

# Open Research Online

---

The Open University's repository of research publications and other research outputs

## The European Data Science Academy: Bridging the Data Science Skills Gap with Open Courseware

Conference or Workshop Item

How to cite:

Mikroyannidis, Alexander; Domingue, John; Phethean, Christopher; Beeston, Gareth and Simperl, Elena (2017). The European Data Science Academy: Bridging the Data Science Skills Gap with Open Courseware. In: Open Education Global Conference 2017, 8-10 Mar 2017, Cape Town, South Africa.

For guidance on citations see [FAQs](#).

© [not recorded]

Version: Accepted Manuscript

Link(s) to article on publisher's website:

<http://conference.oeconsortium.org/2017/presentation/the-european-data-science-academy-bridging-the-data-science-skills-gap-w>

---

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

---

[oro.open.ac.uk](http://oro.open.ac.uk)

# **The European Data Science Academy: Bridging the Data Science Skills Gap with Open Courseware**

Alexander Mikroyannidis and John Domingue, Knowledge Media Institute, The Open University  
{Alexander.Mikroyannidis, John.Domingue}@open.ac.uk  
Christopher Phethean, Gareth Beeston and Elena Simperl, University of Southampton  
{C.J.Phethean, Gareth.Beeston, E.Simperl}@soton.ac.uk

## **Abstract**

As a global society, we are producing data at an incredible rate, fuelled by the increasing ubiquity of the Web, and stoked by social media, sensors, and mobile devices. However, as the amount of produced data continues to increase, so does the demand for practitioners who have the necessary skills to manage and manipulate this data. The European Data Science Academy (EDSA) is looking to bridge the data science skills gap by developing multimodal open courseware tailored to the real needs of data practitioners. The EDSA courseware is implemented as a combination of living learning materials and activities (eBook, online courses, webinars, face-to-face training), produced via a rigorous process and validated by the data science community through continuous feedback.

## **Keywords**

Data Science, Open Courseware, Open Educational Resources, Massive Open Online Courses.

## **Introduction**

An ongoing revolution is occurring lately in higher education, largely driven by the availability of high quality online materials, also known as Open Educational Resources (OERs). OERs can be described as “teaching, learning and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or repurposing by others depending on which Creative Commons license is used” (Atkins, Brown, & Hammond, 2007). The emergence of OERs has greatly facilitated online education through the use and sharing of open and reusable learning resources on the Web. Learners and educators can now access, download, remix, and republish a wide variety of quality learning materials available through open services provided on the Web.

The OER initiative has recently culminated in MOOCs (Massive Open Online Courses), which offer large numbers of students the opportunity to study high quality courses with prestigious universities. These initiatives have led to widespread publicity and also strategic dialogue in the higher education sector. The consensus within higher education is that after the Internet-induced revolutions in communication, business, entertainment, media, amongst others, it is now the turn of universities. Exactly where this revolution will lead is not yet known but some radical predictions have been made including the end of the need for university campuses (Cadwalladr, 2012).

At the same time, the ‘Age of Data’ continues to thrive, with data being produced from all industries at a phenomenal rate that introduces numerous challenges regarding the collection,

storage and analysis of this data. Declared by Harvard Business Review as the “sexiest job of the 21st century” (Davenport & Patil, 2012), data science skills are becoming a key asset in any organisation confronted with the daunting challenge of making sense of information that comes in varieties and volumes never encountered before. The title is typically linked to a number of core areas of expertise, from the ability to operate high-performance computing clusters and cloud-based infrastructures, to the know-how that is required to devise and apply sophisticated Big Data analytics techniques, and the creativity involved in designing powerful visualizations (Magoulas & King, 2014). Moving further away from the purely technical, organizations are more and more looking into novel ways to capitalize on the data they own (Benjamins & Jariego, 2013), and to generate added value from an increasing number of data sources openly available on the Web, a trend which has been coined as “open data”.<sup>1</sup> To do so they need their employees to understand the legal and economic aspects of data-driven business development, as a prerequisite for the creation of product and services that turn open and corporate data assets into decision making insight and commercial value.

Data scientists are, however, still a rare breed. Beyond the occasional data-centric startup and the data analytics department of large corporations, the skills scarcity is already becoming a threat for many European companies and public sector organizations as they struggle to seize Big Data opportunities in a globalized world (Domingue, d'Aquin, Simperl, & Mikroyannidis, 2014). A McKinsey study estimated already in 2011 that the United States will soon require 60 percent more graduates able to handle large amounts of data as part of their daily jobs (James et al., 2011). With an economy of comparable size (by GDP) and growth prospects, Europe will most likely be confronted with a similar talent shortage of hundreds of thousands of qualified data scientists, and an even greater need of executives and support staff with basic data literacy. The number of job descriptions and an increasing demand in higher-education programs and professional training confirm this trend (Glick, 2013), with some EU countries forecasting an increase of almost 100 percent in the demand for data science positions in less than a decade (McKenna, 2012).

This paper introduces the European Data Science Academy (EDSA),<sup>2</sup> a European-funded research project looking to bridge the data science skills gap across sectors and in line with the requirements in each sector, through the development of bespoke open courseware. The remainder of this paper is structured as follows. First, the EDSA project is introduced in terms of its objectives and overall approach. The process established for the production of the curriculum and learning resources of the project is then presented, followed by the best practices acquired from this process. Finally, the paper is concluded and the next steps of this work are outlined.

## **Objectives & approach**

The EDSA project is establishing a virtuous learning production cycle for data science, in order to meet the following objectives:

- Analyse the sector specific skillsets for data analysts across Europe’s main industrial sectors;
- Develop modular and adaptable curricula to meet these data science needs; and

---

<sup>1</sup> <http://okfn.org/opendata/>

<sup>2</sup> <http://edsa-project.eu>

- Deliver training supported by multiplatform and multilingual learning resources based on these curricula.

Throughout the project, the curricula and learning resources are guided and evaluated by experts in both data science and pedagogy to ensure they meet the needs of the data science community. The following paragraphs outline the main activities that are currently being carried out by the project to meet the above objectives.

*Demand Analysis:* EDSA is monitoring trends across the EU to assess the demands for particular data science skills and expertise. We are leveraging a vast network of European data providers, consumers and intermediaries to “track the pulse” of the European data landscape. This allows us to align our criteria with the latest demands of the community.

Using interviews with data science practitioners, an Industrial Advisory Board representing a mix of sectors and automated tools for extracting data about job posts and news articles, we are building dashboards to present the current state of the European data science landscape, with the data feeding into our curricula development.

*Curricula Development:* EDSA is developing a core data science curriculum based on topics extracted from the demand analysis. We are producing high-quality, multilingual and multimodal training materials to cover these topics, utilising existing resources available in the public domain and the internal expertise of the EDSA consortium. The curriculum is constantly updated based on the process model that is adopted by the project for the production of learning resources. This model is driven by a participatory approach that defines a series of iterations in the production of learning materials, with multiple revisions from internal and external stakeholders, in order to ensure high quality in the produced materials.

*Training Delivery and Learning Analytics:* Key parts of our curricula are being delivered through eBooks, MOOCs, webinars, video lectures and face-to-face training. Several members of the EDSA consortium are already established as high-quality training providers in core data science topics. As part of EDSA, these initiatives will be structured into integrated learning pathways, translated into European languages, and expanded to meet the requirements for specific sectors as indicated by our demand analysis.

We are primarily using VideoLectures<sup>3</sup> and FutureLearn<sup>4</sup> – the largest European MOOC platform, founded by The Open University – to maximise outreach and uptake of our materials. Engagement with learners and Big Data stakeholders is key to our training, and therefore monitoring and analysis tools will be utilised to assess learner progress – rather than by simply listening to academics or technology evangelists.

## **Developing open courseware for data science**

A participatory approach has been adopted toward the design and production of the EDSA courseware, which is shown in Figure 1. This approach builds upon and extends the courseware production process established in the EUCLID project, which was focused primarily on the design and delivery of learning resources about Linked Open Data (A. Mikroyannidis,

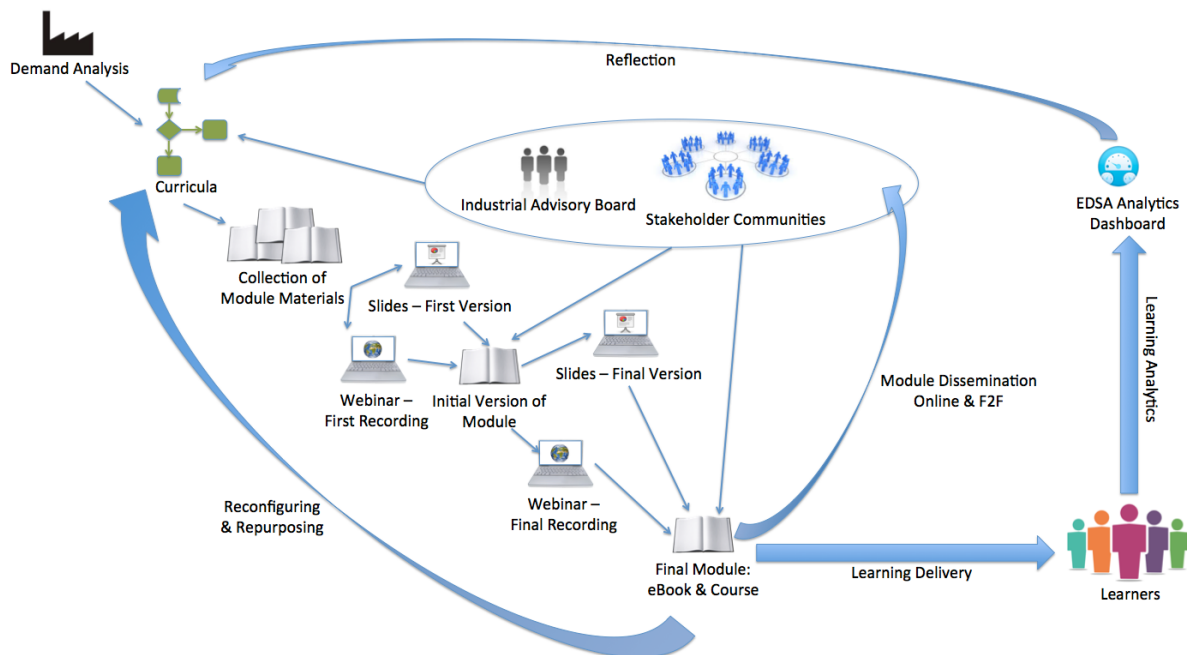
---

<sup>3</sup> <http://videolectures.net>

<sup>4</sup> <https://www.futurelearn.com>

Domingue, Maleshkova, Norton, & Simperl, 2014; Alexander Mikroyannidis, Domingue, Maleshkova, Norton, & Simperl, 2016).

Starting from the results of the demand analysis and input from the Industrial Advisory Board, we are creating relevant data science curricula to meet the outlined training needs. A multidisciplinary course writing team is developing in parallel a repository of relevant source materials, draft modules that will be placed online, as well as materials for webinars. The draft modules are then iteratively revised based on the feedback received from the Industrial Advisory Board, from the face-to-face training activities, as well as from monitoring the main communication channels used by the communities of stakeholders. The analysed feedback is used to restructure and finalise the module content as an eBook and online course, which are then delivered to the stakeholder communities (to support their own training needs) and to target learner communities both online and face-to-face.



**Figure 1. The EDSA production process for open courseware.**

Learning Analytics have been incorporated into our online delivery, allowing us to collect data related to the learning experiences of our users, which feedback into our curricula design. Based upon the Learning Analytics data and the feedback from our stakeholders, we reconfigure and repurpose modules for different learning contexts initiating new cycles of the production process.

The EDSA curricula and learning resources are tested and evaluated during both development and deployment. This evaluation is targeting pedagogical correctness, fit to sector, as well as the overall quality of the learning experience. Throughout the design, development and deployment of our curricula and learning resources, we actively involve pedagogical experts, who provide advice on the design of the curricula and learning resources. Additionally, the Industrial Advisory Board represents relevant industrial sectors and ensures that the developed learning resources are applicable, relevant and at a suitable skill level to meet industry demand.

The EDSA curriculum targets the following 4 themes, which provide the core framework for the development of the EDSA courses:

- Foundations of data science
- Data Storage and Processing
- Data Analysis
- Data Interpretation and Use

Based on the EDSA curriculum, the project is developing a courses portfolio, which includes a wide range of data science learning resources adopting a variety of pedagogical models, as well as employing different delivery channels and formats in order to address different learning contexts and audiences. The EDSA courses cover all types of learning contexts, from the traditional face-to-face pedagogical model, to the more recent trends in online education (MOOCs and OERs):

- *Self-study courses*: These courses consist of self-study learning materials available as Open Educational Resources (OERs). Learners can study them at their own pace, as there is no predetermined start or end date.
- *Massive Open Online Courses (MOOCs)*: These are online courses aimed at unlimited participation and open access on the web. They are available on external MOOC platforms, such as Coursera and FutureLearn.
- *Face-to-face courses*: These courses are taught face-to-face. Face-to-face learning (or in-person learning) is any form of instructional interaction that occurs “in person” and in real time between teachers and students or among colleagues and peers.
- *Blended courses*: These courses are taught in a blended way (face-to-face and online). Blended learning is a formal education program in which a student learns at least in part through delivery of content and instruction via digital and online media with some element of student control over time, place, path, or pace.

The EDSA courses employ different delivery channels and formats in order to maximise the impact of the EDSA learning materials on the community and bring them closer to as many students and practitioners as possible. In particular, the EDSA courses are available:

- Via the EDSA online courses portal<sup>5</sup>
- As an interactive eBook<sup>6</sup>

The EDSA online courses portal hosts the full learning materials (presentations, webinars, text, quizzes, etc.) for the self-study courses, to which learners can enrol and study at their own pace. Learners are not required to register in the portal in order to enrol to a course, but they can login using an existing social media account, such as Google, Facebook or LinkedIn. The portal also lists the other types of courses that the project offers, namely MOOCs, blended and face-to-face courses. All types of courses have been enriched with the following set of metadata:

- Category (e.g. blended, face-to-face, etc.)
- Level (i.e. basic, intermediate, advanced)

---

<sup>5</sup> <http://courses.edsa-project.eu>

<sup>6</sup> <http://courses.edsa-project.eu/mod/page/view.php?id=299>

- Institution offering the course
- Target sector
- Target audience
- Gained skills
- Location
- Language
- License

EUROPEAN DATA SCIENCE ACADEMY ONLINE COURSES

Search courses

Try the EDSA self-study courses

Welcome to the EDSA online courses portal

Here you can find a variety of courses offered by the European Data Science Academy (EDSA). EDSA is a Horizon 2020 project, aiming at bridging the data science skills gap across Europe. [Learn more about EDSA »](#)

### Find courses

Search

Filter by:

Category	Level	Institution	Sector
3 Blended	17 Advanced	11 Fraunhofer	1 Construction
15 Face-to-face	8 Basic	1 JSI	2 Energy
3 MOOC	5 Intermediate	3 KTH	1 Engineering
9 Self-study		4 ODI	1 Market research

Location	Language	License
1 Croatia	24 English	4 CC BY-NC-ND
9 Germany	6 German	4 CC BY-NC-SA
1 Greece		1 Copyrighted work
12 Online		

**Target audience**

application developers architects artists big data system developers business analysts business developers business professionals communications managers computing centres consultants **data analysts** data engineers data leaders data managers **data scientists** data strategists data wranglers database administrators database managers **developers** directors engineers executives graduates ICT architects ICT professionals innovation managers journalists machine learning developers management of building infrastructure **managers** marketing managers performance engineers

**Gained skills**

agent-based software engineering analysis of social media posts audio fingerprinting basics of machine learning methods batch processing **big data** big data architectures big data components big data technologies business analytics business potentials business process management cloud computing commercial big data systems convolutional neural networks data analytics data classification data clustering data engineering data evaluation data features data management data mining data preparation **data processing** data regression data science data storytelling data transformation

**Course categories**

- Self-study courses
- MOOCs
- Blended courses
- Face-to-face courses
- All courses ...

**Latest courses**

- ODI Pre-Summit Training Day
- Deep Learning using TensorFlow Workshop
- Big Data Architecture
- Peer-to-Peer, Cloud Computing and Big Data

**Download the EDSA eBook**

The EDSA eBook is available for the iPad and MacOS (iBooks format) and for all other devices (ePUB format). [Read more »](#)

Figure 2. The EDSA online courses portal.

Additionally, the portal features a faceted search interface (see Figure 2), allowing users to find courses based on a set of search criteria derived from the metadata of the courses. Users, for example, can filter courses by selecting their preferred level of study and the skills they want to acquire from a tag cloud displaying the skills attached to the offered courses.

The EDSA eBook offers an additional delivery medium for the project's courses, targeting primarily tablet devices and mobile phones. In order to widen the audiences reached via different platforms, the EDSA eBook is available both in the iBooks format (supported by iOS and MacOS) and the ePUB format (supported by most desktop and tablet devices). The eBook contains the textual and image/video learning resources of the EDSA self-study courses, as well as self-assessment exercises in the form of quizzes.



Figure 3. Screenshots from the EDSA eBook.

## Best practices for the design and delivery of data science courseware

Feedback acquired so far from the data science community on the EDSA courseware has provided us with a valuable insight into the real needs of data practitioners across a number of sectors. The deployment of the EDSA courseware production process has also led us to identify certain challenges associated with the design and delivery of learning resources specifically for data science. We have thus distilled our experiences and lessons learned into a set of best practices, which is outlined in the following sections.



### ***Best practices for the design of data science courseware***

- *Industry Aligned* – The curriculum is designed in accordance with the expectations of EU industrial sectors connected to data science, providing industry-standard scenarios and tools.
- *Industry Standard Tools* – Our compilation of open source data science tools offer learners experience with tools customary to the industry and their specific sector.
- *Real Data* – Learners utilising this curriculum have access to a number of large-scale open datasets to perform their learned data science skills, enabling real-world data science on real-world data.
- *Open Design* – Our curriculum is designed from user, research, industry and professional recommendations and feedback taken into account from all across the EU, ensuring that the curriculum meets the needs of the industry, academia and the wider market.
- *Expert Provision* – A curriculum that is designed by world-class professional and academic experts in data science.
- *Modular* – The curriculum is flexible and adaptable to educator requirements and the needs of their learners.
- *Transferrable* – Skills learned through the curriculum can be utilised across a range of data science roles, occupations and countries throughout the EU.
- *Concise Learning Goals* – All courses are aligned with clear learning goals depicted by a specific aspect of the data science role. Learning pathways are provided to enable learners to navigate through the content, selecting what is useful to them.
- *Addressing the Whole Data Value Chain* – Data scientists are made aware of the techniques and stages of the whole data science value chain through the use of easily understandable narratives.

### ***Best practices for the delivery of data science courseware***

- *Multilingual* – Learning resources are delivered across a number of European languages in order to extend their reach and enable others to use our curriculum.
- *Multimodal* – Learning resources are provided in a number of modes to suit skill levels and format preferences, such as MOOCs, eBooks and slide decks.
- *Multi-Platform* – Learning resources are delivered via a wide range of platforms in order to remain accessible and available to a large body of data science learners.
- *Reusable* – Learning resources are released under Creative Commons licenses, allowing the community to reuse, repurpose and republish them.
- *Cutting-Edge Quality* – Learning resources are subject to a series of design iterations that encapsulate the latest research and professional practice, prior to their launch.
- *Reflective and Quantified* – Learning resources are delivered with data and analytics in mind, providing all learners quantified measures and analytics to reflect on their aptitude, skills and strengths.

- *Hands-On* – Learning resources are delivered in a way to emphasise a constructivist hands-on approach, meaningfully applying knowledge to real tools and data.

## Conclusion

The EDSA project has established a rigorous process for the production and delivery of open courseware for data science. This process defines a series of iterations in the production of learning resources, with multiple revisions from internal and external stakeholders, in order to ensure high quality in the produced resources. Based on our experiences and lessons learned in designing and implementing the production process, we have also established a set of best practices for the design and delivery of open courseware for data science.

Designing a curriculum that covers data science is an inherently difficult task that faces a number of challenges, most notably the speed at which this field is changing (Hirsh, 2008). Increasing amounts of data lead to challenges around data storage and processing, not to mention increasing complexity in finding the useful story from that data. New computing technologies rapidly lead to others becoming obsolete. New tools are developed which change the data science landscape. These all occur at such a rapid pace that teaching data science requires an agile and adaptive approach that can respond to these changes. In the context of EDSA, we will carry out revisions to the curriculum and the associated learning resources throughout the duration of the project, in order to reflect the most up-to-date needs of the community and the latest cutting-edge techniques for making sense of data. By carrying out rigorous Learning Analytics and sourcing input from the learners and the wider data science community, we aim to ensure that the content on offer from EDSA continues to match these updates.

## Acknowledgement

EDSA is a research project funded by the Horizon 2020 Framework Programme of the European Union, Grant Agreement no. 643937.

## References

- Atkins, D. E., Brown, J. S., & Hammond, A. L. (2007). *A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities*. Retrieved from [http://www.hewlett.org/uploads/files/Hewlett\\_OER\\_report.pdf](http://www.hewlett.org/uploads/files/Hewlett_OER_report.pdf)
- Benjamins, R., & Jariego, F. (2013). Open Data: A ‘No-Brainer’ for all. Retrieved from <http://blog.digital.telefonica.com/2013/12/05/open-data-intelligence/>
- Cadwalladr, C. (2012). Do online courses spell the end for the traditional university? Retrieved from <http://www.theguardian.com/education/2012/nov/11/online-free-learning-end-of-university>
- Davenport, T. H., & Patil, D. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*.
- Domingue, J., d'Aquin, M., Simperl, E., & Mikroyannidis, A. (2014). The Web of Data: Bridging the Skills Gap. *IEEE Intelligent Systems*, 29(1), 70-74. doi:10.1109/MIS.2014.15

- Glick, B. (2013). Government calls for more data scientists in the UK. Retrieved from <http://www.computerweekly.com/news/2240208220/Government-calls-for-more-data-scientists-in-the-UK>
- Hirsh, H. (2008). Data mining research: Current status and future opportunities. *Statistical Analysis and Data Mining*, 1(2), 104-107.
- James, M., Michael, C., Brad, B., Jacques, B., Richard, D., Charles, R., & Angela, H. (2011). Big data: The next frontier for innovation, competition, and productivity. *The McKinsey Global Institute*.
- Magoulas, R., & King, J. (2014). *2013 Data Science Salary Survey: Tools, Trends, What Pays (and What Doesn't) for Data Professionals*: O'Reilly.
- McKenna, B. (2012). Demand for big data IT workers to double by 2017, says eSkills. Retrieved from <http://www.computerweekly.com/news/2240174273/Demand-for-big-data-IT-workers-to-double-by-2017-says-eSkills>
- Mikroyannidis, A., Domingue, J., Maleshkova, M., Norton, B., & Simperl, E. (2014). *Developing a Curriculum of Open Educational Resources for Linked Data*. Paper presented at the 10th annual OpenCourseWare Consortium Global Conference (OCWC), Ljubljana, Slovenia.
- Mikroyannidis, A., Domingue, J., Maleshkova, M., Norton, B., & Simperl, E. (2016). Teaching Linked Open Data Using Open Educational Resources. In D. Mourmtesev & M. d'Aquin (Eds.), *Open Data for Education: Linked, Shared, and Reusable Data for Teaching and Learning* (pp. 135-152). Cham: Springer International Publishing.