Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge

Yuri Demchenko
University of Amsterdam, The Netherlands
y.demchenko@uva.nl

Luca Comminiello, Gianluca Reali
University of Perugia, Italy
lucapio94@gmail.com. gianluca.reali@unipg.it

Abstract— Importance of Data Science education and training is growing with the emergence of data driven technologies and organisational culture that intend to derive actionable value for improving research process or enterprise business using variety of enterprise data and widely available open and social media data. Modern data driven research and industry require new types of specialists that are capable to support all stages of the data lifecycle from data production and input to data processing and actionable results delivery, visualisation and reporting, which can be jointly defined as the Data Science professions family. The education and training of Data Scientists requires multi-disciplinary approach combining wide view of the Data Science and Analytics foundation with deep practical knowledge in domain specific areas. In modern conditions with the fast technology change and strong skills demand, the Data Science education and training should be customizable and delivered in multiple form, also providing sufficient data labs facilities for practical training. This paper discusses approach to building customizable Data Science curriculum for different types of learners based on using the ontology of the EDISON Data Science Framework (EDSF) developed in the EU funded Project EDISON and widely used by universities and professional training organisations.

Keywords—Data Science, Data Scientist Professional, Data Science Ontology, Big Data, EDISON Data Science Framework (EDSF), Data Science Competences Framework, Data Science Body of Knowledge, Data Science Model Curriculum.

I. INTRODUCTION

Sustainable development of the modern data driven economy requires new type of data driven and Data Science and Analytics enabled competences and workplace skills. Fast technology change and new skills demand requires re-thinking and re-designing both traditional educational models and existing courses to reflect multi-disciplinary nature of Data Science and its application domains. At present time most of the existing university curricula and training programs cover a limited set of competences and knowledge areas of what is required for multiple Data Science and general data management professional profiles and organisational roles enacted by research and industry. In conditions of continuous technology development and shortened technology change cycle, Data Science education requires effective combination of theoretical, practical and workplace skills.

The EDISON Data Science Framework (EDSF), which is the products of the EU funded EDSION Project [1, 2], provides a basis for designing customised Data Science curricula based on required competences and intended learning outcome that can be targeted for specific professional profiles and individually constructed based in the learner/trainee competences benchmarking. The detailed definition of the practical Data Science skills in EDSF can provide also recommendations for building effective educational environment combining educational or training components and practical hands on experience with virtual and data labs.

The paper refers to the previous authors' works that researched new approaches to building effective curricula in Cloud Computing, Big Data and Data Science [7, 8, 9, 10] and based on long time practical experience in developing both online and campus based education and training courses. The paper also refers to the experience of the community of practitioners that currently maintains and develops the EDSF which Release 3 has been published in December 2018

The paper is organized as follows. Section II refers to resent studies indicating demand for Data Science related specialists and describes the challenges and specifics in professional education and training of the Data Scientists. Section III describes the EDSF and its components. Sections IV describes the EDSF data model and its ontology definition. Section V discusses how the EDSF ontology can be used for customised curriculum design and provides example of the suggested curriculum structure for two Data Science profiles Data Scientist and Data Steward. Section VII provides summary and suggestions for future work.

II. DEMAND FOR DATA SCIENCE COMPETENCES AND CUSTOMISABLE CURRICULUM

Industry digitalisation and wide use of data driven technologies facilitate demand for Data Science and Analytics enabled professions, this trend is confirmed by multiple European and global market studies. The IDG report 2017 [11] provided deep analysis of the European data market and growing demand for data workers and estimated

the total number of data workers to grow from 6.1 mln in 2016 to 10.4 million in 2020 where the data related skills gap is estimated as 769,000 or 9.8% (2020). Addressing this demand and gap is becoming critical for European economy and challenge for universities.

Business Higher Education Forum (BHEF) has published in 2017 two important reports in cooperation with PriceWaterhouseCoopers, IBM and Burning Glass Technologies [12, 13] that studied Data Science and Analytics (DSA) job market in US and identified a number of actions to be addressed by business, higher education, government and professional organisations to address increased demand and growing gap in demand and supply of skilled DSA workforce capable to effectively work in modern data driven economy.

Recent OECD report [14] confirms the urgent need to address data and general digital skills for all types of workforce and economy sectors. An effective professional education should provide a foundation for future continuous professional self-development and mastering new emerging technologies, that can provide a basis for the life-long learning model adoption. Flexibility in providing education and training curricula and course is a key adopting future skills management and capacity building models.

III. EDISON Data Science Framework (EDSF)

Designing future effective Data Science educational environment will require developing and widely accepted a general framework for Data Science education, curriculum design and competences management that can be based on the proposed

The EDISON Data Science Framework (EDSF), that is the product of the EDISON Project, provides a basis for Data Science education and training, curriculum design and competences management that can customised for specific organisational roles or individual needs. EDSF can be also used for professional certification and career transferability.

Figure 1 below illustrates the main EDSF components and their inter-relations:

- CF-DS – Data Science Competence Framework [3]
- DS-BoK – Data Science Body of Knowledge [4]
- MC-DS – Data Science Model Curriculum [5]
- DSPP - Data Science Professional profiles and occupations taxonomy [6]
- Data Science Taxonomy and Scientific Disciplines Classification

The proposed framework provides the basis for other components and services of the Data Science professional environment such as

- Data Science Education Environment (DSEE) and Virtual Data Labs (that can be cloud based and provisioned on demand)
- Directory of Education and Training Materials
- Data Science Community Portal (CP) that can provide information and community support services, such as individual competences benchmarking and personalized educational path building.

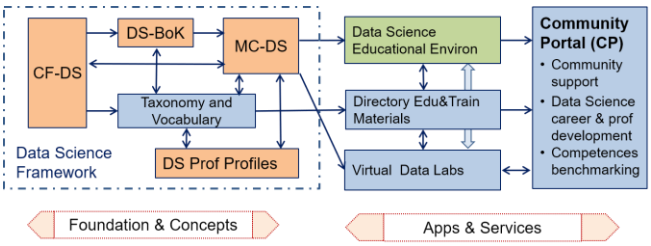


Figure 1 EDISON Data Science Framework components and Data Science Educational environment.

The CF-DS provides the overall basis for the whole framework. The CF-DS includes the core competences required for the successful work of a Data Scientist in different work environments in industry and in research and through the whole career path. The CF-DS is defined using the same approach as e-CFv3.0 [15] (competences defined as abilities supported by knowledge and skills with applied proficiency levels) but have competence structured according to the major identified functional groups (as explained below).

The following core CF-DS competence and skills groups have been identified (refer to CF-DS specification [2] for details):

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others) (DSDA)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools) (DSENG)
- Data Management and Governance (including data stewardship, curation, and preservation) (DSDM)
- Research Methods and Project Methods (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Data Science competences must be supported by knowledge that are defined primarily by education and training, and skills that are defined by work experience correspondingly. The CF-DS defines two types of skills (refer to CF-DS [2] for full definition of the identified knowledge and skills groups):

- Skills Type A which are related to the professional experience and major competences, and
- Skills Type B that are related to wide range of practical computational skills including using programming languages, development environment and cloud based platforms.

CF-DS defines workplace skills, also referred to as “soft” skills or professional attitude skills, which are becoming increasing important in modern data driven and future Industry 4.0 economy. This includes two groups of skills that are increasingly demanded by employers: Data Science Professional skills (Thinking and acting like Data Scientist), and so called the 21st Century skills [??] that comprise a set of workplace skills that include critical thinking,

communication, collaboration, organizational awareness, ethics, and others.

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK is based on ACM/IEEE Classification Computer Science (CCS2012) [16], incorporates best practices in defining domain specific BoK's and provides reference to existing related BoK's. It also includes proposed new KA to incorporate new technologies and scientific subjects required for consistent Data Science education and training.

The MC-DS [4] is built based on DS-BoK and linked to CF-DS where Learning Outcomes are defined based on CF-DS competences (specifically skills type A), and Learning Units are mapped to Knowledge Units in DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles. Practical curriculum should be supported by corresponding educational environment for hands on labs and educational projects development.

The formal DS-BoK and MC-DS definition creates a basis for Data Science educational and training programmes compatibility and consequently Data Science related competences and skills transferability.

IV. EDSF Toolkit and Practical Uses of EDSF

EDSF was developed with the view of multiple practical uses for the whole range of tasks faced by universities, professional training organisations, companies and certification bodies related to Data Science education, training and capacity management. The following are the intended practical applications of EDSF:

- Academic curriculum design for general Data Science education and individual learning path construction for customizable training and career development
- Professional competence benchmarking, including CV or organisational profiles matching
- Professional certification of Data Science Professionals
- Vacancy construction tool for job advertisement (for HR) using controlled vocabulary and Data Science Taxonomy
- Data Science team building and organisational roles specification

The EDSF toolkit has being developed to support mentioned above applications and ensure their compatibility. It contains a number of API, ontologies and datasets representing different components of the EDSF and mapping between them. EDSF Toolkit is an ongoing development and available as Open Source at the EDSF github project [2].

V. EDSF Data Model and Ontology

EDSF data model represents all the complex relations between the EDSF components such as competences, knowledge, skills, professional profiles, proficiency levels, and organisational roles, that exist in real life organisations. Initial EDSF definition followed the 4 parts structure as describes in section III. Initial definition of EDSF was made in a form of Excel workbooks and table what provided good way of documenting but was difficult to use for practical applications [10].

In the current EDSF Release 3 (EDSF2018) [2], the CF-DS and DS-BoK are expressed in a form of ontology that is linked also to DSPP profiles definition. The ontology provides an effective format for representing rich relations between EDSF components in a form of instance, classes and properties, it also allows easy design of APIs and benefitting from existing APIs (e.g. for Python and Java).

CF-DS ontology is a core ontology linking all EDSF entities, classes and properties. It includes ontologies for all individual competences defined for the main competence groups DSDA, DSENG, DSDM, DSRMP (refer to section III) defined as subclasses. Each competence is represented as an instance of the class to which it belongs (e.g. DSDA01 is an instance of DSDA subclass). Each competence instance includes the following properties:

- Knowledge that are required for competences, defined as knowledge topics and linked to Knowledge Units (KU) in the DS-BoK
- Skills related to the knowledge topics (defined in CF-DS as Skills type A)
- Skills related to practical experience including programming, tools and platforms (defined in CF-DS as Skills type B)

Figure 2 illustrates the relation between different data sets and ontologies comprising EDSF, in particular it illustrates example of the DSDA01 competence that is defined as "Effectively use variety of data analytics techniques, such as Machine Learning, Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through the whole data lifecycle". The DSDA01 properties include knowledge topics KDSDA*, Skills Group A SDSDA* and Skills Group B SDSA*.

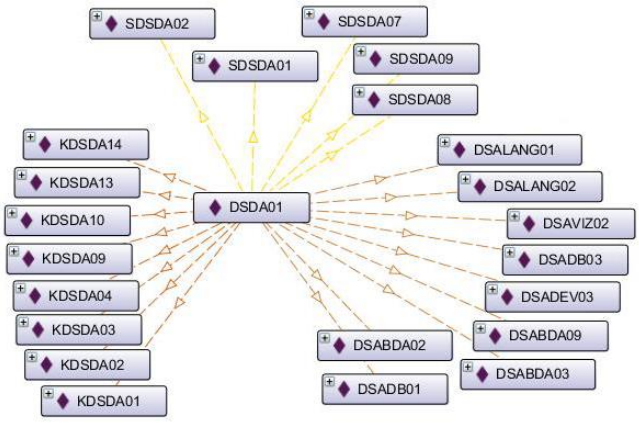


Figure 2. Example DSDA01 Competence and its properties.

The Protégé ontology editor was used for ontology design and management. It allows creating and managing an

ontology through an intuitive graphic interface and permits to export the ontology in a large number of formats. In this project RDF/OWL format is chosen in order to query the ontology using the Python module, OwlReady2.

VI. Data Science Curriculum Design using EDSF Ontology

This section describes the workflow of using EDSF for curriculum design for selected/intended set of competences that are required for (1) a specific Data Science professional profile defined based on DSPP document, or (2) individual training program defined based on competence assessment and identified gaps. The individual competence assessment can be done based on CV matching against intended job position or professional profile. It can be also done based on the certification exam or just self-assessment questionnaire. Outcome of this process is either level of matching or competence gap that can be used for suggesting necessary training program or tailored curriculum. As a part of the EDSF Toolkit development the authors have tested different methods for CV and job vacancy/profile matching using Doc2Vec document embedding and PV-DBOW training algorithms (available in the genism Python libraries) [16, 17].

When a set of required competences is defined together with the required ranking or proficiency level, the set of required knowledge topics can be extracted from individual competences (note, there exist multiple links from competence instances to single knowledge topic) and ordered according to proficiency level and relevance (or benchmark score) for further mapping to DS-BoK Knowledge Areas and Knowledge Units. The set of KAs and KUs defined for a specific competence set define the structure of the curriculum that further can be mapped to the Model Curriculum Learning Units defined as individual courses and KAG related courses groups, otherwise it can be used directly as advice for constructing curriculum by the program or course manager.

At the same time, required proficiency level is scored for each KA and KU, which will define a mastery levels and corresponding learning outcome for the targeted education or training. When using MC-DS as a template for designing customised curriculum, the proficiency levels (using scale 0 to 9) can be easy mapped to 3 mastery levels defined in MC-DS): Familiarity, Usage, Assessment (refer to MC-DS [5]). Collected Skills type B linked to intended competences will provide advice on the required hands on training and practical project development tasks and development platform.

When using EDSF ontology, it is a routine task to extract all required kknowledge topics, map them to KA/KU and define relevance score by querying ontology with a few lines of code using OwlReady2 Python module that allows manipulating ontology classes, instances and properties transparently.

Figure 3 illustrates example of relations between EDSF components when extracting required Knowledge Units for DSDA group of competences for DSP04 – Data Scientist professional profile (refer to DSPP [6] for details). It shows that the following competences are required with the corresponding relevance/weight: DSDA01 = 9; DSD02 = 9; DSDA04 = 7. Required Knowledge Units are defined through the mapping knowledge topics KDSDA* to KU (using DS-BoK) and weighted based on average relevance by competences.

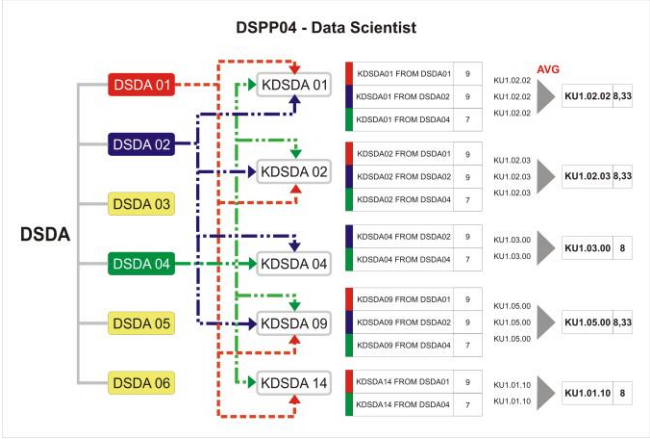
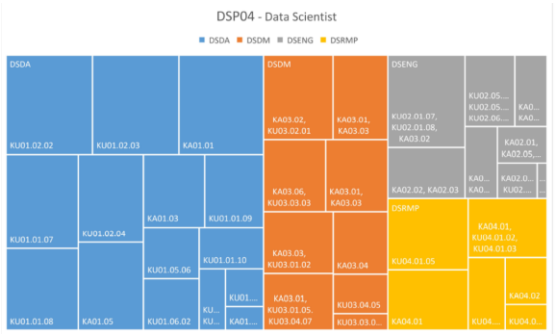
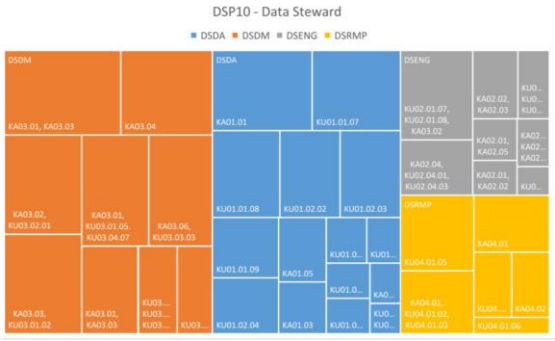


Figure 3. Extracting required Knowledge Units from EDSF ontology.

The same process is applied to other competence groups relevant to specific professional profile or competence gap. Figure 4 (a) and (b) shows example of the suggested curriculum structure for two professional profiles: DSP04 – Data Scientist and DSP10 – Data Steward. The diagrams reflect relative structure of the curriculum where Data Scientist has major part of the Data Analytics courses (DSDA - blue) followed by necessary knowledge in Data Management (DSDM - orange), and Data Steward curriculum must focus on the Data Management courses (DSDM – orange), followed by basic knowledge in Data Analytics (DSDA – blue).



(a) Data Scientist curriculum structure



(b) Data Steward curriculum structure

Figure 4. Example curriculum structure for DSP04 – Data Scientist and DSP10 – Data Steward.

The EDSF Toolkit and its outcome provides advice on the suggested curriculum structure that can be adjusted to real condition of the teaching or training institution depending on the available teaching staff and lab base. It is also important that the courses are correctly ordered and necessary pre-requisite knowledge are specified. When using 3rd party educational platforms providers and cloud based data labs, the presented application can provide a specification for required educational platform.

VII. Conclusion and Further Developments

EDSF provides a common semantic basis for interoperability of all forms of the Data Science curriculum definition and education or training delivery, as well as knowledge assessment based on fully enumerated definition of EDSF components and individual units. Besides defining academic components of the effective and consistent curriculum, EDSF provides also advice on the required Data Science Education Environment to facilitate fast practical knowledge and skills acquisition by students and learners.

Further EDSF Toolkits development will include defining ontologies for MC-DS and DSPP that is intended to be compatible with the ESCO ontologies [20] that is defined as a European standard for competences, skills and occupations definition.

The EDSF and the proposed in this paper its further integration with the Data Science Education Environment will facilitate education and training for highly demanded Data Science and Analytics competences and skills.

Acknowledgment

The research leading to these results has received funding from the Horizon2020 projects FAIRsFAIR (grant number 831558), MATES (grant number 591889) and EDISON (grant n. 675419).

References

[1] EDISON Community wiki. [online] https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome

[2] EDISON Data Science Framework (EDSF). [online] Available at https://github.com/EDISONcommunity/EDSF

[3] Data Science Competence Framework [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-competence-framework

[4] Data Science Body of Knowledge [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-body-of-knowledge

[5] Data Science Model Curriculum [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-model-curriculum

[6] Data Science Professional Profiles [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-professional-profile

[7] Demchenko, Yuri, David Bernstein, Adam Belloum, Ana Oprescu, Tomasz W. Wlodarczyk, Cees de Laat, New Instructional Models for Building Effective Curricula on Cloud Computing Technologies and Engineering. Proc. The 5th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2013), 2-5 December 2013, Bristol, UK.

[8] Demchenko, Yuri, et al, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. Proc. The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 Dec 2014, Singapore.

[9] Manieri, Andrea, et al, Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists, Proc. The 7th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2015), 30 November - 3 December 2015, Vancouver, Canada

[10] Demchenko, Yuri, et all, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, Proc. The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 Dec 2016, Luxembourg.

[11] Final results of the European Data Market study measuring the size and trends of the EU data economy, EC-IDC, March 2017 [online] https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

[12] PwC and BHEF report "Investing in America's data science and analytics talent: The case for action" (April 2017) http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent

[13] Burning Glass Technology, IBM, and BHEF report "The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market" (April 2017) https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF

[14] e-CF3.0, 2016 European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1. Available at http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

[15] CCS, 2012 The 2012 ACM Computing Classification System. Available at http://www.acm.org/about/class/class/2012

[16] Quoc Le and Tomas Mikolov, Distributed Representations of Sentences and Documents

[17] Phillip Lord (2010) Components of an Ontology. Ontogenesis.

[18] Jey Han Lau and Timothy Baldwin - An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation

[19] Matthew Horridge, Simon Jupp, Georgina Moulton, Alan Rector, Robert Stevens, Chris Wroe, A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools.

[20] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at https://ec.europa.eu/esco/portal/#modal-one