R, for purposes of automatically processing data.
- There are numerous sources that seem to be abandoned or in danger of becoming obsolete.
- There is little to none harmonisation with respect to metadata about the training material. This information is seldom available in structured and automatically harvestable form.
- Initial work on defining vocabularies to categorize the material.
- More work will be needed, especially input from other work packages and other stakeholders to define a balanced consistent vocabulary of topics. Part of this work should be also the formulation of a separate vocabulary for tools and technologies (in close collaboration with WP7) and mapping to a taxonomy of scholarly activities.

8 Annex – Overview of sources

List of all sources of training material in the curation application, including extent, curator, status, curated topics and the number of curated items.

Title	Extent	Curator	Status	Curated topics	Cur. items
ACDH-OFAW: GitLab		acdh	Disregarded		0
ACDH-OFAW: HowTo	30 Blogs	acdh	Continuous curation	GIT, Python/Jupyter, Text encoding and TEI	5
CESSDA Data Management Expert Guide	6 Webinars	ukda	Continuous curation	Research data management/FAIR data	7
Courses FOSTER	52 E-learning modules	clarin	Continuous curation	Open Science, Research data management/FAIR data	3
DARIAH-CAMPUS	70 Items	dariah	In development	Citizen science, Data visualization, Digital edition, Open Science, Research data management/FAIR data, Spatial, Text encoding and TEI	3
dariahTeach	10 Items	dariah	Continuous curation	Spatial, Text encoding and TEI	3
Data Carpentry	4 Items	dariah	Continuous curation	Spatial	2
Data Management Plan			Continuous curation	Research data management/FAIR data	0
DigiLex – Legacy Dictionaries Reloaded	23 Items	clarin	Continuous curation	Text encoding and TEI	2
Digital humanities Examples: Digital Humanities Resources for Project Building		ucl	Endangered	GIT, Programming with R, Python/Jupyter, Text encoding and TEI	3

EHRI: Training	1 Course, 8 Units/Lessons	cnr	Continuous curation	Programming with R	3
Harvard University: online courses	200 Items	cnr	Continuous curation		4
Humanities Commons - Open access, open source, open to all	9000 Items	ucl	Continuous curation	Open Science, Research data management/FAIR data	2
Journal of Open Source Education		clarin		Research data management/FAIR data	2
Knowledge Sharing CLARIN ERIC		clarin	Disregarded	Data visualization, Open Science, Quantitative analysis, Research data management/FAIR data, Text encoding and TEI	3
Leadership & Skills Building - LIBER	20 Webinars	acdh	Continuous curation	Copyright, Open Science, Research data management/FAIR data	0
Library Carpentry	4 Items	dariah	Continuous curation	GIT, Programming with R, Python/Jupyter, Research data management/FAIR data	2
Machine Learning for Artists	30 Items	dariah	Continuous curation	Python/Jupyter	3
ONB GitLab	8 Data Notebooks (jupyter)		Continuous curation	Data visualization, Python/Jupyter, Quantitative analysis	0
OpenAIRE: Support	150 Webinars	cnr	Continuous curation	Open Science, Research data management/FAIR data	3
OpenMethods – HIGHLIGHTING DIGITAL HUMANITIES METHODS AND TOOLS			Continuous curation	Research data management/FAIR data, Text encoding and TEI	0
PARTHENOS: Training Suite	32 Videos, 10 Slides, 5	cnr	Finished/support ed	Citizen science, Research data management/FAIR	3

	Webinars			data	
Research data management courses: an overview and gap analysis for the Netherlands	14 Items	ucl	Finished/support ed	Research data management/FAIR data	3
Research Data Mantra	1 course, 9 units	ukda	Finished/support ed	Research data management/FAIR data	1
SERISS: Training overview	10 Webinars, 5 Items	cnr	Continuous curation	Quantitative analysis, Research data management/FAIR data, Survey data	3
Software Carpentry	20 Units/Lessons	acdh	Continuous curation	GIT, Programming with R, Python/Jupyter, Quantitative analysis	0
Standardization Survival Kit - PARTHENOS Project	27 Items	acdh	Continuous curation	Quantitative analysis, Text encoding and TEI	27
Task forces in OpenAIRE Advance			Disregarded	Research data management/FAIR data	0
TeLeMaCo	175 Items	clarin	Endangered	Open Science	2
TeSS (Training eSupport System)	1285 Items, 343 Events		Continuous curation		0
The Programming Historian Programming Historian	129 Items	ucl	Continuous curation	Python/Jupyter, Research data management/FAIR data	3
The Survey of Health, Ageing and Retirement in Europe (SHARE): Special Data Sets			Disregarded	Quantitative analysis	0
Training Courses European Social Survey (ESS)			Deprecated		0
Training material EOOSC Hub	150 Items	dariah	Continuous curation		2

UK Data Service: Manage Data		ukda	Continuous curation	Copyright, Research data management/FAIR data	3
--	--	------	---------------------	---	---

Table 7: Overview on sources including extent, curator and collected items

List of Figures

[Figure 1: Data model for sources and items in the curation application](#)

[Figure 2: Screenshot of the start page of the curation application](#)

[Figure 3: Screenshot of the form for creating a new content type source](#)

List of Tables

[Table 1: Overview on the controlled vocabularies](#)

[Table 2: Overview on the open vocabularies](#)

[Table 3: Discipline tags of sources and items \(only the ones with at least four items\)](#)

[Table 4: Sources and items by intended audience](#)

[Table 5: Format tags applied to sources and items](#)

[Table 6: Most frequent topics split in sources and items \(having at least in overall 5 mentions\)](#)

[Table 7: Overview on sources including extent, curator and collected items](#)