EUROPEAN
DATA SCIENCE
ACADEMY

Project acronym:	**EDSA**
Project full name:	**European Data Science Academy**
Grant agreement no:	**643937**

D5.5 Initial EDSA Data Management Plan

Deliverable Editor:	**Mandy Costello (Open Data Institute)**
Other contributors:	**Simon Bullmore (Open Data Institute)**
	Dr David Tarrant (Open Data Institute)
Deliverable Reviewers:	**Jean-Louis Lievin (ideXlab)**
	Alex Mikroyannidis (Open University)
Deliverable due date:	**31/07/2015**
Submission date:	**31/07/2015**
Distribution level:	**PUBLIC**
Version:	**1.0**

This document is part of a research project funded
by the Horizon 2020 Framework Programme of the European Union

Change Log

Version	Date	Amended by	Changes
0.1	08/06/2015	Mandy Costello	Created document, added initial plan outline
0.2	07/07/2015	Mandy Costello and Simon Bullmore	Amended executive summary and overall policy
0.3	07/07/2015	Mandy Costello	Added dataset templates and descriptions
0.4	08/07/2015	Mandy Costello	Added datasets
0.5	09/07/2015	David Tarrant	Final amendments and review
0.6	22/07/2015	Mandy Costello	Incorporated reviewers comments
0.7	23/07/2015	Mandy Costello	Incorporated reviewers comments
1.0	31/07/2015	Aneta Tumilowicz	Final QA

Table of Contents

List of Tables

1. Executive summary

The European Data Science Academy (EDSA) will participate in the pilot action on open access research data, as defined in Horizon 2020 Work Programme for 2014-15[1].

EDSA will produce an evolving Data Management Plan (DMP), initialised early in the project. Updates of the DMP will be provided at M18 and M36 to incorporate changes to datasets used or generated throughout the life of the project.

Our goals are:

1. To ensure that where possible, data produced by the project is made accessible to anyone interested in using or sharing it.
2. To ensure that data is managed and maintained, so that it is a useful resource.
3. To ensure that data produced by the project is subject to appropriate levels of security.

EDSA will be producing a wide variety of datasets. Our guiding principle is to release data in a format that anyone can access, use or share. The nature of some of the data we will produce means that some datasets, such as interview transcriptions or internal logs of online learning systems, will need to either remain closed or be anonymised. At this stage it is not clear for every dataset what the final position will be.

The EDSA DMP outlines the overall project's policy on:

- Data standards and metadata standards.
- Data sharing.
- Data preservation.

This initial EDSA DMP also outlines the following information for each dataset:

- Dataset name and identifier.
- Description of the dataset including origin, if collected, scale and use.
- Details on data sharing, licensing, and repositories.
- Archiving and preservation, including ongoing management, length of preservation and backup procedures.

At this early stage in the project, an estimation or intention has been outline in some cases where data has not yet been used or generated. Once the data is available, the approach outlined will be evaluated and updated in the second version of the DMP, at M18.

[1] Guidelines on open access - http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

2. EDSA data management policy

This section outlines the generated and used datasets and their current status. We also outline the overall EDSA policies for data standards and metadata standards, data sharing and data preservation.

2.1 Used and generated project datasets

The following tables outline the used and generated datasets for the project. This list will evolve as the project progresses, and will be reflected in the second version of the DMP. Those marked *, have been added since the initial proposal.

Table 1: Used datasets

WP	Lead	Used dataset	Project phase	Status
1	ODI	Corpora of crawled web-based adverts from LinkedIn	M2-M36	On-going
2	SOTON	Linked open data sources	M2-M36	On-going
2	SOTON	Publically available governmental, financial, network and environmental datasets for each course.	M2-M36	On going
2	SOTON	Related course data regarding similar modules and training offerings across the EU*	M6-M36	On-going
3	JSI	Repository statistics on downloads and views of educational resources	M1-M36	On-going
3	JSI	Internal logs of eLearning systems	M1-M36	On-going
3	JSI	Statistics of course registration, participation and completion	M1-M36	On-going
4	SOTON	Web server logs and Google analytics of project website access	M1-M36	On-going
4	SOTON	Generated social media engagement data	M1-M36	On-going

Table 2: Generated datasets

WP	Lead	Generated dataset	Phase	Status
1	ODI	Aggregated statistics of European skill demand based on web-based job adverts	M6-M36	In progress
1	ODI	Individual results from online survey*	M4-M36	In progress
1	ODI	Aggregated results from online survey*	M6-M36	In progress
1	ODI	Recordings/transcriptions of interviews	M4-M36	On-going
1	ODI	ideXlab search platform results*	M6-M36	In progress
1	ODI	Aggregated, anonymous data of interview results*	M6-M36	On-going
2	SOTON	Subsets derived from the existing data repositories produced for exercises in the learning resources	M12-M36	Not yet available
3	JSI	Aggregated statistics of engagement with the developed courses and educational resources	M12-36	Not yet available
4	SOTON	Aggregated statistics of networking and engagement data	M18-M36	Not yet available
4	SOTON	Learning materials access data*	M12-M36	Not yet available
5	ideXlab	List of project exploitation results – collaborations, institutional and geographical beneficiaries,	M36	Not yet available

2.2 Data standards and metadata policy for EDSA

Standardising the project's collection and production of data will ensure the reusability and interoperability of the data within the project, and externally if the data is to be made openly available. Where possible, data will be made available in CSV, JSON or linked data in RDF format, to allow maximum interoperability. Due to the varied nature of data collected, we will attempt to use widely adopted metadata standards for describing the data. Use of generic vocabularies, such as dublincore[2] and DCAT[3] will be used to make datasets easily discoverable and interoperable. Further data packages[4] will be generated with accompanying schema, to describe both the datasets and contents of files. These packages can then be verified using tools such as CSVLint.io[5] and certified using the Open Data Institute's Open Data Certificates[6].

2.3 Data sharing policy for EDSA

To ensure accessibility, where possible, open data will be provided so that others are able to access, use and share the data. This will enable others to evaluate the project's findings and find value in it. This data will be made available under a Creative Commons licence, Creative Commons Attribution (CC BY 4.0), which allows the user to 'copy and redistribute the material in any medium or format' and 'remix, transform, and build upon the material, for any purpose, even commercially' [7]

As data is yet to be generated in some cases, this is an area of the DMP that will be continually revisited, to seek opportunities to make more data available openly, under a Creative Commons licence for reuse.

When it is not possible to publish collected data due to privacy obligations, we will aim to derive anonymous data that can be published openly. For example, in WP1, while we are unable to publish the transcriptions of the interviews conducted, we seek permission from each individual to publish results in an anonymous, aggregated format. The data derived from this will be visualised via the EDSA dashboard, and available as open data under a Creative Commons licence in a repository.

2.3.1 Supporting people who want to use EDSA data

To help users who wish to access data published by EDSA as open data, we will be using the ODI's Open Data Certificate standard to benchmark each dataset[8]. This will enable users to see when the data will be updated, what format the data is in, what support is available and where it came from.

[2] Dublin core metadata initiative - http://dublincore.org/documents/dces/

[3] Data Catalogue Vocabulary - http://www.w3.org/TR/vocab-dcat/

[4] Data package specification - http://data.okfn.org/doc/data-package

[5] CSV Validator - http://csvlint.io/

[6] Open Data Certificates - https://certificates.theodi.org/

[7] Creative Commons - https://creativecommons.org/licenses/by/4.0/

[8] Open Data Certificates - https://certificates.theodi.org/

2.4 Data storage and management policy for EDSA

There are currently three main repositories for EDSA data:

Github

Github is a web-based repository service, which allows easy, open access to the public. It is the world's largest open source community[9]. The open access data from WP1 will be available in the EDSA Dashboard Github repository.

EDSA project website

The EDSA project website, will be a central point for data to be made available openly. The website will host the EDSA Dashboard, which will visualise the research findings, and contain data sources, such as the links to other courses.

Internal institutional repositories

Internal institutional repositories will be used to hold data that will not be made accessible openly at this stage. Examples include University of Southampton and Open University, who will use existing databases to securely hold data that can only be shared with limited partners due to privacy and data protection rights.

2.5 Data preservation and archiving policy for EDSA

Data that is made openly accessible and that is published through Github by EDSA will continue to be accessible beyond the term of the project. Striving for preservation of this data will enable long-term value to be added to the domain beyond the project. It will also prove a valuable resource to a European wide initiative (EDSA) initiated as part of WP5. An explanation of the approach to preservation and archiving for each dataset can be found in the sections below.

[9] GitHub - https://github.com/

3. EDSA data management plan

The following sections of the DMP outline the specifications of each data set within a work package. Details on description, standards, sharing and preservation as required by the guidelines can be found for each data set [10]

3.1 Work package 1 – Demand analysis and advisory board

WP1 will collect and generate data from the demand analysis study. This will include transcriptions and recordings of the one-to-one interviews, online survey responses, aggregated and anonymous data of the results and aggregated data of the results from a specialist search platform.

3.1.1 Corpora of web based job adverts

Table 3: Corpora of web based job adverts

Dataset reference and name	
Dataset identifier	WebSiteHarvest
Data set description	
Generated or collected	Collected
Origin	LinkedIn
Scale	46 terms, 31 languages, 47 countries, 1 harvest per day – total of 2162 data points per day.
Who is this data useful for?	Internal demand analysis. External research into job and skill demand.
Similar existing datasets	Many datasets are collected in this area, however due to the specific nature of this study, collection of new data is required. However, integration with existing datasets will be explored. The value of this dataset comes from the provision of an up-to-date snapshot of current data science skill needs across Europe.
Standards and metadata	
Methodology for data collection/management	Automated harvester developed in PHP. All data collected is translated into CSV format.
Metadata, supporting material	A README.md file is available detailing the data structure and basic usage.
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0).
Data reuse	Data will be available to view on the EDSA dashboard and all files accessible for free in the EDSA dashboard Github repository. A 'Get the data' link on the dashboard will take users to the repository.
Repository for data	Github

[10] Guidelines on data management - http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	As long as Github exists as a minimum. Beyond that a value of maintaining the dataset would have to be evaluated.
Approximate end volume	<1Gb
Who is responsible for data curating and management?	ODI lead data management and curation, other WP1 partners will contribute.
Quality assurance including back up procedures	Data is stored with external providers (Github).
Associated costs for data management	Github is free and public. Approximately 1-day person effort per month to manage the data.

3.1.2 Aggregated statistics of skill demand on web based job adverts

Table 4: Aggregated statistics of skill demand on web based job adverts

Dataset reference and name	
Dataset identifier	WebSiteStatistcs
Data set description	
Generated or collected	Generated
Origin	LinkedIn
Scale	Full scale not yet known.
Who is this data useful for?	Internal demand analysis. External research into job and skill demand.
Similar existing datasets	Many datasets are collected in this area, however due to the specific nature of this study, collection of new data is required However, integration with existing datasets will be explored. The value of this dataset comes from the provision of an up-to-date snapshot of current data science skills needs across Europe.
Standards and metadata	
Methodology for data collection/management	CSV/JSON
Metadata, supporting material	A README.md file is available detailing the data structure and basic usage.
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0)
Data reuse	Data will be available to view on the EDSA dashboard and all files accessible for free in the EDSA dashboard Github repository. A 'Get the data' link on the dashboard will take users to the repository.

Repository for data	Github
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	As long as Github exists as a minimum. Beyond that a value of maintaining the dataset would have to be evaluated.
Approximate end volume	<1Gb
Who is responsible for data curating and management?	ODI lead data management and curation, other WP1 partners will contribute.
Quality assurance including back up procedures	Data is stored with external providers (Github).
Associated costs for data management	Github is free and public. Approximately 1-day person effort per month to manage the data.

3.1.3 Individual results from online surveys

Table 5: Individual results from online surveys

Dataset reference and name	
Dataset identifier	OnlineResponses
Data set description	
Generated or collected	Generated
Origin	N/A
Scale	Not yet known – 13 responses at M6
Who is this data useful for?	Internal demand analysis
Similar existing datasets	A number of surveys exist in this domain but their data is not available to this project. This data will enable EDSA to build up a country-by-country view of current capacity and requirements for data science skills.
Standards and metadata	
Methodology for data collection/management	The online survey collates data automatically and submits it to the private repository where it is aggregated and published to the public Github automatically on a daily basis (see section 3.1.4).
Metadata, supporting material	A README.md file is available detailing the data structure and basic usage.
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Data will be not shared or available for reuse.

Repository for data	Internal ODI repository.
If the data cannot be shared, why?	Data protection of personal data as contact details can be provided as part of the survey.
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	<100Mb
Who is responsible for data curating and management?	ODI lead data management and curation, other WP1 partners will contribute.
Quality assurance including back up procedures	Backed up to an internal ODI repository.
Associated costs for data management	Approximately 1-day person effort per month to manage the data.

3.1.4 Aggregated results from online survey

Table 6: Aggregated results from online survey

Dataset reference and name	
Dataset identifier	OnlineResponsesStatistics
Data set description	
Generated or collected	Generated
Origin	N/A
Scale	Not yet known
Who is this data useful for?	External analysis of results and trends by anyone who wishes to gather survey data in the area of data science. This data will also be used to inform and tailor the activities conducted by the project's activities in WP5, a European wide initiative (EDSA).
Similar existing datasets	There are a number of other surveys that have been aggregated that we can compare our result to and use these results if necessary. This dataset has the same eventual value to others in the area.
Standards and metadata	
Methodology for data collection/management	Aggregated responses from the online survey are automatically generated on a daily basis and published via Github.
Metadata, supporting material	A README.md file is available detailing the data structure and basic usage.
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0)
Data reuse	Data will be available to view on the EDSA dashboard and accessible for free in the EDSA dashboard Github repository.

Repository for data	Data will be available to view on the EDSA dashboard and all files accessible for free in the EDSA dashboard Github repository.
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	As long as Github exists as a minimum. Beyond that a value of maintaining the dataset would have to be evaluated
Approximate end volume	<100Mb
Who is responsible for data curating and management?	ODI lead data management and curation, other WP1 partners will contribute.
Quality assurance including back up procedures	Data is stored with external providers (Github).
Associated costs for data management	Github is free and public. Approximately 1-day person effort per month to manage the data.

3.1.5 Recordings and transcriptions of interviews

Table 7: Recordings and transcriptions of interviews

Dataset reference and name	
Dataset identifier	InterviewTranscriptions
Data set description	
Generated or collected	Generated
Origin	N/A
Scale	Not yet known. 11 recordings and transcriptions at M6.
Who is this data useful for?	Internal demand analysis only
Similar existing datasets	No similar datasets exist that are usable for this project. The interviews provide insights and data points for use in the demand analysis.
Standards and metadata	
Methodology for data collection/management	Qualitative research methodology for collection outlined in D1.1. Recordings are collected in mp3 format and transferred to the subcontracted company via a secure, private dropbox. Transcriptions are provided in a document.
Metadata, supporting material	Supporting documentation includes the interview questions and script saved in a shared Google drive accessible by the project partners.
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.

Data reuse	Data will not be shared or available for reuse. The data collected will be used for internal review to inform the creation of curriculum. Data will only be available publically once anonymous and aggregated via the EDSA Dashboard and via Github.
Repository for data	Internal ODI repository
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	Approximately 30 recordings and transcriptions are anticipated.
Who is responsible for data curating and management?	ODI will lead data management and curation, other WP1 partners will contribute, however, all data will be stored with the ODI.
Quality assurance including back up procedures	Backed up to an internal ODI repository.
Associated costs for data management	Transcription creation as part of the WP1 subcontracting allocation. Approximately 0.5-days of person effort per month to manage the data.

3.1.6 ideXlab search platform results

Table 8: ideXlab search platform results

Dataset reference and name	
Dataset identifier	ExpertIdentification
Data set description	
Generated or collected	Collected
Origin	Research and scientific publications available online.
Scale	Not yet known
Who is this data useful for?	Internal analysis of demand and supply, used as part of WP1, and exploitation plans for WP5. This data will also be informative for curriculum development.
Similar existing datasets	Not in this area. This dataset will provide validation of the demand analysis and form the basis for further insights and exploration of the domain.
Standards and metadata	
Methodology for data collection/management	Sampling approach outlined in D1.2. for data collection. Query containing multiple identified key words used. A list of results is manually created. CSV data can then be exported.
Metadata, supporting material	Internal ideXlab documentation on the platform.
Data Sharing	

Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Raw data will be not be shared or made available for reuse outside of the project.
Repository for data	ideXlab search platform internal repository.
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project
Approximate end volume	Not yet known
Who is responsible for data curating and management?	ideXlab lead data management and curation.
Quality assurance including back up procedures	Backed up to ideXlab repository.
Associated costs for data management	Approximately 2-days person effort per month. No other external costs.

3.2 Work package 2 – Curricula and course development

WP2 will collect data from openly available sources and create subsets of this data to be used in the learning resources produced. Data will also be collected about existing data science courses as part of the recommendations.

3.2.1 Linked open data sources

Table 9: Linked open data sources

Dataset reference and name	
Dataset identifier	N/A
Data set description	
Generated or collected	Collected
Origin	DBLP, GeoNames, others as identified throughout the project.
Scale	Not yet known
Who is this data useful for?	Users of the project’s curricula and learning materials – learners, educators, trainers.
Similar existing datasets	None. The datasets will be used within the learning activities as part of the project’s learning materials.
Standards and metadata	
Methodology for data collection/management	Systematic search and review of available datasets.

Metadata, supporting material	The datasets will be used within learning activities offered as part of the project's learning materials. Supporting material will be produced and included to allow correct interpretation.
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0)
Data reuse	Will be made available via the interactive elements of the project's learning materials and as a resource for members of the EDSA initiative as part of WP5.
Repository for data	DBLP, GeoNames, etc
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	The data will be available after the project ends as part of the project's learning materials.
Approximate end volume	< 1GB
Who is responsible for data curating and management?	OU lead data management and curation. Other WP2 partners will contribute.
Quality assurance including back up procedures	Back up procedures of the external dataset providers.
Associated costs for data management	Approximately 2-days person effort per month for collecting the data.

3.2.2 Publicly available datasets for each course

Table 10: Publicly available datasets for each course

Dataset reference and name	
Dataset identifier	N/A
Data set description	
Generated or collected	Collected
Origin	Government open data platforms such as data.gov.uk
Scale	Not yet known
Who is this data useful for?	Users of the project’s curricula and learning materials – learners, educators, trainers.
Similar existing datasets	None. The datasets will be used within the learning activities as part of the project’s learning materials.
Standards and metadata	
Methodology for data collection/management	Systematic web crawl and review of available datasets.
Metadata, supporting material	The datasets will be used within learning activities offered as part of the project's learning materials. Supporting material will be produced

	and included to allow correct interpretation.
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0)
Data reuse	Will be made available via the interactive elements of the project's learning materials and as a resource for members of the EDSA initiative as part of WP5.
Repository for data	Government open data platforms such as data.gov.uk as these are the sources of datasets and detail licensing and metadata.
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	The data will be available after the project ends as part of the project's learning materials.
Approximate end volume	< 1GB
Who is responsible for data curating and management?	OU lead data management and curation. Other WP2 partners will contribute.
Quality assurance including back up procedures	Back up procedures of the external dataset providers.
Associated costs for data management	Approximately 2-days person effort per month for collecting the dat.a

3.2.3 Related course data regarding similar modules and training offerings across the EU

Table 11: Related course data regarding similar modules and training offerings across the EU

Dataset reference and name	
Dataset identifier	DataScienceCourses
Data set description	
Generated or collected	Collected
Origin	Individual course websites
Scale	Not yet known
Who is this data useful for?	Internal use for development of curricula and learning materials as well as exploitation activities in WP5. External use for identifying useful courses.
Similar existing datasets	None. The data will provide a useful resource for those wishing to understand what courses are available.
Standards and metadata	
Methodology for data collection/management	Systematic web crawl and review of available courses.
Metadata, supporting material	Supporting text for the user, and links to course websites and information.

Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0)
Data reuse	Available on the EDSA project website and as a resource for members of the EDSA initiative as part of WP5.
Repository for data	Internal Southampton institutional repository and EDSA project website.
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	Until the end of the project
Approximate end volume	< 1GB
Who is responsible for data curating and management?	Southampton lead data management and curation
Quality assurance including back up procedures	Backed up remotely and shared on a Google drive between partners.
Associated costs for data management	Approximately 0.5-days person effort per month to maintain data.

3.2.4 Subsets of data from existing data repositories for courses

Table 12: Subsets of data from existing data repositories for courses

Dataset reference and name	
Dataset identifier	EDSAExercisesDatasets
Data set description	
Generated or collected	Generated
Origin	N/A
Scale	Not yet known
Who is this data useful for?	Users of the project's curricula and learning materials - learners, educators, trainers.
Similar existing datasets	None. The datasets will be used within learning activities offered as part of the project's learning materials.
Standards and metadata	
Methodology for data collection/management	Open datasets will be collected from various websites and repositories ready for integrated use where applicable in learning materials.
Metadata, supporting material	Will be made available via the interactive elements of the project's learning materials and as a resource for members of the EDSA

	initiative as part of WP5.
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0).
Data reuse	Will be made available via the interactive elements of the project's learning materials.
Repository for data	Not yet known
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	The data will be available after the project ends as part of the project's learning materials.
Approximate end volume	< 500MB
Who is responsible for data curating and management?	OU lead data management and curation. Other WP2 partners will contribute.
Quality assurance including back up procedures	Backed up remotely and hosted on a shared Google drive.
Associated costs for data management	Approximately 2 days person effort per month for generating the data.

3.3 Work package 3 – Training delivery and learning analytics feedback

WP3 will collect data on the training delivered in the project – face-to-face and online. This will include data on course registration, participation and completion. This will be used to inform best practices for students and educators to improve curricula and content. More data is intended to be collected from WP3 including learning analytics data from the EDSA website, and MOOC platforms. Although in the early stages of planning, Tin Can API (xAPI)[11] and Learning Locker[12] are being explored by the project to collect this data. This section of the DMP will be updated at M18 to incorporate the management elements of this data collection once available.

[11] http://tincanapi.com/

[12] http://learninglocker.net/

3.3.1 Repository statistics on downloads and views of educational resources

Table 13: Repository statistics on downloads and views of educational resources

Dataset reference and name	
Dataset identifier	Statistics
Data set description	
Generated or collected	Collected
Origin	videolectures.net
Scale	Views and comments for each video lecture
Who is this data useful for?	Internal analysis and demand analysis.
Similar existing datasets	None. Provides evidence of resource usage and basis for improving curriculum, content and course structure.
Standards and metadata	
Methodology for data collection/management	JSON is used for the videolectures API.
Metadata, supporting material	Videolectures REST API documentation.
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Aggregated results described as part of WP3 deliverables.
Repository for data	videolectures repository due to proximity to data source.
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	Not yet known
Who is responsible for data curating and management?	JSI lead data management and curation. OU contribute.
Quality assurance including back up procedures	videolectures internal quality assurance & back up procedures.
Associated costs for data management	No additional costs incurred.

3.3.2 Internal log of elearning systems

Table 14: Internal log of elearning systems

Dataset reference and name	
Dataset identifier	InternalLogs
Data set description	
Generated or collected	Collected
Origin	videolectures.net
Scale	20.000 videos, 17.431 lectures, 12.998 authors, 952 events, 579 categories.
Who is this data useful for?	Internal analysis and demand analysis.
Similar existing datasets	None. Provides evidence of resource usage and basis for improving curriculum, content and course structure.
Standards and metadata	
Methodology for data collection/management	JSON is used for the videolectures API.
Metadata, supporting material	Videolectures REST API documentation.
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Aggregated results described as part of WP3 deliverables.
Repository for data	videolectures repository due to proximity to data source.
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	Not yet known
Who is responsible for data curating and management?	JSI lead data management and curation. OU contribute.
Quality assurance including back up procedures	videolectures internal quality assurance & back up procedures.
Associated costs for data management	No additional costs incurred.

3.3.3 Statistics of course registration, participation and completion

Table 15: Statistics of course registration, participation and completion

Dataset reference and name	
Dataset identifier	Statistics
Data set description	
Generated or collected	Collected
Origin	videolectures.net
Scale	Not yet known
Who is this data useful for?	Internal analysis and demand analysis.
Similar existing datasets	None. Provides basis for improving curriculum, content and course structure.
Standards and metadata	
Methodology for data collection/management	JSON is used for the videolectures API.
Metadata, supporting material	videolectures REST API documentation.
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Aggregated results described as part of WP3 deliverables.
Repository for data	videolectures repository due to proximity to data source.
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	< 1GB
Who is responsible for data curating and management?	JSI lead data management and curation. OU contribute.
Quality assurance including back up procedures	videolectures internal quality assurance & back up procedures.
Associated costs for data management	No additional costs incurred.

3.3.4 Aggregated statistics of engagement with developed courses and educational resources

Table 16: Aggregated statistics of engagement with developed courses and educational resources

Dataset reference and name	
Dataset identifier	Aggregated Statistics
Data set description	
Generated or collected	Generated
Origin	videolectures.net
Scale	Not yet known
Who is this data useful for?	Internal analysis and demand analysis.
Similar existing datasets	None. Provides evidence of adoption and basis for improving curriculum, content and course structure.
Standards and metadata	
Methodology for data collection/management	JSON is used for Videolectures API.
Metadata, supporting material	Videolectures REST API documentation.
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Aggregated results described as part of WP3 deliverables.
Repository for data	videolectures repository due to proximity to data source.
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	< 1GB
Who is responsible for data curating and management?	JSI lead data management and curation. OU contribute.
Quality assurance including back up procedures	videolectures internal quality assurance & back up procedures.
Associated costs for data management	No additional costs incurred.

3.4 Work package 4 – Dissemination and community building

WP4 will continuously collect data from web server logs and Google analytics for the project website, as well as social media engagement data from Twitter and LinkedIn. This will allow for monitoring of the projects community building and dissemination. Aggregated statistics of the networking and engagement data will be produced, and included in D4.4 and D4.5.

3.4.1 Web server logs and Google analytics of project website

Table 17: Web server logs and Google analytics of project website

Dataset reference and name	
Dataset identifier	WebsiteAnalytics
Data set description	
Generated or collected	Collected
Origin	http://edsa-project.eu
Scale	Recorded traffic for 1 website.
Who is this data useful for?	Internal analysis for dissemination and community analysis. Secondary use for implicit demand analysis.
Similar existing datasets	None. Provides evidence of engagement and basis for UX improvement.
Standards and metadata	
Methodology for data collection/management	Quantitative recording of website traffic via Google Analytics dashboard, analysed using a variety of analytic tools.
Metadata, supporting material	Description of metric terms
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Analysed data findings will be made available throughout deliverable reports in WP4.
Repository for data	Internal institutional SOTON repositories.
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	< 1GB
Who is responsible for data curating and management?	OU lead data management and curation. Southampton will contribute.
Quality assurance including back up procedures	Backed up remotely by OU and Southampton.
Associated costs for data management	Free storage, approximately 0.5-days person effort per month.

3.4.2 Generated social media engagement data

Table 18: Generated social media engagement data

Dataset reference and name	
Dataset identifier	SocialMediaEngagements
Data set description	
Generated or collected	Collected
Origin	Twitter, LnkedIn
Scale	1 Twitter account, up to 30 LinkedIn community groups.
Who is this data useful for?	Internal analysis for community strength and project dissemination.
Similar existing datasets	None that relate to EDSA. Provides evidence for engagement with project, effectiveness of dissemination activities. Provides basis for understanding what content users find most engaging.
Standards and metadata	
Methodology for data collection/management	Regular access of data from analytics.twitter.com
Metadata, supporting material	Descriptions of data attributes.
Data Sharing	
Licensing, ownership and copyright	Data will be licensed in compliance with each social network's terms and conditions.
Data reuse	Dashboard on EDSA website. Deliverable reports in WP4.
Repository for data	Internal institutional Southampton repositories.
If the data cannot be shared, why?	Data sharing needs to comply with individual site licenses.
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	< 1GB
Who is responsible for data curating and management?	Southampton lead data management and curation.
Quality assurance including back up procedures	Backed up remotely by Southampton.
Associated costs for data management	Free storage. Approximately 1-day person effort per month.

3.4.3 Aggregated statistics of networking and engagement data

Table 19: Aggregated statistics of networking and engagement data

Dataset reference and name	
Dataset identifier	EngagementReports
Data set description	
Generated or collected	Generated
Origin	N/A
Scale	Not yet known
Who is this data useful for?	Internal analysis for dissemination and community building. External analysis to understand the EDSA networks.
Similar existing datasets	None that relate to EDSA. Provides evidence for engagement with project, effectiveness of dissemination activities. Provides basis for understanding the EDSA network.
Standards and metadata	
Methodology for data collection/management	Quantitative analysis of engagement data.
Metadata, supporting material	Supportive documentation on events (location, presentations, speakers).
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0)
Data reuse	Deliverable reports in WP4.
Repository for data	EDSA Dashboard.
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	< 500MB
Who is responsible for data curating and management?	Southampton lead data management and curation.
Quality assurance including back up procedures	Backed up remotely by Southampton.
Associated costs for data management	Free storage. Approximately 2-days person effort per month.

3.4.4 Learning materials access data

Table 20: Learning materials access data

Dataset reference and name	
Dataset identifier	LearningMaterialsAccess
Data set description	
Generated or collected	Collected
Origin	Various sources: MOOCs (Futurelearn, Coursera), project website, iBook Store.
Scale	Not yet known.
Who is this data useful for?	Internal analysis for dissemination and engagement with learning materials.
Similar existing datasets	Repository statistics on downloads and views of educational resources, and Statistics of course registration, participation and completion from WP3. These can be aggregated and integrated.
Standards and metadata	
Methodology for data collection/management	Quantitative recording of web server logs and page views.
Metadata, supporting material	Description of terms.
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0)
Data reuse	Deliverable reports in WP4. Dashboard on EDSA website.
Repository for data	EDSA Dashboard. The dashboard provides basis for the project to engage with learners.
If the data cannot be shared, why?	N/A
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	< 1GB
Who is responsible for data curating and management?	OU lead data management and curation. Southampton contribute.
Quality assurance including back up procedures	Backed up remotely by OU.
Associated costs for data management	Approximately 1-day person effort per month.

3.5 Work package 5 – Exploitation

WP5 will generate an ongoing list of established collaboration initiatives and institutions benefiting from the project and geographical regions using the project's results.

3.5.1 Project exploitation results

Table 21: Project exploitation results

Dataset reference and name	
Dataset identifier	ProjectExploitation
Data set description	
Generated or collected	Generated
Origin	Project partners
Scale	Not yet known
Who is this data useful for?	Internal analysis of impact of project and opportunities for continuation of work.
Similar existing datasets	None. Provides data on dissemination activity, network and results of the project
Standards and metadata	
Methodology for data collection/management	Report detailing results from interviews and exploitation activities.
Metadata, supporting material	N/A
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
Data reuse	Deliverable reports in WP5.
Repository for data	Shared drive between project partners.
If the data cannot be shared, why?	Privacy
Archiving and preservation	
How long should the data be preserved?	Until the end of the project.
Approximate end volume	< 500MB
Who is responsible for data curating and management?	ideXlab data management and curation.
Quality assurance including back up procedures	Backed up remotely by ideXlab.
Associated costs for data management	Free storage. Approximately 1-day person effort per month.