Recognition of Formal and Non-formal Training in Data Science

Ernestina Menasalvas, Nik Swoboda, Ana Moreno, Andreas Metzger, Aristide Rothweiler, Niki Pavlopoulou, and Edward Curry

Abstract The fields of Big Data, Data Analytics and Data Science, which are key areas of current and future industrial demand, are quickly growing and evolving. Within Europe, there is a significant skills gap which needs to be addressed. A key activity is to ensure we meet future needs for skills and align the supply of educational offerings with the demands from industry and society. In this chapter, we detail one step in this direction, a programme to recognise Data Science skills. The chapter introduces the data skills challenge and the importance of formal and non-formal education. It positions data skills within a framework for skills and education, and it reviews key projects which have advanced the data skills agenda. It then introduces recognition frameworks for formal and non-formal Data Science training, and it details a methodology to achieve consensus between interested stakeholders in both academia and industry, and the platforms needed to be deployed for the proposal. Finally, we present a case study of the application of recognition frameworks within an online educational portal for students.

Keywords Big Data · Data skills recognition · Skill badges · Skill labels · Education hub

E. Menasalvas · N. Swoboda · A. Moreno (✉)
Universidad Politécnica de Madrid, Madrid, Spain
e-mail: emenasalvas@fi.upm.es; nswoboda@fi.upm.es; ammoreno@fi.upm.es

A. Metzger · A. Rothweiler
paluno, University of Duisburg-Essen, Essen, Germany

N. Pavlopoulou · E. Curry
Insight SFI Research Centre for Data Analytics, NUI, Galway, Ireland

1 Introduction

Nowadays, fields like Big Data, Data Analytics and Data Science have drawn a considerable amount of attention from industry. In order to boost the data-driven economy in Europe, the data needs required by industry keep growing; therefore, the main challenge is bridging the gap between these industrial needs and the availability of skilled data scientists.

The popularity of data-oriented fields has an impact on the creation of a plethora of degrees in universities and online courses that offer a wide range of skill sets to aspiring data scientists. Therefore, the data skills needed by industry can be acquired through formal learning (e.g. undergraduate or graduate university degrees) or non-formal learning (e.g. e-learning or professional training).

Nevertheless, the availability of a plethora of resources does not suggest a direct link between industry and future data scientists, resulting in a range of challenges for the gap to be bridged, defined below:

- Given the constant technological and societal changes, the needs may also quickly change; therefore, it is vital to identify the current industrial needs or trends and adjust the educational offerings according to those altered needs.
- Given the plethora of available formal and non-formal programmes, there is a need to provide a platform and living repository that will give more targeted and filtered access to these resources to potential data scientists or professionals that want to enhance their skills.
- A programme needs to be defined that will be able to provide recognition of skills of data scientists acquired through both formal and non-formal education.
- A framework needs to be defined that will align the current industrial needs with the Data Science curricula and skills provided by formal and non-formal institutions.

This chapter explores the ways in which Europe could build a strong and vibrant big data economy by tackling the challenges above through the enhancement of the benefits that educational institutions and existing skills recognition initiatives have to offer. Specifically, some directions towards the desirable result involve the creation of the Big Data Value Education Hub (EduHub) and the Big Data Value (BDV) Data Science Badges and Labels.

The EduHub is a platform that provides access to Data Science and Data Engineering programmes offered by European universities as well as on-site/online professional training programmes. The aim of the platform is to facilitate knowledge exchange on educational programmes and meet current industrial needs.

BDV Data Science Badges and Labels are skills recognition programmes for skills acquired by formal and non-formal education, respectively. The initial stage of the badges contained the types and requirements for the system by leveraging

existing work by the European Data Science Academy¹ (EDSA) and EDISON² projects, which were European Union (EU) projects related to Data Science skills. Later, the programmes were enhanced by gathering feedback from academia and industry and by proposing methodologies to bring together interested stakeholders (from both academia and industry) for the design and deployment of the badges and labels, as well as their evaluation and feedback.

This chapter also explores a practical view of how this platform and the skills recognition programme can work in isolation as well as together in order to bridge the industry with academia. This is presented via a pilot of the BDV Data Science Analytics Badge that is currently issued by two universities and the way the badges as well as the educational programmes which issue them can be accessed in the EduHub.

1.1 The Data Skills Challenge

In order to leverage the potential of BDV, a key challenge for Europe is to ensure the availability of highly and correctly skilled people who have an excellent grasp of the best practices and technologies for delivering BDV within applications and solutions (Zillner et al. 2017). In addition to meeting the technical, innovation and business challenges as laid out in this chapter, Europe needs to systematically address the need to educate people so that they are equipped with the right skills and are able to leverage BDV technologies, thereby enabling best practices. Education and training will play a pivotal role in creating and capitalising on BDV technologies and solutions.

There was a need to jointly define the appropriate profiles required to cover the full data value chain. One main focus should be on the individual needs linked to company size. Start-ups, SMEs and big industries have individual requirements in Data Science. We distinguish between three different profiles, (1) to cover the hardware- and software-infrastructure-related part, (2) the analytical part and (3) the business expertise.

The educational support for data strategists and data engineers is, however, far too limited to meet the industry's requirements, mainly due to the spectrum of skills and technologies involved. By transforming the current knowledge-driven approach into an experience-driven one, we can fulfil industry's needs for individuals capable of shaping the data-driven enterprise. Current curricula are furthermore highly siloed, leading to communication problems and suboptimal solutions and implementations. The next generation of data professionals needs this wider view in order to deliver the data-driven organisation of the future:

¹<http://edsa-project.eu/>

²<http://edison-project.eu/>

- **Data-intensive engineers:** Successful data-intensive engineers control how to deal with data storage and management. They are experts on distributed computing and computing centres; hence they are mostly at the advanced system administrator levels. They have the know-how to operate large clusters of (virtual) machines, configure and optimise load balancing, and organise Hadoop clusters, and know about Hadoop Distributed File System and Resilient Distributed Datasets, etc.
- **Data scientists:** Successful data scientists will require solid knowledge in statistical foundations and advanced data analysis methods, combined with a thorough understanding of scalable data management, with the associated technical and implementation aspects. They will be the specialists that can deliver novel algorithms and approaches for the BDV stack in general, such as advanced learning algorithms and predictive analytics mechanisms. They are data-intensive analysts. They need to know statistics and data analysis; they need to be able to talk to data-intensive engineers, but should be relieved from system administrator problems; and they need to understand how to transform problems into appropriate algorithms which may need to be modified slightly. Data scientist benchmarks select and optimise these algorithms to reach a business objective. They also need to be able to evaluate the results obtained, following sound scientific procedures. A data scientist curriculum would ideally provide enough insight into the Data Engineering discipline to steer the selection of algorithms, not only from a business perspective but also from an operational and technical perspective. For this, Europe needs new educational programmes in Data Science as well as ideally a network between scientists (academia) and industry that will foster the exchange of ideas and challenges.
- **Data-intensive business experts:** These are the specialists that develop and exploit techniques, processes, tools and methods to develop applications that turn data into value. In addition to technical expertise, data-intensive business experts need to understand the domain and the business of the organisations. This means they need to bring in domain knowledge and are thus working at the intersection of technology, application domains and business. In a sense, they thereby constitute the link between technology experts and business analysts. Data-intensive business experts will foster the development of big data applications from an “art” into a disciplined engineering approach. They will thereby allow the structured and planned development and delivery of customer-specific big data solutions, starting from a clear understanding of the domain, as well as the customer’s and user’s needs and requirements.

In order to successfully meet the skills challenge, it is critical that industry works with both higher education institutes and education providers to identify the skill requirements that can be addressed with the establishment of:

- New educational programmes based on interdisciplinary curricula with a clear focus on high-impact application domains.
- Professional courses to educate and re-skill/up-skill the current workforce with the specialised skillsets needed to be data-intensive engineers, data scientists and

data-intensive business experts. These courses will stimulate lifelong learning in the domain of data and in adopting new data-related skills.

- Foundational modules in Data Science, Statistical Techniques, and Data Management within related disciplines such as law and the humanities.
- A network between scientists (academia) and industry that leverages innovation spaces to foster the exchange of ideas and challenges.
- Datasets and infrastructure resources, provided by industry, that enhance the industrial relevance of courses.

1.2 Formal and Non-formal Learning

To provide a more enhanced educational support to tackle the skills challenges defined above, both formal³ and non-formal⁴ learning can be considered as they contribute to the lifelong learning of data scientists – the continual training of data scientists throughout their careers. While formal systems are often focused on initial training, a lifelong learning system must include a variety of formal and non-formal learning together. This is necessary to meet the individual's need for continuous and varied renewal of knowledge and the industry's need for a constantly changing array of knowledge and competences.

Here, we will consider non-formal education to include any organised training activity outside of formal education (undergraduate or graduate university degrees). Non-formal training includes both e-learning and traditional professional training. These courses can be of widely different durations and include training provided by employers, traditional educational institutions and other third parties.

Therefore, in Data Science non-formal education plays a crucial role and complements formal training, by allowing practitioners to up-skill and re-skill to adapt to new Data Science requirements.

³“Education that is institutionalised, intentional and planned through public organisations and recognised private bodies and, in their totality, make up the formal education system of a country. Formal education programmes are thus recognised as such by the relevant national educational authorities (UNESCO)” (<http://uis.unesco.org/en/glossary-term/formal-education>).

⁴“Education that is institutionalised, intentional and planned by an education provider. The defining characteristic of non-formal education is that it is an addition, alternative and/or a complement to formal education within the process of the lifelong learning of individuals (UNESCO)” (<https://unevoc.unesco.org/home/TVETipedia+Glossary/filt=all/id=185>).

2 Key Projects on Data Skills

Previous EU projects have already worked on Data Science skills. The two main initiatives in this context have been the EDISON project and the EDSA project analysed below.

2.1 *The EDISON Project*

The EDISON project defined the EDISON Data Science Framework (EDSF). The definition of the whole framework was based on the results of extensive surveys. Its four components are as follows:

- The Data Science Competence Framework (CF-DS) provides the definition of Competences for Data Science according to the e-CF 3.0. These competences are represented in five competence groups:
 - Data Science Analytics
 - Data Science Engineering
 - Domain Knowledge and Expertise
 - Data Management
 - Research Methods

For each of these groups, several component competences are given at three levels of proficiency (associate, professional, expert). For example, for the Data Science Analytics competence group, six component competences have been defined. Two of them are:

- DSDA01: Effectively use a variety of Data Analytics techniques, such as machine learning (including supervised, unsupervised, semi-supervised learning), data mining and prescriptive and predictive analytics for complex data analysis through the whole data lifecycle.
 - DSDA02: Apply designated quantitative techniques, including statistics, time series analysis, optimisation and simulation to deploy appropriate models for analysis and prediction.
- The Data Science Body of Knowledge (DS-BoK) provides, for each competence group, the identification of knowledge areas and knowledge units.
- The Data Science Model Curriculum (MC-DS) provides, for each competence group and individual competence, the learning outcomes required to obtain the competence. These outcomes are given for each of the three levels of proficiency.
- The Data Science Professional Profiles (DSPP) provides a listing of 22 professional profiles in Data Science grouped in 6 categories: managers, professional (data handling/management), professional (database), technicians and associate professionals, and clerical support workers (general and keyboard workers). The

framework also identifies the relevance of each competence group for each professional profile.

2.2 *The EDSA Project*

One of the aims of the EDSA project was to propose a curriculum for Data Science. That curriculum was based upon what the EDSA consortium identified as core Data Science knowledge rather than the skills that might be needed for a particular job in Data Science. This curriculum was validated through various surveys.

The EDSA curriculum consists of 15 core Data Science topics. Each of these topics has learning objectives, descriptions as well as resources and materials, which were also produced as part of the EDSA project. The 15 topics that make up the core EDSA curriculum were divided into 4 stages: Foundations, Storage and Processing, Analysis, and Interpretation and Use. Table 1 shows an example of the documentation provided by EDSA for a topic, in this case for the Data-Intensive Computing Topic.

3 The Need for the Recognition of Data Skills

With the development of new technologies and the digital transformation of our economy, the labour market has also evolved. Nowadays, applicants for a job are no longer asked to submit a traditional paper résumé; this information is presented digitally, that is, recruiters and headhunters search the Internet (on an international level) for candidates who have the required skills, and some assessment of candidates can be done online. Moreover, the labour market is constantly evolving, and the required skills and qualifications change rapidly over time. Adequately adapting to these changes is essential for the success of employers, learning institutions and governmental agencies related to education. In this section, we will discuss mechanisms for recognising skills in the EU, with a focus on the internationalisation, digitalisation and flexibility of these credentials and their application to Data Science. We begin with a brief review of the main challenges we hope to address.

How Can We Standardise Credentials Throughout Europe?

Although political institutions in the EU have strived to coordinate and standardise diplomas and other forms of credentialing in higher education, the variety of educational systems in the EU and the lack of an adequate system to recognise learning and skills have contributed to great differences in the economic and social outcomes of the member states. The many different educational and training systems in Europe make it difficult for employers to assess the knowledge of potential employees. There is no automatic EU-wide recognition of academic diplomas; students can only obtain a “statement of comparability” of their university degree. The statement of comparability details how the student’s diploma compares to the

Table 1 Material developed by EDSA for a data-intensive computing-related course^a**Scalable machine learning and deep learning**

The course studies the fundamentals of distributed machine learning algorithms and the fundamentals of deep learning. It covers the basics of machine learning and introduces techniques and systems that enable machine learning algorithms to be efficiently parallelised. The course complements courses in machine learning and distributed systems, with a focus on both deep learning and the intersection between distributed systems and machine learning. The course prepares the students for master's projects and Ph.D. studies in the area of Data Science and distributed computing.

The main objective of this course is to provide the students with a solid foundation for understanding large-scale machine learning algorithms, in particular deep learning and their application areas.

Intended learning outcomes

Upon successful completion of the course, the student will:

Be able to re-implement a classical machine learning algorithm as a scalable machine learning algorithm

Be able to design and train a layered neural network system

Syllabus and topic descriptions

Main topics:

Machine learning (ML) principles

Using scalable data analytics frameworks to parallelise machine learning algorithms

Distributed linear regression

Distributed logistic regression

Distributed principal component analysis

Linear algebra, probability theory and numerical computation

Convolutional networks

Sequence modelling: recurrent and recursive nets

Applications of deep learning

Detailed content

Introduction:

Brief history and application examples of deep learning and large-scale machine learning: at Google and in industry, ML background, brief overview of deep learning, understanding deep learning systems, linear algebra review, probability theory review

Distributed ML and linear regression:

Supervised and unsupervised learning, ML pipeline, classification pipeline, linear regression, distributed ML, computational complexity

Gradient descent and Spark ML:

Optimisation theory review, gradient descent for least squares regression, the gradient, large-scale ML pipelines, feature extraction, feature hashing, Apache Spark and Spark ML

Logistic regression and classification:

Probabilistic interpretation, multinomial logistic classification, classification example in Tensorflow, quick look in Tensorflow

Feedforward neural nets and backprop:

Numerical stability, neural networks, feedforward neural networks, feedforward phase, backpropagation

Regularisation and debugging:

A flow of deep learning, techniques for training deep learning nets, regularisation, why does deep learning work?

.....

Existing courses:

Scalable machine learning and deep learning at the Royal Institute of Technology, KTH

Scalable machine learning, edX, <https://courses.edx.org/courses/BerkeleyX/CS190.1x/1T2015/info>

(continued)

Table 1 (continued)

Distributed machine learning with Apache Spark, edX <https://www.edx.org/course/distributed-machine-learning-apache-uc-berkeleyx-cs120x>

Deep learning systems, University of Washington, <http://dlsys.cs.washington.edu/>

Scalable machine learning, University of Berkeley, <https://bcourses.berkeley.edu/courses/1413454/>

Existing materials:

Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep learning, MIT Press

Spark ML pipelines, <http://spark.apache.org/docs/latest/ml-pipeline.html>

Spark ML overview, <https://www.infoq.com/articles/apache-sparkml-data-pipelines>

^a<https://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2017-P-D23-FINAL.pdf>

diplomas of another EU country.⁵ Something similar happens with the recognition of professional qualifications as the mobility of Europeans between member states of the EU often requires the full recognition of their professional qualifications (training and professional experience). This is accomplished through an established procedure in each European country.⁶

Directives 2005/36/EC and 2013/55/UE on the recognition of professional qualifications establish guidelines that allow professionals to work in another EU country different from the one where they obtained their professional qualification, on the basis of a declaration.

These directives provide three systems of recognition:

- Automatic recognition – for professions with harmonised minimum training conditions, i.e. nurses, midwives, doctors (general practitioners and specialists), dental practitioners, pharmacists, architects and veterinary surgeons
- General system – for other regulated professions such as teachers, translators and real estate agents
- Recognition on the basis of professional experience – for certain service providers such as carpenters, upholsterers, beauticians, etc.

Additionally, the European professional card (EPC) has been available since 18 January 2016 for five professional areas (general care nurses, physiotherapists, pharmacists, real estate agents and mountain guides). It is an electronic certificate issued via the first EU-wide fully online process for the recognition of qualifications. Unfortunately, these existing mechanisms do not easily accommodate many professions including that of Data Science.

⁵http://europa.eu/youreurope/citizens/education/university/recognition/index_en.htm

⁶http://europa.eu/youreurope/citizens/work/professional-qualifications/recognition-of-professional-qualifications/index_en.htm

How Can Data Science Credentials Be Digital, Verifiable, Granular and Quickly Evolving?

Traditionally, skills and credentials were conveyed via a résumé on paper and other paper-based credentials. Nowadays, this information can be shared via the Internet in web pages, on social media and in many other forms. The digitalisation of credentials not only allows easier access but also offers new possibilities like:

- The online verification of the validity of the credentials
- Greater granularity in the definition of the credentials
- The expiration of credentials requiring their periodic renewal, which could take into account changes in the demands for skills
- Access to the evidence used in the awarding of credentials

Future schemes for the recognition of skills need to adapt to and accommodate these new demands.

How Can Non-formal Learning in Data Science Be Recognised?

The educational landscape is rapidly changing. The great emphasis which was previously placed on formal university training is slowly eroding. The role of both informal and non-formal learning is increasing, and skills recognition schemes need to contemplate these changes. The BDVe⁷ proposed BDV Data Science Badges as a skills recognition tool for formal education and BDV Data Science Labels for non-formal education.

As mentioned, our work on data skills recognition aimed to address these challenges. To do so, the needs of the different stakeholders participating in the process, formal and non-formal education providers, as well as students and industry also play a very relevant role.

4 BDV Data Science Badges for Formal Education

4.1 Methodology

The recognition strategy proposed by the BDVe for formal education science is based on the use of Open Badges.

Open Badges are images that can be included in a curriculum, uploaded to platforms like LinkedIn and shared on social media. They contain metadata to allow:

- The online verification of their authenticity and ownership
- Reviewing information regarding requirements to receive the badge
- Access to details regarding the organisation who issued the badge
- Viewing when the badge was issued and when it expires
- Downloading evidence of the acquisition of skills

⁷<https://www.big-data-value.eu/>

Table 2 Key aspects of the BDV badge recognition schema

Who defines the requirements for the badge?	A collection of experts, including representatives from industry and academia, establish both the types and the requirements of the badges included in the programme. They also define the process for applying to issue badges. This group of experts will periodically meet to review the programme. Based upon progress reports, they can propose changes and improvements.
Who issues the badge?	Interested institutions/educators can apply to issue a badge. This application includes submitting evidence that shows that they provide their students with the skills required by the badge.
Who decides if an institution can issue badges?	A group of experts defined by representatives from industry/academia evaluate applications received to issue badges. Reviewers are assigned applications to assess and decide whether the applicant programme meets the established standards to issue badges. Applications can be rejected, accepted conditionally for 1 year or accepted for 4 years.
How does an institution issue a badge?	Students in a Data Science programme authorised to issue BDV Badges acquire and demonstrate their Data Science skills through their studies. Students in the programme can submit an application to their programme to receive a badge. The programme reviews badge applications, and if the applicant has met the requirements of the badge, then it issues the student that badge. Badges are individualised and contain metadata including the requirements to earn the badge and evidence of the student’s achievements.
How are the badges displayed?	Students can display their badges online: in their CV, on social networks, etc. Interested employers can verify that a badge is valid and use its metadata to access relevant information regarding the earners of a badge.

Fig. 1 BDV Badges – application and issuing process

The key aspects of the Open Badges recognition schema proposed by the BDVe are detailed in Table 2.

Figure 1 represents graphically the BDV Badge programme proposed. The badges will be designed by a committee of experts from both industry and academia. Institutions will be responsible for issuing the badges (once a review process has been successfully passed) to their students, and they will be able to display their badges online, so employers will have access to the content and thereby verify the Data Science knowledge of the students.

4.2 Badge Overview

Based on the EDISON framework, we initially proposed the creation of one group of badges for each competence group, with each group of badges having three levels of proficiency (basic, intermediate and expert). To make the proposal more accessible to a wider audience, we chose to use the term “required skills” in place of “learning outcomes”.

Thus, the following is the initial collection of BDV Data Science Badges:

- Data Science Analytics Badge
- Data Engineering Badge
- Data Science Management Badge
- Business Process Management Badge
- Data Science Research Method and Project Management Badge

With the aim of verifying the comprehensibility and utility of this proposal, we conducted an evaluation process which involved both industry and academia. In order to get detailed feedback and make this assessment process effective, in the initial stage, we focused only on the Data Science Analytics Badge. We obtained feedback from 12 companies from industry. The aims were to obtain information about the relevance of the different required skills to their hiring practices and to ensure that the descriptions of the required skills were easy to understand. Fifteen universities were contacted to participate in several rounds of the evaluation. The aim was to get feedback about the review process (specifically the kinds of material to be requested of badge applicants) and about the requirements of the badge. Additionally, the members of the Big Data Value Association (BDVA) Skills and Education Task Force provided feedback on the initial version of the badges as well as on the comments gathered from industry and academia.

Based on the results of the assessment process, the three levels of proficiency (basic, intermediate and expert) were replaced by two levels (academic and professional) having the same required skills. The academic level requires knowledge and training which can be acquired in an academic context, while the professional level requires real professional practice.

Table 3 BDV Data Science Analytics Badge skills

Data Science Analytics Badge v1.0
Required skills
DSA.1. Identify existing requirements to choose and execute the most appropriate data discovery techniques to solve a problem depending on the nature of the data and the goals to be achieved
DSA.2. Select the most appropriate techniques to understand and prepare data prior to modelling to deliver insights
DSA.3. Assess, adapt and combine data sources to improve analytics
DSA.4. Use the most appropriate metrics to evaluate and validate results, proposing new metrics for new applications if required
DSA.5. Design and evaluate analysis tools to discover new relations in order to improve decision-making
DSA.6. Use visualisation techniques to improve the presentation of the results of a Data Science project in any of its phases

Fig. 2 Data Science Analytics Badges with academic and professional levels (v1.0)

The description of some of the requirements was also modified, providing the final version of the BDV Data Science Analytics Badge shown in Table 3. Images of both the academic and professional badges are shown in Fig. 2.

Figure 3 shows how the Data Science Analytics Badge of one student could be visualised.

4.3 Platform

As mentioned, the proposed recognition framework works with Open Badges. In this section, we address the badge-issuing platform selected. First, we will consider some details of v2.0 of the Open Badge Standard.

The most recent version of the technical specifications for Open Badges (v2.0) was published on 12 April 2018.⁸ An Open Badge must contain three pieces of linked metadata in JSON-LD:

⁸<https://www.imsglobal.org/sites/default/files/Badges/OBv2p0Final/index.html>

academic setting. This badge includes skills related to Machine Learning, Data Mining, Statistical Methods and Algorithms.

This badge was issued by an educational program which meets the educational standards established by the BDV Data Science Skills Recognition Program.

Required Skills

To earn this badge, a student must acquire and demonstrate their competence in the following data science skills in an academic setting:

- DSA.1 Identify existing requirements to choose and execute the most appropriate data discovery techniques to solve a problem depending on the nature of the data and the goals to be achieved.
- DSA.2 Select the most appropriate techniques to understand and prepare data prior to modeling to deliver insights.
- DSA.3 Assess, adapt, and combine data sources to improve analytics.
- DSA.4 Use the most appropriate metrics to evaluate and validate results, proposing new metrics for new applications if required.
- DSA.5 Design and evaluate analysis tools to discover new relations in y order to improve decision-making.
- DSA.6 Use visualization techniques to improve the presentation of the results of a data science project in any of its phases.

Credits

This work was partially funded by the project ID: 732630 under: H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies.

The description, requirements and types of badges used in the BDV Data Science Skills Recognition Program were based upon the Edison EDSF and include changes suggested by the data science community.

(Additional Criteria can be added here.)

EVIDENCE

sampleEvidenceFile.pdf

<https://openbadgepassport.com/file>

BDV DATA SCIENCE ANALYTICS - ACADEMIC LEVEL

Issued by:

Created by: Big Data Value eCcosystem Project

Issued on: 2020

Expires on: 2024 (days)

Earned:

Data Science Analytics is use of techniques to gather, organize and process a collection of raw data to discover new and useful information and support the making of conclusions and decisions based upon that data. Earners of this badge have demonstrated their ability to apply Data Science Analytics techniques in an academic setting.

[Check this badge...](#)

CRITERIA

[Open criteria...](#)

BDV Data Science Analytics Badge v1.0 - Academic Level

General Description

This badge is part of the BDV Data Science Skills Recognition Program organized in a collaboration between the BDV Project and the BDVA Skills Task Force.

The BDV Data Science Analytics badge serves to recognize an academic level of understanding in Data Analytics. Earners of this badge have shown that they can use appropriate techniques to discover new relations in a collection of data and thereby give insight into that data set to support decision making in an

Fig. 3 Data Science Analytics Badge of one student

Table 4 Requirements defined for platforms issuing BDV Data Science Badges

Functional requirements	FR1 – Badges are designed by a central body and then shared only with approved issuers (who cannot modify the shared badges) FR2 – Approved institutions can issue instances of the badges without the intervention of the central body FR3 – Approved institutions can only access the personal information of their own students FR4 – The hosting platform offers a badge application process for students at an approved institution. This application includes the upload of evidence FR5 – The hosting platform provides space to store student evidence files FR6 – The central body can view statistics regarding all issued BDV Badges (but not individual personal data) FR7 – The central body can modify or revoke existing badges as well as design new badges FR8 – The central body can revoke permission to issue a badge
Platform requirements	TR1 – Open Badge 2.0 certified https://www.imsglobal.org/cc/statuschart/openbadges https://www.imsglobal.org/sites/default/files/Badges/OBv2p0Final/cert/index.html TR2 – Meets requirements of GDPR TR3 – Company and/or infrastructure located in Europe TR4 – Perpetual badge hosting agreement (badges are not deleted if the issuing contract is discontinued) TR5 – Issued badges hashed emails and are signed by each approved institution

- Issuer Profile – This resource describes who issued the badge. Usually, one profile is created for each organisation, but it is possible to have multiple issuers (e.g. different departments within the same university).
- BadgeClass – This resource contains information regarding the badge itself and must include information such as the issuer, a description of the badge and the criteria used to issue the badge.
- Assertion – This represents one particular badge (a BadgeClass) issued to one particular person. People can be identified in a number of ways (telephone number, URL), but many badge platforms only accept email identifiers.

From this standard and other considerations specific to the BDV Badge programme, we developed two lists of requirements for the badge-issuing platform. These are summarised in Table 4.

Finally, all Open Badge v2.0-certified badge-issuing platforms were evaluated according to the previous requirements. The issuing platforms assessed were those listed at <https://www.imsglobal.org/cc/statuschart/openbadges> on 1 February 2019. From them, one that is based in the EU was chosen, which also fulfils the previous criteria.

Fig. 4 Example of the application of the UK guidelines for Front of Pack Labels (Source: (Department of Health 2016)). (Public sector information licensed under the Open Government Licence v3.0.)

5 BDV Data Science Labels for Non-formal Education

5.1 Methodology

In recent years the offerings of non-formal training in Data Science in the form of online courses, massive open online courses, in-company training, etc., from both official academic institutions and other non-academic institutions, have greatly increased.

Though the needs of stakeholders in the Data Science ecosystem when considering non-formal education are similar to those of formal education, there are a few issues worth highlighting:

- Students interested in Data Science training can quickly find a huge variety of options, and therefore face difficulties when trying to pick from this overwhelming supply. Which courses are more highly valued by the industry and what is the right course for their experience and expectations are just a couple of issues that arise.
- Employers that need to evaluate non-formal training also face the problem of how to compare the wide variety of different types of courses. For example, how rigorous are the different programmes in terms of duration, quality, evaluation of the students, identity verification during assessment activities, etc.?
- Educators offering these courses also face difficulties related, for example, to how to stand out from other courses, that is, how to clearly communicate their offer, attract students, ensure the quality of their training, etc.

In other contexts, standardised labelling systems are used to systematically provide information to help to characterise and compare different products in the same category. For example, Fig. 4 shows the UK guidelines for Front of Pack Labels, which could be used to, for example, compare different kinds of breakfast cereal.

With this idea of a standardised nutritional labelling system as an inspiration, a labelling system for characterising non-formal training in Data Science was proposed. The aim is to provide a labelling system to highlight educational value, which

can be useful for the different stakeholders involved in the process (students, industry and course providers).

To develop this proposal, we have followed a process similar to that used for formal training, in the sense of obtaining a consensus from the stakeholders involved in the process about the content of the labels. For that aim, we have gathered feedback through different activities, such as an online seminar for BDVA members, internal feedback collected from BDVe members and feedback from course providers. This process has led us to define the content of the criteria to be included in the label, as we will explain in the next section.

5.2 *Label Overview*

The labelling system for non-formal training aims to promote and encourage the recognition of Data Science skills acquired through non-formal training. This new system is designed to achieve the following goals:

- Increase transparency – The labelling system should provide an easy-to-understand representation of the most relevant aspects of a Data Science course. The labels should assist students in the filtering of the vast offer to identify the courses best suited for their needs. The labels should also help employers to assess the relevancy of a course to a position. Lastly, the labels should encourage educational providers to readily provide the information which students and employers most need.
- Simplify the comparison – By standardising the presentation of the essential features of courses, the side-by-side comparison of different courses should be easier for everyone.
- Encourage practices which contribute to quality – By highlighting in the label the principal aspects of non-formal education which contribute to quality training, both students and employers can more easily assess the quality of a course. Also, training providers will be encouraged to adopt practices which increase the quality of their offerings.
- Ensure the alignment of training with industrial needs – The label should encourage educational providers to contrast their training with current industrial needs. This should help students and employers assess whether the training meets their needs. It should also promote changes in the content offered by educators to meet the needs of the industry.

From initial interviews with educational providers and employers, a list of criteria which could be used as the basis for the label has been identified.

Table 5 contains an example of these criteria for an imaginary online course containing the preliminary set of criteria which we are proposing. The appropriate graphical design will need to be produced. Then, the corresponding educational label can be provided along with the course information. Note that this labelling system

Table 5 Preliminary criteria of an online course for BDV Data Science Labels

Course title: introduction to data analytics		
C1 – In what language or languages is the course offered	English	
C2 – Total cost of the course	1000 euros	
C3 – Main audience of the course (programmers, data analytics, manager)	Programmers	
C4 – Expertise level expected for new students (beginner, intermediate, advanced)	Beginners	
C5 – What kind of training is provided; specify the hours of training in each of the following groups	In person: 0 h Online with the possibility of interacting with the trainer: 30 h Online with static (non-interactive) content: 120 h	
C6 – What kind of assessment is performed (no testing, in person, online testing/with or without identification verification)	Online testing without identification verification	
C7 – What is the content of the course – specify the number of hours of training and homework dedicated to acquiring each of the Data Science analytics skills	DSA.1. Identify existing requirements to choose and execute the most appropriate data discovery technique to solve a problem depending on the nature of the data and the goals to be achieved.	20
	DSA.2. Select the most appropriate techniques to understand and prepare data prior to modelling deliver insights.	50
	DSA.3. Asses, adapt, and combine data sources to improve analytics.	70
	DSA.4. Use the most appropriate metrics to evaluate and validate results, proposing new metrics for new applications if required.	0
	DSA.5. Design and evaluate analysis tools to discover new relations in order to improve decision-making.	0
	DSA.6. Use visualisation techniques to improve the presentation of the results of a data science project in any of its phases.	10

does not require any platform to be implemented, as it consists only of an image with the corresponding educational information.

6 Pilot and Use Case

To showcase how the skills recognition methodologies proposed above can be applied to bridge industry with academia, the BDVe conducted a pilot of the Data Science Analytics Badge with the results displayed on the EduHub, which is a platform that contains information about educational programmes as well as their offered BDV Badges.

6.1 BDV Badge Pilot

A pilot of the entire Data Science Analytics Badge application process was conducted in order to validate the process to be followed by the universities applying to issue the badge, as well as the review process. Institutions aiming to issue the badge must provide evidence to show that their students have acquired the corresponding skills. Table 6 shows for the first skill of the Data Science Analytics Badge the information to be provided, so reviewers can check the degree to which this skill is acquired by the students.

Each application form must be reviewed by two reviewers. A final decision is made if the recommendations of the two reviewers coincide. If the two reviewers are not able to reach a consensus, a third reviewer is asked to participate in the process. Each reviewer provides recommendations. The reviewer can recommend that the applicant programme be able to issue the badge for 4 years, that the badge-issuing period of the programme be limited, and that the programme will be required to resubmit another application to issue badges in the following year or that the institution is not able to issue the badge as major drawbacks have been found regarding the acquisition of the required skills.

Reviewers participating in this process must agree to the Code of Conduct for Badge Issuing Application Reviewers, available at <https://www.big-data-value.eu/>

Table 6 Extract of the application form with information about DSA.1

DSA.1. Identify existing requirements to choose and execute the most appropriate data discovery techniques to solve a problem depending on the nature of the data and the goals to be achieved					
Name of the course or activity	Optional/ mandatory	The average number of hours that students spend in the course/ activity acquiring the skill	Evidence used to evaluate the acquisition of the skill (exam, assignment,...)	Name of the evidence file	The minimum level required to demonstrate the acquisition of the skill
Course 1		X hours	Exam 1, questions 2-3		
			Assignment 4		
			...		
			Add as many rows as needed		
Course 2		Y hours			
Add as many rows as needed					
Minimum number of hours that the student dedicates to acquiring the skill					
Comments (optional)					

skills/skills-recognition-program/call-for-academic-level-data-science-analytics-badge-issuers/?et_fb=1&PageSpeed=off.

The pilot resulted in two institutions being able to issue the Data Science Analytics Badge: one application was accepted, and another application was accepted with comments regarding improvements that could be submitted within the following year.

The institutions and programmes that were granted the right to issue the badge were:

- M.Sc. in Big Data Analytics, Universitat Politècnica de València (Spain)
- Data and Web Science M.Sc. Programme, Aristotle University of Thessaloniki (Greece)

6.2 *BDV Education Hub*

The BDV Education Hub (EduHub) is designed to help users find the right programme of study or special training course among the many education and training opportunities in the big data area.

Accessible via <http://bigdataprofessional.eu/>, the EduHub is an online platform that offers a living repository for knowledge about European educational offerings related to big data. The EduHub covers programmes of all areas of the BDV Reference Model (see Chap. 3), including data processing, data management, data analytics, data visualisation and data protection.

The EduHub inventories European master's and Ph.D. programmes, as well as European training programmes (both online and on-site) in the field of Big Data and Data-Driven AI. At the time of writing, the EduHub included over 360 European educational offerings (217 European M.Sc. programmes, 12 European Ph.D. programmes as well as 133 professional trainings). The programmes are carefully selected to reflect their focus on BDV, thereby helping interested students and professionals to find the matching skilling and up-skilling offerings. While the master's programmes are targeted for undergraduate students and the Ph.D. programmes for graduate students, the professional training is targeted for professionals looking for reconversion towards Data Science, as well as employees/employers looking for up-skilling opportunities.

The EduHub reflects the intention of the BDVA to promote the education of European citizens in this important key area (Zillner et al. 2017). The European Digital Skills and Jobs Coalition recognises these efforts and lists the EduHub as part of the European Digital Skills and Jobs Coalition's Pledge Viewer, a tool for creating, viewing and managing pledges reflecting an organisation's commitment to equip Europeans with the skills they need for life and work in the digital age.

The EduHub also serves as a platform to advertise and make visible the BDV Badges that are awarded to university programmes (see above). Figure 5 shows an example of how the badges are shown together with the key information about the university programme.

Fig. 5 Screenshot of the BDV EduHub showing awarded BDV Badges

7 Conclusion

Given the considerable amount of attention drawn lately to fields like Big Data, Data Analytics and Data Science, there is an ever-growing need for skilled data scientists by the industry. However, in order to create a vibrant data-driven economy in Europe, it is vital to find ways to bridge the gap between the industrial needs and skills offered by formal or non-formal education. This chapter explored how challenging this goal is as the current knowledge-driven approaches need to be transformed into experience-driven ones via re-definition of the roles and skills of data professionals. This could be achieved by the collaboration of industry and educational providers (formal or non-formal) to define the necessary skills requirements that need to be obtained by future data professionals. The chapter explored steps in that direction that involve the creation of an education platform and a skills recognition programme. Specifically, the EduHub was described, which is a platform that provides access to Data Science and Data Engineering programmes offered by European universities as well as on-site/online professional training programmes, and its aim is to facilitate knowledge exchange on educational programmes and meet current industrial needs. Additionally, the BDV Data Science Badge and Label recognition programmes were analysed for skills acquired by formal and

non-formal training, respectively. The aim of the programmes is not only to provide a form of skills recognition but also to align the current industrial needs with the Data Science curricula and skills. Finally, a more practical view was given on how the EduHub and the skills recognition programmes can work in isolation as well as together by demonstrating a pilot on the Data Science Analytics Badge that is currently issued by two universities, and how the badges as well as the educational programmes to which they are issued can be accessed in the EduHub.

Acknowledgements Research leading to these results received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 732630 (BDVe). This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

References

- Department of Health. (2016). Guide to creating a Front of Pack (FoP) nutrition label for pre-packed products sold through retail outlets. In *Food standards agency*.
- Zillner, S., Curry, E., Metzger, A., Auer, S., & Seidl, R. (Eds.). (2017). *European big data value strategic research & innovation agenda*. Retrieved from Big Data Value Association website: www.bdva.eu

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.