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

D5.7 Final Data Management Plan

Deliverable Editor:	**David Tarrant (ODI)**
Other contributors:	**Mandy Costello (ODI)**
	Ryan Goodman (ODI)
Deliverable Reviewers:	**Inna koval (JSI)**
	Simon Scerri (Fraunhofer)
Deliverable due date:	**31/01/2018**
Submission date:	**31/01/2018**
Distribution level:	**P**
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	01/11/2017	Ryan Goodman	Created document, added initial plan outline
0.2	06/11/2017	Ryan Goodman	Drafting Data Summary
0.3	21/11/2017	Ryan Goodman	Updating datasets relevant to their WP
0.4	15/12/2017	Mandy Costello	Drafting sections
0.5	05/01/2018	David Tarrant	Changes and updates to the data summary
0.6	22/01/2018	Simon Scerri	Review
0.7	30/01/2018	David Tarrant	Implement review comments
1.0	31/01/2018	Alexander Mikroyannidis	Final QA

Table of Contents

List of Tables

1. Executive Summary

Data is collected, generated and used within the EDSA project to support the goals of better understanding and disseminating information on the data science skills gap in Europe, producing learning resources and materials to support skill development in identified areas and to recommend best practice in learning delivery through innovative learning analytics.

EDSA collected and generated a number of datasets throughout the life of the project. The data is varied in type and topic and several methodological approaches were implemented across these.

At a high level data collected and used can be split into three main types:

1. Demand analysis data
2. Learning/Course supply data
3. Supplementary data (such as that used in courses)

This categorisation is newly introduced in this final version of the Data Management Plan (DMP) in order to clearly differentiate the long term preservation requirements of each dataset.

Demand analysis data was key during the first phases of the project to establish what data science skills gaps exist and what curriculum was required to address these gaps. The project collected a large amount of original data during this stage that could be very useful to other analysts for both research and application purposes. Thus where possible, the anonymised demand analysis data (such as that from individual surveys) owned by EDSA will be deposited into a trusted digital repository, alongside the deliverables, for long term access and preservation. For this purpose the project has selected Zenodo as this platform. This also offers the benefit of providing Digital Object Identifiers (DOIs) for the data as well as following the FAIR principals.

The Learning/Course supply data will not be archived or made openly available in its raw format due to data privacy and limitations of use, which have been addressed in previous DMPs. This data is also closely tied to instances of courses run by EDSA partners and thus the reuse potential is lower. Aggregated statistics have been included within project deliverables that will be archived in Zenodo.

Finally, supplementary data is listed in the DMP to ensure that rights for usage are clear, however no preservation action will be taken with this data.

2. Introduction

This report describes the final Data Management Plan (DMP) for the European Data Science Academy (EDSA) project, which is funded by the European Union's Horizon 2020 Programme under Grant Agreement number 643937.

The objective of the final DMP is to outline EDSA's data management policies for the datasets generated by the project inline with the H2020 pilot guidance on making data findable, accessible, interoperable and reusable (FAIR) and to give details of the datasets standards, metadata, sharing, archiving and preservation. This is the third and final version of the EDSA DMP (D5.7). As the final

DMP, there is specific focus given to archiving and preservation of data generated throughout the life of the project.

All information about the datasets generated as a result of the EDSA project can be accessed through the project's live register, which has been updated throughout the project to reflect changes and additions. The EDSA register is published under a CC-BY-4.0 Creative Commons license. It can be accessed through the EDSA website at http://edsa-project.eu/resources/datasets/.

There have been no significant additions to the dataset registry since the previous data management plan. The main purpose of this final data management plan is to focus on the long term preservation and archiving of original data created as part of the EDSA project and provide an update on the continued use and management of third party data services, such as that used on the dashboards.

3. Data Summary

In total, there are 29 datasets recorded in the EDSA dataset register, including the register itself, which has been published openly. During the demand analysis phase of the project, four new datasets were generated and published with a further two collected from third parties that cannot be published due to terms of use. During the analysis phase of the project 11 datasets were used to evaluate engagement with the project and its outputs and curricular and courses. Finally the project makes use of a number of datasets within courses themselves which are also recorded in the data register.

In the following sections, we summarise these data categories in more detail.

3.1 Data Summary - Category 1- Demand analysis data

Demand analysis data, including that used on the dashboard forms a key part of the project and is used in many deliverables to evidence our approach, such as "D1.2 Study Evaluation Report 1". As such, the data has been prioritised for open release where possible to allow others to repeat our or perform their own analysis.

The demand analysis data is directly tied to Work Packages 1 and 2 surrounding the demand analysis and development of curricular.

As part of the demand analysis nearly 500 survey responses were collected. This dataset, along with other related datasets has been anonymised and published openly for others to reuse, where permission was obtained. Additionally this dataset will be deposited into a partners institutional repository, alongside the analysis documents, for long term preservation.

As the project developed as did the scope and volume of data analysed. The project dashboard has been developed throughout the project to provide a "live" view on the demand and supply from data science throughout Europe. Data for analysis is provided under license by a third party provider for use in the project. Due to licensing conditions the project is not permitted to re-publish the data (see D5.6, D5.7). In these cases the provider of the raw data have been identified and summary data, such as that shown via the graphs on the dashboard can be downloaded allowing users to take a snapshot of the dashboards data for use and reuse. These snapshots of data from the dashboard, will be archived

and preserved alongside the deliverables in which the data was used. Data will be made available in a machine readable open format allowing others to reproduce the results of the deliverables.

Where possible, demand analysis data will be deposited in an institutional repository and given a Digital Object Identifier, thus maximising reuse.

Appendix 1 gives more details on the demand analysis data that includes:

- Individual results from demand analysis (not openly available)
- De-identified survey responses from demand analysis.
- Recordings and transcriptions of interviews (not openly available).
- Related course data regarding similar modules and training offerings across the EU.

Appendix 1 gives more details on third party sourced data which includes:

- Corpora of crawled web-based adverts from LinkedIn
- Training and job adverts sourced from the JSI QMiner platform including:
    - Expert identification results (ideXlab search platform results)
    - Aggregated statistics of European skill demand based on web-based job adverts

WP/dataset	Lead	Dataset	Project Phase	Status	Archived in open repository
WP1 (1)	**ODI**	Individual results from demand analysis	M2-M18	Published	Not permitted
WP1 (2)	**ODI**	Raw anonymised data from demand analysis	M2-M18	Published	In progress
WP1 (3)	**ODI**	Recordings and transcriptions of interviews	M2-M18	Published	Not permitted
WP2 (4)	**ODI**	Related course data regarding similar modules and training offerings across the EU	M18	Published	In progress
WP1 (-)	**ODI**	Corpora of crawled web-based adverts from LinkedIn	M6-M18	No longer used	Not permitted
WP1 (5)	**ODI**	ideXlab search platform results	M6-M36	Ongoing	Not permitted
WP1 (6)	**ODI**	Aggregated statistics of European skill demand based on web-based job adverts	M6-M18	Ongoing	Not permitted

Table 1: Demand Analysis Data

Dataset (2) and (4) in the table above were created by the EDSA project, are thus owned by the project and can be published openly, Datasets (1) and (3) contain information about individuals and cannot be placed in an open repository. Datasets (5) and (6) are sourced from third parties whose terms of use forbid verbatim republishing of the source data, however aggregated and enriched statistics will be archived alongside the relevant deliverables.

3.2 Data Summary - Category 2- Learning/Course supply data

Learning/Course supply data has been used to evaluate the success of EDSA resources against Key Performance Indicators (KPIs) and in order to derive best practice in training delivery. Again, much of the data is related to individual interactions with learning materials. In this case, appropriate and complete aggregate statistics are included within tables of the deliverable documents.

The learning analytics data will not be preserved for long term archiving for a number of key reasons:

1. **Very limited or no reuse potential**: Beyond the use to evaluate the effectiveness of the project against a number of KPIs the data has very little reuse potential. The majority of the data is specifically tied to instances of courses and appropriate aggregated statistics are presented in the related deliverables.
2. **High privacy impact**: The majority of the data related to learning analytics is highly personal and cannot be shared outside of the project. Additionally, data privacy policies mean that users data has to be removed after a period of time. Management of this process has been retained by the data owners.
3. **Third party licenses**: Many of the learning platforms (such as FutureLearn) have specific privacy and data usage policies that prevent the wider sharing of data.

A description of the datasets that form part of the analysis that will not be archived are included in the table below:

WP/Dataset	Lead	Dataset	Project Phase	Status
WP2 (7)	**TU/e**	Event log from a municipality process	M12-M36	Finished
WP3 (8)	**OU**	Learning Analytics data generated from the EDSA Online Courses portal	M12-M36	Finished
WP3	**JSI**	Repository statistics on downloads and views of educational resources	M12-M36	Finished
WP3	**JSI**	Internal logs of elearning systems	M12-M36	Finished
WP3	**JSI**	Statistics of course registration, participation and completion	M12-M36	Finished
WP3	**JSI**	Aggregated statistics of engagement with the developed courses and educational resources	M12-M36	Finished
WP3	**TU/e**	Recorded behavior of students following the first session of the process mining MOOC	M12	Finished
WP3	**TU/e**	FutureLearn course run data 'Introduction to process mining with ProM'	M36	Finished
WP4	**SOTON**	Web server logs and Google analytics of project website access	M12-M36	Ongoing
WP4	**SOTON**	Generated social media engagement data	M12-M36	Ongoing
WP5	**ideXlab**	List of project exploitation results – collaborations, institutional and geographical beneficiaries,	M18-M36	Ongoing

Table 2: Learning/Course supply datasets

3.3 Data Summary - Category 3 – Supplementary Data such as that used in courses

Supplementary data, such as that used in the delivery of EDSA courses is also recorded in the data management plan for reference. While this data is published, its usage as a learning resource means that it does not conform to a particular data management policy, however, it must be open for others to use.

Work Package	Lead	Dataset	Project Phase	Status
WP2	**Persontyle**	Datasets for course examples and exercises	M6-M36	Ongoing
WP2	**ODI**	Monthly Rainfall (mm) Totals for Selected Stations in Tanzania, 2014	M6-M36	Published
WP2	**ODI**	BBC RSS Feed	M6-M36	Published
WP2	**ODI**	Health Facility list ratings Tanzania	M6-M36	Published
WP2	**ODI**	Louisiana Secretary of State Officials	M6-M36	Published
WP2	**ODI**	US Projects Dataset	M6-M36	Published
WP2	**ODI**	UK GP Earnings	M6-M36	Published

Table 3: Supplementary datasets

3.4 The EDSA Data register

In addition to the data management plan the EDSA dataset register is also available as a live dataset for people to reference. The register holds records of all datasets used in EDSA and outlines information about the dataset standards, licensing and sharing details. The EDSA register dataset is hosted by the ODI, through Github and available through the EDSA website.

Work Package	Lead	Dataset	Project Phase	Status
WP5	**ODI**	EDSA register	M6-M36	Published

Table 4: The EDSA data register

4. Policy

In D5.5 we outlined the overall EDSA policies for data standards and metadata standards, data sharing and data preservation, in line with best practice for establishing a DMP.[1]

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'.

4.1 Data Standards and metadata policy for EDSA

Where it is possible to publish data, the project enforces the application of the FAIR data principles to make data findable, accessible, interoperable and reusable.

Standardising the project's collection and production of data ensures reusability and interoperability within the project, and externally if openly available.

4.1.1 FAIR data (Findability, Accessibility, Interoperability and reusability)

Each dataset is recorded in the EDSA dataset register once it is created. It is given an identifier and information is logged about the datasets standards, metadata, sharing, archiving and preservation by the partner responsible for its generation. The EDSA register dataset is hosted by the ODI, through Github and available through the EDSA website.

Where possible, data is made available in CSV, JSON or linked data in RDF format, to allow maximum interoperability if publishing openly or for interoperability amongst consortium partners. Due to the varied nature of data collected, we will use widely adopted metadata standards for describing the data. Use of generic vocabularies, such as dublincore2 and DCAT will be used to make datasets easily discoverable and interoperable. Further data packages 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.

Due to the variety of data used and generated through the project, EDSA partners adhere to their institutions' naming conventions.

As an example, the following is taken from the EDSA demand analysis summary data. This data is published using the Open Data Institutes OctoPub tool that creates a data package conforming to the data package specification. A Data Package is a simple container format used to describe and package a

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

collection of data. The format provides a simple contract for data interoperability that supports frictionless delivery, installation and management of data.

The extensible data package metadata schema is delivered in the JSON format which both describes the data and the way it is packaged. Some of the metadata fields used are shown below:

Metadata field	**Description**	**Example**
name	a-unique-human-readable-and-url-usable-identifier	edsa-demand-analysis-summary-data
title	A nice title	EDSA demand analysis summary data
description	A long description to help the consumer understand the data.	“As part of the demand analysis - to identify the skills gap and data science training needs in Europe - 584 quantitative surveys and 108 qualitative...
version	A version number	1.0
keywords	A comma separated list of keywords	Survey results, EDSA, Data Science, Demand analysis
licenses	The data licence	"url": "https://creativecommons.org/licenses/by/4.0/", "name": "Creative Commons Attribution 4.0", "id": "cc-by 4.0"
sources	The sources of the data of from a third party.	“name”: “LinkedIn”, “web”: “https://www.linkedin.com/”
contributors	Contributors to the dataset	“name”: “The Open Data Institute”, “web”: “https://theodi.org”
maintainers	Maintainers of the dataset	“name”: “The Open Data Institute”, “web”: “https://theodi.org”
publishers	Publishers of the dataset	“name”: “The Open Data Institute”, “web”: “https://theodi.org”
dependencies	Any dependencies to obtain a complete dataset (e.g. if monthly).	"data-package-name": ">=1.0"
resources	Links and descriptions of the data files that are included in the package.	"name": "EDSA file 1", "mediatype": "text/csv", "description": "Sample by country.", "path": "data/data1.csv"
[custom field]	Adding your own custom fields	...

Table 5: Example of metadata fields

4.1.2 Making data openly accessible

Research data created in this project is owned by the partner who generates it. Each partner must disseminate its results as soon as possible unless there is legitimate reason to protect the results. For example, the project has generated several datasets which will not be made openly available as they contain identifiable information about individuals. See section on 'Increase data re-use (through clarifying licences)' for more information on EDSA's licensing policy.

When it is not possible to publish collected data due to privacy or license obligations, we will list the use of the dataset within the project registry. We will also look to publish derived or aggregated statistics of data where possible.

Information about sensitive and restricted data will be published in the project registry with details about if, when and how the dataset will become available included in the metadata.
Where a restriction on open access to data is necessary, attempts will be made to make data available under controlled conditions, such as through including contact details for the data owner and holder.

Due to the diverse nature of the data collected as part of the EDSA project it has been decided to take a mixed methods approach to publication. This ensures that data can be disseminated via the most appropriate channels and remain closely tied to deliverables as well as the interactive dashboards created as part of the project.

There are four main repositories for EDSA data:

1. **Open access repositories**: We are following a policy of 'open by default.' If there is no reason why the data cannot or should not be published openly, then our policy is that it should be published under an open licence. Open data about individuals should be de-identified, and only published with the consent of the individuals concerned. The data should also be unrestricted by terms of use.
2. **The EDSA project website**: The aim is that all of the data that is published openly will be made available via the EDSA website. This is to ensure that the data is findable by as wide an audience as possible. Data that is openly licensed but difficult to discover is not widely considered to be open data. The EDSA website also displays data that cannot be published openly, often due to restrictions in terms of use. This allows users to view the data, or aggregations of the data.
3. **Internal institutional and organisational repositories:** Some datasets in the dataset registry are hosted in repositories of the organisation responsible for that data. While some of these are internal, hosted in Consortium partners' internal repositories, some datasets used in course materials are hosted on external repositories therefore that organisation is responsible for maintaining the data. Datasets hosted in internal repositories cannot be published, usually due to restrictions of use of personal data.
4. **Trusted digital repositories for preservation**: In addition to open access repositories, the final phase of the project will see the deposit of both project deliverables and related datasets created and owned by the project into a long term trusted digital repository. The project has selected the Zenodo as the place for long term preservation. Zenodo provides a repository for EC funded research, including publications and datasets.

4.1.3 Making data interoperable

Throughout the project we have used the register to collect and document the data in a standardised way to ensure that the datasets can be understood, interpreted and shared in isolation alongside accompanying metadata and documentation. Best practices have been followed to link deliverables to their source data such as others can reproduce project results.

Generated data will be preserved on a number of platforms both during and after the completion of the project.

A metadata file will be created and linked within each dataset conforming to the data package standard from table 5.

Alongside this general metadata, each dataset will be accompanied with a document that describes the structure of the data e.g. definitions of column titles and how the dataset should and should not be interpreted. Where applicable data sets should conform to widely used metadata vocabularies and ontologies.

4.1.4 Increase data re-use (through clarifying licences)

Datasets created in the project are owned by the partner who generated it. To ensure accessibility, where possible, open data will be provided so that others are able to access, use and share the data. 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'[2]

The consortium will use the ODIs Open Data Certificates[3] as a certification framework to indicate to users the quality of the data and documents accessibility aspects of the dataset.

4.2 Allocation of resources

The lead partner who generates a dataset through the project is responsible for maintenance of the dataset, including version control and updates. The partner is also responsible for ensuring the dataset complies with EDSA data management policies, as outlined in this DMP.

The EDSA dataset registry outlines the resources approximated for each partner to maintain the data during the life of the project.

[2] https://creativecommons.org/licenses/by/4.0/

[3] http://certificates.theodi.org/en/

As part of establishing the EDSI (see D5.2) partners have agreed on a commitment to continue maintenance of the main project outputs, including datasets, which will ensure the results of the project are sustained. During the post-project period, the lead consortium partner (OU) will be assigned to monitor and update the European Data Science Academy Register.

4.3 Data Security

For the duration of the project, datasets will be stored with the appropriate partner, following their institutions policies on data management, security and storage. Individual partners are responsible for their own data management and security. The table below shows the selected services used by each project partner.

Partner reference	Data Storage
ODI	Github & Google Cloud Storage
SOTON	Internal institutional repository
OU	EDSA Learning Locker
ideXlab	Internal ideXlab storage
JSI	JSI server
KTH	KTH internal course repository
Tue	Local storage at TU/e, to store privacy sensitive data.
Fraunhofer	Dydra
Persontyle Group Limited	CRAN

Table 6: Internal partner repository

4.4 Ethical aspect

EDSA partners will comply with guidance in Article 34 of the Grant Agreement regarding ethical standards and principles.

Over the course of the project, a number of datasets that have been generated that will not be shared or opened in their original or raw format. This includes original copies of interviews from the demand analysis (WP1) or data containing information about individuals as part of the project's learning analytics (WP3)

The consortium must ensure compliance with the principles set out in Article 39 of the Grant Agreement, which states that the beneficiaries must inform the personnel whose personal data are

collected and processed[4]. The consortium's policy is to inform individuals of data collection, and the intention for its use. No identifiable information about individuals will be published. Any aggregated or de-identified data will first be assessed using a Privacy Impact Assessment framework (such as the one from the Information Commissioner's Office[5]).

4.5Appendix 1 – Demand analysis data in detail

Table 7: Individual results from demand analysis

Dataset reference and name	
Dataset identifier	IndividualResponses
Dataset description	
Generated or collected	Generated
Origin	Guided surveys and online responses
Scale	584 surveys 108 interviews
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	Data collection methods outlined in D1.4. Translated into CSV format.
Metadata, supporting material	Data will be not available for reuse or accessible by anyone outside of the project. The data collected will be used for internal analysis to inform the creation of curriculum. Anonymised data will be publicly available.
Status and location of metadata	Metadata is not publically available
Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
If the data cannot be published openly, why?	Data protection of personal data
How will the data be shared?	Data will be not shared or available for reuse
Data repository	Internal ODI repository
Dataset Link	There is no external link
Archiving and preservation	
How long should the data be preserved?	Until the end of the project

[4] Grant Agreement number: 643937 — EDSA — H2020-ICT-2014/H2020-ICT-2014-1

[5] https://ico.org.uk/media/for-organisations/documents/1595/pia-code-of-practice.pdf

Approximate end volume	<100Mb
Who is responsible for data curation 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	Negligible
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Contains personal data
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	No

Table 8: Raw anonymised data from demand analysis

Dataset reference and name	
Dataset identifier	DeidentifiedResponses
Dataset description	
Generated or collected	Generated
Origin	Online Survey http://edsa-project.eu/resources/survey/
Scale	496 survey results
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
Similar existing datasets	There are a number of other surveys that have been aggregated that we can compare our result too 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	Data collection methods outlined in D1.4. Translated into CSV format.
Metadata, supporting material	A README.md file is available detailing the data structure and basic usage.
Status and location of metadata	http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis-/
Data Sharing	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/
If the data cannot be published openly, why?	The data is published openly
How will the data be shared?	Data will be available to view on the EDSA dashboard and accessible for free in the EDSA dashboard Github repository.

Data repository	Github/ EDSA Dashboard on website
Dataset Link	http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis-/
Archiving and preservation	
How long should the data be preserved?	As long as Github exists as a minimum. Beyond that a value judgement would have to be made.
Approximate end volume	<100Mb
Who is responsible for data curation and management?	ODI lead data management and curation, other WP1 partners will contribute
Quality assurance including back up procedures	Stored in external repositories - EDSA website and Github
Associated costs for data management	Stored in external repositories - EDSA website and Github
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 9: Recordings and transcriptions of interviews

Dataset reference and name	
Dataset identifier	InterviewTranscipts
Dataset description	
Generated or collected	Generated
Origin	Interviews
Scale	108 transcripts 108 recordings
Who is this data useful for?	Internal demand analysis
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 and quantitative research methodology for collection outlined in D1.4
Metadata, supporting material	Data will be not available for reuse or accessible by anyone outside of the project. The data collected will be used for internal analysis to inform the creation of curriculum.
Status and location of metadata	Metadata is not publically available

Data Sharing	
Licensing, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
If the data cannot be published openly, why?	Data protection of personal data
How will the data be shared?	Data will be not shared or available for reuse
Data repository	Internal ODI repository
Dataset Link	There is no external link
Archiving and preservation	
How long should the data be preserved?	Until the end of the project
Approximate end volume	<3GB
Who is responsible for data curation and management?	ODI lead data management and curation
Quality assurance including back up procedures	Backed up to an internal ODI repository
Associated costs for data management	As part of the subcontracting costs of WP1
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Raw data including personal data
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	No

Table 10: Related course data regarding similar modules and training available across the EU

Dataset Reference and Name	
Dataset Identifier	DataScienceCourses
Dataset description	
Generated or collected	Collected
Origin	Course websites
Scale	459 courses (0.5Mb)
Who is this useful for?	Internal use for development of curricula and learning materials.
Similar existing datasets	None. The data will provide a useful resource as part of the demand analysis.
Standards and metadata	
Methodology for data collection/management	Systematic search and review of available data science courses. The search terms were Data Science, Big Data, Data Analytics, Business Analytics, Machine Learning, Distributed Computing, Advanced Computing Data Science Stream, Data Analytics stream.

Metadata, supporting material	Metadata has been published alongside the data
Status and location of metadata	
Data sharing	
Licensing, data protection, ownership and copyright	The data is licensed under a Creative Commons CC-BY 4.0 licence
If the data cannot be published openly, why?	The data is published openly
How will the data be shared?	GitHub/EDSA website
Data repository	GiHub. Also available via the EDSA website
Dataset Link	https://theodi.github.io/data-science-courses-in-europe-2016/
Archiving and preservation	
How long should the data be preserved?	Until the end of the project
Approx end volume	< 1GB
Who is responsible for the data management and curation?	ODI lead data management and curation
Quality assurance including back up procedures	Backed up to an internal ODI repository
Associated costs for data management	As part of the subcontracting costs of WP1. No ongoing costs.
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 11: Corpora of crawled web-based adverts from LinkedIn

Dataset Reference and Name	
Dataset Identifier	WebSiteHarvest
Dataset description	**Dataset description**
Generated or collected	Collected
Origin	LinkedIn
Scale	46 terms 31 languages 47 countries

	1 harvest per day 2162 data points per day
Who is this useful for?	Internal demand analysis and to inform curriculum development.
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 and integration with existing datasets not viable. The value of this dataset comes from the provision of an up-to-date snapshot of current data science skills needs across the EU.
Standards and metadata	**Standards and metadata**
Methodology for data collection/management	All data collected is translated into CSV format.
Metadata, supporting material	Data will be not available for reuse or accessible by anyone outside of the project. The data collected will be used for internal analysis to inform the creation of curriculum.
Status and location of metadata	Metadata is not publically available
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	Usage of the LinkedIn service is bound by the user agreement
If the data cannot be published openly, why?	The terms of the LinkedIn user agreement **now** forbid harvesting and collection of data without express permission. When the data was collected, this was not the case. https://www.linkedin.com/legal/user-agreement?trk=hb_ft_userag
How will the data be shared?	Data will be not shared or available for reuse
Data repository	Internal ODI Repository
Dataset Link	There is no external link
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	Until the end of the project
Approx end volume	<1Gb
Who is responsible for the data management and curation?	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 effort per month
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can	Yes

have an impact on sharing this data? Y/N	
What are the ethical or legal issues, if any, that can occur from sharing this data?	Internal demand analysis only
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	NA

Table 12: ideXlab search platform results

Dataset Reference and Name	
Dataset Identifier	ExpertIdentification
Dataset description	**Dataset description**
Generated or collected	Collected
Origin	Research publications
Scale	Not yet known as collection is ongoing
Who is this useful for?	Internal demand analysis and to inform curriculum development. Provides insights into offer side of skills analysis.
Similar existing datasets	Not in this area. This dataset will provide validation of the demand analysis and form the basis for further insights.
Standards and metadata	**Standards and metadata**
Methodology for data collection/management	The ideXlab search engine will use the sampling approach outlined in D1.2. for data collection. CSV data will be created
Metadata, supporting material	Data will be not available for reuse or accessible by anyone outside of the project. The data collected will be used for internal analysis to inform the creation of curriculum.
Status and location of metadata	Accompanying document to explain data structure. This will not be made open.
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
If the data cannot be published openly, why?	Data protection of personal data
How will the data be shared?	The data will not be shared due to restrictions on the use of personal data.
Data repository	ideXlab search platform
Dataset Link	There is no external link
Archiving and preservation	**Archiving and preservation**

How long should the data be preserved?	Until the end of the project
Approx end volume	Est. 1000 returns
Who is responsible for the data management and curation?	ideXlab lead data management and curation, other WP1 partners will contribute
Quality assurance including back up procedures	Backed up to an internal ideXlab repository
Associated costs for data management	Approx 2 person days per month. No other external costs
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Raw data including personal data
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	No

Table 13: Aggregated statistics of European skill demand based on web-based job adverts

Dataset Reference and Name	
Dataset Identifier	WebSiteStatistics
Dataset description	**Dataset description**
Generated or collected	Collected
Origin	Adzuna API[6] Trovit[7]
Scale	Varied
Who is this useful for?	Populating the dashboard, internal demand analysis and to inform curriculum development.
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 and integration with existing datasets not viable. The value of this dataset comes from the provision of an up-to-date snapshot of current data science skills needs across the EU.
Standards and metadata	**Standards and metadata**
	All data collected is translated into CSV format.

[6] https://developer.adzuna.com/
[7] https://www.trovit.com/

Methodology for data collection/management	
Metadata, supporting material	The Adzuna data is accessible via the Adzuna API. The Trovit data will be not available for reuse or accessible by anyone outside of the project.
Status and location of metadata	Metadata is not publically available
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	The data will be available for use via the EDSA dashboard However it will not be available to download as this contravenes Trovit's terms and conditions.
If the data cannot be published openly, why?	Trovit's terms of use prohibit the use of their data. The research exception allows us to use the data but not to make it available in raw format for others to consume for commercial purposes.
How will the data be shared?	Via the EDSA dashboard
Data repository	In an internal JSI repository
Dataset Link	N/A
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	Until the end of the project
Approx end volume	<1Gb
Who is responsible for the data management and curation?	ODI lead data management and curation, other WP1 partners will contribute
Quality assurance including back up procedures	Backed up to an internal JSI repository
Associated costs for data management	Approximately 1 day effort per month
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Data can only be used for research purposes
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	NA

4.6Appendix 2 – Learning/Course supply data

4.6.1 Work Package 3

WP3 collected data on the training delivered in the project – face-to-face and online.

This includes data on course registration, participation and student retention rate. We use this data to inform best practices for students and educators, and to improve the curricula and content. This is still a lot to be explored around the learning analytics data, especially as we continue to create more online modules. Different partners have created modules using different software. - for example Coursera[8], Tin Can API (xAPI)[9], Learning Locker[10].

Table 14: Event log from municipality process

Dataset Reference and Name	**Dataset Reference and Name**
Dataset Identifier	a07386a5-7be3-4367-9535-70bc9e77dbe6
Dataset description	**Dataset description**
Generated or collected	Collected
Origin	Dutch municipality
Scale	200 KB
Who is this useful for?	Users interested in real life event logs.
Similar existing datasets	Large collection of real life event logs at http://data.3tu.nl/repository/collection:event_logs_real
Standards and metadata	**Standards and metadata**
Methodology for data collection/management	Management through 3TU datacentre
Metadata, supporting material	Includes number of traces, events, attributes, timespan, etc.
Status and location of metadata	http://data.3tu.nl/repository/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	Own licence (Attribution, non-commercial) http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/General_terms_of_use_3TU.Datacentrum.pdf
If the data cannot be published openly, why?	The data is available publicly. As there are restrictions of use with the licence, this cannot be considered 'open data'
How will the data be shared?	Via 3TU Datacentre
Data repository	3TU Datacentre
Dataset Link	http://data.3tu.nl/repository/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	past project end

[8] https://www.coursera.org
[9] http://tincanapi.com/
[10] http://learninglocker.net/

Approx end volume	200 KB
Who is responsible for the data management and curation?	3TU
Quality assurance including back up procedures	Reliant on third party. If the dataset becomes unavailable we will use a similar one in the online module.
Associated costs for data management	None
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	No

Table 15: Repository statistics on downloads and views of educational resources

Dataset Reference and Name	
Dataset Identifier	RepositoryStatistics
Data set description	
Generated or collected	Collected
Origin	videolectures.net
Scale	Views and comments for each video lecture
Who is this useful for?	Internal analysis, curriculum development, external 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	CSV is used for Videolectures API
Metadata, supporting material	Videolectures REST api documentation. An MD Readme file is available for download
Status and location of metadata	https://github.com/innanoval/edsa-videolectures-statistics-dataset-1/tree/gh-pages/data
Data sharing	**Data sharing**

Licensing, data protection, ownership and copyright	The data is published under a CC-BY licence.
If the data cannot be published openly, why?	N/A
How will the data be shared?	Available to see at videolectures website; described as part of WP3 deliverables; published on Github
Data repository	Github/videolectures repository. Proximity to data source.
Dataset Link	https://github.com/innanoval/edsa-videolectures-statistics-dataset-1/tree/gh-pages/data
Archiving and preservation	**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
Approx end volume	< 1GB
Who is responsible for the data management and curation?	JSI lead data management and curation. OU contribute
Quality assurance including back up procedures	videolectures - relying on internal quality assurance & back up procedures
Associated costs for data management	Approximately 1 day per month during the project's lifetime
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 16: Learning Analytics data generated from the EDSA Online Courses portal

Dataset Reference and Name	
Dataset Identifier	EDSAOnlineCoursesLA
Data set description	**Data set description**
Generated or collected	Generated
Origin	http://courses.edsa-project.eu
Scale	Small (<100Mb)

Who is this useful for?	Course producers can get an understanding of how their courses are being used. Learners can monitor their learning progress.
Similar existing datasets	Not many Learning Analytics datasets are publicly available. The OU has recently published a similar dataset: https://analyse.kmi.open.ac.uk/open_dataset
Standards and metadata	
Methodology for data collection/management	The xAPI specification is used for expressing the data; the open source Learning Locker software is used for storing and visualising the data.
Metadata, supporting material	Introduction to the xAPI (or Tin Can API): https://tincanapi.com/overview/. Introduction to Learning Locker: https://learninglocker.net
Status and location of metadata	https://tincanapi.com/overview/ https://learninglocker.net https://alexmikro.github.io/learning-analytics-dataset-from-the-edsa-online-courses-portal/
Data sharing	
Licensing, data protection, ownership and copyright	Creative Commons Attribution (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/
If the data cannot be published openly, why?	The data is published openly.
How will the data be shared?	Via the EDSA website / Github
Data repository	We have setup a dedicated EDSA Learning Locker. This was chosen for the reasons outlined in https://learninglocker.net/benefits/
Dataset Link	https://alexmikro.github.io/learning-analytics-dataset-from-the-edsa-online-courses-portal/
Archiving and preservation	
How long should the data be preserved?	At least until the end of project
Approx end volume	Not yet known
Who is responsible for the data management and curation?	OU lead data management and curation.
Quality assurance including back up procedures	Relying on the backup procedures of the OU, as the dataset is hosted on an OU server.
Associated costs for data management	Server storage has already been purchased. Effort for analysing the data has been allocated in Task 3.4.

Data Ethics	Data Ethics
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 17: Internal log of eLearning systems

Data set description	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 useful for?	Internal 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 Videolectures API
Metadata, supporting material	Videolectures REST api documentation
Status and location of metadata	N/A
Data sharing	
Licensing, data protection, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
If the data cannot be published openly, why?	Privacy. Data requires anonymisation and/or aggregation, and at the moment the use case for anonymised data is not clear.
How will the data be shared?	Available to see at videolectures website; described as part of WP3 deliverables
Data repository	videolectures repository. Proximity to data source.

Dataset Link	There is no external link
Archiving and preservation	
How long should the data be preserved?	at least until the end of project
Approx end volume	N/A
Who is responsible for the data management and curation?	JSI lead data management and curation. OU contribute
Quality assurance including back up procedures	Videolectures - relying on internal quality assurance & back up procedures
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Privacy
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	NA

Table 18: Statistics of course registration, participation and completion

Dataset Reference and Name	
Dataset Identifier	StatisticsForCourses
Data set description	**Data set description**
Generated or collected	Collected
Origin	videolectures.net
Scale	for videolectures - available per videolecture, per viewer
Who is this useful for?	Internal 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 Videolectures API
	Videolectures REST api documentation

Metadata, supporting material	
Status and location of metadata	N/A
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
If the data cannot be published openly, why?	Privacy. Data requires anonymisation and/or aggregation. It is intended that this data will be published before the end of the project.
How will the data be shared?	Available to see at videolectures website; described as part of WP3 deliverables
Data repository	videolectures repository. Proximity to data source.
Dataset Link	N/A
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	at least until the end of project
Approx end volume	< 1GB
Who is responsible for the data management and curation?	JSI lead data management and curation. OU contribute
Quality assurance including back up procedures	videolectures - relying on internal quality assurance & back up procedures
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	NA

Table 19: Aggregated statistics of engagement with the developed courses and educational resources

Dataset Reference and Name	
Dataset Identifier	AggregatedStatistics
Data set description	**Data set description**

Generated or collected	Generated
Origin	videolectures.net
Scale	for videolectures - available per videolecture, per viewer
Who is this useful for?	internal analysis, demand analysis
Similar existing datasets	None. Provides evidence of adoption and basis for improving curriculum, content and course structure.
Standards and metadata	**Standards and metadata**
Methodology for data collection/management	JSON is used for Videolectures API
Metadata, supporting material	Videolectures REST api documentation
Status and location of metadata	N/A
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
If the data cannot be published openly, why?	Privacy. Data that does not contain privacy issues might be publishable
How will the data be shared?	Available to see at videolectures website; described as part of WP3 deliverables
Data repository	videolectures repository. Proximity to data source.
Dataset Link	N/A
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	at least until the end of project
Approx end volume	< 1GB
Who is responsible for the data management and curation?	JSI lead data management and curation. OU contribute
Quality assurance including back up procedures	videolectures - relying on internal quality assurance & back up procedures
Associated costs for data management	Approximately 1 day of effort p**er** month
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No

What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 20: Recorded behavior of students following the first session of the process mining MOOC

Dataset Reference and Name	
Dataset Identifier	CourseraMOOCprocmin001
Data set description	**Data set description**
Generated or collected	collected
Origin	coursera.org
Scale	several large tables
Who is this useful for?	learning analytics within EDSA
Similar existing datasets	Every Coursera course has this data recorded
Standards and metadata	**Standards and metadata**
Methodology for data collection/management	Data collection is managed by Coursera
Metadata, supporting material	There is no external link to the metadata
Status and location of metadata	There is no external link to the metadata
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	Raw data is managed by TU/e and cannot be shared due to Coursera restrictions of use.
If the data cannot be published openly, why?	Restrictions of use from the data provider
How will the data be shared?	This data will not be published openly
Data repository	The data is collected by and stored on a Coursera repository.
Dataset Link	There is no external link to the data.
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	N/A
Approx end volume	around 1 GB

Who is responsible for the data management and curation?	Joos Buijs
Quality assurance including back up procedures	N/A
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Student identifiable data
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	No

Table 21: FutureLearn course run data 'Introduction to process mining with ProM'

Dataset Reference and Name	
Dataset Identifier	FLMOOC-procmin1
Data set description	
Generated or collected	Collected
Origin	https://www.futurelearn.com/admin/courses/process-mining/1/
Scale	<10Mb
Who is this useful for?	Learning analytics within EDSA
Similar existing datasets	One of several datasets of FutureLearn course behavior data, interesting for comparison between sessions and runs.
Standards and metadata	
Methodology for data collection/management	Data collection is managed by FutureLearn
Metadata, supporting material	https://partners.futurelearn.com/data/datasets/
Status and location of metadata	N/A
Data sharing	**Data sharing**
Licensing, data protection, ownership and copyright	Raw data is owned by TU/e and cannot be shared outside the EDSA project due to FutureLearn restrictions of use.
If the data cannot be published openly, why?	N/A

How will the data be shared?	User privacy. The data can be aggregated and published under an open license
Data repository	Local storage at TU/e, to store privacy sensitive data.
Dataset Link	N/A
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	N/A
Approx end volume	<10MB
Who is responsible for the data management and curation?	JSI lead data management and curation. OU contribute
Quality assurance including back up procedures	Relying on FutureLearn
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Student identifiable data
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	No

4.6.2 Work Package 4 – Dissemination and community building

WP4 has continued to 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 allows for monitoring of the projects community building and dissemination.

Table 22: Web server logs and Google analytics of project website access

Dataset Reference and Name	
Dataset Identifier	WebsiteAnalytics
Dataset description	
Generated or collected	Collected
Origin	http://edsa-project.eu
Scale	1 website

Who is this 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	Sessions, Page views, Demographics, User Flow, Bounce rate.
Status and location of metadata	There is no metadata publically available as the data is not openly published All sections that will be used are within https://analytics.google.com/
Data sharing	
Licensing, data protection, ownership and copyright	Raw data will be owned by the project and unlicensed. It will not be available for reuse.
If the data cannot be published openly, why?	User privacy. The data can be aggregated and published under an open licence. A judgement call will have to be made on whether this is worth it.
How will the data be shared?	Analysed data will be made available throughout deliverable reports in WP4.
Data repository	Internal institutional Soton/OU repositories
Dataset Link	There is no external link
Archiving and preservation	**Archiving and preservation**
How long should the data be preserved?	At least until the end of project
Approx end volume	< 1GB
Who is responsible for the data management and curation?	OU lead data management and curation. Soton contribute
Quality assurance including back up procedures	Backed up remotely
Associated costs for data management	Free storage. 0.5 day per month
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Google analytics and web log data contain information which could be used to identify

	individuals
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 23: Generated social media engagement data

Dataset Reference and Name	
Dataset Identifier	SocialMediaEngagements
Dataset description	
Generated or collected	Collected
Origin	Twitter
Scale	1 Twitter Account
Who is this 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	Tweets, Impressions, Profile Visits, Followers, Mentions
Status and location of metadata	https://analytics.twitter.com/user/edsa_project/home
Data sharing	
Licensing, data protection, ownership and copyright	Data will be licensed in compliance with each social network's terms and conditions
If the data cannot be published openly, why?	Data sharing needs to comply with individual site licenses. However the majority of social networks do not permitted collection, harvesting and republication of data
How will the data be shared?	Available via EDSA website. Deliverable reports in WP4.
Data repository	Internal institutional Soton repositories
Dataset Link	There is no external link as the terms and conditions have not yet been checked.

Archiving and preservation	
How long should the data be preserved?	Until the end of the project
Approx end volume	< 1GB
Who is responsible for the data management and curation?	Soton lead data management and curation.
Quality assurance including back up procedures	Backed up remotely
Associated costs for data management	Free storage. 1 day per month
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	Yes
What are the ethical or legal issues, if any, that can occur from sharing this data?	Subject to license conditions from the social network
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

4.7 Appendix 3 – Supplementary data

4.7.1 Work Package 2

WP2 used a number of datasets in the delivery of curricular. This includes unclean data for exercises on data management, schemas and cleaning as well as datasets which can analysed in tools such as R and Python.

Table 24: Dataset for course examples and exercises

Dataset Reference and Name	
Dataset Identifier	Using namespace notation to specify R packages: sml::poly4, sml::poly4b, sml::kmeans, sml::seeds, car::Duncan, car::Davis, datasets::car, datasets::HairEyeColor, datasets::Airquality, datasets::swiss, bestGLM::zprostate, MASS::menarche
Dataset description	
Generated or collected	Both
Origin	Third party R packages students download from CRAN. Some in an author developed package hosted on CRAN
Scale	12 small datasets. <1MB
Who is this useful for?	Students in the "Essentials of Data Analytics and Machine Learning" course.

Similar existing datasets	Datasets are archived in CRAN. Used in course examples and exercises.
Standards and metadata	
Methodology for data collection/management	None
Metadata, supporting material	The datasets will be used within learning activities offered as part of the "Essentials of Data Analytics and Machine Learning" course. They are stored in the sml R package.
Status and location of metadata	Package documentation (except, currently, for those in the sml package)
Data sharing	
Licensing, data protection, ownership and copyright	GNU GPL V3 http://www.gnu.org/licenses/gpl-3.0.en.html
If the data cannot be published openly, why?	The data is published openly
How will the data be shared?	Via R packages, searchable online.
Data repository	CRAN
Dataset Link	https://vincentarelbundock.github.io/Rdatasets/datasets.html
Archiving and preservation	
How long should the data be preserved?	As long as the owners do not remove them. If the datasets are no longer accessible, other similar datasets will be used in the module.
Approx end volume	< 1MB
Who is responsible for the data management and curation?	Persontyle lead data management and curation, third parties for collected data
Quality assurance including back up procedures	Relying on CRAN
Associated costs for data management	None
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 25: Monthly Rainfall (mm) Totals for Selected Stations in Tanzania, 2014

Dataset Reference and Name	
Dataset Identifier	Tanzania_Rainfall
Dataset description	
Generated or collected	Collected
Origin	http://training.theodi.org/InPractice/inpractice1/course/en/exercises/Tanzania_Rainfall.pdf
Scale	<66KB
Who is this useful for?	Anyone interested in understanding the exercises within the finding stories curriculum
Similar existing datasets	None.
Standards and metadata	
Methodology for data collection/management	None
Metadata, supporting material	The datasets will be used within learning activities offered as part of the "Finding stories in Data" course.
Status and location of metadata	Modules 4 - Gathering Data
Data sharing	
Licensing, data protection, ownership and copyright	This dataset is published on Github, under a CC-BY licence.
If the data cannot be published openly, why?	N/A
How will the data be shared?	Via Github and via the EDSA website (http://courses.edsa-project.eu/course/view.php?id=52)
Data repository	Github, EDSA website
Dataset Link	http://training.theodi.org/InPractice/inpractice1/course/en/exercises/Tanzania_Rainfall.pdf
Archiving and preservation	
How long should the data be preserved?	N/A
Approx end volume	<66KB
Who is responsible for the data management and curation?	David Tarrant, ODI
Quality assurance including back up procedures	Github
Associated costs for data management	N/A

Data Ethics	Data Ethics
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 26: BBC RSS Feed

Dataset Reference and Name	
Dataset Identifier	BBCnews
Dataset description	
Generated or collected	Collected
Origin	http://feeds.bbci.co.uk/news/rss.xml
Scale	1 Twitter account
Who is this useful for?	Anyone interested in understanding the exercises within the finding stories curriculum
Similar existing datasets	None
Standards and metadata	
Methodology for data collection/management	None
Metadata, supporting material	The datasets will be used within learning activities offered as part of the "Finding stories in Data" course.
Status and location of metadata	Modules 4 - Gathering Data
Data sharing	
Licensing, data protection, ownership and copyright	This data is publicly available from an external source
If the data cannot be published openly, why?	N/A
How will the data be shared?	Via Github and via the EDSA website (http://courses.edsa-project.eu/course/view.php?id=52)

Data repository	N/A
Dataset Link	http://feeds.bbci.co.uk/news/rss.xml
Archiving and preservation	
How long should the data be preserved?	N/A
Approx end volume	Unknown
Who is responsible for the data management and curation?	David Tarrant, ODI
Quality assurance including back up procedures	N/A
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 27: Health Facility list ratings Tanzania

Dataset Reference and Name	
Dataset Identifier	healthfacilitiy
Dataset description	
Generated or collected	Collected
Origin	https://drive.google.com/file/d/0B1VBoooQ3X5jeEQycHo4OG4tclE/view
Scale	<22KB
Who is this useful for?	Anyone interested in understanding the exercises within the finding stories curriculum
Similar existing datasets	None
Standards and metadata	
Methodology for data collection/management	None
	The datasets will be used within learning activities

Metadata, supporting material	offered as part of the "Finding stories in Data" course.
Status and location of metadata	Modules 4 - Gathering Data
Data sharing	
Licensing, data protection, ownership and copyright	This dataset is published on Github, under a CC-BY licence.
If the data cannot be published openly, why?	N/A
How will the data be shared?	Via Github and via the EDSA website (http://courses.edsa-project.eu/course/view.php?id=52)
Data repository	Github, EDSA website
Dataset Link	https://drive.google.com/file/d/0B1VBoooQ3X5jeEQycHo4OG4tclE/view
Archiving and preservation	
How long should the data be preserved?	N/A
Approx end volume	<22KB
Who is responsible for the data management and curation?	David Tarrant, ODI
Quality assurance including back up procedures	Github
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 28: Louisiana Secretary of State Officials

Dataset Reference and Name	
Dataset Identifier	Dataset1
Dataset description	

Generated or collected	Collected
Origin	http://www.sos.la.gov/tabid/136/Default
Scale	>2.5MB
Who is this useful for?	Anyone interested in understanding the exercises within the finding stories curriculum
Similar existing datasets	None
Standards and metadata	
Methodology for data collection/management	None
Metadata, supporting material	The datasets will be used within learning activities offered as part of the "Finding stories in Data" course.
Status and location of metadata	Module 6 - Cleaning Data
Data sharing	
Licensing, data protection, ownership and copyright	This dataset is published on Github, under a CC-BY licence.
If the data cannot be published openly, why?	N/A
How will the data be shared?	Via Github and via the EDSA website (http://courses.edsa-project.eu/course/view.php?id=52)
Data repository	Github, EDSA website
Dataset Link	http://training.theodi.org/resources/dataset1.xls
Archiving and preservation	
How long should the data be preserved?	N/A
Approx end volume	>2.5MB
Who is responsible for the data management and curation?	David Tarrant, ODI
Quality assurance including back up procedures	Github
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data?	No

Y/N	
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 29: Projects Dataset

Dataset Reference and Name	
Dataset Identifier	Dataset2
Dataset description	
Generated or collected	Collected
Origin	http://www.itdashboard.gov/data_feeds
Scale	<2MB
Who is this useful for?	Anyone interested in understanding the exercises within the finding stories curriculum
Similar existing datasets	None
Standards and metadata	
Methodology for data collection/management	None
Metadata, supporting material	The datasets will be used within learning activities offered as part of the "Finding stories in Data" course.
Status and location of metadata	Module 6 - Cleaning Data
Data sharing	
Licensing, data protection, ownership and copyright	This dataset is published on Github, under a CC-BY licence.
If the data cannot be published openly, why?	N/A
How will the data be shared?	Via Github and via the EDSA website (http://courses.edsa-project.eu/course/view.php?id=52)
Data repository	Github, EDSA website
Dataset Link	http://training.theodi.org/resources/dataset2.csv

Archiving and preservation	
How long should the data be preserved?	N/A
Approx end volume	<2MB
Who is responsible for the data management and curation?	David Tarrant, ODI
Quality assurance including back up procedures	Github
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A

Table 30: UK GP Earnings

Dataset Reference and Name	
Dataset Identifier	Dataset3
Dataset description	
Generated or collected	Collected
Origin	http://data.gov.uk/dataset/gp-earnings-and-expenses-2009-10
Scale	<2MB
Who is this useful for?	Anyone interested in understanding the exercises within the finding stories curriculum
Similar existing datasets	None
Standards and metadata	
Methodology for data collection/management	None
	The datasets will be used within learning activities offered

Metadata, supporting material	as part of the "Finding stories in Data" course.
Status and location of metadata	Module 6 - Cleaning Data
Data sharing	
Licensing, data protection, ownership and copyright	This dataset is published on Github, under a CC-BY licence.
If the data cannot be published openly, why?	N/A
How will the data be shared?	Via Github and via the EDSA website (http://courses.edsa-project.eu/course/view.php?id=52)
Data repository	Github, EDSA website
Dataset Link	http://training.theodi.org/resources/dataset2.csv
Archiving and preservation	
How long should the data be preserved?	N/A
Approx end volume	<2MB
Who is responsible for the data management and curation?	David Tarrant, ODI
Quality assurance including back up procedures	Github
Associated costs for data management	N/A
Data Ethics	**Data Ethics**
Are there any ethical or legal issues that can have an impact on sharing this data? Y/N	No
What are the ethical or legal issues, if any, that can occur from sharing this data?	None
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data? Y/N/NA	N/A