Designing and Delivering a Curriculum for Data Science Education across Europe

Alexander Mikroyannidis and John Domingue
The Open University, Milton Keynes, United Kingdom
{Alexander.Mikroyannidis, John.Domingue}@open.ac.uk

Christopher Phethean, Gareth Beeston and Elena Simperl
University of Southampton, Southampton, United Kingdom
{C.J.Phethean, Gareth.Beeston, E.Simperl}@soton.ac.uk

Abstract—Data is currently being produced at an incredible rate globally, fuelled by the increasing ubiquity of the Web, and stoked by social media, sensors, and mobile devices. However, as the amount of available data continues to increase, so does the demand for professionals who have the necessary skills to manage and manipulate this data. This paper presents the European Data Science Academy (EDSA), an initiative for bridging the data science skills gap across Europe and training a new generation of world-leading data scientists. The EDSA project has established a rigorous process and a set of best practices for the production and delivery of curricula for data science. Additionally, the project's efforts are dedicated to linking the demand for data science skills with the supply of learning resources that offer these skills. In order to achieve this, EDSA is offering interactive tools for finding learning resources and building personalised learning pathways towards acquiring the skills that are currently in demand.

Keywords—Data Science, Curricula, Courseware, Skills, Demand Analysis, Personalised Learning Pathways.

1 Introduction

Data has the potential to revolutionise business, government, and society. Data Science methods offer a variety of instruments to create economic and social value, so that decisions are informed by insights and inferences gained from data analyses, while products and services are optimally designed and delivered to meet customer needs. By combining both internally owned and external sources of data, organisations can learn about their processes, accurately plan and target operations, and achieve significant cost savings and productivity gains. This applies to any type of organisation, public or private. For instance, the large amount of data published openly by governmental agencies enables reductions of labour costs and lean management, as well as a better level of accountability and transparency towards citizens and other organisations.

The 'Age of Data' is currently thriving, 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" [3], 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 [8]. Moving further away from the purely technical, organizations are more and more looking into novel ways to capitalize on the data they own [2], 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".[1] 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.

Nevertheless, data scientists are 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 [4]. 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 [7]. 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 [5], with some EU countries forecasting an increase of almost 100 percent in the demand for data science positions in less than a decade [9].

Training data scientists and designing curricula that cover data science pose a number of challenges, most notably due to the speed at which this field is changing [6]. 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.

The European Data Science Academy (EDSA)[2] aims at bridging the data science skills gap across Europe. EDSA has established a virtuous learning production cycle for data science in order to analyse the sector specific skillsets for data analysts across Europe's main industrial sectors; develop modular and adaptable curricula to meet

[1] http://okfn.org/opendata/
[2] http://edsa-project.eu

these data science needs; and deliver training supported by multiplatform and multilingual learning resources based on these curricula.

The remainder of this paper is structured as follows. First, the EDSA approach for bridging the data science skills gap is presented, with an emphasis on the methodology for the development of curricula and courseware, as well as linking demand with supply. We then discuss the EDSA best practices for the design and delivery of data science curricula. Finally, the paper is concluded and the next steps of this work are outlined.

2 Bridging the data science skills gap

2.1 Methodology for the development of curricula and courseware

In order to address the demand for data science skills, a participatory approach has been adopted by EDSA for the design and production of bespoke curricula and courseware (see Figure 1). This approach builds upon and extends the courseware production process established in the EUCLID project,[3] which was focused primarily on the design and delivery of learning resources about Linked Open Data [10, 11].

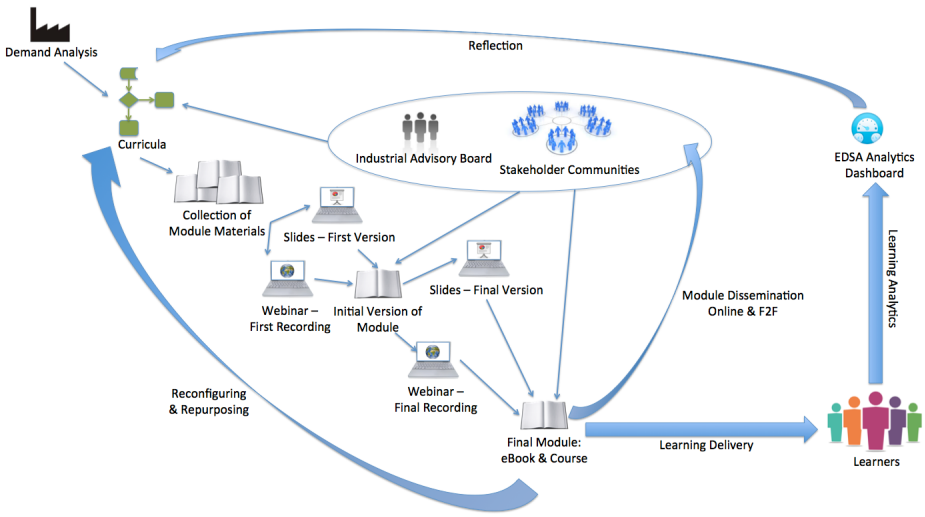

Figure 1. The EDSA production process for curricula and courseware.

EDSA is monitoring trends across Europe in order to assess the demands for particular data science skills and expertise, using automated tools for the extraction of data science job posts, as well as interviews with data science practitioners. The project has also established an Industrial Advisory Board representing a mix of

[3] http://www.euclid-project.eu

sectors to ensure that project activities continue to meet changes in the demands on data science across Europe.

Starting from the results of this 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.

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.

Based on the EDSA curricula, 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 more recent trends in online education:

- *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 FutureLearn[4] and Coursera.[5]
- *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.
- *Online courses:* These courses are taught online via Learning Management Systems (LMSs) like Moodle or Sakai. A subset of these courses consists of self-

[4] https://www.futurelearn.com
[5] https://www.coursera.org

study learning materials available as Open Educational Resources (OERs) [1], which learners can study at their own pace, as there is no predetermined start or end date.

- *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 courses portal[6] and as an interactive eBook.[7]

The courses portal is the EDSA hub for courses offered both by the project consortium, as well as by external organisations. The portal features a faceted search interface, 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.

2.2 From demand to supply

Linking the demand for data science skills with the supply of learning resources that offer these skills is crucial for bridging the data science skills gap. Towards this goal, EDSA is developing an interactive dashboard that will enable its users to explore both the current data science skills demand and supply. Users of this dashboard will be able not only to explore the current demand in the data science market, but also find learning materials and training relevant to the skills they will need to secure a specific job position. We are deploying automated tools for extracting data about job posts and news articles in order to present the current state of the European data science landscape. Additionally, users are supported in building personalised learning pathways, consisting of courses and learning materials that will help them reach their learning goals.

In particular, the EDSA dashboard will enable users to:

- View the current demand for data science jobs and skills across Europe.
- Filter demand by required skills and region.

[6] http://courses.edsa-project.eu

[7] http://courses.edsa-project.eu/mod/page/view.php?id=299

- View trends and statistics regarding data science jobs and skills for a given timeframe.
- Explore the current supply of courses and learning materials that will help them acquire certain skills.
- Build personalised learning pathways towards acquiring certain skills.

Figures 2 and 3 show mock-ups of the EDSA dashboard, which is currently being developed. Figure 2 shows the default view which is deliberately kept as simple as possible. In this view, queries typed into the search box at the top result in a simple list of related data science jobs. Selecting any job results in additional details of the post being displayed. The toolbar below the query entry box allows users to add or remove additional views. In Figure 3, a map and courses view have been selected. Google maps are used for the map view incorporating zoom facilities. Selecting any anchor point in the map brings up details about the job. The courses view shows recommended courses related to the query, which are offered by the EDSA project consortium and external organisations. Additional work on these dashboard views will explore how simple filter mechanisms can be added, allowing users to focus, for example, on a single country, role or industrial sector.

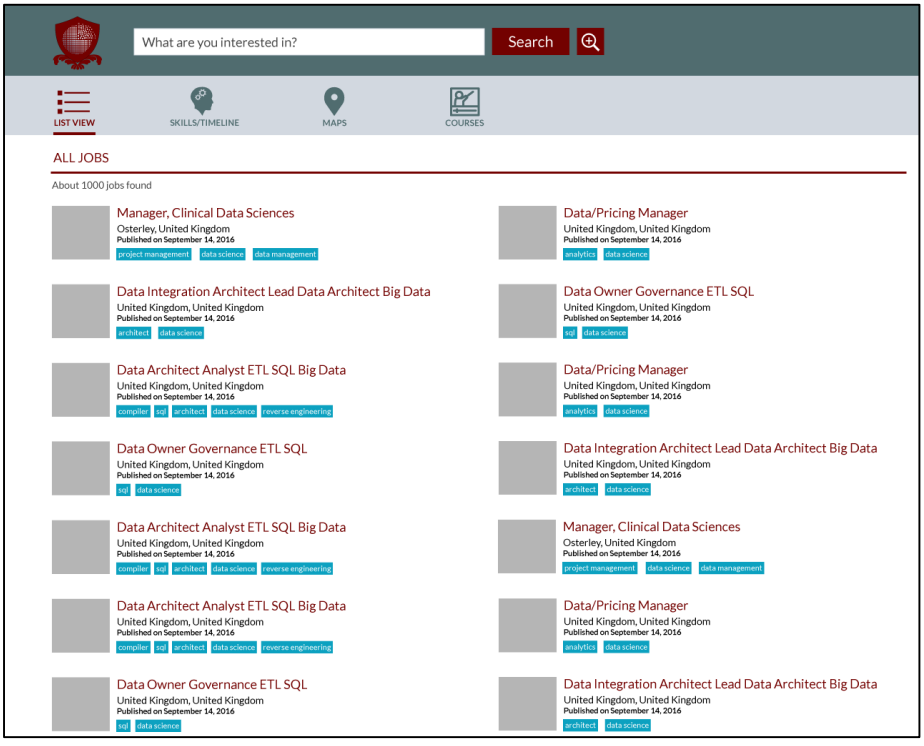

Figure 2. Mock-up of the EDSA dashboard default view.

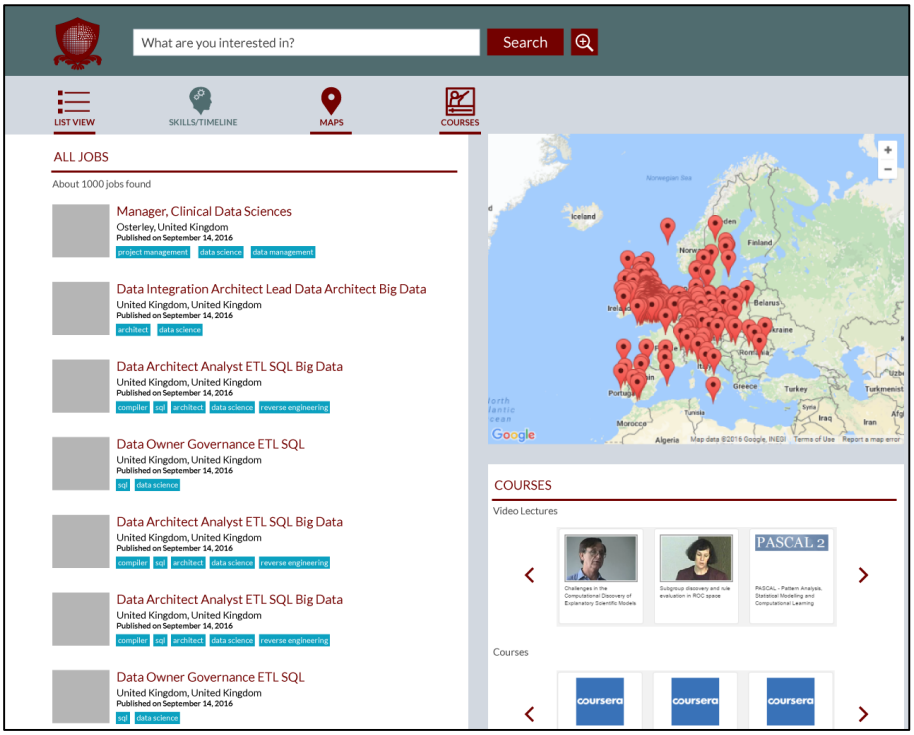

Figure 3. Mock-up of the EDSA dashboard showing a combination of the list view of jobs, a map view of jobs and a view of related courses.

In order to build their personalised pathways, users of the dashboard will start by identifying the job position that they are after. Based on their selection, the dashboard will present to users the skills that are required for the job position and will recommend courses for acquiring these skills. Following these recommendations, users will then be able to start building their learning pathway towards gaining the required skills.

Figure 4 shows the list of data science learning pathways currently offered by the EDSA courses portal.[8] These pathways are based on the EDISON Data Science Framework[9] and consist of recommended data science topics, as well as courses for acquiring certain sets of skills related to these topics. Users can use these pathways as templates in order to build their own pathways by adding courses, monitoring their progress towards completing their pathways, as well as reflecting on the contents of the pathways and on what they have learned, as shown in Figure 5.

[8] http://courses.edsa-project.eu/course/view.php?id=70
[9] http://edison-project.eu/edison/edison-data-science-framework-edsf

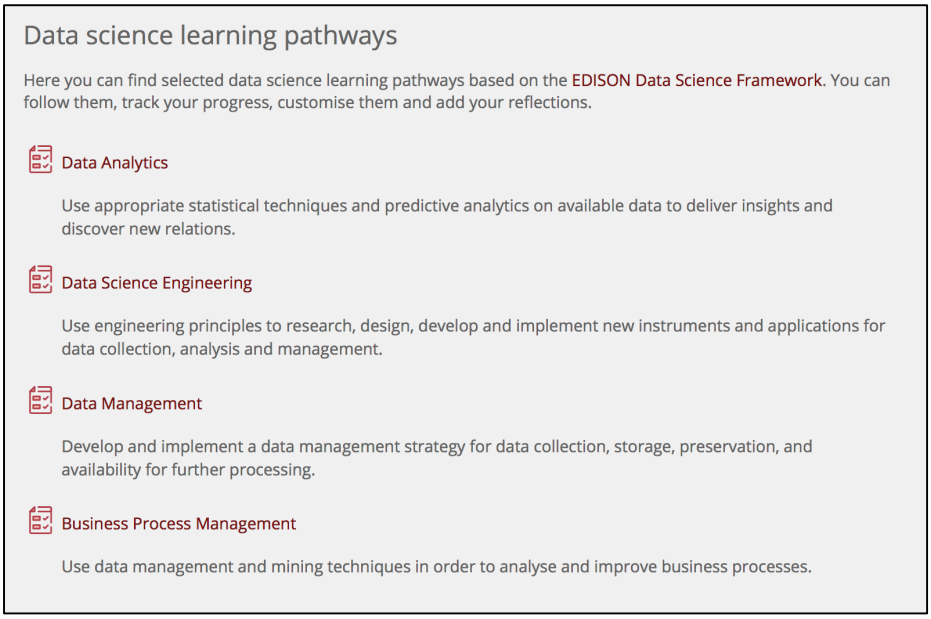

Figure 4. Data science learning pathways in the EDSA courses portal.

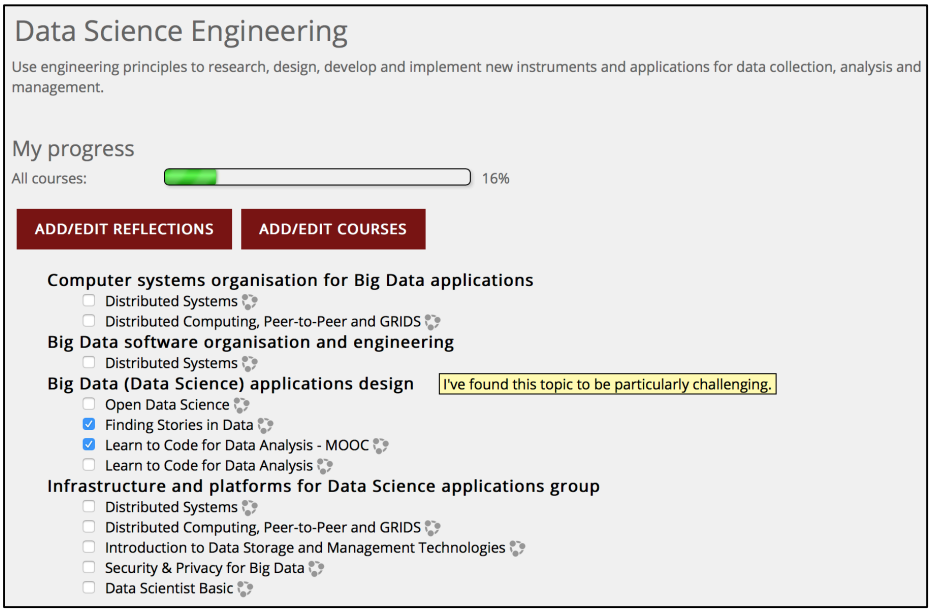

Figure 5. Building a personalised learning pathway.

3 Best practices for the design and delivery of data science curricula

Feedback acquired so far from the data science community on the EDSA curriculum has provided us with a valuable insight into the real needs of data practitioners across different sectors. The deployment of the EDSA curriculum and 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.

3.1 Best practices for the design of data science curricula

- *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.

3.2 Best practices for the delivery of data science curricula

- *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.

4 Conclusions and future work

The EDSA project has established a rigorous process for the production and delivery of curricula and 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 established a set of best practices for the design and delivery of curricula for data science. We are also working towards linking the demand for data science skills with the supply of learning resources that offer these skills via tools for exploring the current demand and building personalised learning pathways for acquiring the skills in demand.

The next steps of this work will involve regular reviews and updates of the EDSA curricula and the associated learning resources and learning pathways, so that these outputs reflect the latest trends in the data science market and the current needs of the data science community. By carrying out rigorous Learning Analytics and sourcing input from learners and the wider data science community, we aim to ensure that the learning content and tools on offer from EDSA continue to match the latest demand.

5 Acknowledgement

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

6 References

[1] Atkins, D. E., Brown, J. S., and Hammond, A. L. A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities. The William and Flora Hewlett Foundation 2007.

[2] Benjamins, R. and Jariego, F. Open Data: A 'No-Brainer' for all. Telefónica Innovation Hub, http://blog.digital.telefonica.com/2013/12/05/open-data-intelligence/, 2013.

[3] Davenport, T. H. and Patil, D. *Data scientist: The sexiest job of the 21st century*. Harvard Business Review, 2012.

[4] Domingue, J., d'Aquin, M., Simperl, E., and Mikroyannidis, A. *The Web of Data: Bridging the Skills Gap*. IEEE Intelligent Systems, 2014. **29**(1): p.70-74.

[5] Glick, B. Government calls for more data scientists in the UK. Computer Weekly, http://www.computerweekly.com/news/2240208220/Government-calls-for-more-data-scientists-in-the-UK, 2013.

[6] Hirsh, H. *Data mining research: Current status and future opportunities*. Statistical Analysis and Data Mining, 2008. **1**(2): p.104-107.

[7] James, M., Michael, C., Brad, B., Jacques, B., Richard, D., Charles, R., and Angela, H. *Big data: The next frontier for innovation, competition, and productivity*. The McKinsey Global Institute, 2011.

[8] Magoulas, R. and King, J. *2013 Data Science Salary Survey: Tools, Trends, What Pays (and What Doesn't) for Data Professionals.* O'Reilly, 2014.

[9] McKenna, B. Demand for big data IT workers to double by 2017, says eSkills. Computer Weekly, http://www.computerweekly.com/news/2240174273/Demand-for-big-data-IT-workers-to-double-by-2017-says-eSkills, 2012.

[10] Mikroyannidis, A., Domingue, J., Maleshkova, M., Norton, B., and Simperl, E. *Developing a Curriculum of Open Educational Resources for Linked Data.* In *Proceedings of the 10th annual OpenCourseWare Consortium Global Conference (OCWC)*. 2014. Ljubljana, Slovenia.

[11] Mikroyannidis, A., Domingue, J., Maleshkova, M., Norton, B., and Simperl, E., "Teaching Linked Open Data Using Open Educational Resources", in *Open Data for Education: Linked, Shared, and Reusable Data for Teaching and Learning*, Mouromtsev, D. and d'Aquin, M., Eds., Cham Springer International Publishing, 2016, p. 135-152.