



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

## **D5.6 Updated EDSA Data Management Plan**

Deliverable Editor: **Emily Vacher (ODI)**  
Other contributors:  
  
Deliverable Reviewers: **R. Brochenin (Tu/e) / Shatha Jaradat (KTH)**  
Deliverable due date: **31/07/2016**  
Submission date: **29/07/2016**  
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

<b>Version</b>	<b>Date</b>	<b>Amended by</b>	<b>Changes</b>
0.1	27/05/2016	Emily Vacher	Created document, added initial plan outline
0.2	31/05/2016	Emily Vacher	Added datasets and WP descriptions
0.3	10/06/2016	Emily Vacher	Incorporated amendments
0.4	23/06/2016	Emily Vacher	Updated with new published data
0.5	28/06/2016	Emily Vacher	Incorporated amendments
0.6	27/07/2016	Elena Simperl	Scientific Review
1.0	29/07/2016	Aneta Tumilowicz	Final QA

## Table of Contents

Change Log.....	2
Table of Contents.....	3
List of Tables .....	4
List of Figures .....	5
1 Executive Summary .....	6
1.1 Lessons learnt.....	6
1.2 Updates from Initial DMP.....	7
2 Policy .....	10
2.1 Data standards and metadata policy for EDSA.....	10
2.2 Data sharing policy for EDSA.....	10
2.3 Supporting people who want to use EDSA data .....	11
2.4 Data storage and management policy for EDSA .....	12
2.5 Data preservation and archiving policy for EDSA.....	13
3 Challenges and decisions.....	14
3.1 Informed consent .....	14
3.2 Anonymisation of personal data .....	14
3.3 Third party licences.....	15
4 Data Management Plan .....	16
4.1 Summary.....	16
4.2 The EDSA Register .....	16
4.2.1 Introduction.....	16
4.3 Work package 1 - Demand analysis and advisory board .....	17
4.3.1 Corpora of crawled web-based adverts from LinkedIn.....	17
4.3.2 Aggregated statistics of European skill demand based on web-based job adverts .....	18
4.3.3 Individual results from demand analysis .....	20
4.3.4 Summary data from surveys and interviews .....	21
4.3.5 De-identified survey responses from demand analysis.....	22
4.3.6 Recordings and transcriptions of interviews.....	23
4.3.7 ideXlab search platform results.....	24
4.4 Work package 2 – Curricula and course development .....	26
4.4.1 Related course data regarding similar modules and training available across the EU ...	26
4.4.2 Dataset for course examples and exercises .....	27
4.4.3 Event log from a municipality process .....	29
4.5 Work package 3 – Training delivery and learning analytics feedback.....	30
4.5.1 Repository statistics on downloads and views of educational resources.....	30

4.5.2	Learning Analytics data generated from the EDSA Online Courses portal.....	32
4.5.3	Internal log of eLearning systems.....	33
4.5.4	Statistics of course registration, participation and completion .....	34
4.5.5	Aggregated statistics of engagement with the developed courses and educational resources.....	36
4.5.6	Recorded behavior of students following the first session of the process mining MOOC37	
4.6	Work package 4 – Dissemination and community building.....	38
4.6.1	Web server logs and Google analytics of project website access .....	38
4.6.2	Generated social media engagement data.....	40
4.7	Work package 5 – Exploitation .....	41
4.7.1	List of project exploitation results - collaborations, institutional and geographical beneficiaries.....	41
4.7.2	The EDSA Register.....	42

## List of Tables

Table 1:	Entries in the Data Management Plan-----	8
Table 2:	Four Levels of Certificates-----	11
Table 3:	Corpora of crawled Web-based adverts from LinkedIn-----	17
Table 4:	Aggregated Statistics of European skill demand on web-based job adverts-----	18
Table 5:	Individual results from demand analysis-----	20
Table 6:	Summary data from surveys and interviews-----	21
Table 7:	De-identified survey responses from demand analysis-----	22
Table 8:	Recordings and transcriptions of interviews-----	23
Table 9:	IdeXlab search platform results-----	24
Table 10:	Related course data regarding similar modules and training available across the EU-----	26
Table 11:	Dataset for course examples and exercises-----	27
Table 12:	Event log from a municipality process-----	29
Table 13:	Repository statistics on downloads and views of educational resources-----	30
Table 14:	Learning analytics data generated from EDSA online courses portal-----	32
Table 15:	Internal log of elearning systems-----	33
Table 16:	Statistics of course registration, participation and completion-----	34
Table 17:	Aggregated statistics of engagement with the developed courses and educational resources-----	36
Table 18:	Recorded behaviour of students following the first session of the process mining MOOC	37
Table 19:	Web server logs and Google analytics of project website access-----	38
Table 20:	Generated social media engagement data-----	40



Table 21: List of project exploitation results - collaborations, institutional and geographical beneficiaries-----	41
Table 22: The EDSA Register -----	42

## List of Figures

Figure 1 Entries in the Data Management Plan -----	8
Figure 2 The Data Spectrum -----	10

# 1 Executive Summary

The European Data Science Academy (EDSA) is participating in the pilot action on open access research data, as defined in the guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020<sup>1</sup>.

EDSA data includes data that is used, generated and collected by the project. The data management plan (DMP) is key to tracking these datasets, and identifying which of them have been or can be published under an open licence. It is not always appropriate to publish data as open data; the data management plan allows us to clearly see what data is not published openly and the reason for that.

This is the second iteration of the DMP. The first was included in D5.5 at Month 6 of the project. The original DMP was an outline of the data we anticipated, whereas this DMP includes data that we have started collecting, generating and using in the project. The final version of the DMP is due at Month 36 of the project.

The EDSA DMP includes the following information for each dataset:

- Dataset reference and name
- Dataset description
- Standards and metadata
- Data sharing
- Archiving and preservation

Specifically, our goals are to:

- Manage and maintain data, where applicable, to ensure quality and to make the data usable.
- Ensure that all data produced by the project is subject to appropriate levels of security and privacy.
- Publish data produced by the project under an open licence, where possible.

At the halfway stage of the project, this DMP addresses some of the key challenges and lessons that the Consortium has learnt. These are outlined below. The plan also outlines how individual datasets will be maintained during and after the project. We will continue to update datasets where appropriate, collect and generate new datasets and publish as much of our data as open as possible, using the lessons that we have learnt.

## 1.1 Lessons learnt

The management of EDSA data has provided us with some useful lessons for further iterations of the DMP. These lessons can be divided into two categories: licensing and managing risk.

---

<sup>1</sup> Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 (2016) [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-pilot-guide\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf) [accessed 29/06/2016]



### Licensing

- Check third party licences at regular intervals to ensure that we are adhering to their terms as licence changes are generally not communicated by data providers.
- Encourage the use of Creative Commons licences<sup>2</sup>, specifically CC-BY, within the EDSA project to make it as easy as possible to reuse our data.

### Managing risk

- Anticipate changes in data use where possible.
- Get informed consent for use of personal data

We will add to these lessons as the project continues and we face more challenges with our data use.

## 1.2 Updates from Initial DMP

The EDSA DMP is not a static log of the project datasets but an evolving resource, that reflects the changing nature of the data the project collects or generates. This DMP reflects the progress the project has made in the past 12 months of the project - incorporating further datasets and adhering to our guiding policies on open publication. D5.5 provided an initial snapshot of the data we managed in the early stages of the project and the data that we anticipated would be collected or generated over the coming months. This DMP reflects the current status of the project's datasets in more extensive detail.

Figure 1 shows the datasets which have been collected or generated by the Consortium between months 6 and 18 of the project.

There are two new entries to the DMP:

- De-identified survey responses from the demand analysis research
- Learning Analytics data generated from the EDSA Online Courses portal

The de-identified survey responses have replaced the aggregated results from the online survey dataset as the way the data is presented has been changed. For a more detailed discussion on this topic see the section on 'Challenges and decisions', or D1.4.

Learning Analytics from the EDSA Online Courses portal is a new entry to this DMP. It has been published openly via Github<sup>3</sup>, for others to benefit from, as there are no restrictions with the third party.

The entries below have been removed from this DMP as they have either been replaced by specific dataset entries, or are no longer expected to be collected or generated as the project has progressed.

The removed entries are:

---

<sup>2</sup> Please note that there are multiple Creative Commons licences, which are outlined on their website: <https://creativecommons.org/licenses/>

<sup>3</sup> <https://alexmikro.github.io/learning-analytics-dataset-from-the-edsa-online-courses-portal/>

- Aggregated results from the online survey (as above)
- Aggregated statistics of networking and engagement data (datasets now explicitly stated)
- Linked open data sources, such as the DBLP Computer Science Bibliography<sup>4</sup> and GeoNames Ontology<sup>5</sup> (datasets now explicitly stated)
- Publically available governmental, financial, network and environmental datasets for each course (datasets now explicitly stated)

**Table 1: Entries in the Data Management Plan**

Work Package	Lead	Dataset	Project Phase	Status	New entry to DMP D5.6
WP1	ODI	Corpora of crawled web-based adverts from LinkedIn	M6-M18	Finished	No
WP1	ODI	Aggregated statistics of European skill demand based on web-based job adverts	M6-M18	Ongoing	No
WP1	ODI	Individual results from demand analysis	M2-M18	Finished	No
WP1	ODI	DemandAnalysisSummary	M18	Finished	Yes
WP1	ODI	De-identified data from demand analysis	M2-M18	Finished	Yes
WP1	ODI	Recordings and transcriptions of interviews	M2-M18	Finished	No
WP1	ODI	ideXlab search platform results	M6-M36	Ongoing	No
WP2	ODI	Related course data regarding similar modules and training offerings across the EU	M18	Finished	No
WP2	Persontyle	Datasets for course examples and exercises	M6-M36	Ongoing	No
WP2	TU/e	Event log from a municipality process	M12-M36	Ongoing	No
WP3	OU	Learning Analytics data generated from the EDSA Online Courses portal	M12-M36	Ongoing	Yes
WP3	JSI	Repository statistics on downloads and views of educational resources	M12-M36	Ongoing	No
WP3	JSI	Internal logs of elearning systems	M12-M36	Ongoing	No
WP3	JSI	Statistics of course registration, participation and completion	M12-M36	Ongoing	No

<sup>4</sup> <http://dblp.uni-trier.de/>

<sup>5</sup> <http://www.geonames.org/ontology/documentation.html>





<b>WP3</b>	<b>JSI</b>	Aggregated statistics of engagement with the developed courses and educational resources	M12-M36	Ongoing	No
<b>WP3</b>	<b>TU/e</b>	Recorded behavior of students following the first session of the process mining MOOC	M12	Finished	No
<b>WP4</b>	<b>SOTON</b>	Web server logs and Google analytics of project website access	M12-M36	Ongoing	No
<b>WP4</b>	<b>SOTON</b>	Generated social media engagement data	M12-M36	Ongoing	No
<b>WP5</b>	<b>ideXlab</b>	List of project exploitation results – collaborations, institutional and geographical beneficiaries,	M18-M36	Ongoing	No
<b>WP5</b>	<b>ODI</b>	EDSA register	M6-M36	Ongoing	Yes

## 2 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<sup>6</sup>.

### 2.1 Data standards and metadata policy for EDSA

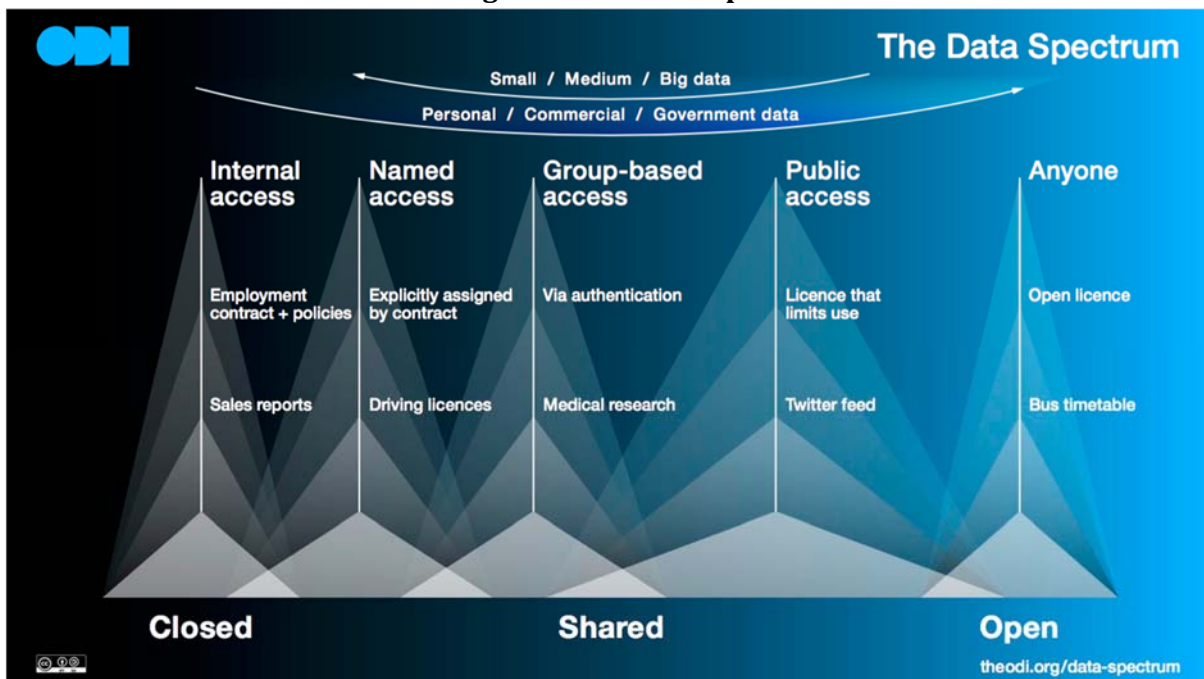
Standardising the project's collection and production of data ensures reusability and interoperability within the project, and externally if openly available. Where possible, data is made available in CSV, JSON or linked data in RDF format

### 2.2 Data sharing policy for EDSA

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 Attribution licence (CC BY 4.0)

The Open Data Institute has produced a data spectrum asset to explain frequently used, but frequently misinterpreted terms, such as open data, closed data, personal data, and big data. The most useful categorisation of data is through the licence and access rights. Data exists on a spectrum<sup>7</sup>, which ranges from closed to shared, to open.

Figure 1: The Data Spectrum



CC-BY The [Open Data Institute](https://theodi.org/)

<sup>6</sup> [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-pilot-guide\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf)

<sup>7</sup> <https://theodi.org/data-spectrum>



The survey and interview data from the demand analysis provides us with a good example of data across the spectrum. The list of names and email addresses of participants is closed. This is currently only held by one member of the Consortium and will only be used if required for audit purposes. The individual recordings and transcripts are an example of shared data. These are only available to members of the consortium who have been given named access. The de-identified survey results from the demand analysis are open. This data is published on Github under a CC-BY licence<sup>8</sup>.





## 2.3 Supporting people who want to use EDSA data

We use the ODI's Open Data Certificate standard to benchmark each open dataset. 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.

Where we have published data openly, we have used the Open Data Institute's certification process to demonstrate to potential reusers that it is quality open data.

There are four levels of Certificates<sup>9</sup>:

**Table 2: Four Levels of Certificates**

Bronze 	The data is openly licensed, available with no restrictions, accessible and legally reusable.
Silver 	The data satisfies the Bronze requirements, the data is documented in a machine readable format, is reliable and offers ongoing support from the publisher via a dedicated communication channel.
Gold 	The data satisfies the Silver requirements, is published in an open standard machine readable format, has guaranteed regular updates, offers greater support, documentation, and includes a machine readable rights statement.
Platinum 	The data satisfies the Gold requirements, has machine readable provenance documentation, uses unique identifiers in the data, the publisher has a communications team offering support. This is an exceptional example of an information infrastructure.

CC-BY [The Open Data Institute](https://theodi.org/)

<sup>8</sup> <http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis/>

<sup>9</sup> <https://certificates.theodi.org/en/>

Currently, not all of the data that has been published has been certified, although this is our aim and in progress.

EDSA Datasets that currently have a certificate:

- The EDSA Register has a bronze certificate because it is:
  - Openly licensed and legally reusable (= 'open')
  - Accessible on the web
- De-identified survey responses from the demand analysis has a silver certificate because it is:
  - Openly licensed and legally reusable (= 'open')
  - Accessible on the web
  - Published in a machine readable format
  - Offers ongoing support from the publisher via Github

It is important to note that there is no issue in being at Bronze level, as the data is still published openly to a level that meets user needs. Higher is not always appropriate for the EDSA data as there is not always a mechanism in place for ongoing discussion of the data (Silver), and it may not be updated regularly, especially beyond the length of the project (Gold). Over 99% of all ODI certificates are to Bronze standard<sup>10</sup>.

## 2.4 Data storage and management policy for EDSA

There are currently three main types of repository for EDSA data: open access repositories (for example Github), the EDSA project website and internal institutional and organisational repositories for securely holding data.

The criteria for determining where data is stored is as follows:

### **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.

### **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.

### **Internal institutional and organisational repositories:**

Some datasets in the data management plan are hosted in repositories of the organisation responsible for that data. While some of these are internal, hosted in Consortium partners' internal repositories,

---

<sup>10</sup> <https://certificates.theodi.org/status>



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.

## **2.5 Data preservation and archiving policy for EDSA**

Striving for preservation of 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 work package 5. Although the aim of the project is to preserve as much of the data as possible, data published in external open repositories is reliant on that system. As a project, EDSA are yet to determine a policy regarding archiving of datasets. This will be decided prior to the final Data Management Plan (M36).

## 3 Challenges and decisions

Creating and maintaining a DMP for the project has ensured that we have been able to highlight potential data management and usage challenges and make informed decisions within the Consortium on how they should be addressed. In this section, we highlight some of the main topics of consideration when managing project datasets. These provide us with useful lessons for further iterations of the DMP.

### 3.1 Informed consent

It is a legal requirement to inform people how their personal data is going to be used and to retrieve their informed consent. Whilst there are exceptions to this requirement, such as national security or for services in the public interest, these exceptions do not apply to this project. This is area that we have addressed in the project over the last 12 months.

The intended use of the demand analysis survey data changed over the course of the 18-month data collection, due to the length of the study. At the start of the project, we planned to release summary statistics of the quantitative survey data through our skills dashboard. Accordingly, participants were informed that data would be made accessible in an anonymous, aggregated form.

Following discussions during the evaluation of the pilot study, we established that releasing de-identified survey responses would add value to the project's outputs. This data would provide access to responses on an individual basis, thus adding much greater detail and utility to potential reusers of the data. Consequently, the data would no longer be aggregated. In month 9 of the project, we therefore changed the wording of the informed consent section of the survey, stating that data could later be made publicly available in de-identified formats, using an open licence.

Due to this change in use, we have not used the data collected before the permissions change on the dashboard, or in the open dataset. The data collected from early study participants has been included in the aggregated analysis in D1.4, but are not available in the open data. We have also not published data from people who had withheld consent for this use of their data.

### 3.2 Anonymisation of personal data

It was important to ensure that the Learning Analytics data was anonymised before it could be published openly and that no individual user could be identified. The Consortium came to an agreed policy which will be applied where appropriate to further datasets. We will publish data openly if the data has been de-identified, and individual users cannot be recognised. De-identified data will also be published alongside a Privacy Impact Assessment which identifies potential risks and how they have been managed.



### 3.3 Third party licences

There are many types of open licences. Several times, the Consortium faced challenges when scouring the terms and conditions of third party sites, such as LinkedIn<sup>11</sup>, Adzuna API<sup>12</sup>, Learning Locker<sup>13</sup>, to ensure the terms of use are adhered to. For the EDSA data, the Consortium is encouraged to use Creative Commons licences to ensure that people wishing to use our data can clearly find how they can use it. Alternative open licences used in the project are the open source GNU General Public Licence V3<sup>14</sup> and the 3TU.Datacentrum General Terms of Use<sup>15</sup>.

When data is collected from a third party website, it is vital to track the terms of use, as these can change. At the beginning of the project, the Consortium collected and published data from a third party website, LinkedIn. At the time, the terms allowed such use, but during the project the licence provided by LinkedIn changed. We did not keep a record of the original licence. There was debate about whether the project could keep the data openly available, as at the time of collection this was permitted. To avoid risk we decided to remove that data, especially as the links to the licence that we had used specifically stated that we could not use it in the way we had planned.

Another notable challenge was the use of data from Trovit<sup>16</sup>, a website that aggregates job advertisements from across Europe. This data populates the jobs dashboard. The terms of the licence did not allow us to use the data, however after careful consideration we decided as a Consortium that the UK text and data mining (TDM) exception for research purposes<sup>17</sup> allowed the use of the data as long as the data itself was not accessible by others. The UK law follows guidance from the EU Database Directive (96/9/EC), and discussions on an EU-wide TDM exception are likely to take place in 2016.

Restrictions on data use frequently prevent individuals from maximising the value of that data. If the data was open, and anyone could access, use and share it for any purpose, Trovit would ultimately benefit from increased coverage and traffic, via attribution. If a company does not want anyone to benefit financially from their work, a non-commercial licence such as CC-BY-NC 4.0<sup>18</sup> would still enable others to use the data and link back to Trovit.

---

<sup>11</sup> <https://developer.linkedin.com/legal/api-terms-of-use>

<sup>12</sup> <https://developer.adzuna.com/>

<sup>13</sup> <http://learninglocker.net/>

<sup>14</sup> <http://www.gnu.org/licenses/gpl-3.0.en.html>

<sup>15</sup> [http://researchdata.4tu.nl/fileadmin/editor\\_upload/pdf/General terms of use 3TU.Datacentrum.pdf](http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/General_terms_of_use_3TU.Datacentrum.pdf)

<sup>16</sup> <http://jobs.trovit.co.uk/>

<sup>17</sup> <http://www.legislation.gov.uk/ukdsi/2014/9780111112755> p6 (accessed 30/06/2016)

<sup>18</sup> <https://creativecommons.org/licenses/by-nc/4.0/>

## 4 Data Management Plan

### 4.1 Summary

When creating the EDSA DMP, we took guidance from best practice, online tools such as DMPTool<sup>19</sup> and DMPonline<sup>20</sup>. DMPonline allows you to select the specific project category, in this case Horizon 2020 pilot action on open access research data, and therefore ensure that we captured all the necessary metadata.

The datasets are organised by work package. Each table represents one dataset generated or collected by the EDSA project. Each table includes the following information:

- Dataset reference and name
- Dataset description
- Standards and metadata
- Data sharing
- Archiving and preservation

We created a dataset of all the datasets generated or collected by the EDSA project. Details of this can be found below.

### 4.2 The EDSA Register

#### 4.2.1 Introduction

The EDSA Register is published under a CC-BY 4.0 Creative Commons licence<sup>21</sup>. It is published on GitHub<sup>22</sup> and can also be accessed via the EDSA website at <http://edsa-project.eu/resources/datasets/>. The dataset has been certified as Bronze using the Open Data Certificates<sup>23</sup>.

This dataset is updated every three months by the ODI with information from the Work Package leads. The next update is due at month 21 of the project.

With the Consortium work package leads, we explored the datasets for each work package, enabling discovery of what data could be published openly. We look to share best practices and to ensure a high quality of open data. Best practices include:

- Publishing in a machine readable format, e.g. CSV
- Providing supporting documentation or metadata
- Using a clear open licence, preferably Creative Commons Attribution 4.0 licence<sup>24</sup> for consistency.

---

<sup>19</sup> <https://dmp.cdlib.org/>

<sup>20</sup> <https://dmponline.dcc.ac.uk/>

<sup>21</sup> <https://creativecommons.org/licenses/>

<sup>22</sup> <https://theodi.github.io/european-data-science-academy-register/>

<sup>23</sup> <https://certificates.theodi.org>

<sup>24</sup> <https://creativecommons.org/licenses/>





We used a Google Sheet as our data management tool, which we update every three months. It is this Google Sheet which is embedded on the EDSA website<sup>25</sup>. Any edits made on the Google Sheet show on the website in real time.

### 4.3 Work package 1 - Demand analysis and advisory board

WP1 has collected and generated data from the demand analysis study. This includes recordings and transcriptions of the interviews, survey responses and anonymised results of the surveys and interviews

#### 4.3.1 Corpora of crawled web-based adverts from LinkedIn

**Table 3: Corpora of crawled Web-based adverts from LinkedIn**

<b>Dataset Reference and Name</b>	
Dataset Identifier	WebSiteHarvest
<b>Dataset description</b>	
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 is 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.
<b>Standards and metadata</b>	
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

<sup>25</sup> <http://edsa-project.eu/datasets>

<b>Data sharing</b>	
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 <b>now</b> forbid harvesting and collection of data without express permission. When the data was collected, this was not the case.  <a href="https://www.linkedin.com/legal/user-agreement?trk=hb_ft_userag">https://www.linkedin.com/legal/user-agreement?trk=hb_ft_userag</a>
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
<b>Archiving and preservation</b>	
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

### 4.3.2 Aggregated statistics of European skill demand based on web-based job adverts

Table 4: Aggregated Statistics of European skill demand on web-based job adverts

<b>Dataset Reference and Name</b>	
Dataset Identifier	WebSiteStatistics
<b>Dataset description</b>	
Generated or collected	Collected



Origin	Adzuna API <sup>26</sup> Trovit <sup>27</sup>
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 is 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.
<b>Standards and metadata</b>	
Methodology for data collection/management	All data collected is translated into CSV format.
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
<b>Data sharing</b>	
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
<b>Archiving and preservation</b>	
How long should the data be preserved?	Until the end of the project
Approx end volume	<1Gb

<sup>26</sup> <https://developer.adzuna.com/>

<sup>27</sup> <https://www.trovit.com/>

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

### 4.3.3 Individual results from demand analysis

**Table 5: Individual results from demand analysis**

<b>Dataset reference and name</b>	
Dataset identifier	IndividualResponses
<b>Dataset description</b>	
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.
<b>Standards and metadata</b>	
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. De-identified data will be publicly available, where possible.
Status and location of metadata	Metadata is not publically available
<b>Data Sharing</b>	
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
<b>Archiving and preservation</b>	
How long should the data be preserved?	Until the end of the project
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	Approximately 1 day effort per month

#### 4.3.4 Summary data from surveys and interviews

**Table 6: Summary data from surveys and interviews**

<b>Dataset reference and name</b>	
Dataset identifier	DemandAnalysisSummary
<b>Dataset description</b>	
Generated or collected	Generated
Origin	Guided surveys and online responses
Scale	584 surveys 108 interviews
Who is this data useful for?	External analysis of respondents who took the surveys and interviews.
Similar existing datasets	None
<b>Standards and metadata</b>	
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	<a href="https://theodi.github.io/edsa-demand-analysis-summary-data/">https://theodi.github.io/edsa-demand-analysis-summary-data/</a>
<b>Data Sharing</b>	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0) <a href="https://creativecommons.org/licenses/by/4.0/">https://creativecommons.org/licenses/by/4.0/</a>

If the data cannot be published openly, why?	The data is published openly
How will the data be shared?	Data will be available to access from the EDSA website and the ODI's Github repository.
Data repository	Github/ EDSA website
Dataset Link	<a href="https://theodi.github.io/edsa-demand-analysis-summary-data/">https://theodi.github.io/edsa-demand-analysis-summary-data/</a>
<b>Archiving and preservation</b>	
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

### 4.3.5 De-identified survey responses from demand analysis

**Table 7: De-identified survey responses from demand analysis**

<b>Dataset reference and name</b>	
Dataset identifier	DeidentifiedResponses
<b>Dataset description</b>	
Generated or collected	Generated
Origin	Online Survey <a href="http://edsa-project.eu/resources/survey/">http://edsa-project.eu/resources/survey/</a>
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.
<b>Standards and metadata</b>	
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	<a href="http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis/">http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis-/</a>
<b>Data Sharing</b>	
Licensing, ownership and copyright	Creative Commons Attribution (CC BY 4.0) <a href="https://creativecommons.org/licenses/by/4.0/">https://creativecommons.org/licenses/by/4.0/</a>
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	<a href="http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis/">http://davetaz.github.io/quantitative-data-from-edsa-demand-analysis-/</a>
<b>Archiving and preservation</b>	
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

### 4.3.6 Recordings and transcriptions of interviews

**Table 8: Recordings and transcriptions of interviews**

<b>Dataset reference and name</b>	
Dataset identifier	InterviewTranscripts
<b>Dataset description</b>	
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.
<b>Standards and metadata</b>	
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
<b>Data Sharing</b>	
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
<b>Archiving and preservation</b>	
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

### 4.3.7 ideXlab search platform results

**Table 9: IdeXlab search platform results**

<b>Dataset Reference and Name</b>	
Dataset Identifier	ExpertIdentification
<b>Dataset description</b>	
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.
<b>Standards and metadata</b>	
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.
<b>Data sharing</b>	
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?	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
<b>Archiving and preservation</b>	
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

## 4.4 Work package 2 – Curricula and course development

WP2 has collected data from openly available sources and created subsets of this data to be used in the learning resources produced. Data has also been collected about existing data science courses as per earlier recommendations.

### 4.4.1 Related course data regarding similar modules and training available across the EU

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

<b>Dataset Reference and Name</b>	
Dataset Identifier	DataScienceCourses
<b>Dataset description</b>	
Generated or collected	Collected
Origin	Course websites
Scale	Not yet known
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.
<b>Standards and metadata</b>	
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	<a href="https://theodi.github.io/data-science-courses-in-europe-2016/">https://theodi.github.io/data-science-courses-in-europe-2016/</a>
<b>Data sharing</b>	
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	GitHub. Also available via the EDSA website
Dataset Link	<a href="https://theodi.github.io/data-science-courses-in-europe-2016/">https://theodi.github.io/data-science-courses-in-europe-2016/</a>



<b>Archiving and preservation</b>	
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.

#### 4.4.2 Dataset for course examples and exercises

**Table 11: Dataset for course examples and exercises**

<b>Dataset Reference and Name</b>	
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
<b>Dataset description</b>	
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.
<b>Standards and metadata</b>	
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)
<b>Data sharing</b>	
Licensing, data protection, ownership and copyright	GNU GPL V3 <a href="http://www.gnu.org/licenses/gpl-3.0.en.html">http://www.gnu.org/licenses/gpl-3.0.en.html</a>
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	<a href="https://vincentarelbundock.github.io/Rdatasets/datasets.html">https://vincentarelbundock.github.io/Rdatasets/datasets.html</a>
<b>Archiving and preservation</b>	
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



### 4.4.3 Event log from a municipality process

Table 12: Event log from a municipality process

<b>Dataset Reference and Name</b>	
Dataset Identifier	<a href="https://data.3tu.nl/repository/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6">a07386a5-7be3-4367-9535-70bc9e77dbe6</a>
<b>Dataset description</b>	
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 <a href="http://data.3tu.nl/repository/collection:event%20logs%20real">http://data.3tu.nl/repository/collection:event logs real</a>
<b>Standards and metadata</b>	
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	<a href="http://data.3tu.nl/repository/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6">http://data.3tu.nl/repository/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6</a>
<b>Data sharing</b>	
Licensing, data protection, ownership and copyright	Own licence (Attribution, non-commercial) <a href="http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/General%20terms%20of%20use%203TU.Datacentrum.pdf">http://researchdata.4tu.nl/fileadmin/editor_upload/pdf/General terms of use 3TU.Datacentrum.pdf</a>
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	<a href="http://data.3tu.nl/repository/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6">http://data.3tu.nl/repository/uuid:a07386a5-7be3-4367-9535-70bc9e77dbe6</a>
<b>Archiving and preservation</b>	
How long should the data be preserved?	past project end

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

## 4.5 Work package 3 – Training delivery and learning analytics feedback

WP3 has started collecting data on the training delivered in the project – face-to-face and online - and will continue to collect data as more training is created and delivered.

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<sup>28</sup>, Tin Can API (xAPI)<sup>29</sup>, Learning Locker<sup>30</sup>.

### 4.5.1 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	RepositoryStatistics
Data set description	
Generated or collected	Collected
Origin	<a href="http://videolectures.net">videolectures.net</a>
Scale	Views and comments for each video lecture
Who is this useful for?	Internal analysis, curriculum development, external demand analysis

<sup>28</sup> <https://www.coursera.org>

<sup>29</sup> <http://tincanapi.com/>

<sup>30</sup> <http://learninglocker.net/>



Similar existing datasets	None. Provides evidence of resource usage and basis for improving curriculum, content and course structure.
<b>Standards and metadata</b>	
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	<a href="https://github.com/innanoyal/edsa-videolectures-statistics-dataset-1/tree/gh-pages/data">https://github.com/innanoyal/edsa-videolectures-statistics-dataset-1/tree/gh-pages/data</a>
<b>Data sharing</b>	
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	<a href="https://github.com/innanoyal/edsa-videolectures-statistics-dataset-1/tree/gh-pages/data">https://github.com/innanoyal/edsa-videolectures-statistics-dataset-1/tree/gh-pages/data</a>
<b>Archiving and preservation</b>	
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

## 4.5.2 Learning Analytics data generated from the EDSA Online Courses portal

Table 14: Learning analytics data generated from EDSA online courses portal

Dataset Reference and Name	
Dataset Identifier	EDSAOnlineCoursesLA
Data set description	
Generated or collected	Generated
Origin	<a href="http://courses.edsa-project.eu">http://courses.edsa-project.eu</a>
Scale	Not yet known
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: <a href="https://analyse.kmi.open.ac.uk/open_dataset">https://analyse.kmi.open.ac.uk/open_dataset</a>
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): <a href="https://tincanapi.com/overview/">https://tincanapi.com/overview/</a> . Introduction to Learning Locker: <a href="https://learninglocker.net">https://learninglocker.net</a>
Status and location of metadata	<a href="https://tincanapi.com/overview/">https://tincanapi.com/overview/</a> <a href="https://learninglocker.net">https://learninglocker.net</a> <a href="https://alexmikro.github.io/learning-analytics-dataset-from-the-edsa-online-courses-portal/">https://alexmikro.github.io/learning-analytics-dataset-from-the-edsa-online-courses-portal/</a>
Data sharing	
Licensing, data protection, ownership and copyright	Creative Commons Attribution (CC BY 4.0) <a href="https://creativecommons.org/licenses/by/4.0/">https://creativecommons.org/licenses/by/4.0/</a>
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 <a href="https://learninglocker.net/benefits/">https://learninglocker.net/benefits/</a>





Dataset Link	<a href="https://alexmikro.github.io/learning-analytics-dataset-from-the-edsa-online-courses-portal/">https://alexmikro.github.io/learning-analytics-dataset-from-the-edsa-online-courses-portal/</a>
<b>Archiving and preservation</b>	
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.

### 4.5.3 Internal log of eLearning systems

**Table 15: Internal log of elearning systems**

<b>Data set description</b>	
Generated or collected	Collected
Origin	<a href="http://videolectures.net">videolectures.net</a>
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.
<b>Standards and metadata</b>	
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

<b>Data sharing</b>	
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
<b>Archiving and preservation</b>	
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

#### 4.5.4 Statistics of course registration, participation and completion

Table 16: Statistics of course registration, participation and completion

<b>Dataset Reference and Name</b>	
Dataset Identifier	StatisticsForCourses
<b>Data set description</b>	
Generated or collected	Collected
Origin	<a href="http://videolectures.net">videolectures.net</a>
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.
<b>Standards and metadata</b>	
Methodology for data collection/management	JSON is used for Videlectures API
Metadata, supporting material	Videlectures REST api documentation
Status and location of metadata	N/A
<b>Data sharing</b>	
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 videlectures website; described as part of WP3 deliverables
Data repository	videlectures repository. Proximity to data source.
Dataset Link	N/A
<b>Archiving and preservation</b>	
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	videlectures - relying on internal quality assurance & back up procedures
Associated costs for data management	N/A

## 4.5.5 Aggregated statistics of engagement with the developed courses and educational resources

**Table 17: Aggregated statistics of engagement with the developed courses and educational resources**

<b>Dataset Reference and Name</b>	
Dataset Identifier	AggregatedStatistics
<b>Data set description</b>	
Generated or collected	Generated
Origin	<a href="https://videlectures.net">videlectures.net</a>
Scale	For videlectures - 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.
<b>Standards and metadata</b>	
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
<b>Data sharing</b>	
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 videlectures website; described as part of WP3 deliverables
Data repository	videlectures repository. Proximity to data source.
Dataset Link	N/A
<b>Archiving and preservation</b>	



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	Videlectures - relying on internal quality assurance & back up procedures
Associated costs for data management	Approximately 1 day of effort per month

#### 4.5.6 Recorded behavior of students following the first session of the process mining MOOC

**Table 18: Recorded behaviour of students following the first session of the process mining MOOC**

<b>Dataset Reference and Name</b>	
Dataset Identifier	CourseraMOOCprocmin001
<b>Data set description</b>	
Generated or collected	Collected
Origin	<a href="https://coursera.org">coursera.org</a>
Scale	several large tables
Who is this useful for?	learning analytics within EDSA
Similar existing datasets	Every Coursera course has this data recorded
<b>Standards and metadata</b>	
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
<b>Data sharing</b>	
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.
<b>Archiving and preservation</b>	
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

## 4.6 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. Aggregated statistics of the networking and engagement data will be produced and included in D4.4 and D4.5.

### 4.6.1 Web server logs and Google analytics of project website access

**Table 19: Web server logs and Google analytics of project website access**

<b>Dataset Reference and Name</b>	
Dataset Identifier	WebsiteAnalytics
<b>Dataset description</b>	
Generated or collected	Collected
Origin	<a href="http://edsa-project.eu">http://edsa-project.eu</a>
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.
<b>Standards and metadata</b>	
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 <a href="https://analytics.google.com/">https://analytics.google.com/</a>
<b>Data sharing</b>	
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
<b>Archiving and preservation</b>	
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

## 4.6.2 Generated social media engagement data

**Table 20: Generated social media engagement data**

<b>Dataset Reference and Name</b>	
Dataset Identifier	SocialMediaEngagements
<b>Dataset description</b>	
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.
<b>Standards and metadata</b>	
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	<a href="https://analytics.twitter.com/user/edsa_project/home">https://analytics.twitter.com/user/edsa_project/home</a>
<b>Data sharing</b>	
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 permit collection, harvesting and republication of data
How will the data be shared?	Dashboard on 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.
<b>Archiving and preservation</b>	
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

## 4.7 Work package 5 – Exploitation

WP5 will generate an on-going list of established collaboration initiatives, institutions benefiting from the project and geographical regions using the project's results. The EDSA Register is an additional dataset that comes under this work package.

### 4.7.1 List of project exploitation results - collaborations, institutional and geographical beneficiaries

**Table 21: List of project exploitation results - collaborations, institutional and geographical beneficiaries**

<b>Dataset Reference and Name</b>	
Dataset Identifier	ProjectExploitation
<b>management description</b>	
Generated or collected	Generated
Origin	Project partners
Scale	Variable
Who is this useful for?	Internal analysis for results to be exploited and targets
Similar existing datasets	None. Provides data on dissemination activity, network and results.
<b>Standards and metadata</b>	
Methodology for data collection/management	Report detailing results from interviews and exploitation activities
Metadata, supporting material	This data will be internal only
Status and location of metadata	This data will be internal only
<b>Data sharing</b>	

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?	Confidentiality
How will the data be shared?	Deliverable reports in WP5.
Data repository	Google docs shared document
Dataset Link	This data will be internal only
<b>Archiving and preservation</b>	
How long should the data be preserved?	Until the end of the project
Approx end volume	< 500MB
Who is responsible for the data management and curation?	ideXlab lead data management curation
Quality assurance including back up procedures	Backed up remotely
Associated costs for data management	Free storage. 1 day per month

## 4.7.2 The EDSA Register

**Table 22: The EDSA Register**

<b>Dataset Reference and Name</b>	
Dataset Identifier	EDSARegister
<b>Dataset description</b>	
Generated or collected	Generated
Origin	Project partners
Scale	<500KB
Who is this useful for?	Anyone interested in understanding the datasets used within the EDSA project. Internal management tool.
Similar existing datasets	None.
<b>Standards and metadata</b>	



Methodology for data collection/management	Project partners update every three months until the end of the project. ODI responsible for conversion to CSV and publication as open data.
Metadata, supporting material	A README.md file is available detailing the data structure and basic usage.
Status and location of metadata	<a href="https://theodi.github.io/european-data-science-academy-register/">https://theodi.github.io/european-data-science-academy-register/</a>
<b>Data sharing</b>	
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 ( <a href="http://edsa-project.eu/resources/datasets/">http://edsa-project.eu/resources/datasets/</a> )
Data repository	Github
Dataset Link	<a href="https://theodi.github.io/european-data-science-academy-register/">https://theodi.github.io/european-data-science-academy-register/</a>
<b>Archiving and preservation</b>	
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.
Approx end volume	<500KB
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	Stored in external repositories - EDSA website and Github
Associated costs for data management	Stored in external repositories - EDSA website and Github; approximately 2 days per month effort for maintenance.