



EDISON Data Science Framework: Part 5. EDSF Use Cases and Applications (EDSF-UCA) Release 4 (ESDSF04 or EDSF2022)

EDISON Community Initiative
(Maintaining the H2020 EDISON project outcome)

Release Date	31 December 2022
Document Editor/s	Yuri Demchenko
Version	Release 4, v01
Status	Working document, request for comments



This work is licensed under the Creative Commons Attribution 4.0 International License.
To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>



Document Version Control

Version	Version	Date	Change Made (and if appropriate reason for change)	Contributors and Editors (initials)
Release 4	01	17/12/2020	Initial draft	YD
Release 4	02	31/12/2022	First public version	YD



This work is licensed under the Creative Commons Attribution 4.0 International License (CC BY).
 To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>
 This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation.

Contributors

Document Editors: Yuri Demchenko		
Author Initials	Name of Contributor	Institution
YD	Yuri Demchenko (editor)	University of Amsterdam
TW	Tomasz Wiktorski	University of Stavanger
JJCG	Juan Cuadrado	Alcala University
LC	Luca Comminiello	Erasmus+ Master, University of Perugia
MM	Mathijs Mayer	University of Amsterdam
OC	Oleg Chertov	Sikorski National Technical University of Ukraine "Kyiv Polytechnic Institute"
	Ernestina Menasalvas, Ana M. Moreno, and Nik Swoboda	Universidad Politécnica de Madrid, Madrid, Spain

Acknowledgement

Partners of the EDISON, FAIRsFAIR, MATES projects.

Executive summary

The initial definition of the EDISON Data Science Framework (EDSF) was done in the Horizon2020 Project EDISON (Grant 675419) that produced Release 1 in 2016 and published Release 2 in 2017. Currently, EDSF is maintained by the EDISON Community initiative that is coordinated by the University of Amsterdam. The new EDSF Release 4 is the product of the wide community of academicians, researchers and practitioners that are practically involved in Data Science and Data Analytics education and training, competences and skills management in organisations, and standardisation in the area of competences, skills, occupations and digital technologies. In particular, the current release incorporates revisions to competences proposed during the Data Stewardship Professional Competence Framework (CF-DSP) definition by the FAIRsFAIR project (Grant 831558).

The EDISON Data Science Framework (EDSF) includes the following components: Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS), Data Science Professional Profiles (DSPP), EDSF Use cases and applications (EDSF-UCA). The EDSF provides a conceptual basis for the Data Science Profession definition, targeted education and training, professional certification, organizational capacity building, and organisation and individual skills management and career transferability.

The current document EDSF Use cases and applications (EDSF-UCA) includes the following use cases of using EDSF in curriculum design, skills management, and workplace alignment for new technology/digital transformation:

- Definition of the Data Stewardship Professional Competence Framework (DSP-CF) by the FAIRsFAIR project;
- Use of EDSF for CV assessment and job profiles matching;
- Customised curriculum design based on EDSF ontology search for selected job profiles;
- Designing customised Data Science Education Environment;
- Big Data Infrastructure Technologies and Tools for Data Analytics course design and teaching experience
- Education and training on Research Data Management for research institutions and Enterprise Data Management for industry and business;
- Transversal and 21st Century skills for data driven economy and Industry 4.0;
- Digital and data competences definition and training profiles

The EDSF Part 5 Use cases and applications is intended to collect known use cases by the authors and other projects and practitioners that can be used as examples and guidelines for practical use by universities, training organisations, data management and data steward team to define their Data Science curricula and courses selection, on the one hand, and for companies to better define a set of required competences and skills for their specific industry domains in their search for Data Science and related talents, on the other hand.

The EDSF documents are available for public discussion at the EDISON Community initiative at <https://github.com/EDISONcommunity/EDSF/wiki/EDSFHome>

TABLE OF CONTENTS

1	Introduction.....	8
2	EDISON Data Science Framework (EDSF)	9
	References.....	10
3	Data Scientist Professional Revisited: Competences Definition and Assessment, Curriculum and Education Path Design	11
3.1	Introduction.....	11
3.2	Data Scientist Professional Definition.....	12
	3.2.1 Data Scientist Definition Evolution	12
	3.2.2 Importance of continuous and self-education	13
3.3	EDISON Data Science Framework (EDSF).....	13
3.4	EDSF Practical uses and EDSF Toolkit.....	15
3.5	Data Science competences assessment.....	16
	3.5.1 Pre-processing steps	16
	3.5.2 Implementation	17
3.6	Building Learning path for the Designed Curricula using Bloom’s Taxonomy	20
	3.6.1 Customised Curriculum Design using EDSF Ontology	20
	3.6.2 Defining Knowledge Units to include into the curriculum	22
	3.6.3 Applying Bloom’s Taxonomy to curriculum structuring and course planning	22
3.7	Conclusion and Further Developments	23
	References.....	23
4	Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge	25
4.1	Demand for Data Science Competences and Customisable Curriculum	25
4.2	EDSF Toolkit and Practical Uses of EDSF	25
4.3	EDSF Data Model and Ontology.....	26
	4.3.1 EDSF Data Model	26
	4.3.2 Definition of the EDSF Ontology	27
4.4	Data Science Curriculum Design using EDSF Ontology	28
4.5	Cloud based DSEE and Virtual Data Labs: IDE, Tools and Datasets	30
4.6	Conclusion and Further Developments	31
	References.....	31
5	Big Data Platforms and Tools for Data Analytics in the Big Data and Data Science Curricula	33
5.1	Introduction.....	33
5.2	Data Science Engineering BoK and Model Curriculum	33
	5.2.1 DSENG Model Curriculum Components	33
	5.2.2 DSENG/BDIT - Big Data infrastructure technologies course content	34
	5.2.3 Data Management and Data Stewardship in the Big Data and Data Science Curriculum	34
5.3	Platforms for Big Data Processing and Analytics	34
	5.3.1 Essential Hadoop Ecosystem Components	34
	5.3.2 Hadoop Programming Languages	35
	5.3.3 Cloud based Big Data Platforms	36
5.4	Example BDIT4DA courses and experience	36
	5.4.1 BDIT4DA Course for Big Data Engineering Masters	36
	5.4.2 Course for Data Science Masters	38
	5.4.3 Big Data Infrastructure Technologies (BDIT) Course for MBA in Big Data	39
5.5	Conclusion and Recommendations	40
6	Courses to Facilitate Data Science Professional Skills	42
6.1	Professional Issues in Data Science	42
6.2	Data Science and Analytics Foundation (DSAF)	42
6.3	Research Methods and Process Management	43
	6.3.1 Data Science Process Management Frameworks	43
	6.3.2 CRISP-DM, CRoss-Industry Standard Process for Data Mining	43
	6.3.3 ASUM, Analytics Solutions Unified Method (IBM)	44
	6.3.4 TDSP, Team Data Science Process (Microsoft)	45
	6.3.5 KNIME Model Factory (KMF)	45

6.4	ML Model Formats.....	46
	References.....	46
7	Data Stewardship Professional Competence Framework (DSP-CF)	47
7.1	Introduction.....	47
7.2	Research Data Management and Data Stewardship	48
7.3	Data Stewardship and FAIR competence Frameworks.....	49
	7.3.1 EOSCpilot FAIR4S Framework	49
	7.3.2 ELIXIR Data Stewardship Competency Framework	50
	7.3.3 DeIC Data Stewardship curricula recommendations/principles	50
	7.3.4 GO FAIR Metadata Management Requirements and FAIR Data Maturity Model	51
	7.3.5 DAMA DMBOK: Data Governance and Stewardship	52
	7.3.6 EDISON Data Science Framework and Data Steward Professional Profile Definition	53
7.4	Job market analysis for demanded key competence.....	56
	7.4.1 Method and context	56
	7.4.2 EDISON methodology to collect and analyse job market and competence related data [1, 2]	57
	7.4.3 Collected data	57
	7.4.4 Identified competences, skills and knowledge and their mapping to CF-DS	58
	7.4.5 Outcome of the job vacancies analysis and further steps	60
	7.4.6 Technological and organisational aspects of the FAIR data principles implementation	60
7.5	Defining a Competence Framework for Data Stewardship and FAIR Data Principles (CF-DSP)	62
	7.5.1 Data Management and Governance competence group (DSDM)	62
	7.5.2 Data Engineering competence group (DSENG)	64
	7.5.3 Research Methods and Project Management competence group (DSRMP)	65
	7.5.4 Domain related competence (DSDK/DSBA)	66
7.6	Data Steward professional and transversal skills.....	68
7.7	Comparing/Mapping CF-DSP to other Competence Frameworks	68
	7.7.1 Summary on defining the Data Stewardship and FAIR competence framework for Higher Education	73
7.8	Defining Data Stewardship and FAIR Body of Knowledge	73
	7.8.1 Data Science Body of Knowledge Areas and Knowledge Units	73
	7.8.2 Defining a DSP BoK profile	74
	7.8.3 Using CF-DSP and DSP-BoK for Data Stewardship curriculum definition	75
	References.....	76
8	Data Management and Data Stewardship for Industry, Research and Academia	77
8.1	Research Data Management and Stewardship (RDMS)	77
8.2	FAIR Teaching Handbook: Curricula Topic 15: Research Data Management: Overview and Best Practices	78
8.3	Data Management and Governance (enterprise scope)	79
8.4	FAIR Teaching Handbook: Curricula Topic 16: Data Management and Governance in Industry and Research	80
9	EDISON Data Science Framework (EDSF): Addressing Demand for Data Science and Analytics Competences for the Data Driven Digital Economy.....	82
9.1	Introduction.....	82
9.2	Demand for Data Science and Data Skills	83
9.3	EDISON Data Science Framework (EDSF).....	83
	9.3.1 Data Science Competence Framework (CF-DS)	84
	9.3.2 Workplace skills	84
	9.3.3 Data Science Professional Profiles (DSPP)	85
	9.3.4 Data Science Body of Knowledge and Model Curriculum	85
9.4	Example Curricula to Facilitate Digital Transformation	86
	9.4.1 Data Management and Data Governance	86
	9.4.2 Data Science and Analytics Foundation (DSAF)	86
	9.4.3 Professional Issues in Data Science	87
	9.4.4 Cloud based DSEE and Virtual Data Labs	87
9.5	Conclusion and Further Developments	87
10	Transversal Skills required by Emerging Industry 4.0 Transformation	89

10.1	Introduction	89
10.2	Digital Competences and Data Literacy	89
10.3	Suggested Knowledge and curriculum topics	91
10.3.1	Digital and Data Literacy Topics	91
11	Big Data Value Data Science Badges.....	92
11.1	Introduction	92
11.2	The Strategic Framework for Education and Training	93
11.2.1	New Skills Agenda for Europe	94
11.2.2	New Europass framework	94
11.2.3	European Qualifications Framework for lifelong learning	95
11.2.4	European Skills, Competences, Qualifications and Occupations	96
11.2.5	Highlights and common threads in these initiatives	96
11.3	Education and Training Recognitions.....	97
11.3.1	A Survey of Recognitions	97
11.3.2	A Comparison of Recognitions	99
11.3.3	Recommendations for Data Science skills recognition	100
11.3.4	The Logistics of the BDV Data Science Badge Program	101
11.4	BDV Data Science Badges Types and Requirements.....	102
11.4.1	Refining and Evaluating this Initial Proposal	103
11.5	Conclusions	104
12	Part 5 Conclusion and further document extension	105
	Acronyms	106
	Appendix A. EDSF reviews, citation and references in other research and projects	107
	A.1. EDSF reviews and references	107
	A.2. Projects Using or contributing to the EDSF development.....	108
	A.3. Publications by Authors	108

1 Introduction

EDSF Part 5 is a part of the EDISON Data Science Framework (EDSF) that comprises the following documents: Data Science Competence Framework (CF-DS) [1], Data Science Body of Knowledge (DS-BoK) [2], Model Curriculum (MC-DC) [3], and Data Science Professional Profiles (DSPP) [4], EDSF Use cases and applications (EDSF-UCA) [5].

The current document EDSF Use cases and applications (EDSF-UCA) includes the following use cases of using EDSF in curriculum design, skills management, and workplace alignment for new technology/digital transformation:

- Definition of the Data Stewardship Professional Competence Framework (DSP-CF) by the FAIRsFAIR project;
- Use of EDSF for CV assessment and job profiles matching;
- Customised curriculum design based EDSF ontology research for selected job profiles;
- Designing customised Data Science Education Environment;
- Big Data Infrastructure Technologies and Tools for Data Analytics course design and teaching experience
- Education and training on Research Data Management for research institutions and Enterprise Data Management for industry and business;
- Transversal and 21st Century skills for data driven economy and Industry 4.0;
- Digital and data competences definition and training profiles

The document also contains the experience of the Big Data Value Association on the Data Science Badges initiatives and implementation.

The assembled in this part materials is based on the published papers and project reports. Appendix A contains references to all contributed sources and also links to EDSF derivatives, reviews, citations and references.

2 EDISON Data Science Framework (EDSF)

The EDISON Data Science Framework provides a basis for the definition of the Data Science profession and enables the definition of the other components related to Data Science education, training, organisational roles definition and skills management, as well as professional certification.

Figure 2.1 below illustrates the main components of the EDISON Data Science Framework (EDSF) and their inter-relations that provides the conceptual basis for the development of the Data Science profession:

- CF-DS – Data Science Competence Framework (this document [1])
- DS-BoK – Data Science Body of Knowledge [2]
- MC-DS – Data Science Model Curriculum [3]
- DSPP - Data Science Professional profiles and occupations taxonomy [4]
- Data Science Taxonomy and Scientific Disciplines Classification

The proposed framework provides a basis for other components of the Data Science professional ecosystem¹, such as

- EDISON Online Education Environment (EOEE)
- Education and Training Directory and Marketplace
- Data Science Community Portal (CP) that also includes tools for individual competences benchmarking and personalized educational path building
- Certification Framework for core Data Science competences and professional profiles

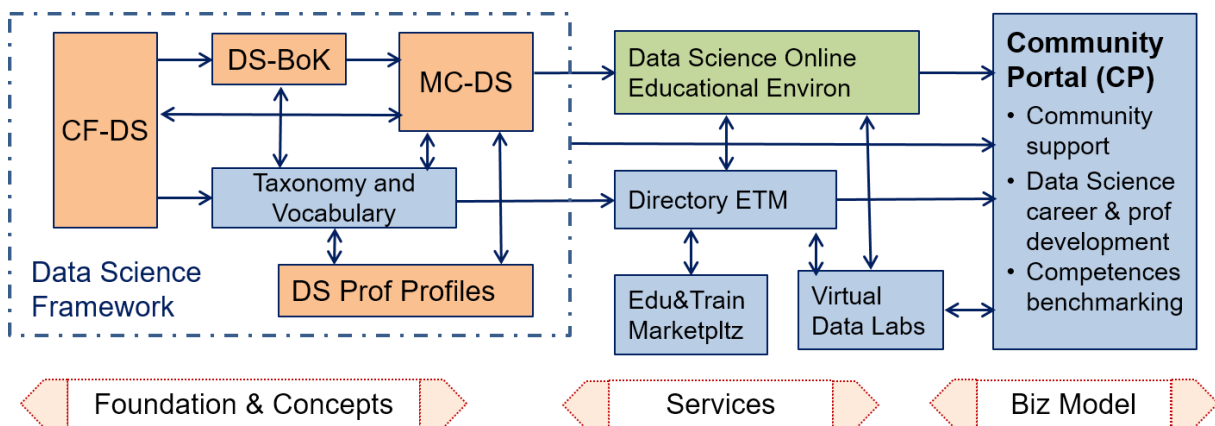


Figure 2.1 EDISON Data Science Framework components and Data Science professional ecosystem.

The EDSF Release 4 includes Part 5 EDSF Use cases and applications [5] which describes a few uses of using EDSF by universities and professional education and training organisations as well as subject domain communities; the guidelines part provides recommendations on using EDSF for practical cases of defining new domain specific competence profiles, knowledge areas and model curricula.

The CF-DS provides the overall basis for the whole EDSF. The core CF-DS includes common competences required for the successful work of a Data Scientist in different work environments in industry and in research and throughout the whole career path. The future CF-DS development may include coverage of the domain specific competences and skills by involving domain and subject matter experts, which may be published as separate CF-DS profiles².

¹ The described Data Science ecosystem components are defined and piloted in the EDISON project and constitute the project legacy that can be re-used and followed by the community.

² Data Stewardship Professional Competence Framework (CF-DSP) has been developed by the FAIRsFAIR project by extending CF-DS with the Data Stewardship and FAIR related competences and skills and published as a separate document referring to the core EDSF documents [6]

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. Knowledge Areas are composed of a number of Knowledge Units (KU) which are currently the lowest component of the DS-BoK. DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs and KUs defined where possible based on the Classification Computer Science (CCS2012) [7], components taken from other BoKs and proposed new KAs/KUs to incorporate new technologies used in Data Science and their recent developments.

The MC-DS is built based on CF-DS and DS-BoK where Learning Outcomes (LO) are defined based on CF-DS competences, 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. The proposed Learning Outcomes are enumerated to have a direct mapping to the enumerated competences in CF-DS.

The DSPP professional profiles are defined as an extension to the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy [8] using the ESCO top classification groups. DSPP definition provides an important instrument to define effective organisational structures and roles related to Data Science positions and can also be used for building individual career paths and corresponding competences and skills transferability between organisations and sectors.

The Data Science Taxonomy and Scientific Disciplines Classification will serve to maintain consistency between four core components of EDSF: CF-DS, DS-BoK, MC-DS, and DSP profiles. To ensure consistency and linking between EDSF components, all individual elements of the framework are enumerated, in particular: competences, skills, and knowledge topics in CF-DS, knowledge groups, areas and units in DS-BoK, learning outcomes and learning units in MC-DS, and professional profiles in DSPP.

It is anticipated that successful acceptance of the proposed EDSF and its core components will require standardisation and interaction with the European and international standardisation bodies and professional organisations. This work is being done as a part of the EDSF sustainability support by the EDISON community initiative provided by the University of Amsterdam³.

The EDISON Data Science professional ecosystem illustrated in Figure 2.1 shows how the core EDSF components may be related to the potential services that can be offered for the professional Data Science community and provide basis for sustainable Data Science competences and skills management by organisations, in particular in conditions of emerging Industry 4.0, growing digitalisations and Artificial Intelligence development. As an example of practical use, CF-DS and DS-BoK can be used for individual competences and knowledge benchmarking and play an instrumental role in constructing personalised learning paths and professional (up/re-) skilling programs based on MC-DS.

References

- [1] Data Science Competence Framework, EDSF Part 1 [online] <https://github.com/EDISONcommunity/EDSF/tree/master/data-science-competence-framework>
- [2] Data Science Body of Knowledge, EDSF Part 2 [online] <https://github.com/EDISONcommunity/EDSF/tree/master/data-science-body-of-knowledge>
- [3] Data Science Model Curriculum, EDSF Part 3 [online] <https://github.com/EDISONcommunity/EDSF/tree/master/data-science-model-curriculum>
- [4] Data Science Professional Profiles, EDSF Part 4 [online] <https://github.com/EDISONcommunity/EDSF/tree/master/data-science-professional-profile>
- [5] [EDSF Use cases and applications](#), EDSF Part 5 [online] <https://github.com/EDISONcommunity/EDSF/tree/master/data-science-edsf-use-cases-applications>
- [6] FAIR Competence Framework for Higher Education (Data Stewardship Professional Competence Framework), FAIRsFAIR Project Deliverable D7.3, February 2021 [online] <https://zenodo.org/record/4562089#.Y6uactnbMK38>

³ EDISON Community Initiative website <https://edisoncommunity.github.io/EDSF/>

3 Data Scientist Professional Revisited: Competences Definition and Assessment, Curriculum and Education Path Design⁴

Data Science is maturing as a scientific and technology domain and creates a basis for new emerging technologies and data driven application domains. Educated and/or trained Data Scientist is becoming a critical component of the whole data driven science and technology ecosystem. It is important to revisit the Data Scientist Professional definition propose/identify effective approaches to Data Science competences and skills assessment that would allow developing customisable education and training curricula that would support organisational capacity building (effective HR management) and individual career development. The paper discusses how the EDISON Data Science Framework can be used to solve these problems. New approaches to Data Science competences assessment is proposed that introduces the concept of acquired competence which is calculated based on the practitioner career path. An important aspect in targeted education and training for professionals is correct and effective education path building based on initial competences and knowledge assessment. The paper proposes a new approach in customised curriculum building by applying Bloom's Taxonomy to training courses sequence and timing/scheduling.

3.1 Introduction

Big Data technologies and the availability of scalable cloud based Big Data platforms and tools that can be provisioned and used on demand created new opportunities to work with a variety of data produced by human activity and technological processes. Data driven technologies development facilitated the emergence of Data Science as a scientific and technology domain focused on different aspects of data analysis to support data driven technologies and applications in all scientific, industry and human activity domains. Modern data driven research and industry created strong demand for new types of specialists that are capable of supporting all stages of the data lifecycle from data production and input to data processing and actionable results delivery, visualisation and reporting, as well and technological processes control and automation. With the growing importance of data in modern economy, data is becoming an important asset, understanding of the importance to create the whole ecosystem for data management and governance is growing. Organisations moving to agile data driven model, need to redefine many organisational role and introduce new data related roles, in addition to the commonly accepted importance of Data Scientists, which can be jointly defined as the Data Science professions family. Continuous technology evolution imposes new challenges to modern data driven organisations in technology change management and in managing organisational human/capacity resources in related data driven technologies. Effective Data Science education must combine theoretical and practical skills, while developing right attitude to continuous professional (self-) education. The fact that modern technologies are led by large technology companies who are interested in their technologies adoption, should be recognised and motivate universities and research community to cooperate with technology leaders in enriching academic education with using new available technology platform, especially in Big Data and cloud computing.

The proposed approaches to competences assessment and customisable educational and career path building are based on the EDISON Data Science Framework (EDSF), which was developed in the EU funded EDISON Project and is currently maintained by the EDISON Initiative [1, 2]. Since the first EDSF release in 2016, the framework has undergone significant development, EDSF Release 3 (2018) summarised the experience of numerous practitioners and educators that contributed to the definition of EDSF components. The new EDSF2020 (Release 4) incorporated recent technology developments that confirmed the Data Science as a central component in the whole ecosystem of data intensive and data driven technologies that include Machine Learning, Artificial Intelligence, Digital Twins, immersive technologies and IoT.

This paper presents new results in continuous research by the authors to improve the Data Science competences assessment based on the professional career path, and optimal learning path definition based on the competences gap. The paper refers to the previous authors' works that researched new approaches to building

⁴ Based on paper Yuri Demchenko, Mathijs Majjer, Luca Comminiello, Data Scientist Professional Revisited: Competences Definition and Assessment, Professional Development and Education Path Design, International Conference on Big Data and Education (ICBDE2021), February 3-5, 2021, London, United Kingdom (PDF) Yuri Demchenko, Lennart Stoy, Research Data Management and Data Stewardship Competences in University Curriculum, In Proc. Data Science Education (DSE), Special Session, EDUCON2021 – IEEE Global Engineering Education Conference, 21-23 April 2021, Vienna, Austria

effective curricula in Cloud Computing, Big Data and Data Science [7, 8, 9, 10, 11] and based on a long time practical experience in developing both online and campus based education and training courses.

The paper is organized as follows. Section II revisits popular definitions of Data Scientists and refers to important technologies related to Data Science to create a context for the proposed definition of the Data Scientist Professional and related Data Science competences. Section III describes the EDSF, its components and application domains. Section IV describes the different uses of EDSF and the functionality of the EDSF Toolkit. Section V describes the Data Science competences assessment, and describes the proposed method to assess acquired competences based on the career path. Section VI discusses how the EDSF can be used for designing customised curriculum based on competences assessment (or professional profile) while using Bloom's taxonomy and competences ranking for educational path construction, and Section VII provides summary and suggestions for future work.

3.2 Data Scientist Professional Definition

3.2.1 Data Scientist Definition Evolution

There are multiple definitions of the Data Science discipline and technology, given in different contexts, that stress/put in the centre one of the four aspects of data analysis: Data Analytics, Data Science, Machine Learning/Deep Learning, and Artificial Intelligence:

- *Data Analytics* is a process of inspecting, transforming and modelling data with the goal to discover trends, patterns, relations that describe observable real-life phenomena and can be used for informed decision-making.
- *Data Science* makes the systematic study of the structure and behaviour of data in order to understand past and current occurrences, as well as predict the future behaviour of that data. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.
- *Machine Learning* deals with the development of algorithms, some of them based on statistical models, with the objective that their computational implementation allows the computer not only to carry out the tasks without supervision but learn of the results for continuous improvement. Within Machine Learning, *Deep Learning* is the set of predictive methodologies that use artificial neural networks to progressively extract higher level features from unstructured raw data. This class of methods is particularly effective for making predictions from big data generated by real life behavioural processes or sensors. Machine Learning and Deep Learning are considered as subfields of Data Science focused on specific tasks, while Data Science provides a general methodology for working with a wide variety of data using different methods and tools.
- *Artificial Intelligence* is a machine or application with the capability to autonomously execute upon predictions it makes from data, where prediction is made based on Data Science and analytics methods. Artificial Intelligence is strongly connected to Digital Twins and robotics which increase importance of the consistent industrial data management and quality assurance.

It is important to clarify the relation of Data Science to other closely related scientific disciplines and technology domains such as Big Data, Artificial Intelligence, Machine Learning, and Statistics. Despite the fact that some authors may refer to historical facts of mentioning these terms 10s of years ago [12], we refer to the current data driven technologies development that made Data Science a central component of all other data related and data driven technologies development. We identify such technology fusion and consolidation took place in 2011-2013 with advents of Cloud Computing and Big Data what also aligned with the National Institute of Standards and Technologies, NIST definition of the Cloud Computing in 2011 [13] and Big Data definition in 2013 [14].

Big Data serves as a technology platform to allow the Data Science and Analytics solutions and applications to work with modern data, which are of the *Big Data 3V scale: Volume, Velocity, and Variety*. Big Data technology platform includes large scale computation, storage and network facilities, typically cloud based, such as Hadoop, Spark, NoSQL databases, data lakes, and others.

In the whole digital economy ecosystem, Data Science integrates all multiple components from other scientific and technology domains to drive data intensive research and emerging digital technologies development. It is important to give Data Science definition as a scientific discipline to become a foundation for academic research and curricula development:

Data Science is a complex discipline that uses conceptual and mathematical abstractions and models, statistical methods, together with modern computational tools to obtain knowledge/derive insight from data to uncover correlations and causations in business data and support decision making in scientific research and business activity.

Data Scientist is defined as a professional practicing Data Science. Starting from the first years of the Data Science and Analytics technologies adoption, there were many Data Scientist definitions proposed by practitioners in the new domain that reflected their personal professional development. The following competence areas and skills were included in the Data Scientist competence profile: mathematics, statistics, computer skills, domain knowledge, and also hacking skills as the ability to understand (undocumented) functionality of software and algorithms and effectively use them for practical purposes.

The experience of the EDISON Data Science Framework development and practical implementation supported by wide research and educational community discussions allowed us to propose an actionable definition of the Data Scientist Professional, which is based on the NIST definition and extended with organisational role of the Data Scientist [14]:

A Data Scientist is a practitioner who has sufficient competences and knowledge in the overlapping regimes of expertise in data analytics skills, domain knowledge, business needs, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle, till the delivery of an expected scientific and business value to science or industry.

3.2.2 Importance of continuous and self-education

It is commonly recognised that in such dynamically developing area as Big Data, Data Science, continuous education and self-study plays a critical role. The proposed EDSF definition provides a good basis for defining Data Scientist professional development path, including knowledge acquisition, skills development, and career path building.

The recent OECD report [15] confirms the urgent need to address data and general digital skills for all types of workforce and economy sectors. 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 courses is key to adopting future skills management and capacity building models.

3.3 EDISON Data Science Framework (EDSF)

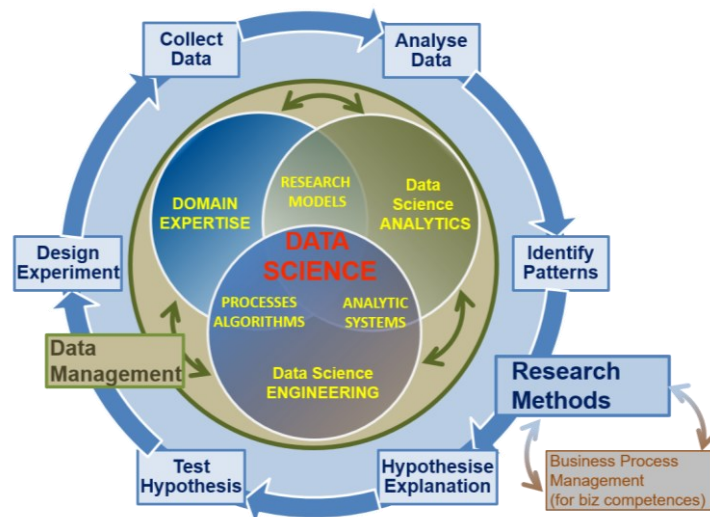
The EDISON Data Science Framework (EDSF), that is the product of the EDISON Project, provides a basis for Data Science education and training, curriculum design and competences management that can be customised for specific organisational roles or individual needs. EDSF can also be used for professional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

The main EDSF components include:

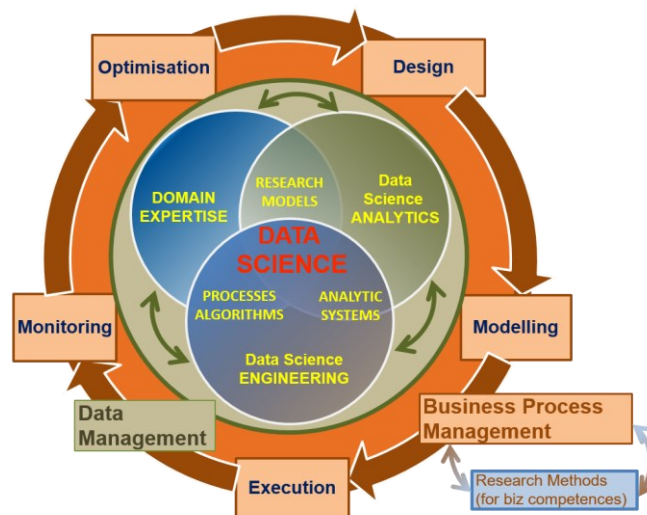
- CF-DS – Data Science Competence Framework [3]
- DS-BoK – Data Science Body of Knowledge [4]
- MC-DS – Data Science Model Curriculum [5]
- DSPP - Data Science Professional profiles and occupations taxonomy [6]
- Data Science Taxonomy and Scientific Disciplines Classification

The 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 [16] (competences defined as abilities supported by knowledge and skills with applied proficiency levels) but has competences structured according to the major identified functional groups (as explained below).

Figures 1 (a) and (b) provide a graphical presentation of relations between identified competence groups as linked to Research Methods or to Business Process Management. The figure illustrates the importance of Data Management competences and skills and Research Methods or Business Process Management knowledge for all categories and profiles of Data Scientists.



(a) Data Science competence groups for general and research-oriented profiles.



(b) Data Science competence groups for business oriented profiles.

Figures 1. Relations between identified Data Science competence groups for (a) general or research oriented and (b) business oriented professions/profiles.

The Research Methods typically include the following stages (see Appendix C for reference to existing Research Methods definitions):

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

An important part of the research process is theory building, but this activity is attributed to the domain or subject matter researcher. The Data Scientist (or related role) should be aware of domain related research methods and theory as a part of their domain related knowledge and team or workplace communications. See an example of the Data Science team building in the Data Science Professional Profiles definition provided as a separate document [6].

The following core CF-DS competence and skills groups are identified (refer to CF-DS specification [3] for details):

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

In total, CF-DS includes 30 enumerated competences, 6 competences for each of competence groups. The Data Science competences must be supported by the 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 [3] for the full definition of the identified knowledge and skills groups):

- Skills Type A which are built based on practicing major competences acquired based on education and training; depend on years of working as a Data Scientist or related roles,
- Skills Type B that are related to a wide range of practical computational skills, including using programming languages, development environment, and cloud based platforms.

The DS-BoK defines the Knowledge Areas (KA) and Knowledge Units (KU) 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. It is important to note that the CF-DS defines knowledge topics linked to specific competences that can be mapped to KU and KA in the DS-BoK. The DS-BoK is based on ACM/IEEE Classification Computer Science (CCS2012) [17] and incorporates best practices in defining domain specific BoK's. It provides a reference to related existing BoK's and includes proposed new KA to incorporate new technologies and scientific subjects required for consistent Data Science education and training.

The MC-DS [5] 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. The practical curriculum should be supported by a corresponding educational environment for hands-on labs and educational projects development.

The DSPP [6] defines a number of Data Science professional profiles in accordance with existing classifications, such as European standards ESCO [18] or EN 16234-1 "e-Competence Framework" [19]. The DSPP includes important parts of the competences relevance (scores) to each defined profile in a scale from 0 to 9 (from low to high) that be used defining targeted education and training and building an effective career path.

3.4 EDSF Practical uses and EDSF Toolkit

The EDSF toolkit has been developed to support multiple practical applications for Data Science competences and skills management and to ensure their compatibility. It primarily contains enumerated competences, skills and knowledge topics/units definition supported by corresponding ontologies. Ongoing development includes API definition and creation of the reference datasets representing different components of the EDSF to support applications development. EDSF Toolkit is a community effort and available as an Open Source at the EDSF github project [2].

The following are the intended practical applications of EDSF facilitated by the EDSF toolkit:

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

EDSF provides an example of the integrated competence and skills management framework that is being used in other technology domains and economy sectors, which examples include education and training framework for

digital and data skills in maritime industry (as part of the MATES project [20]), Data Stewardship competences and curriculum definition (as part of the FAIRsFAIR project [21]), reference in the German Ministry of Economics study on Data Science competences and resources [22].

3.5 Data Science competences assessment

The CF-DS and DSPP components provide a basis for the novel improved Data Science competences assessment in a quantified manner that takes into account the practitioner or candidate career path. This assessment is helpful to support the increasing demand for Data Scientists. For example, using CF-DS and DSPP, a desired profile can be constructed against which a curriculum vitae (CV) of a Data Scientist can be tested. Based on such a test, competence gaps can be identified between the desired profile and the assessed CV, which can help with the decision making if the candidate can be hired, on which position or role, and also predict their possible career path.

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) [23, 24].

This section describes how the document similarity techniques combined with regular expressions were used on the CVs of Data Scientists to create insights into the competencies of the CVs' creators (later referred also as a job/vacancy candidate). This is done by computing three components: a timeline that indicates the career path of the data scientist, a graph showing the competency scores based on document similarity, and a graph showing competency scores based on the career path. This method can then be deployed as a tool that could, for example, be used by recruiters or practitioners that wish to assess their own competencies and to identify a potential path for professional development. It can also be used for a Data Science team composition and management as well as identification what kind of training the team requires. This method was used to develop a tool in the form of a web-application to provide easy access [25, 26], recent development has improved the gap identification to integrate it better with the customised curriculum design.

3.5.1 Pre-processing steps

To create the three components mentioned that can be used to assess individual Data Science competences, several steps should be taken. First, to acquire the competency scores, both the CVs and the documents containing information about the required job competences (e.g. job vacancy typically using the CF-DS vocabulary) should be pre-processed, especially normalised. The punctuations are removed, all letters are transformed to lower case, and the Porter algorithm [27] is applied to remove suffixes of the words. This process is applied to compare similar words that appear in different forms; the steps are shown in figure 2. To extract the career path from a CV, a list of job names with corresponding DSPP classifications, a DSPP competence relevance scores list, and a CF-DS competence labels list were prepared. These lists were then linked to each other, so that every job found in the CV is classified with a profile that is listed in the DSPP list. This classification can then be used to obtain competence relevance scores belonging to that professional profile of the job. The relationship between the components is shown in figure 3.

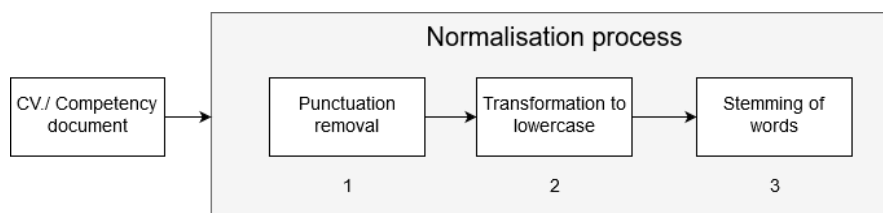


Figure 2. Processing steps of the CV and competency documents text

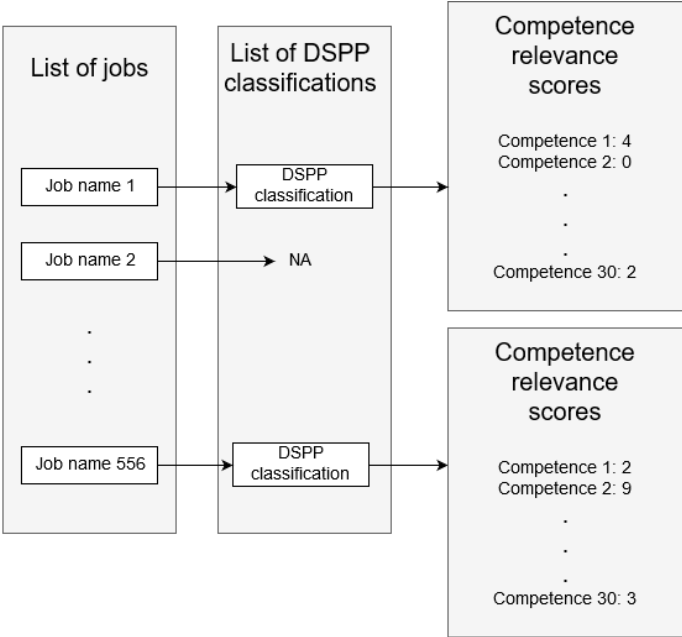


Figure 3. Relations between data components used for career path extraction

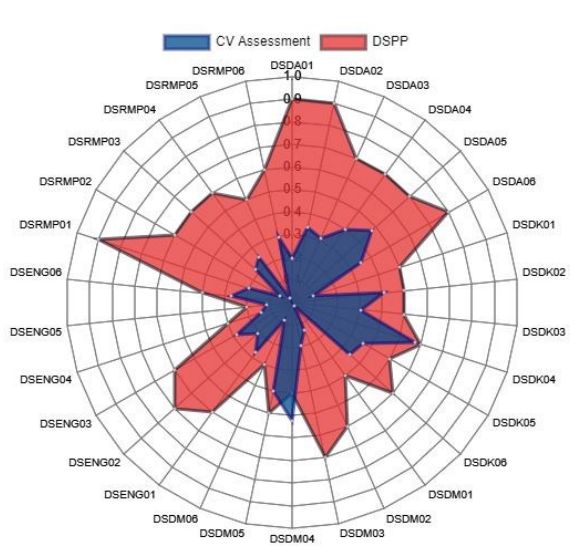
3.5.2 Implementation

3.5.2.1 CV matching and competence gap definition

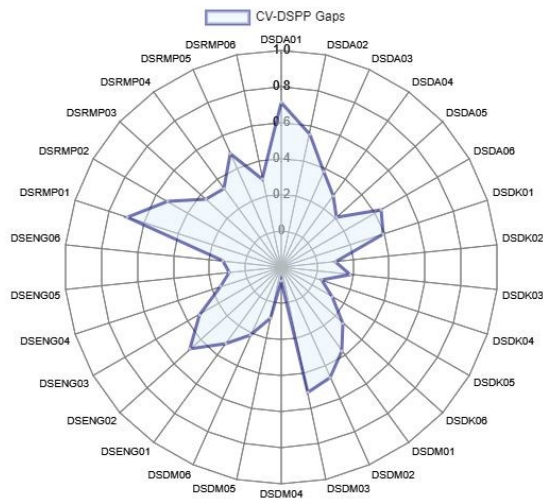
The graph showing the competency scores extracted from a CV was acquired using multiple steps. After pre-processing the CVs and the thirty competence documents representing the controlled vocabulary of the CF-DS, TF-IDF features were gathered. A matrix of TF-IDF features is acquired and can then be used to calculate cosine similarity between the CV and the competence documents. This is done by taking the dot product of the row-normalized vectors, which gives a similarity matrix that contains the similarity score. This score ranges from 0 to 1, however, because a CV will never contain the exact same text as a competence document, a score of 0.7 is seen as the maximum achievable value of a CV. These values can then be used to create a spider chart, to give an overview of the scores per competence.

When the competence scores have been acquired using a CV, it can be compared to a desired data scientist profile to identify gaps in lacking skills or knowledge. The competence relevance scores for a DSPP classification, also shown in Figure 3, can be mapped from a 0 – 9 range to a 0 – 1 range. After the mapping, we call these values rates, which then can be used to subtract the CV scores for each competence from the rates. If a negative result is acquired after the subtraction, the data scientist is deemed to have sufficient knowledge for that competence, and the result is set to 0.

After the subtractions, the differences are multiplied by the corresponding rates again, in order to weigh the proficiency level. Finally, the results are evaluated: all the competences that have scored a grade greater than 0.5 are identified as gaps. Figure 4 shows two graphs, where the first graph shows how an example CV can be matched to a DSPP or a vacancy profile. As an example, the profiles DSP04 – Data Scientist and fictitious candidate CV are used. The second graph shows how the gaps are identified using the process described above.



a) Matching CV and Vacancy profiles



b) Identified competence gaps

Figure 4. Comparing candidate’s competences and target professional profile DSPP04 Data Scientist: (a) DSPP04 and candidates competences assessed; (b) competence gap.

3.5.2.2 Career path extraction

Using described above CV and job profile matching based on both documents similarity doesn’t take into account acquired experience by a candidate. This can be resulted that two candidates using the same CV template or a CV design tool (popular service by online job search agencies), although having significantly different job experience, can be scored equally. To avoid this situation and make the CV and job matching correct, we introduced the acquired competence concept that defines the candidate’s/professional’s real competence as acquired competence amplified by years of working in relevant positions/roles.

To assess the acquired competences, next to defining the competence scores from a CV using document similarity techniques, the career path is also extracted from the CV. This information is then used to create another scores vector and the graph showing the competence relevance scores calculated together with a timeline indicating the career path of the candidate’s CV. The first step in this process is to extract all the mentioned jobs in the CV. Then, for each job, the position level is identified and years of work are extracted, a DSPP classification was made, and based on this, the DSPP competence relevance scores are assigned, and finally, the job is tagged as relevant or not. Different jobs in a CV are identified by testing whether each word of the CV appears in a list with known jobs using regular expressions. Multiple mentions of one job in a CV were separately handled to get a better overview.

Then, for all identified jobs, the position levels were acquired by looking at different position level names in the CV text. In this case, five different levels were used with their common alternatives: entry-, intermediate-, senior-, principal-, and lead-level. The location of where the job name was found is used to look for the job level by looking at words that appear near to the job name. Afterward, the amount of experience, the time that someone has practiced a job, is acquired based on the CV/career timeline. This is achieved using regular expressions to look for different date patterns near the occurrences of the job names mentioned in the CV. Multiple small regular expressions were used in combination with each other to be able to find date patterns in multiple formats, as there is no formal manner in which every CV lists its experience. The regular expressions are constructed to identify different date separators, months, time span indicators, years, and day patterns. These can be combined to identify dates in multiple formats that are often used. When data has been found for a job, the start date is subtracted from the end date, which is then divided by 365 and rounded down to get the amount of years that someone has practiced a job. These jobs are then classified using the DSPP. Then to tag whether identified jobs are relevant, they must have a DSPP classification, as well as other attributes such as start and end date, or a profession/job level.

To create a graph showing competence scores based on the extracted career path, two assumptions were made, namely: 1) when someone has practiced a job for a longer time, he/she becomes more competent at the relevant competencies that are listed, and 2) when someone quits a job, the competencies to perform relevant tasks from that job are not lost. Then, to calculate the competence scores using the career path, equation (1) was used:

$$Competence_i \leftarrow \min\left(\sum_{j \in J} c_{ij} \cdot multiplier_j, 100\right) \quad (1)$$

where i is the current competence, j is a job, J are all the jobs that were extracted, c_{ij} is the competence relevance score for the current competence and job, and the multiplier is the multiplier that is used for the job.

For every competence in the CF-DS, the minimum of 100 and the sum of the values calculated using all the relevant jobs for the current competence is taken. The multiplier is acquired by either using the direct amount of job experience in years, or by using a common mapping from the position level to the amount of experience associated with the level. This will ensure final competence scores that have a range of 0 – 100.

Figure 5 provides an example of the candidate’s competence profile with 8 years experience in positions related to Data Scientist calculated using simple documents similarity (like in Figure 4) and using the proposed acquired competence algorithm explained in this section. This can then be directly used in the same manner as shown in Figure 5, after remapping the range back to 0 – 1, which is in accordance with the earlier computed document similarity scores.

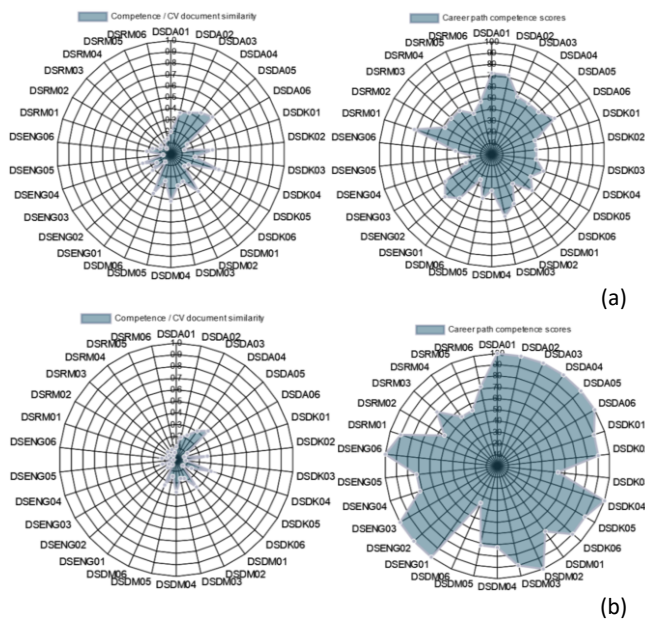


Figure 5. Comparison of competences calculated based on simple document similarity (left diagram) and using acquired competence concept (right diagram) for (a) Data Scientist with 8 years of experience, and (b) Hadoop developer with 15 years of experience.

3.6 Building Learning path for the Designed Curricula using Boom’s Taxonomy

The next step after defining the set of intended competences (that can be either a selected professional profile or an assessed competence gap) is to design an effective curriculum and education or training path. It is rather a well defined process and a routing process to create the curriculum for the whole academic programme that should be delivered by an educational institution when a whole set of required competences and corresponding learning outcomes provide input to the curriculum definition. The general purpose curriculum can be created in this way. However, as technology develops fast and correspondingly required competences change, there is a need for educational and training institutions to react fast and offer as much as possible a customised curriculum design. This section presents the proposed approach that extends the curriculum design methods using EDSF ontology described in the previous authors’ paper [11] by using knowledge units ranking and Bloom’s taxonomy learning levels for customised learning path building.

3.6.1 Customised Curriculum Design using EDSF Ontology

The input for the (customised) curriculum design is the intended competences set together with competences relevance or ranking for the desirable/target professional profile or job position. For individual learning path building, the individual competences can be assessed based on CV matching against the intended job position or professional profile, certification exam, or just a self-assessment questionnaire.

When a set of required competences is defined together with the relevance scores and required proficiency levels, the set of required knowledge topics can be extracted from individual competences (note, there exist multiple links from competence instances to a single knowledge topic) and ordered according to required proficiency level and relevance for further mapping to DS-BoK Knowledge Areas and Knowledge Units. The set of KAs and KUs defined for a specific competence set specifies the structure of the curriculum that can further 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 programme or course manager.

At the same time, the required proficiency level is scored for each KA and KU, which will define mastery levels and corresponding learning outcomes for the targeted education or training curriculum. When using EDSF ontology, it is a routine task to extract all required knowledge 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.

The MC-DS provides a set of templates for designing general purpose curricula composed of specified Learning Units, together with mastery levels defined for different types of programmes: Introductory, bachelor. 3 master levels are defined based on Bloom’s Taxonomy: Familiarity, Usage, and Assessment (refer to MC-DS [5] for details). When using MC-DS for customised curricula design, the competence scores/relevance defined in DSPP using a scale 0 to 9 can be easily mapped to MC-DS mastery levels [10]. 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.

The EDSF Toolkit and its outcome provide advice on the suggested curriculum structure that can be adjusted to the 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 platform providers and cloud based data labs, the presented approach can provide a specification for the required educational platform.

Table 1 below provides an example of KUs scores grouped into four DS-BoK Knowledge Area Groups for the competences defined for the DSP04 Data Scientist (only KU with the highest scores are included). Figure 6 illustrates the whole set of required KUs presented in the visual form of the tree map.

Table 1. Core KUs identified for DSP04 Data Scientist

<i>Data Analytics and Machine Learning (core)</i>		
KU01.02.02	Supervised Machine Learning	48

KU01.02.03	Unsupervised Machine Learning	48
KA01.01	Statistical methods for data analysis	47
KU01.01.07	Quantitative analytics	38
KU01.01.08	Qualitative Analytics	33
KU01.02.04	Reinforced learning	32
KA01.05	Predictive Analytics	32
KU01.01.09	Data preparation and preprocessing	24
<i>Data Management (core)</i>		
KA03.02	Data management systems	33
KU03.02.01	Data architectures (OLAP, OLTP, ETL)	33
KA03.01	General principles and concepts in Data Management and organisation	26
KA03.03	Data Management and Enterprise data infrastructure	26
KU03.01.02	Data Lifecycle Management	24
<i>Data Science Engineering (core)</i>		
KU02.01.07	NoSQL databases	40
KA02.02	Infrastructure and platforms for Data Science applications	40
KA02.03	Cloud Computing technologies for Big Data and Data Analytics	22
<i>Research Methods and Project Management (core)</i>		
KU04.01.05	Use cases analysis: research infrastructures and projects	32
KA04.01	Research Methods	27
KU04.01.02	Modelling and experiment planning	26
KU04.01.03	Data selection and quality evaluation	26

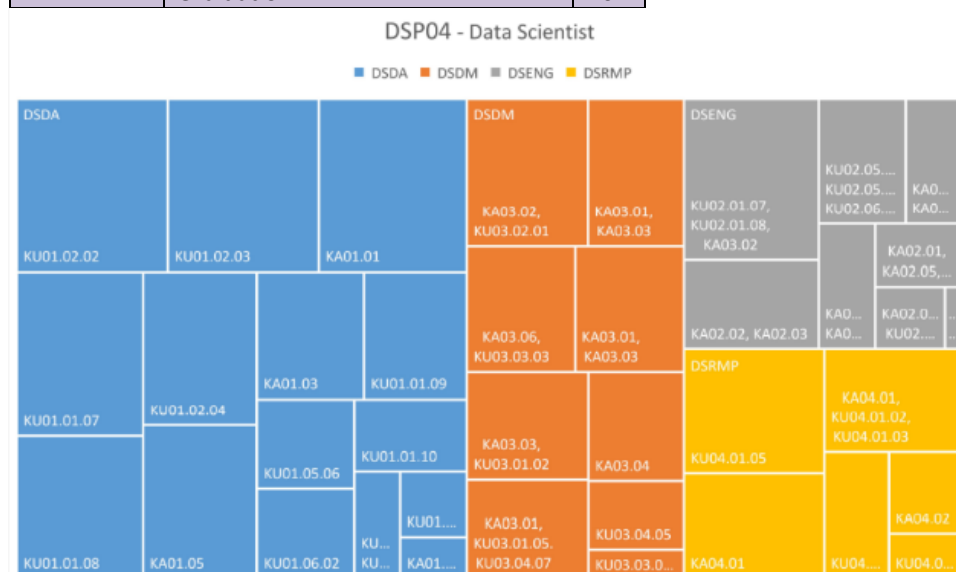


Figure 6. Example curriculum structure for DSP04 – Data Scientist

3.6.2 Defining Knowledge Units to include into the curriculum

Once the suggested Knowledge Units have been obtained, it is possible to combine them into educational courses, map them to courses defined in MC-DS or to existing courses, which are typically defined according to DS-BoK Knowledge Areas or Knowledge Units. EDSF ontology defines for these purposes the Course class, whose instances are directly connected to the KUs through the object property course. This allows for collecting relative scores for all KUs linked to the required curriculum.

When moving to a practical curriculum and courses design, it is important to define the courses relevance and their priority or sequence. The suggested courses content can be defined by KUs grouping based on their ontological similarity and difference. In a simple view, this defines the courses that need to be attended to achieve intended learning outcome and collect the necessary number of credits, in a classical education model. However, this doesn't solve the problem of the efficient programme planning or learning path design, what is especially important when designing a curriculum for workplace training, vocational education or self-education.

The Course class in the EDSF ontology can be used to calculate the course weight based on the integral score of the component KUs. This can be done by querying the ontology that will produce the list of associated courses for the required competence profile, sorted in descending order by weight. The course weight is calculated based on collecting all individual KU's scores linked to required competences, given the multiple relations and mapping between competences, knowledge topics in CF-DS, Knowledge Units in DS-BoK, and Learning Units in MC-DS. The course weights are normalized to 0-9 scale and aligned with the related competences relevance.

3.6.3 Applying Bloom's Taxonomy to curriculum structuring and course planning

Data Science programme structuring and course planning is an important stage in the practical curriculum implementation. The EDSF Toolkit supports the interactive curriculum design approach and courses planning. At this stage, the course weights are used to assign credit points to the planned courses in an academic curriculum. The course design application (programmed in Python) uses external csv files that contain the mapping between the courses and the related credit points. The association between weights and credits is assigned as follows, using the calculated course weight: low priority is given to courses that have weight less than 3, medium priority is given to courses whose weight is between 3 and 6, the higher priority is assigned to course intended for key competences. Figure 7 puts the candidate's competence profile and gaps (vertical bars) into the context of the target DSPP profile or vacancy (solid line); main competence gaps are marked with the circles.

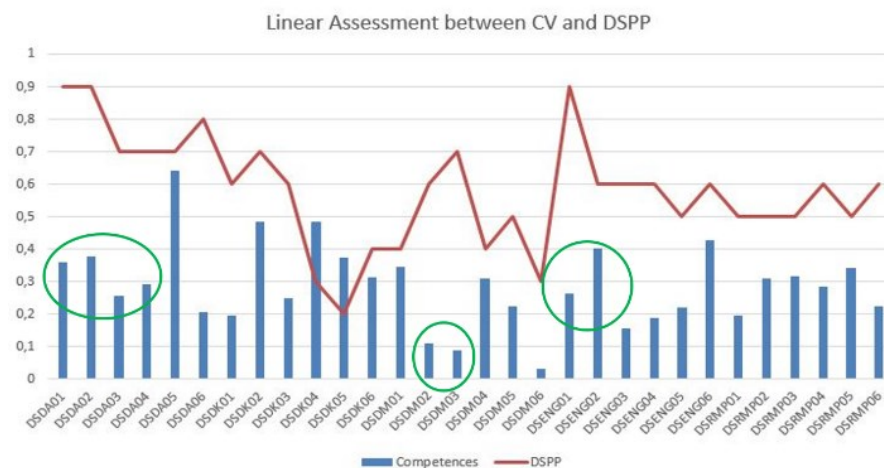


Figure 7. Candidate competences gap in the context of the target DSPP04 profile (in a linear coordinate comparing to spider diagram). Main competence gaps are marked with the circles.

At this stage, the Bloom's Taxonomy learning (cognitive) levels are applied to the courses duration and planning [28]. Courses that correspond to both larger identified gaps and higher priority are suggested to have longer duration or even are recommended to split into multiple periods (in practice 2-or 3) to allow for the learners' reflection and practical skills acquisition, which are time dependent.

The number of credits for a course is linked to the course weight but limited to 3 or 6 credits. If the course has a weight greater than 6, then it is split into two parts, whose sum of credits is equivalent to the total expected, and it is ensured that the two modules appear in two consecutive semesters in case of two years master programme. Furthermore, in defining the learning path, the number of courses for each semester is limited to a maximum of four.

Figure 8 provides an example of the two years Data Science master curriculum design that incorporates described above approach and methodology supported by the EDSF Toolkit. The first-year courses are targeted to create a strong background for core Data Science and Analytics courses. Core courses such as Data Mining and Machine Learning are split into two semesters and taught in two periods. Courses planning for vocational education and professional training will benefit from a similar approach, however using different time span.

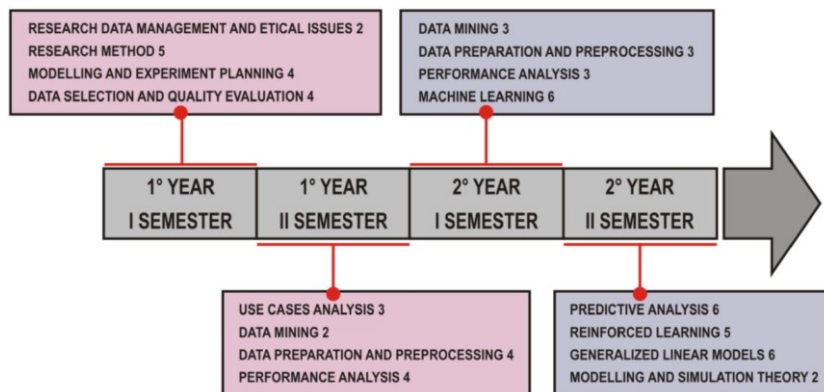


Figure 8. Example curriculum planning based on implied courses duration and DSPP profile proficiency level.

3.7 Conclusion and Further Developments

EDSF is a continuously evolving framework maintained by the community of educators and practitioners in Data Science and other data related technologies. EDSF provides a basis for defining Data Science competences, Body of Knowledge and Model Curriculum that can be used for designing customised curriculum for target competence profiles. With the publishing the new EDSF Release 4 (also referred to as EDSF2020), the framework became a mature product and currently counts multiple practical uses and is cited in multiple studies. The four EDSF parts describe Data Science Competence Framework, Body of Knowledge, Model Curriculum, and Data Science Professional Profiles. The new EDSF Part 5 (since Release 4) is intended to provide a practical guidance for universities, training organisations, data management and data steward team, and practitioners to define their Data Science curricula and courses selection, on the one hand, and for companies to better define a set of required competences and skills for their specific industry domain in their search for Data Science talents, on the other hand.

The authors are involved in multiple ongoing developments related to Data Science programs definitions and courses development, such as Vodafone Ukraine Data Science Academy, Data Science MBA programme for the Amsterdam Business School, as well as digital and data skills framework for the maritime industry in the framework of the EU funded MATES project. All such activities contribute to further EDSF development and will facilitate the EDSF Toolkit development.

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, 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.
- [8] 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
- [9] Demchenko, Yuri, et al, 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.
- [10] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong.
- [11] Yuri Demchenko, Luca Communiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, ISBN978-1-4503-6186-6/19/03.
- [12] David Donoho, 50 Years of Data Science, Journal of Computational and Graphical Statistics, Volume 26, 2017, Issue 4, pp 745-766, Published online: 19 Dec 2017 [online] <https://doi.org/10.1080/10618600.2017.1384734>
- [13] SP 800-145, The NIST Definition of Cloud Computing, NIST 2011 [online] <https://csrc.nist.gov/publications/detail/sp/800-145/final>
- [14] NIST SP 1500-1 NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions, September 2015 [online] <http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf>
- [15] OECD Skills Outlook 2019, Thriving in a Digital World, Published on May 09, 2019 [online] <https://www.oecd.org/education/oecd-skills-outlook-2019-df80bc12-en.htm>
- [16] 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
- [17] CCS, 2012 The 2012 ACM Computing Classification System. Available at <http://www.acm.org/about/class/class/2012>
- [18] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at <https://ec.europa.eu/esco/portal/#modal-one>
- [19] EN 16234-1 “e-Competence Framework”, CEN Standard 2019.
- [20] MATES Project: Maritime Alliance for fostering the European Blue Economy through a Marine Technology Skilling Strategy [online] <https://www.projectmates.eu/>
- [21] FAIRsFAIR Project: Fostering FAIR data practices in Europe [online] <https://www.fairsfair.eu/>
- [22] Data Science Lern- und Ausbildungsinhalte, Gesellschaft für Informatik, BMBD und Plattform Lernende Systeme, Arbeitspapeir, Dezember 2019 [online] https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/GI_Arbeitspapier_Data-Science_2019-12_01.pdf
- [23] Quoc Le and Tomas Mikolov, Distributed Representations of Sentences and Documents
- [24] Phillip Lord (2010) Components of an Ontology. Ontogenesis.
- [25] Matching CVs based on EDISON Data Science Competencies (CF-DS) [online] <https://github.com/EDISONcommunity/EDSFapps/tree/edsfcv>
- [26] Maijer, Mathijs, Matching CVs based on EDISON Data Science Competencies (CF-DS) and advanced text analysis methods, Project report, 2018. [online] <https://esc.fnwi.uva.nl/thesis/centraal/files/f1532411291.pdf>
- [27] Martin F Porter. “An algorithm for suffix stripping”. In: Program 14.3 (1980), pp. 130-137.
- [28] Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company.

4 Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge⁵

The 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 require a multi-disciplinary approach combining a 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 forms, also providing sufficient data labs facilities for practical training. This section discusses an approach to building a 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.

4.1 Demand for Data Science Competences and Customisable Curriculum

Sustainable development of the modern data driven economy requires a 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 the multi-disciplinary nature of Data Science and its application domains. At the 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 an effective combination of theoretical, practical and workplace skills.

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 a 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 the European economy and a challenge for universities.

Business Higher Education Forum (BHEF) published in 2017 two important reports in cooperation with PriceWaterhouseCoopers, IBM and Burning Glass Technologies [12, 13] that studied the 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.

A recent OECD report [14] confirms the urgent need to address data and general digital skills for all types of workforce and economy sectors. 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 the courses is a key to adopting future skills management and capacity building models.

4.2 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:

⁵ Based on paper Yuri Demchenko, Luca Communiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, ISBN978-1-4503-6186-6/19/03.

- 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 been 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].

4.3 EDSF Data Model and Ontology

The 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 described in section III. The initial definition of EDSF was made in the form of Excel workbooks and tables which provided a good way of documenting but was difficult to use for practical applications [10].

4.3.1 EDSF Data Model

Currently, EDSF toolkit contains a number of datasets representing different components of the EDSF and mapping between them. Future EDSF development will formally define the ontologies related to the EDSF components and related dictionaries.

Figure 1 illustrates the relation between different data sets and ontologies comprising EDSF. The CF-DS is structured along four dimensions (similar to European e-Competence Framework e-CFv3.0 [15]) that include (1) competence groups, (2) individual competences definition, (3) proficiency levels, and (4) corresponding knowledge and skills. In this context, each individual competence includes a set of required knowledge topics and a set of skills type A and skills type B. Such CF-DS structure allows for competence based curriculum design where competences can be defined based on the professional profile (see DSPP [6] for mapping between professional profiles and competences) or target learners group when designing a full curriculum, or based on competence benchmarking for tailored training to address identified competences and knowledge gaps.

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 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 can further be mapped to Model Curriculum Learning Units defined as individual courses and KAG related courses groups.

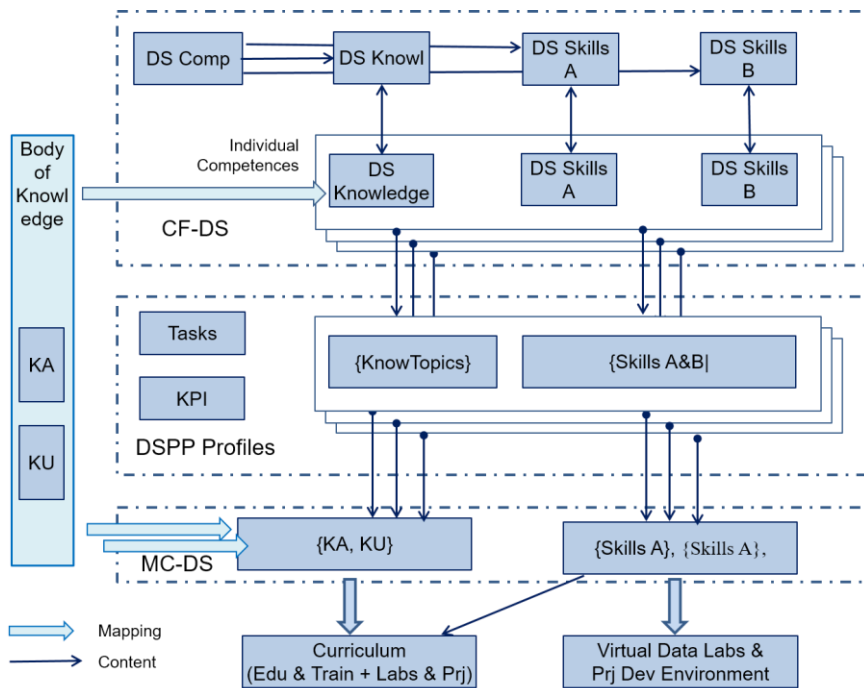


Figure 1: EDSF Data Model and customised curriculum design for target professional group(s)

At the same time, the required proficiency level is scored for each KA and KU, which will define mastery levels and corresponding learning outcomes for the targeted education or training. The following mastery levels are defined (using workplace terminology that can be easy mapped to mastery levels defined in MC-DS):

A - Awareness

- 1) Understand Terminology
- 2) Understand the Principles
- 3) Apply principles
- 4) Understand the Methods

U - Use/Application

- 5) Apply basics
- 6) Supervised use
- 7) Unsupervised Use

P - Professional/Expert

- 8) Development of applications using a wide range of technologies
 - 9) Supervise project development, a team of professionals,
- where borderline mastery levels 4 and 7 actually belong to both higher level and lower level groups.

4.3.2 Definition of the EDSF Ontology

In the new EDSF Release 3 (EDSF2018) [2], the CF-DS and DS-BoK are expressed in the form of ontology that is also linked to DSPP profiles definition. The ontology provides an effective format for representing rich relations between EDSF components in the form of instances, classes and properties, it also allows the 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 the 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 an example of the DSDA01 competence that is defined as “Effectively use a 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 KSDSA*, Skills Group A SDSDA* and Skills Group B SDSA*.

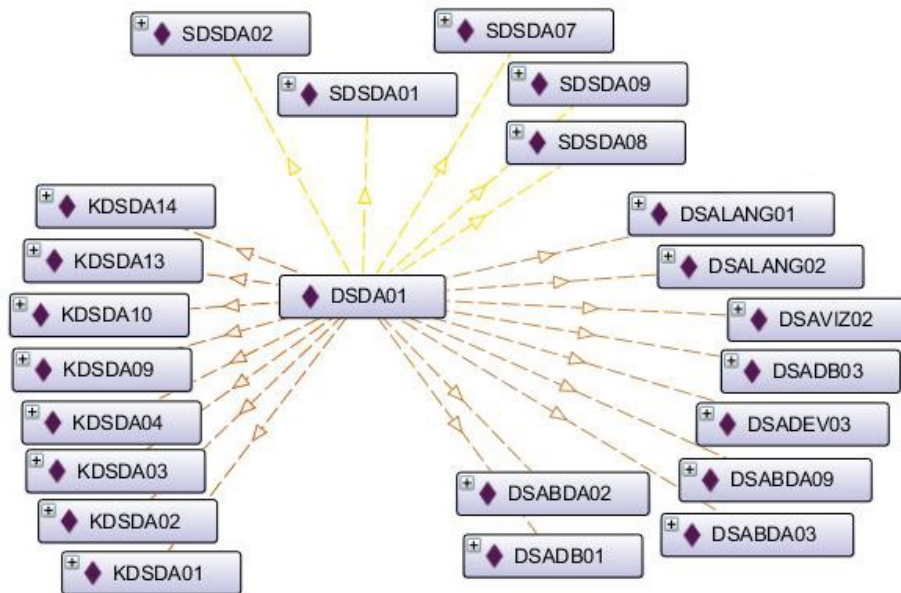


Figure 2: Example DSDA01 Competence and its properties.

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.

4.4 Data Science Curriculum Design using EDSF Ontology

This section describes the workflow of using EDSF for curriculum design for a selected/intended set of competences that are required for (1) a specific Data Science professional profile defined based on the DSPP document, or (2) an individual training program defined based on competence assessment and identified gaps. The individual competence assessment can be done based on CV matching against the intended job position or professional profile. It can also be done based on the certification exam or just a self-assessment questionnaire. The outcome of this process is either a level of matching or a competence gap that can be used for suggesting necessary training program or tailored curricula. 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, the required proficiency level is scored for each KA and KU, which will define mastery levels and corresponding learning outcomes for the targeted education or training. When using MC-DS as a template

for designing customised curriculum, the proficiency levels (using a scale 0 to 9) can be easily 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 knowledge 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 an example of relations between EDSF components when extracting required Knowledge Units for the 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.

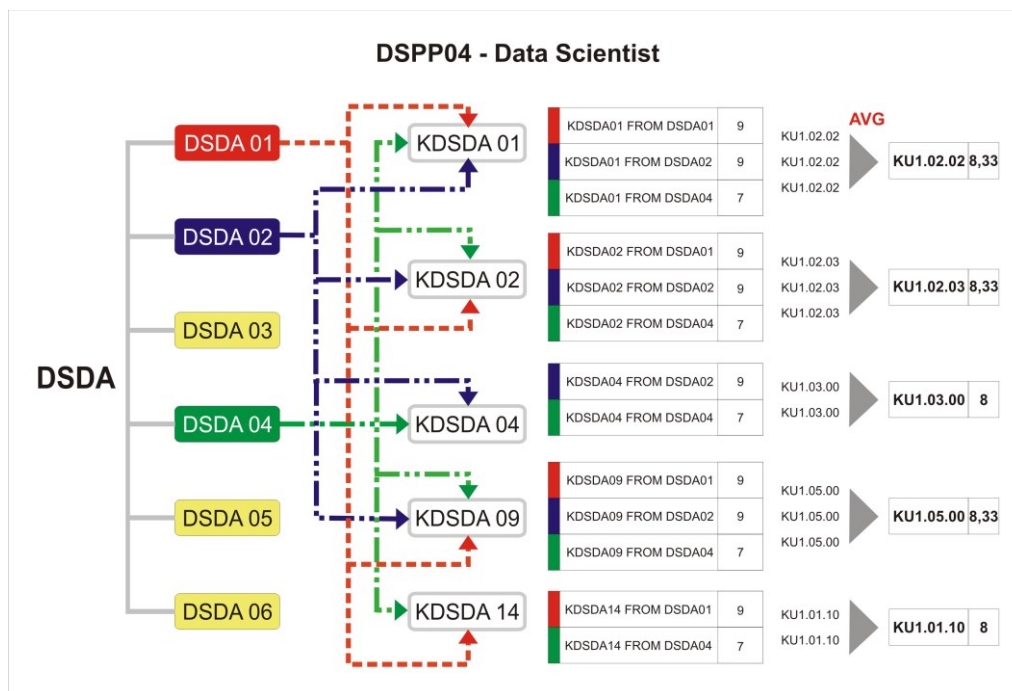
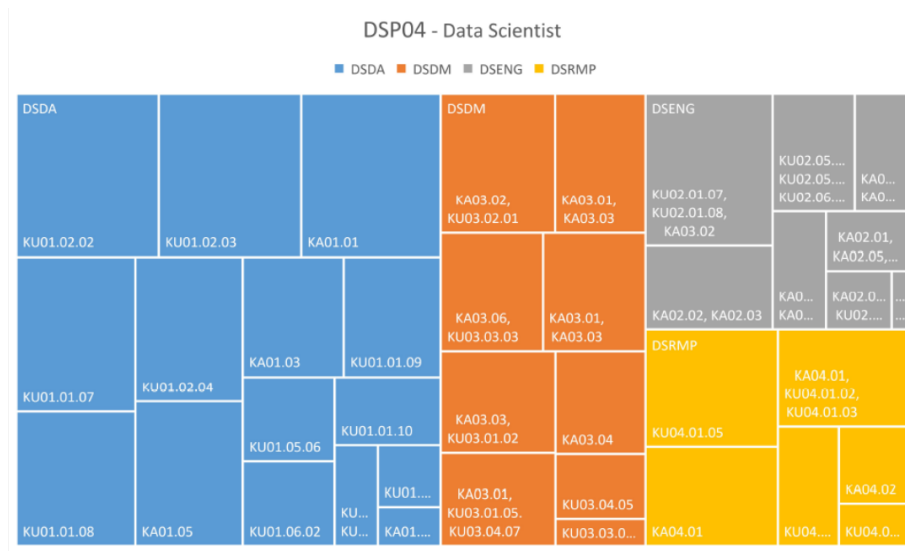
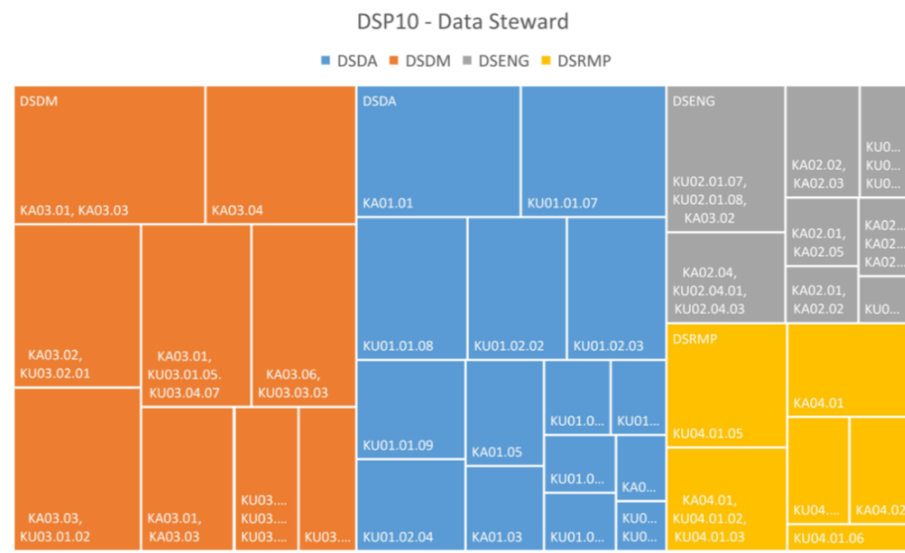


Figure 3: Extracting required Knowledge Units from EDSF ontology.

The same process is applied to other competence groups relevant to specific professional profiles or competence gaps. Figure 4 (a) and (b) shows an example of the suggested curriculum structure for two professional profiles: DSP04 – Data Scientist and DSP10 – Data Steward. The diagrams reflect the relative structure of the curriculum where Data Scientist has the major part of the Data Analytics courses (DSDA - blue) followed by necessary knowledge in Data Management (DSDM - orange), and the 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 provide advice on the suggested curriculum structure that can be adjusted to the 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 platform providers and cloud based data labs, the presented application can provide a specification for the required educational platform.

4.5 Cloud based DSEE and Virtual Data Labs: IDE, Tools and Datasets

As outcome of the curriculum design, the application may provide suggestions on the set of Skills type A will define Learning Outcomes and Skills type B will provide advice on the required hands on training and practical project development environment and platform (refer to CF-DS document). As an example, the Data Scientist curriculum should include the following elements to achieve the necessary skills Type B:

- Python (or R) and corresponding data analytics libraries
- NoSQL and SQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, MS SQL, My SQL, PostgreSQL, etc.)
- Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
- Real time and streaming analytics systems (Flume, Kafka, Storm)
- Kaggle competition, resources and community platform

- Visualisation software (D3.js, Processing, Tableau, Julia, Raphael, etc.)
- API management and web scrapping
- Git versioning system as a general platform for software development
- Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others
- Cloud based Big Data and data analytics platforms and services, including large scale storage systems.

Using cloud resources to build an effective and up-to-date professional Data Science education environment is inevitable with current fast technology development and required computational performance that can be requested on-demand.

Major Cloud Service Providers (CSP) provides a wide range of data analytics and business analytics services and platforms that can be equally used by big, small and medium companies and individuals on the pay-per-use basis. In addition to the possibility to use the same resources for education and training purposes, the major CSPs also provide designated education and self-training resources that are, in many cases, supported also educational grants for students and teachers.

The following cloud based resources from the major cloud providers can be used to build hybrid DSEE and VDLabs (in addition to regular compute and storage resources):

- Microsoft Azure Data Lakes Analytics, Power BI, HDInsight Hadoop as a Service, others
- AWS Elastic MapReduce (EMR), QuickSight, Kinesis and a wide collection of open datasets
- IBM Data Science Experience, Data Labs, Watson Analytics.

An important component of Data Science education is educational datasets that often need to be provided with their specific applications. While many educational datasets are available from mentioned above cloud platforms, from community run Kaggle (<https://www.kaggle.com/>) and UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/index.php>), the use of cloud based VDLabs allows to instantiate the whole experimental setup or environment together with used data sets in case of specific domain focused education or training.

4.6 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 a fully enumerated definition of EDSF components and individual units. Besides defining academic components of the effective and consistent curriculum, EDSF also provides 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.

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 al, 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] Going Digital in a Multilateral World, OECD Report on Meeting of the OECD Council at Ministerial Level, Paris, 30-31 May 2018 [online] [https://one.oecd.org/document/DSTI/CDEP/GD\(2018\)2/en/pdf](https://one.oecd.org/document/DSTI/CDEP/GD(2018)2/en/pdf)
- [15] CCS, 2012 The 2012 ACM Computing Classification System. Available at <http://www.acm.org/about/class/class/2012>
- [16] Phillip Lord (2010) Components of an Ontology. Ontogenesis.
- [17] 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.
- [18] Quoc Le and Tomas Mikolov, Distributed Representations of Sentences and Documents
- [19] Jey Han Lau and Timothy Baldwin - An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation
- [20] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at <https://ec.europa.eu/esco/portal/#modal-one>

5 Big Data Platforms and Tools for Data Analytics in the Big Data and Data Science Curricula⁶

5.1 Introduction

Modern Data Science and Business Analytics applications extensively use Big Data infrastructure technologies and tools which are commonly cloud based and are available at all major cloud platforms. Knowledge and ability to work with modern Big Data platforms and tools to effectively develop and operate data analytics applications is required from the modern Data Science practitioners. Including Big Data Infrastructure topics into the general Data Science curriculum will help the graduates to be easily integrated into the future workplace.

This paper refers to and effectively uses the EDISON Data Science Framework (EDSF), initially developed in the EDISON Project (2015-2017) and currently maintained by the EDISON community [1, 2]. The EDSF provides a general framework for Data Science education, curriculum design and competences management what has been discussed in the author's previous works [3, 4, 5]. Big Data Infrastructure Technologies (BDIT) is a part of the defined in EDSF the Data Science Engineering Body of Knowledge (DSENG-BoK) and Model Curriculum (MC-DSENG) described in details below.

This paper is focused on the definition of the Data Science Engineering Body of Knowledge and Big Data Infrastructure Technologies for Data Analytics (BDIT4DA) course. The paper provides a brief overview of the Big Data infrastructure technologies and existing cloud based platforms and tools for Big Data processing and data analytics which are relevant to the BDIT4DA course. The focus is given on the cloud based Big Data infrastructure and analytics solutions and, in particular, on understanding and using the Apache Hadoop ecosystem as the major Big Data platform, its main functional components MapReduce, Spark, HBase, Hive, Pig, and supported programming languages Pig Latin and HiveQL.

Knowledge and basic experience with the major cloud service providers (e.g., Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform GCP) as well as the Cloudera Hadoop Cluster or Hortonworks Data Platform are important to develop the necessary knowledge and strong practical skills. These topics need to be included into both lecture course and hands on practice.

5.2 Data Science Engineering BoK and Model Curriculum

5.2.1 DSENG Model Curriculum Components

Data Science Engineering Knowledge Group builds the ability to use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management. It includes Knowledge Areas that cover: software and infrastructure engineering, manipulating and analysing complex, high-volume, high- dimensionality data, structured and unstructured data, cloud based data storage and data management.

Data Science Engineering includes software development, infrastructure operations, and algorithms design with the goal of supporting Big Data and Data Science applications in and outside the cloud. The following are commonly defined Data Science Engineering Knowledge Areas (as part of KAG02-DSENG):

- KA02.01 (DSENG/BDI) Big Data infrastructure and technologies, including NoSQL databased, platforms for Big Data deployment and technologies for large-scale storage;
- KA02.02 (DSENG/DSIAPP) Infrastructure and platforms for Data Science applications, including typical frameworks such as Spark and Hadoop, data processing models and consideration of common data inputs at scale;
- KA02.03 (DSENG/CCT) Cloud Computing technologies for Big Data and Data Analytics;
- KA02.04 (DSENG/SEC) Data and Applications security, accountability, certification, and compliance;
- KA02.05 (DSENG/BDSE) Big Data systems organization and engineering, including approach to big data analysis and common MapReduce algorithms;

⁶ Based on paper Yuri Demchenko, Big Data Platforms and Tools for Data Analytics in the Data Science Engineering Curriculum, Proc 2019 3rd International conference on Cloud and Big Data (ICCBDC 2019), August 28-30, 2019, Oxford, UK

- KA02.06 (DSENG/DSAPPD) Data Science (Big Data) application design, including languages for big data (Python, R), tools and models for data presentation and visualization;
- KA02.07 (DSENG/IS) Information Systems, to support data-driven decision making, with a focus on data warehouse and data centers.

The DS-BoK provides the mapping of the DS-BoK to existing classifications and BoKs: ACM Computer Science BoK (CS-BoK) selected KAs [7], Software Engineering BoK (SWEBoK) [8], and related scientific subjects from CCS2012 [6]: Computer systems organization, Information systems, Software and its engineering.

5.2.2 DSENG/BDIT - Big Data infrastructure technologies course content

Big Data infrastructures and technologies shape many Data Science applications. Systems and platforms behind Big Data differ significantly from traditional ones due to specific challenges of volume, velocity, and variety of data that need to be supported by data storage and transformation. Data Lakes and SQL/NoSQL databases must be included in the DSENG curriculum

Deployment of Data Science applications is usually tied to one of the most common platforms, such as Hadoop or Spark, hosted either on private or public clouds. The applications workflow must be linked to a whole data processing pipeline, including ingestion and storage for a variety of data types and sources. Data Scientists should have a general understanding of data and application security aspects in order to properly plan and execute data-driven processing in the organization. This module should provide an overview of the most important security aspects, including accountability, compliance and certification.

5.2.3 Data Management and Data Stewardship in the Big Data and Data Science Curriculum

Data Management and Governance (DMG) [9, 10], although belonging to different KAG4-DSDM, must accompany the DSENG courses and short overview of the DMG common practices must be included in the BDIT curriculum. This should also include the introduction of the FAIR data principles (data must be Findable, Accessible, Interoperable, Reusable) [11] that are growingly adopted by the research community and recognised by the industry. Data Stewardship is a DMG application domain that combines general and subject domain data management, ensuring the FAIR principles are incorporated into the organisational practice.

5.3 Platforms for Big Data Processing and Analytics

This section describes what platforms can be used for teaching the BDIT4DA course and other courses in the Data Science Engineering curricula requiring processing Big Data. The section describes the Hadoop Ecosystem and its main components and functionalities and provides information about cloud based Big Data Infrastructure and analytics platforms from the major cloud providers.

5.3.1 Essential Hadoop Ecosystem Components

Hadoop is commonly used as a main platform for Big Data processing, it includes multiple components and applications developed by the Apache Open Source Software community, with rich functionality to support all processes and stages in the data processing workflow/pipeline. Giving a general understanding and basic experience with the Hadoop applications and tools is an important part of the practical activity and assignments in the BDIT4DA course. Figure 1 below illustrates the Hadoop main components and a few other popular applications for data processing [12, 13].

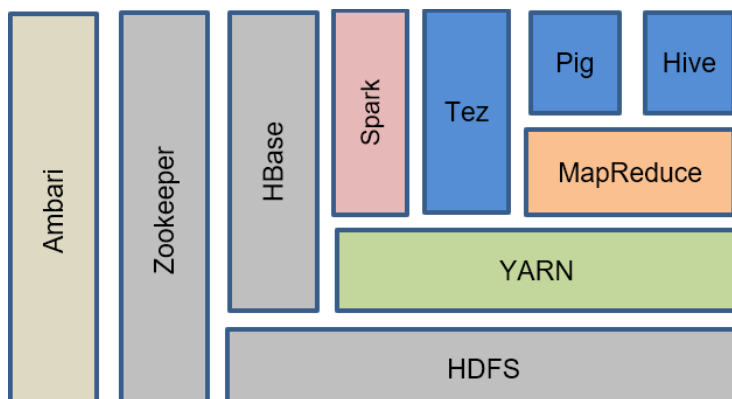


Figure 1: Main components of the Hadoop ecosystem

The following main Hadoop applications constitute the foundation of the Hadoop ecosystem and provide basis for other applications.

- HDFS: Hadoop Distributed File System** optimized for large scale storage and processing of data on commodity hardware.
- MapReduce:** A YARN-based system for parallel processing of large data sets.
- YARN:** A framework for job scheduling and cluster resource management.
- Tez:** A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.

Other Hadoop-related projects at Apache that provide a rich set of functionalities for data processing during the whole data lifecycle:

- Hive:** A data warehouse system that provides data aggregation and querying.
- Pig:** A high-level data-flow language and execution framework for parallel computation.
- HBase:** A distributed column oriented database that supports structured data storage for large tables
- Spark:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- Mahout:** A scalable machine learning and data mining library.
- Solr:** Open source enterprise search platform that uses lucene as indexing and search engine.
- Oozie:** Server-based workflow scheduling system to manage Hadoop jobs.
- Ambari:** A web-based tool for provisioning, managing, and monitoring YARN jobs and Apache Hadoop clusters
- Hue:** A user graphical interface providing full functionality for programming Hadoop applications, including dashboard, data upload/download, visualisation.

5.3.2 Hadoop Programming Languages

Introducing multiple Hadoop programming options is essential to allow future integration of the Hadoop platform and tools into research and business applications. Hadoop is natively programmed in Java, with current support for Scala by many applications. There is also support for Hadoop API calls from many popular programming and data analytics IDE and tools for R, Python, C, .NET. Specific for Hadoop are query languages to work with HBase, Hive, Pig.

Introducing multiple Hadoop programming options is essential to allow future integration of the Hadoop platform and tools into research and business applications. Hadoop is natively programmed in Java, with current support for Scala by many applications. There is also support for Hadoop API calls from many popular programming and data analytics IDE and tools for R, Python, C, .NET. Specific for Hadoop are query languages to work with HBase, Hive, Pig as shown in Figure 2.

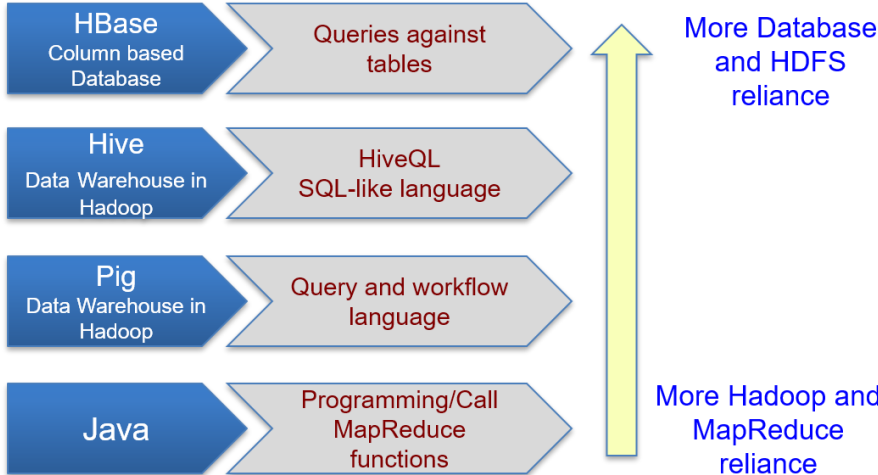


Figure 2: Query languages for Hadoop

Hive Query Language (HiveQL or HQL) [14]: Provides higher-level data processing language, used for Data Warehousing applications in Hadoop. Query language is HiveQL, variant of SQL, tables are stored on HDFS as flat files. HiveQL facilitates large-data processing that compiles down to Hadoop jobs.

Pig Latin [15] is a scripting language used for large-scale data processing system to describe a data processing flow. In fact, Pig Latin has similarity to HiveQL query commands with additional flow control commands. Similar to HiveQL, it compiles down to Hadoop jobs and relies on MapReduce or Tez for execution.

5.3.3 Cloud based Big Data Platforms

Major cloud platforms Amazon Web Services (AWS) [16], Microsoft Azure [17], Google Cloud Platform (GCP) [18] provide a rich set of Big Data services and applications.

AWS Big Data stack includes such services as Elastic MapReduce (EMR) which is a hosted Hadoop platform for Data Analytics, Amazon Kinesis is a managed service for real-time processing of streaming big data, Amazon DynamoDB - highly scalable NoSQL data stores, Amazon Aurora - a scalable relational database, and Amazon Redshift - fully-managed petabyte-scale data warehouse. Separately provided is the Machine Learning stack with a number of services. All services and tools are accessible from the AWS Console and can be programmed via Command Line Interface (CLI), where the former provides all necessary functionality to program, deploy and operate complex business applications by integrating all necessary components into one data processing pipeline.

Microsoft Azure provides well integrated and supported with the development tools the Big Data and Analytics stack that includes such services as HDInsight which is Hortonworks based Hadoop platform, Data Lake Storage and Data Lake Analytics, CosmosDB multi-format NoSQL database, and other services.

Google Cloud provides general cloud services and a set easy configured Big Data services such as BigQuery column based NoSQL database, Google Spanner Big SQL database, and Machine Learning stack with well-defined API that support the whole data analytics

5.4 Example BDIT4DA courses and experience

This section provides an example of three Big Data Infrastructure and Technologies for Data Analytics courses that can be adjusted to different academic or training programmes. BDIT4DA includes lectures, practice/hands-on labs, projects and such engaging activities as literature study and seminars. The courses should beneficially include a few guest lectures to expose the students to external experts and real practices.

5.4.1 BDIT4DA Course for Big Data Engineering Masters

5.4.1.1 BDIT4DA Lectures

Lectures must provide a foundation for understanding the whole BDIT4DA technology domain, available platforms, tools and link other course activities. However, form and technical level must be adjusted to the incumbent programme, for example distinguishing Computer Science and MBA programs. The same should be related to the selection of practical assignments and used tools and programming environment.

The following example is the set of lectures that have been developed and taught by the authors. Depending on programme configuration and scheduling, the mentioned below topics can be delivered in the form of sessions that can combine lectures (2-3 hrs.), practice (2-4 hrs.), and interactive activities such as literature review, and project progress presentation.

Lecture 1 Cloud Computing foundation and economics.

Cloud service models, cloud resources, cloud services operation, multitenancy. Virtual cloud datacenter and outsourcing enterprise IT infrastructure to cloud. Cloud use cases and scenarios for the enterprise. Cloud economics and pricing model.

Lecture 2 Big Data architecture framework, cloud based Big Data services

Big Data Architecture and services. Overview of major cloud based Big Data platforms: AWS, Microsoft Azure, Google Cloud Platform (GCP). MapReduce scalable computation model. Overview Hadoop ecosystem and components.

Lecture 3 Hadoop platform for Big Data analytics

Hadoop ecosystem components: HDFS, HBase, MapReduce, YARN, Pig, Hive, Kafka, others.

Lecture 4 SQL and NoSQL Databases

SQL basics and popular RDBMS. Overview NoSQL databases types. Column based databases and their use (e.g. HBase). Modern large scale databases AWS Aurora, Azure CosmosDB, Google Spanner.

Lecture 5 Data Streams and Streaming Analytics

Data streams and stream analytics. Spark architecture and components. Popular Spark platforms, DataBricks. Spark programming and tools, SparkML library for Machine Learning.

Lecture 6 Data Management and Governance and Stewardship

Enterprise Big Data Architecture and large scale data management. Data Governance and Data Management. FAIR Principles in data management.

Lecture 7 Big Data Security and Compliance.

Big Data Security challenges, Data protection. cloud security models. Cloud compliance standards and cloud provider services assessment. CSA Consensus Assessment Initiative Questionnaire (CAIQ) and PCI DSS cloud security compliance.

Figure 3 illustrates in the form of 2D map relations between the proposed lecture and practice topics and the Big Data systems and applications infrastructure components. Using this kind of illustration provides good guidance for designing courses with better practical orientation and at the same time provides advice to students for future self-study.

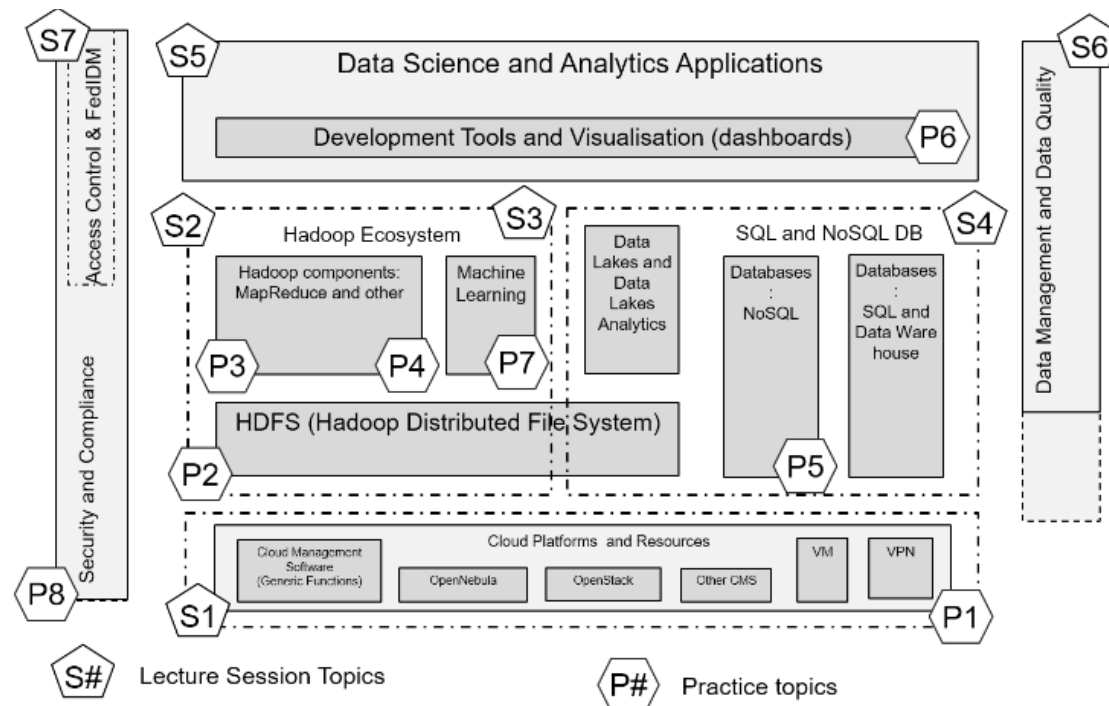


Figure 3: BDIT4DA course lectures and practices topics map.

5.4.1.2 BDIT4DA Practice

Recommended practice includes working with the main Hadoop applications and programming simple data processing tasks. Different Hadoop platforms can be used for running practical assignments using either dedicated Hadoop cluster installations (e.g. Cloudera Hadoop Cluster [19], Hortonworks Data Platform [20], or cloud based AWS Elastic MapReduce (EMR), or Azure HDInsight platform). Students can also be recommended to install personal single host Hadoop cluster using either Cloudera Starter edition or Hortonworks Sandbox which are available for both VirtualBox and VMware.

The following are example topics for practice and hands-on assignments.

Practice 1: Getting started with the selected cloud platform. For example, Amazon Web Services cloud; cloud services overview EC2, S3, VM instance deployment and access.

Practice 2: Understanding MapReduce, Pregel, and other massive data processing algorithms. Wordcount example using MapReduce algorithm (run manually and with Java MapReduce library).

Practice 3: Getting started with the selected Hadoop platform. Command line and visual graphical interface (e.g. Hue), uploading, downloading data. Running simple Java MapReduce tasks.

Practice 4: Working with Pig: using simple Pig Latin scripts and tasks. Develop Pig script for programming Big Data workflows. This can also be done as a part of the practical assignment on Pig.

Practice 5: Working with Hive: Run a simple Hive script for querying Hive data base. Import external SQL database into Hive. Develop Hive script for processing large datasets. This can also be a part of the practical assignment on Hive.

Practice 6: Streaming data processing with Spark, Kafka, Storm. Using simple task to program Spark jobs and using Kafka message processing. The option for this practice can also use Databricks platforms that provides a good tutorial website.

Practice 7: Creating a dashboard and data visualisation. Using tools available from the selected Hadoop platform to visualise data, in particular using results from Practice 5 or 6 that deal with large datasets where a dashboard is necessary

Practice 8: Cloud compliance practicum. This practice is important for the students to understand the complex compliance issues for applications run on the cloud. Using Consensus Assessment Initiative Questionnaire (CAIQ) tools.

5.4.2 Course for Data Science Masters

In contrast to the Big Data Engineer example, a course for Data Scientists spends more time on the algorithm design aspect. All basic tool and concepts are introduced, but less time is spent on topics related to security and governance.

First five lectures have corresponding laboratory sessions. Afterwards, students begin working on group projects on datasets of their choice, applying concepts, technologies, and tools from lectures. Progress in projects is presented at plenary presentations sessions, in the middle and at the end of the course.

To further motivate and guide students, 1 or 2 guest lectures with practitioners from the industry are organized. The can be scheduled any time after Lecture 5. In some cases, it might also be scheduled together with Lecture 1.

5.4.2.1 Lectures

Lecture 1: Introduction to data intensive systems and use cases. Data as 4th paradigm of science. Increasing focus on data collection, data architectures, data centers. Use cases in search, commerce, healthcare, energy.

Lecture 2: Hadoop 101 and Functional abstraction. Introductory but fully functioning MapReduce program in Python with execution from command line.

Lecture 3: MapReduce. A detailed description of file splitting, mapping, combining, shuffling, reducing, and storage of results.

Lecture 4: Hadoop Architecture. Resource management, permanent and temporary storage, batch processing, real-time processing, higher-level tools.

Lecture 5: MapReduce algorithms and patterns. Counting, summing, and averaging. Processing multiline input. Random sampling. Search Assist. Inverted index.

Lecture 6: HBase and other NoSQL databases. Alternative permanent storage for big data. CAP/PACELC theorems. Interaction between Hadoop/MapReduce and NOSQL databases.

Lecture 7: First project presentation. Focused on choice of dataset, data preprocessing, identification of interesting problems.

Lecture 8: Spark (RDD based). Data model. Programming model, actions, transformations, other operations. Architecture.

Lecture 9: Spark (SQL/other structures/Mllib). Alternative programming models, advantages and drawbacks. Incorporating existing libraries in the programming workflow.

Bonus: 1-2 industrial guest lectures. Usually focusing on data quality and data workflow in the industry.

Lecture 10: Final project presentations. Focused on MapReduce implementation of identified problems. Performance tuning.

5.4.2.2 Practice and project development

Lab 1: refresh Bash knowledge, setup Docker and Hortonworks Sandbox. Ensures that students have a working test environment on their laptops.

Lab 2: recreate steps from the lecture (system setup, file copying, running ready MRJob and Hive examples). Ensures that students can correctly execute examples in the book/lecture.

Lab 3: introduce modifications to MRJob based program on the Sandbox. Ensures that students understand basic concepts related to MapReduce programming.

Lab 4: setup Hadoop from scratch on a VM (not Sandbox). Ensures that students understood Hadoop architecture.

Lab 5: in-depth analysis of typical algorithms and patterns in groups. Ensures that students understand the details of MapReduce programming

After five laboratory sessions, students work on **group projects**. They are still encouraged to come on a regular basis to laboratory sessions where they can discuss and get support with any technical problems they meet.

5.4.3 Big Data Infrastructure Technologies (BDIT) Course for MBA in Big Data

Big Data and Data Analytics tools are important part of the business supporting infrastructure and services which are growingly cloud based. The specifics of the MBA DataScience groups is the diverse background of the students from economics and business to Computers Science and engineering. The main goal of the BDIT course is to provide knowledge sufficient for future business managers to make assessments and advice development of necessary services in their future organisations. The practical work is entirely based on using cloud based applications and tools. The course also includes project where the students working in groups need to deliver the design of the cloud based Big Data infrastructure supporting the business processes of their hypothetical company.

5.4.3.1 Lectures

BDIT lectures include a subset of topics outlined in section 5.4.1 but are enriched with examples and closely linked to practices and labs.

Lecture 1 Cloud Computing foundation and cloud economics: Provides basic for understanding and working with clouds.

Lecture 2 Big Data architecture framework, cloud based Big Data services: Overview of cloud based Big Data platforms and tools, including AWS, Azure and Google Cloud Platform.

Lecture 3 MapReduce and Hadoop platform: Introduce the Hadoop ecosystem and main components; an example of use **Lecture 4** Spark and Streaming Analytics: Including data structure, programming with Scala.

Lecture 5 SQL and NoSQL Databases: Database classification and types, Cloud based big data bases, Hadoop based HBase, Hive

Lecture 6 Data Management and Governance: based on DAMA DMBOK, extended with FAIR and QA methods

Lecture 7 Big Data Security and Compliance: Cloud data security services, access control, CSA Compliance framework.

5.4.3.2 Practice and Project Development

The practice covers major aspects of working with two main cloud platforms AWS and Microsoft Azure, starting with AWS as presenting a more generic cloud services model, and following with Microsoft Azure as providing better alignment with Business Analytics processes. The following topics were included in the course:

Practice 1: Getting started with Amazon Web Services cloud

Practice 2: AWS services EC2, S3 deployment and access.

Practice 3: Amazon Elastic MapReduce (EMR). Running MapReduce wordcount example manually and using EMR.

Practice 4. AWS Aurora scalable SQL database, deployment and simple query exercises.

Practice 4. Getting started with Microsoft Azure cloud, Storage and Compute services, instances deployment

Practice 5. Azure HDInsight business analytics platform, deployment and Hadoop cluster visual interface. Running simple Spark examples.

Practice 6: Cloud compliance practicum using CSA Consensus Assessment Initiative Questionnaire (CAIQ) tools.

5.5 Conclusion and Recommendations

The described above BDIT4DA course has been taught by the authors in different programmes and different installations: campus face-to-face teaching, part-time evening lecturing and practice, and remote lecturing. Experience confirms that in general lecture materials can be used the same, given that there is no single textbook for the course. However, practice must be adopted to the hosting master programme, student background and lecture-practice scheduling. An important aspect of this course and any other course related to Data Science is to develop in students a kind of data-centric approach and thinking. This aspect is related to the Data Science professional skills (“Thinking and acting like Data Scientist”), which are defined in the EDISON Data Science Framework [2] and discussed in the authors’ publications [4, 5, 21, 22].

The presented in this paper general approach and practical experience in teaching the Big Data Infrastructure Technologies for Data Analytics is based in the EDISON Data Science Framework, which is widely used by universities, professional training organisations and certification organisations, providing valuable feedback for further framework development and continuous courses evolution. The presented work is also based on the long author’s experience in teaching cloud computing technologies [22] that provide computational and infrastructure basis for Big Data technologies. The Cloud Computing curriculum design used the technology maps and linked professional profiles, together with the Bloom’s Taxonomy, to design the customisable curriculum. Such an approach has been developed into the EDSF and its main components.

Academic education or professional training must provide strong basis for graduates and trainees to continue their further self-study and professional development in conditions of the fast developing technologies and agile business environment adopted by a majority of modern companies. To achieve this, the Data Science curriculum needs to be supported by professional skills development courses such as to develop general 21st Century skills and specific Data Science workplace skills.

One of the general skills for data workers is considered Data Management and Governance and specifically Research Data Management and Stewardship adopting FAIR data principles, which is a part of the authors’ cooperation in the FAIRsFAIR project [23].

The EDSF maintenance and continuous development as well as collection of the best practices in Data Science education and training is supported and coordinated by the EDISON community, in cooperation with national and EU projects as well as supported by the Research Data Alliance (RDA) Interest Group on Education and Training on Handling Research Data (IG-ETHRD) [24]. Participation and contribution to both the IG-ETHRD and EDSF Community Initiative are open and free.

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] Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, ISBN978-1-4503-6186-6/19/03
- [4] Demchenko, Yuri, et al, 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.
- [5] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, Proc. The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong.
- [6] The 2012 ACM Computing Classification System [online] <http://www.acm.org/about/class/class/2012>
- [7] ACM and IEEE Computer Science Curricula 2013 (CS2013) [online] <http://dx.doi.org/10.1145/2534860>
- [8] Software Engineering Body of Knowledge (SWEBOK) [online] <https://www.computer.org/web/swebok/v3>
- [9] Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [online] <http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf>
- [10] Data Maturity Model (DMM), CMMI Institute, 2018 [online] <https://cmmiinstitute.com/data-management-maturity>
- [11] Barend Mons, et al, The FAIR Guiding Principles for scientific data management and stewardship [online] <https://www.nature.com/articles/sdata201618>
- [12] Apache Hadoop [online] <https://hadoop.apache.org/>

- [13] Hadoop Ecosystem and Their Components – A Complete Tutorial [online] <https://data-flair.training/blogs/hadoop-ecosystem-components/>
- [14] Apache Hive Tutorial [online] <https://cwiki.apache.org/confluence/display/Hive/Tutorial>
- [15] Apache Pig Tutorial [online] <https://data-flair.training/blogs/hadoop-pig-tutorial/>
- [16] Amazon Web Services (AWS) [online] <https://aws.amazon.com/>
- [17] Microsoft Azure [online] <https://docs.microsoft.com/en-us/azure/architecture/data-guide/>
- [18] Google Cloud Platform [online] <https://cloud.google.com/>
- [19] Cloudera Hadoop Cluster (CDH) [online] <https://www.cloudera.com/documentation/other/reference-architecture.html>
- [20] Hortonworks Data Platform [online] <https://hortonworks.com/products/data-platforms/hdp/>
- [21] Demchenko, Yuri, Emanuel Gruengard, Sander Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. 1st IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science, in Proc. The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 December 2014, Singapore.
- [22] 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.
- [23] FAIRsFAIR Project [online] <https://www.fairsfair.eu/>
- [24] Research Data Alliance (RDA) Education and Training on Handling of Research Data interest Group (IG-ETHRD) [online] <https://www.rd-alliance.org/groups/education-and-training-handling>
- [25]

6 Courses to Facilitate Data Science Professional Skills

This section provides examples of supportive courses to facilitate the Data Science professional competences and skills that can be also used as a basis for general digital competences and data literacy training. All described courses have been taught and are currently in the development phase answering the demand for research and industry to enable an effective Agile Data Driven Enterprise model (ADDE).

6.1 Professional Issues in Data Science

The goal of this course is to equip students and practitioners with the knowledge and skills for further focused study of more specific Data Science and Analytics areas and courses.

- Data Science Competences and Skills management and capacity building, EDISON Data Science Framework.
- Data Science professional skills (“Act and think as Data Scientist”) and 21st Century Skills.
- Data Science and Analytics methods and technologies overview.
- Research Methods; Business processes management.
- Project management; Agile development process and best practices, PMI process groups.
- Data Management in research, industry and personal: standards and best practices.
- FAIR (Findable, Accessible, Interoperable, and Re-usable) principles in Open Data and enterprise data management.
- Privacy enabling technologies.
- Ethical and legal principles and regulations.

It is also beneficial to supply this course with guided/tutored groups and/or individual training on essential professional skills such as complex problem solving, critical thinking, creativity, etc. defined as critical for Industry 4.0 workforce.

6.2 Data Science and Analytics Foundation (DSAF)

The goal of this course is to introduce students to the whole spectrum of Data Science and Analytics technologies and at the same time provide strong statistical background for future mastering the core data analytics methods and Machine Learning technologies. This defines the main emphasis in the DSAF course on statistical methods, probability theory, hypothesis testing, data preparations, methods of qualitative and quantitative analytics. The primary analytics platform for this course is recommended to have a low programming threshold to enable fast learning. The RapidMiner visual data analytics environment (<https://rapidminer.com/>) was identified as a good alternative (in the same cases, preferable choice) against typically used R or python tools, to introduce the trainees to the key data analytics methods and enable active experimentation. Exercises in the course should also be available for an Open-Source Python-based data analytic stack.

The following topics are suggested for the DSAF curriculum:

- Introduction and course overview: Data Science and Big Data technologies, Data Science competence and skills, Research Methods in Data Science, Machine Learning and Data Mining overview.
- Statistical methods and Probability theory
- Data description and Statistical Data Analysis
- Data preparation: data loading, data cleaning, data pre-processing, parsing, transforming, merging, and storing data
- Qualitative and Quantitative data analysis
- Classification: methods and algorithms
- Cluster analysis basics and algorithms
- Performance of data analytics algorithms and tools
- Building engaging visualizations of data analysis
- Organizing data analysis following CRISP-DM and Data Science Process
- Open Data repositories, test datasets, developer communities
- Data Management, FAIR Data Principles

It is important to provide such a course at the beginning of the Bachelor or Master Data Science programme and repeat key components in the more specialised courses, such as Machine Learning or Data Mining to allow

gradual knowledge building that benefits from repetition. DSAF course can also be recommended for other applied Data Science programmes, Elements of the DSAF course can be included into the Data Literacy training. To support students with limited programming background we suggest taking a preparatory course that refreshes key programming techniques, elements of Python language, and basic database operations.

6.3 Research Methods and Process Management

Research Methods and Project Management is one of the Data Science Competence group and DS-BoK Knowledge Area Group [3]. Data Science uses research methods in its foundation to drive experiment based hypothesis validation, taking an insight into or extracting value from the business or industrial data.

Besides the critical importance of understanding and applying different research methods and experiment design, a Data Scientist must have comprehensive knowledge and the ability to effectively use existing process models that represent best practices in solving practical Data Analytics tasks and provide a basis for the project organisation and management. A well-defined process model should answer the following questions:

- What is the whole process of obtaining information from the data? What are stages?
- How are datasets transformed at each stage? How are datasets organised, accessed and stored at each stage?
- What is done in each stage? What roles and activities are involved?
- Which techniques must be applied?
- Which technology and tools must be employed?
- How are the results evaluated and quality ensured?

The following presents the reference and a short description of the popular data analysis process models.

6.3.1 Data Science Process Management Frameworks

6.3.2 CRISP-DM, CRoss-Industry Standard Process for Data Mining

The CRISP-DM was the first model to formalise the data mining process and its relevance to the business processes and intended actionable outcome. The following 6 sequential phases are defined which can be also organised into iterative/continuous development and improvement cycle:

Business understanding: Understanding business processes, problem definition, required outcome.

Data understanding: Obtaining available data, data inspection, data preparation, data observation and initial hypothesis formulation.

Data preparation: Prepare datasets for analysis by extracting necessary data from the raw data.

Modelling: Create one or several data models, select corresponding techniques, and possibly test and experiment with several to assess performance.

Evaluation: Ensure that the solution fulfills the business objectives established in the first phase, and to obtain this certainty a in deep evaluation of the model must be performed during this phase. Evaluated if the knowledge obtained with the model has the desired value for the customer.

Deployment: Solution deployment in a production environment for the customer, providing the results in easy understandable form to the customer.

Figure 1 and Figure 2 illustrate CRISP-DM phases and Tasks correspondingly.

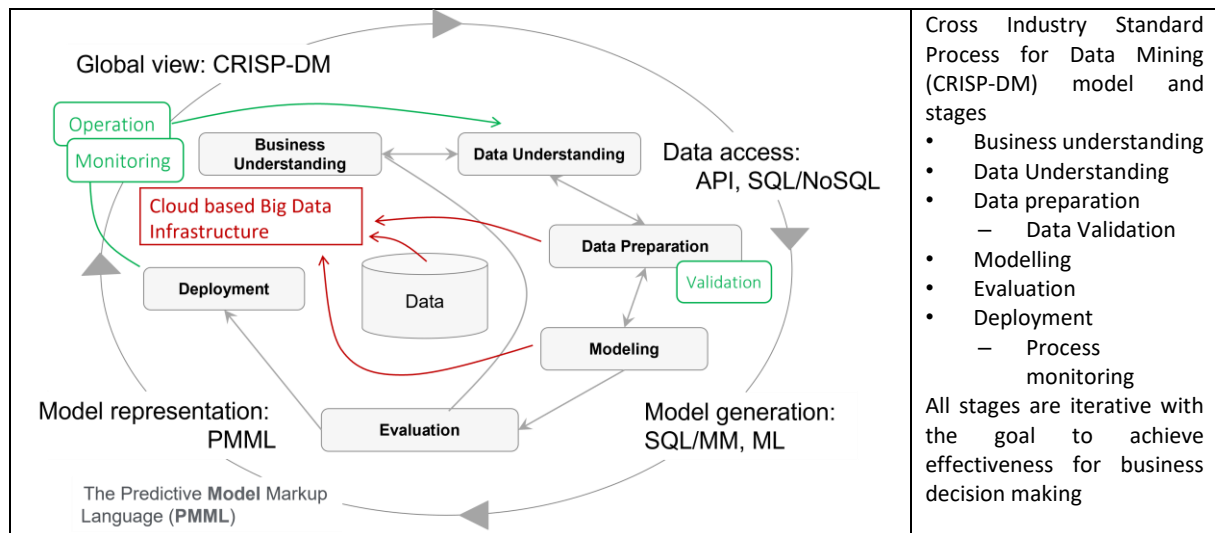


Figure 1. CRISP-DM process model extended with infrastructure operation and data management processes.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives Background Business Objectives Business Success Criteria Situation Assessment Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits Determine Data Mining Goal Data Mining Goals Data Mining Success Criteria Produce Project Plan Project Plan Initial Assessment of Tools and Techniques	Collect Initial Data Initial Data Collection Report Describe Data Data Description Report Explore Data Data Exploration Report Verify Data Quality Data Quality Report	Data Set Data Set Description Select Data Rationale for Inclusion / Exclusion Clean Data Data Cleaning Report Construct Data Derived Attributes Generated Records Integrate Data Merged Data Format Data Reformatted Data	Select Modeling Technique Modeling Technique Modeling Assumptions Generate Test Design Test Design Build Model Parameter Settings Models Model Description Assess Model Model Assessment Revised Parameter Settings	Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models Review Process Review of Process Determine Next Steps List of Possible Actions Decision	Plan Deployment Deployment Plan Plan Monitoring and Maintenance Monitoring and Maintenance Plan Produce Final Report Final Report Final Presentation Review Project Experience Documentation

Figure 2. Detailed Phases and Tasks in CRISP-DM.

6.3.3 ASUM, Analytics Solutions Unified Method (IBM)

ASUM is a hybrid of agile and traditional process models. It has five (5) sequential phases and a set of processes to manage and monitor the progress and maintenance of the project:

Analyze. Requirements specified and agreed; contract or services agreement is signed.

Design. Define all components of the solution and their relationships and dependencies, identify necessary resources.

Configure and Build. The solution is developed all components are integrated and configured.

Deploy. Create a plan to run and maintain the developed solution, including configuration management and migration plan if necessary.

Operate and Optimize. The solution is operational is monitoring data are collected and maintained.

Figure 3 illustrates the ASUM process model that includes the overall cycle and internal ML model building cycle.

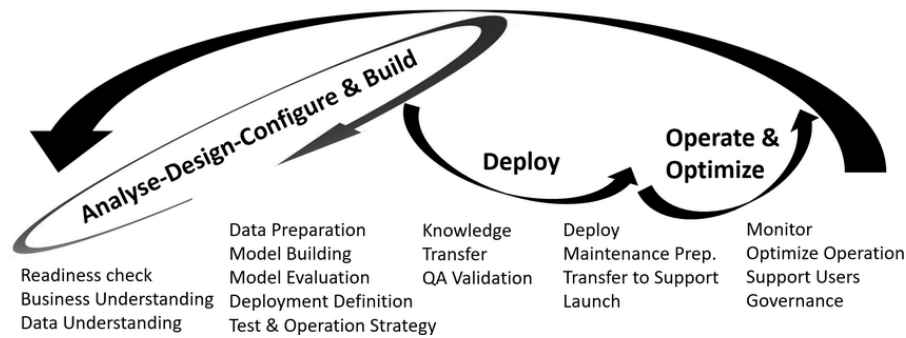


Figure 3. ASUM process model including overall cycle and ML model building cycle.

6.3.4 TDSP, Team Data Science Process (Microsoft)

Team Data Science Process includes the following components:

- Data science lifecycle definition
- Standardized project structure
- Infrastructure and resources recommended for data science projects
- Tools and utilities recommended for project execution

TDSP is an agile and iterative process model. It has five sequential lifecycle phases: Business Understanding, Data acquisition and understanding, Modeling, Deployment, Customer acceptance. TDSP framework document provides valuable information for Data Science application developers for planning and managing the whole process from development to deployment and operation.

Figure 5. illustrates the TDSP process model.

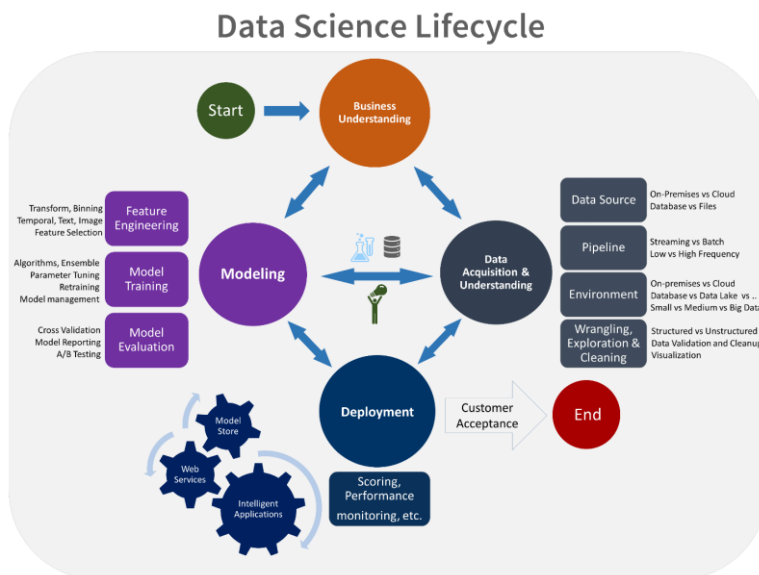


Figure 5. TDSP process model.

6.3.5 KNIME Model Factory (KMF)

KMF defines the process model specifically adopted to Data Mining and Machine Learning processes that includes similar stages such as Init, Load, Transform, Learn, Score, Deploy.

Figure

6.4 ML Model Formats

After the model is built and trained, it needs to be deployed on the operational platform. The following portable ML model formats are used:

- **Predictive Models Markup Language (PMML)** that have benefits of transferring a developed model to production, access to coefficients
- **Portable Format for Analytics (PFA)**, an emerging standard for statistical models and data transformation engines to ease portability across systems with algorithmic flexibility by defining composable models, pre-processing, and post-processing functions that can be built into complex workflows
- **ONNX (Open Neural Network Exchange)** - an open format built to represent machine learning models. ONNX defines
 - Common set of operators - the building blocks of machine learning and deep learning models, and
 - Common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers.
- **TensorFlow Model**
 - SavedModel
 - TF1 Hub format
 - TFLite format
 - TFJS format

Figure 4 illustrates comparative properties of the different model formats.

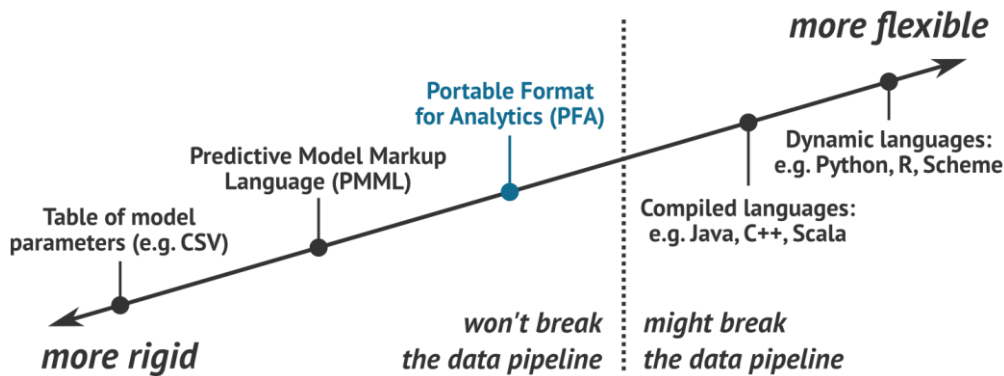


Figure 4. Comparative properties of the different model formats.

References

- [1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0, Cross Industry Standard Process for Data Mining Reference Model, <http://edison...> (2000).
- [2] IBM, ASUM, Analytics solutions unified method [online] https://www01.ibm.com/marketing/iwm/iwm/web/pick.do?source=swerpba-basimext&lang=en_US
- [3] TSDP, Team Data Science Process, <https://docs.microsoft.com/en-us/azure/machinelearning/team-data-science-process/overview>

7 Data Stewardship Professional Competence Framework (DSP-CF)⁷

Skills for data governance and management are critical for the wide adoption of Open Science practices and effective use of the data in research, industry, business and other economic sectors. The FAIR (Findable – Accessible – Interoperable - Reusable) data management principles and data stewardship provide a foundation for effective research data management. The 2018 “Turning FAIR into Reality” report and other documents recommend that data skills should be more widely included in university curricula and that a concerted effort should be made to coordinate and accelerate the pedagogy for professional data roles. Throughout Europe and beyond, many organisations, projects and initiatives work on providing training on FAIR data competences. However, wider adoption of the FAIR data culture can be achieved by including FAIR competences into university curricula. This paper presents the ongoing work of the FAIRsFAIR project to develop a Data Stewardship competence framework and to provide recommendations for implementing this framework in university curricula by means of defining the Data Stewardship Body of Knowledge Model Curricula. The proposed approach and identified competences and knowledge topics are supported by a job market analysis. The presented work is actively using the EDISON Data Science Framework as a basis for the Data Stewardship competences definition and methodology for linking competences, skills, knowledge, and intended learning outcomes when designing curricula.

The information presented in this section is based on the FAIRsFAIR project Deliverable D7.3⁸ and the EDUCON2021 paper⁹ published by authors.

7.1 Introduction

The growing importance of data in the modern data driven economy, research and industry, requires special attention to including data management and governance related topics in university education. The future specialists should understand the role and value of data in research and industry and be able to derive actionable value from data collected from research, technological process or business/social activity, and be able to use open data and public data. Modern data-driven research and industry require new types of specialists capable of supporting all stages of the data lifecycle from data production to data processing and actionable results delivery, visualisation and reporting, which can be jointly defined as the Data Science professions family [1, 2].

Data Management and Data Analytics are critical aspects of digital transformation, however it requires a change of the whole organisational culture, which is often referred to as data literacy. The research community has responded to this with the formulation of the FAIR data principles that suggest data must be Findable, Accessible, Interoperable, and Reusable [3].

The education and training of Data Stewards should not be limited to general data management or FAIR data principles. The presented research identified a number of competences, skills, and knowledge areas covering technology and data management that are required from the Data Stewards for successful work in their future organisation. Besides data-related competences and knowledge, the Data Stewards are required to have an understanding of project management and organisational processes (research or business, depending on organisation).

At the present time, most of the existing university curricula and training programs cover a limited set of competences and knowledge and fall short of what is required for multiple Data Science and Data Stewardship professional profiles and organisational roles within research and industry. In conditions of continuous technology development and shortened technology change cycles, Data Science and Data Stewardship education requires an effective combination of theoretical, practical and workplace skills.

Recent European initiatives and projects such as the European Open Science Cloud (EOSC) [4] and the Research Data Alliance (RDA) [5] facilitate the implementation of the FAIR (Findable, Accessible, Interoperable, Reusable)

⁷ Based on FAIRsFAIR project WP7 “FAIR Competences for Higher Education” work that produced Deliverables D7.3 “FAIR Competence Framework for Higher Education (Data Stewardship Professional Competence Framework)” and D7.4 “FAIR Competences Handbook for Universities”

⁸ “FAIR Competence Framework for Higher Education (Data Stewardship Professional Competence Framework)”, Deliverable D7.3, FAIRsFAIR project, February 2021

⁹ Yuri Demchenko, Lennart Stoy, Research Data Management and Data Stewardship Competences in University Curriculum, In Proc. Data Science Education (DSE), Special Session, EDUCON2021 – IEEE Global Engineering Education Conference, 21-23 April 2021, Vienna, Austria

data principles [6, 7]. They aim to allow for a more effective data exchange and integration across scientific domains, making scientific data a valuable resource and a growth factor for the whole digital economy and society.

The proposed work has been done in the framework of the EU-funded FAIRsFAIR project [8] that recognises the importance of establishing the new profession of the Data Steward and introducing FAIR principles and culture at the early stage of professional education and careers by including FAIR principles into university curricula. The FAIR competences and corresponding Knowledge Areas can be introduced as a special course and/or a part of other courses typically taught at universities such as Research Methods, Research Data Management, or Professional Issues [9]. Research Data Management and FAIR principles are currently attributed to the emerging profession of the Data Steward.

The proposed Data Stewardship Professional Competence Framework (CF-DSP) is based on the EDISON Data Science Framework (EDSF) [2] and defines the main competences required from the Data Steward in their work in different organisations. CF-DSP is also complemented by the DSP Body of Knowledge (DSP-BoK) that is defined as a subset of the Data Science Body of Knowledge. This allows reusing the whole EDSF toolkit developed for customised curriculum design [10].

The paper refers to the previous authors' works on defining the EDISON Data Science Framework (EDSF) [10] and its application of individual competences management and customised curricula design based on required competences and intended learning outcomes that can be targeted for specific professional profiles including Data Stewards [12].

The paper is organized as follows. Section II provides a reference to European and international initiatives related to research data management and the growing demand for the Data Stewardship profession. Section III summarises the job market analysis for Data Steward and Data Management vacancies to identify demanded competences, skills and knowledge. Section IV provides an overview of existing frameworks defining Data Stewardship and related competences, including EDSF. Section V discusses the proposed definition of the Data Stewardship Professional Competence Framework (CF-DSP) as an extension to EDSF. Section VI provides suggestions about new knowledge topics to be included in the DSP Body of Knowledge. A conclusion in section VII provides a summary and refers to ongoing and future developments.

7.2 Research Data Management and Data Stewardship

The importance of data and research information sharing has been central in a number of European-wide initiatives and projects addressing Open Access, Open Data, and Open Science in general. The Research Data Alliance (RDA) that was created in 2012 jointly by the National Science Foundation of USA (NSF) and the European Commission, became a key community coordination body to exchange and develop best practices in research data management.

To facilitate research data sharing and implementation of the FAIR principles, the European Commission introduced the Open Research Data (ORD) Pilot [13]. EU-funded projects by default are required to develop and implement the Data Management Plan (DMP) at the initial stage of the project [14]. Data produced in projects must be made as 'open as possible' and deposited in data repositories (operated by the project or using national or European data archive services). Metadata must be published, and the quality of data must be ensured, in particular through compliance with the FAIR principles. The DMP template provided by the Commission is structured to ensure that the data produced by funded research are open and FAIR [15].

The FAIR data principles and Data Stewardship are among the key objectives of the European Open Science Cloud (EOSC) initiative started in 2016 as the part of the "European Cloud Initiative - Building a competitive data and knowledge economy in Europe" [16], which targeted to capitalise on the data revolution. Under this initiative, EOSC federates existing and emerging e-Infrastructures to provide European science, industry, and public authorities with world-class data infrastructure connected to high performance computers (HPC).

The EOSC goals are to enable the Open Science Commons [17] and achieve FAIRness in research data management and in the services provided. At the present time, the EOSC projects created the foundation for

research data interoperability and integration for European IRs. The Minimum Viable EOSC (MVE) achieved by the end of 2020, will create a starting point for future EOSC development [18].

7.3 Data Stewardship and FAIR competence Frameworks

This chapter provides a short summary of the existing frameworks and standards that essentially contributed to the proposed definition of FAIR4HE and are required for understanding the FAIR4HE alignment with other frameworks and developments.

The proposed summary is based on the FAIRsFAIR Deliverable 7.2 “Briefing on FAIR Competences and Synergies” that provided an overview of various FAIR-data related competence frameworks and training initiatives, forming a basis for the definition of the FAIR4HE framework. The following frameworks are analysed to extract and map identified competences, skills, knowledge topics:

- EOSCpilot FAIR4S Data Stewardship Competence Framework
- ELIXIR Data Stewardship Competence Framework (DSP4LS)
- DelC and DM Forum: Report on National Coordination of Data Steward Education in Denmark
- FOSTER Open Science Learning outcomes
- GO FAIR Data Principles and Maturity Framework
- DAMA BoK (2007) DAMAI Data Management Body of Knowledge
- EDISON Data Science Framework (EDSF) and EDISON Community Initiative

7.3.1 EOSCpilot FAIR4S Framework

The EOSCpilot project defines data stewardship as a shared responsibility of professional groups involved in different data management activities: data management and curation, data science and analytics, data services engineering and domain research. The EOSCpilot deliverable “D7.5: Strategy for Sustainable Development of Skills and Capabilities”¹⁰ describes the comprehensive FAIR4S framework that defines six skill profiles grouped around the research data lifecycle stages and four professional groups (researchers, data scientists, data advisors, and data services providers) involved into different aspects of data management, data curation and related services provisioning. The defined FAIR4S is primarily focused on the EOSC services as they were defined in the EOSCpilot project 2017-2019.

The total 31 individual competences and capabilities that are defined in FAIR4S are grouped into the following groups around typical processes and stages in the research data lifecycle¹¹:

- Plan and design: Plan stewardship and sharing of FAIR outputs,
- Capture and process: Reuse data from existing sources,
- Integrate and analyse: Use or develop FAIR research tools/services,
- Appraise and preserve: Prepare and document data/code to make outputs FAIR,
- Publish and release: Publish FAIR outputs on recommended repositories,
- Expose and discover: Recognise, cite and acknowledge contributions.

The FAIR4S framework defined two templates for describing the Skills profiles and Role profiles. The Skills profile template includes knowledge, skills and attitude (that can also be treated as aptitude) for three levels of proficiency Basic, Intermediate, and Expert. The template also includes a list of professional groups and roles to which the competence group applies. The Role profile includes the list of suggested skills, an explanation of their importance and suggestions for where these skills can be learned.

Applicability and use for FAIR4HE

The FAIR4S framework provided a valuable analysis of the FAIR competences for Data Stewardship from the point of view of the EOSC projects. It defines competences as a combination of *knowledge*, *skills* and *attitude*, an approach that has been used in other frameworks and also used in the proposed FAIR4HE/CF-DSP. The definition of the three levels of proficiency is important for defining learning outcomes when developing academic and training curricula.

¹⁰ EOSCpilot Deliverable D7.5 Strategy for sustainable development of skills and capabilities [online <https://eoscipilot.eu/content/d75-strategy-sustainable-development-skills-and-capabilities>]

¹¹ The intermediate EOSCpilot deliverable D7.3 (2018) contained 59 individual competences that included both data lifecycle groups and general activities groups.

7.3.2 ELIXIR Data Stewardship Competency Framework

The ELIXIR Data Stewardship Competency Framework for life sciences¹² (hereafter referred to as DSP4LS – Data Steward Profession for Life Sciences) is the most complete of the reviewed frameworks. It defines the competencies, skills and knowledge related to Data Stewardship as a distinct profession in the modern data driven science ecosystem and the life sciences in particular. The framework allows translating the Data Stewards organisational responsibilities and tasks, together with required knowledge, skills and abilities into practical learning objectives that provide a basis for developing tailored training. In this way, the framework provides a strong foundation for professionalizing Data Stewardship.

The DSP4LS starts by defining the Data Steward Roles and Competence Profiles in the following three areas:

- Policy: institute and policy focused
- Research: project and research focused
- Infrastructure: data handling and e-infrastructure focused

For all Data Steward roles, the eight competence areas are defined: Policy/strategy; Compliance; Alignment with FAIR data principles; Services; Infrastructure; Knowledge management; Network; Data archiving. In the extended definition, for each competence the following attributes are defined:

- Activities and tasks (in the organisational context)
- Knowledge, Skills and Abilities
- Learning Objectives (LO) formulated as “*after successful completing training you will be able to [..]*”

Applicability and use for FAIR4HE

The DSP4LS provides the complete definition of the Data Stewardship competences for three profiles defining the main responsibilities and organisational roles of the Data Stewards with focus on Policy, Research, and Infrastructure. The defined eight competence areas reflect the whole spectrum of the activities conducted by Data Stewards in organisations and research processes. The presented detailed definition of the Learning Objectives can be directly used for the Data Stewardship curriculum definition.

7.3.3 DeIC Data Stewardship curricula recommendations/principles

The Danish e-Infrastructure Cooperation (DeIC) and Danish National Forum for Research Data Management (DM Forum) Report on National Coordination of Data Steward Education in Denmark¹³ provided valuable recommendations on defining Data Stewardship curricula, primarily aligned with the Danish research environment. The report is based on the strong evidence base derived from the LinkedIn profiles analysis (74 profiles analysed during March 2019) and Job vacancies database in Denmark analysis (119 vacancies of Data Scientists and Data Stewards analysed during March-April 2019) and an extensive overview and analysis of existing competence frameworks and educational programmes for Data Science and Data Stewardship. In addition, the community feedback was collected via a Questionnaire that received 86 complete responses (and 42 partial responses).

The Data Stewardship competences are defined in six competence groups comprising 22 competences related to: Open Science, Data Collection and Data Processing, Data publishing and data preservation, and competences related to research data lifecycle phases: Planning phase, Active research phase, and Dissemination/publication phase.

The report defined the four roles for Data Stewards: Administrator; Analyst; Developer; Agent of change.

The report proposed three modes for Data Stewards education (based on the prospective student/learner background and entry level):

- Student with Bachelor degree
- Student with PhD and equivalent
- Continuing and professional education

¹² Towards FAIR Data Steward as profession for the Life Sciences, Final report ZonMw & ELIXIR-NL projects (Oct 3, 2019) [online] <https://doi.org/10.5281/zenodo.3471707>

¹³ The Danish e-Infrastructure Cooperation (DeIC) and Danish National Forum for Research Data Management (DM Forum) Report on National Coordination of Data Steward Education in Denmark [online] https://www.deic.dk/sites/default/files/Data%20Steward%20Education%20in%20Denmark_0.pdf

The competences identified for the Data Stewardship curricula are grouped into three groups reflecting the main Data Stewards responsibility areas in organisations:

- i) Open Science
 - Open Science policies
 - Data management plans
 - Rights, licenses
- ii) Data collection and data processing
 - Data and Source Search and Data Collection
 - Data storage (in connection with data collection, data storage and storage of active data in the project process)
 - Data processing
 - Open Reproducible Research
- iii) Data publishing and data preservation
 - Data archiving (finished data) and long-term storage
 - Data publishing
 - "Scientific publishing / scholarly communication"
 - Open Access publishing

and other competences related to research data lifecycle phases: The planning phase, Active research phase, and Dissemination/publication phase.

Applicability and use for FAIR4HE

The proposed Data Stewardship education format and curriculum design approaches provide a valuable example of the knowledgeable community approach to defining Data Stewardship competences and designing curriculum profiles for selected groups of learners such as master students, PhD students and practitioners.

7.3.4 GO FAIR Metadata Management Requirements and FAIR Data Maturity Model

The GO FAIR initiative¹⁴, which is devoted to promoting and sustainable adoption of the FAIR data principles, provides recommendations on FAIR metadata management¹⁵ that can be used for linking between general requirements to FAIR implementation and underlying technology and infrastructure and consequently for defining technical expertise areas. These requirements are compiled in Table 1 and later discussed when defining the required FAIR competences, skills and knowledge topics.

Table 1. FAIR metadata requirements

Findable	<ul style="list-style-type: none"> ● F1 (meta)data are assigned a globally unique and persistent identifier; ● F2 data are described with rich metadata; ● F3 metadata clearly and explicitly includes the identifier of the data it describes; ● F4 (meta)data are registered or indexed in a searchable resource.
Interoperable	<ul style="list-style-type: none"> ● I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation; ● I2. (meta)data use vocabularies that follow FAIR principles; ● I3. (meta)data include qualified references to other (meta)data.
Accessible	<ul style="list-style-type: none"> ● A1 (meta)data are retrievable by their identifier using a standardized communications protocol; <ul style="list-style-type: none"> ○ A1.1 the protocol is open, free, and universally implementable; ○ A1.2 the protocol allows for an authentication and authorization procedure, where necessary; ● A2 metadata are accessible, even when the data are no longer available.
Reusable	<ul style="list-style-type: none"> ● R1 meta(data) are richly described with a plurality of accurate and relevant attributes; ● R1.1 (meta)data are released with a clear and accessible data usage license; ● R1.2 (meta)data are associated with detailed provenance; ● R1.3 (meta)data meet domain-relevant community standards.

¹⁴ GO FAIR Initiative [online] <https://www.go-fair.org/go-fair-initiative/>

¹⁵ EOSCpilot deliverable "D7.5: Strategy for Sustainable Development of Skills and Capabilities" The FAIR Guiding Principles for scientific data management and stewardship, March 2016, Scientific Data 3(160018 (2016)) DOI: 10.1038/sdata.2016.18

The FAIR Data Maturity Model¹⁶, which was developed and is maintained by the RDA community¹⁷ provides a set of compliance indicators to assess the level of implementation of the FAIR principles and can be used for defining policy, research and infrastructure related competences in Data Stewardship and data management.

Applicability and use for FAIR4HE

The GO FAIR definition of metadata requirements and FAIR Data Maturity Model provides valuable insight into required infrastructure technologies and technical competences needed for consistent implementation of the FAIR data principles in research, industry and business. This information has been used in defining the necessary engineering competences for Data Stewards.

7.3.5 DAMA DMBOK: Data Governance and Stewardship

The Data Management Body of Knowledge (DMBOK) Framework by Data Management Association International (DAMAI) is an industry standard summarizing best practices in Data Management¹⁸. It is a valuable document that provides a basis for setting up organisational policy and structure to ensure consistent data management and governance. The DMBOK is directly used for training and certification of several data management and governance professions and roles. It goes into depth about the Knowledge Areas that make up the overall scope of data management.

The DMBOK defines 11 main Knowledge Areas and several additional areas related to technologies used. Each Knowledge Area is provided with a detailed context diagram that includes: Definition, Goals, Inputs, Activities, Deliverables, Suppliers, Participants, Consumers, Tools, Technics and Metrics – that can be used as direct guidance for organisations setting up their data management and governance structure.

The Data Governance and Stewardship Knowledge Area is defined as central for the whole DMBOK. The DMBOK also explains the relationship between Data Governance and Data Management, where Data Governance is focused on a Legal and Judicial view (“Do right things”) and Data Management deals with Executive issues (“Do things right”). This also defines staffing of the Data Governance Office: Chief Data Steward, Executive Data Steward, Coordinating Data Steward, Business Data Steward roles. Data Management functions are performed by the Chief Information Officer office that includes Data Architects, Data Analysts, Coordinating Data Stewards and technical Data Steward roles.

Data Management principles, according to DMBOK, provide a good summary of best practices that can be included in data management curricula and training:

- Data is an asset with unique properties
- The value of data can and should be expressed in economic terms
- Managing data means managing the quality of data
- It takes Metadata to manage data
- It takes planning to manage data
- Data management requirements must drive Information Technology decisions
- Data management is cross-functional; it requires a range of skills and expertise
- Data management requires an enterprise perspective
- Data management must account for a range of perspectives
- Data management is lifecycle management
- Different types of data have different lifecycle characteristics
- Managing data includes managing the risks associated with data
- Effective data management requires leadership commitment

Data Steward organisational roles

The Data Steward is a core role to execute the organisational Data Governance and Data Management Policy: define, implement, embed. They typically belong to the Chief Data Officer office. The DMBOK defines the core Data Steward activity as follows:

¹⁶ FAIR Data Maturity Model [online] https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v0.90.pdf

¹⁷ RDA Data maturity model Working Group [online] <https://www.rd-alliance.org/groups/fair-data-maturity-model-wg>

¹⁸ Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [online] <http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf>

Creating and managing core Metadata: Definition and management of business terminology, valid data values, and other critical Metadata. Documenting rules and standards: Definition/documentation of business rules, data standards, and data quality rules. High quality data are often formulated in terms of rules rooted in the business processes that create or consume data. Stewards help surface these rules and ensure their consistent use.

Managing data quality issues: Stewards are often involved with the identification and resolution of data related issues or in facilitating the process of resolution.

Executing operational data governance activities: Stewards are responsible for ensuring that day-to-day and project-by-project data governance policies and initiatives are adhered to. They should influence decisions to ensure that data is managed in ways that support the overall goals of the organization.

The importance of having a devoted Data Steward in the organisation is recognised by a remark in the first version of the DMBOK1 (2009): “Best Data Steward is not made but found.”

Applicability and use for FAIR4HE

The DMBOK provides a reference model and approach for defining the baseline Research Data Management and Data Stewardship Body of Knowledge and Knowledge Areas. DMBOK is used in the EDISON Data Science Framework for defining Data Management and Governance Knowledge Area Group (KAG) that is extended with the topics related to Research Data Management; further extension should include FAIR data related knowledge topics. It is important to align the definition of the FAIR4HE and Data Stewardship competence framework with the DMBOK as an industry standard, understanding that the majority of university graduates will work in the industry. The industry best practices can also provide a contribution to improving the definition of Data Stewardship for the research area.

7.3.6 EDISON Data Science Framework and Data Steward Professional Profile Definition

7.3.6.1 Components of EDISON Data Science Framework (EDSF Release 4)¹⁹

The EDISON Data Science Framework provides a basis for the definition of the Data Science profession and enables the definition of the other components related to Data Science education, training, organisational roles definition and skills management, as well as professional certification.

Figure 1 below illustrates the main components of the EDISON Data Science Framework (EDSF) and their inter-relations that provide a conceptual basis for the development of the Data Science profession:

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- DSPP - Data Science Professional Profiles and occupations taxonomy
- Data Science Taxonomy and Scientific Disciplines Classification

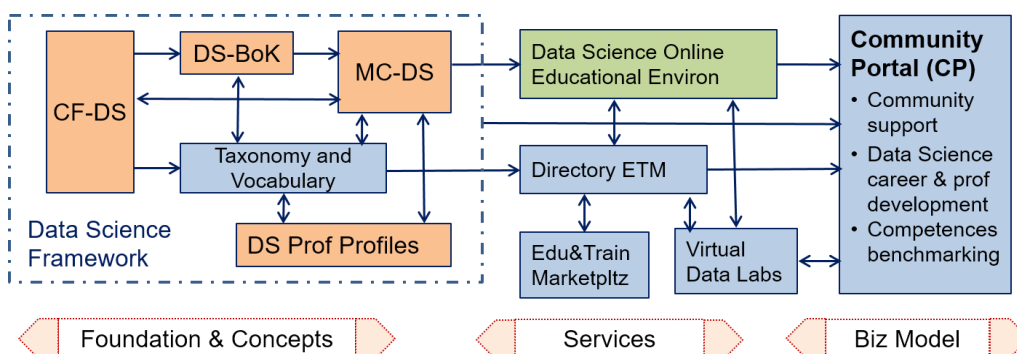


Figure 1 EDISON Data Science Framework components and related services.

The CF-DS provides the overall basis for the whole EDSF. The core CF-DS includes the common competences required for Data Scientists in different work environments in industry and in research and through the whole career path. The ongoing CF-DS development includes coverage of domain specific competences and skills based on the contribution of domain and subject matter experts.

¹⁹ EDISON Data Science Framework (EDSF). [online] Available at <https://github.com/EDISONcommunity/EDSF>

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. The DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. Knowledge Areas are composed of a number of Knowledge Units (KU) which are currently the lowest component of the DS-BoK. The DS-BoK incorporates best practices in Computer Science and domain specific BoKs and includes KAs and KUs defined where possible based on the Classification Computer Science (CCS2012)²⁰, components taken from other BoKs and proposed new KAs/KUs to incorporate new technologies used in Data Science and its recent developments.

The Data Science Model Curriculum (MC-DS) is built based on the CF-DS and DS-BoK, where Learning Outcomes (LO) are defined based on CF-DS competences and Learning Units are mapped to Knowledge Units in the DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome²¹ to allow for flexible curricula development and profiling for different Data Science professional profiles. The proposed Learning Outcomes are enumerated to have a direct mapping to the enumerated competences in CF-DS.

The Data Science Professional Profiles and occupations taxonomy (DSPP) is defined as an extension to the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy²² using the ESCO top classification groups. The DSPP definition provides an instrument to define effective organisational structures and roles related to Data Science positions and can also be used for building individual career paths and corresponding competences and skills transferability between organisations and sectors.

The Data Science Taxonomy and Scientific Disciplines Classification will serve to maintain consistency between four core components of EDSF: CF-DS, DS-BoK, MC-DS, and DSP profiles. To ensure consistency and linking between EDSF components, all individual elements of the framework are enumerated, in particular: competences, skills, and knowledge topics in CF-DS, knowledge groups, areas and units in DS-BoK, learning outcomes and learning units in MC-DS, and professional profiles in DSPP.

7.3.6.2 Data stewards and data management related professional profiles

The definitions of the Data Science Professional Profiles, including a set of Data Management profiles and that of 'Data Steward' in particular, is one of the EDSF components described in a separate document EDSF Part 4²³. The DSPP are defined in accordance with and as a proposed extension to the ESCO Taxonomy which is a European standard for European Skills, Competences and Occupations. The DSPP definition can be instrumental in defining organisational roles in Data Science and Data Management. It can also be used for defining education and training profiles for students and for practitioners to acquire the necessary competences and knowledge for specific professional profiles or occupations. When linked to the Competence Framework and Body of Knowledge, it can be used for professional certification or career path building.

Figure 2 illustrates the existing ESCO hierarchy and the proposed new Data Science classification groups and corresponding new Data Science related profiles. The table in the figure illustrates what CF-DS competence groups are relevant to each profile by indicating competence relevance from 0 to 5 (0 – not relevant, 5 – very important).

The Data Science occupation groups are placed in the following top level ESCO hierarchies:

- Managers (for managerial roles);
- Professionals (for Data Science and Analytics, Data Management and Stewardship, infrastructure and data centre engineering roles);
- Technicians and associate professionals (for operators, facility administrators and technicians)
- Optionally, some data management occupations can also be placed into the General and Keyboard Clerks group such as data entry clerks and user support workers.

Correspondingly, the following 3rd level occupation groups are proposed in DSPP:

²⁰ CCS2012, The 2012 ACM Computing Classification System. Available at <http://www.acm.org/about/class/class/2012>

²¹ Refer to the EDSF documentation for full information about MC-DS and DS-BoK

²² European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at <https://ec.europa.eu/esco/portal/#modal-one>

²³ EDISON Data Science Framework, Part 4. Data Science Professional Profiles. Available at https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DSPP-release3-v07.pdf

- Data Science Services/Infrastructure Managers
- Data Science Professionals
- Data handling/management professionals that include Data Stewards, Digital Data Curators, Data Librarians
- Database and infrastructure professionals
- Technicians and associate professionals
- Data and information entry and access

A group of occupations related to Data Stewardship, data curation, data archives and libraries are currently placed in the 3rd proposed group of professionals in the ESCO hierarchy:

Professionals > Information and communications technology professionals > Data Science technology professionals > Data handling professionals not elsewhere classified

Recognising the importance of the Data Steward in a typical research institution, the DSPP provides the following definition of the Data Steward professional profile:

Data Steward is a data handling and management professional whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation. The Data Steward creates a data model for domain specific data, supports and advises domain scientists/ researchers during the whole research cycle and data management lifecycle.

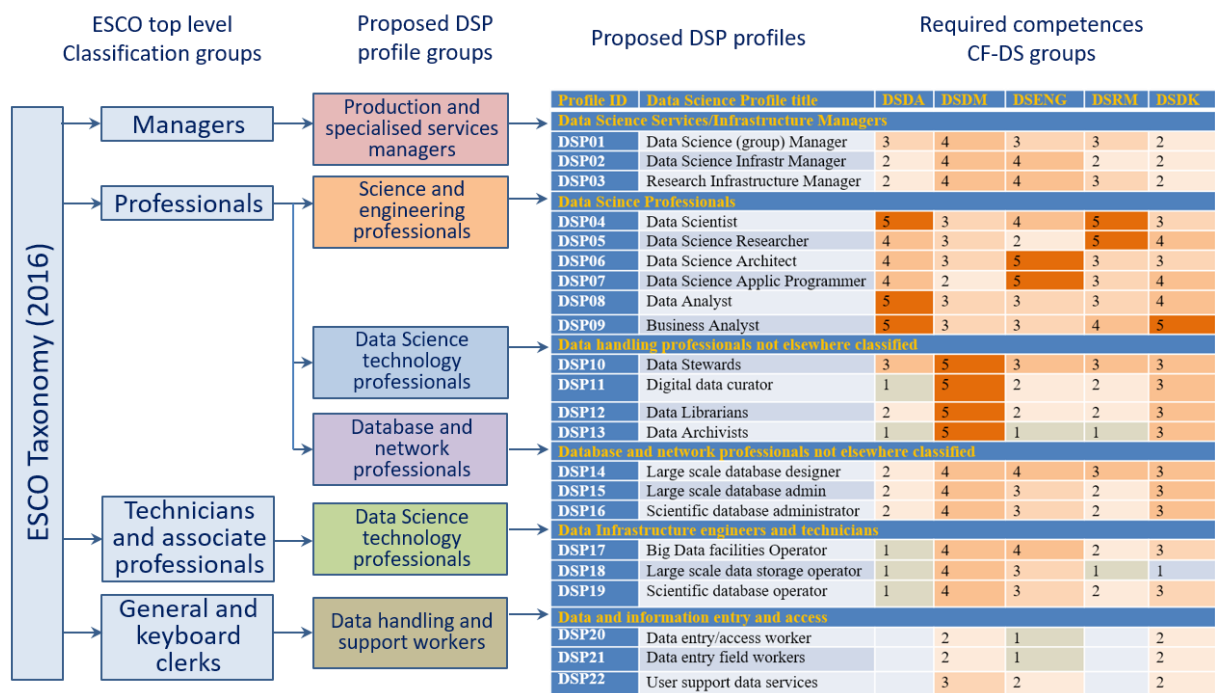


Figure 2. Proposed Data Science related extensions to the ESCO classification hierarchy and corresponding DSPP by classification groups (EDSF Part 4 [14]).

7.3.6.3 Using EDSF as a basis for defining FAIR4HE and aligning with other frameworks

The EDSF provides a complete framework and methodology for defining the whole ecosystem of Data Science competences, professional profiles, Body of Knowledge and Model Curriculum where Data Stewardship is one of the professional profiles with defined competences, skills, knowledge, and recommended Learning Objectives and Learning Units. EDSF can be used for defining an organisational structure for Data Science and Data Stewardship tasks and activities management using the definition of the Data Science Professional Profiles that support the whole data lifecycle management in a research organisation.

The EDSF and its Data Stewardship profile can be used for aligning existing Data Stewardship and FAIR competences and consolidating them into the intended FAIR4HE competence framework. It is also foreseen that future EDSF releases can accommodate the FAIR4HE competence framework as a special profile.

The following chapters describe the proposed Data Stewardship and FAIR4HE Competence framework in detail.

7.4 Job market analysis for demanded key competence

A preliminary analysis was done of data collected from job advertisements on popular job search and employment portals indeed.com, IEEE Jobs portal and LinkedIn Jobs, where indeed.com provided the largest number of advertised Data Steward vacancies. The collected data were used to extract information on competences, skills and knowledge demanded from prospective Data Steward candidates. The following sections explain what approach was used for the analysis of vacancies and how the extracted information was mapped to the structure of the competences definition.

7.4.1 Method and context

The EDSF was used as a basis for defining the initial set of Data Stewardship competences, with the following revision and extension of the individual competences specific to Data Stewardship and FAIR data principles identified from the collected data. The full EDISON/EDSF methodology used for the initial definition of Data Science competences is explained in Appendix A.

When applying this methodology to the current analysis of Data Stewardship competences, the initial identification of the competence groups was not required.

The assumption was that the Data Steward competences would have the same structure as the whole Data Science Professional family, namely the competence groups Data Science Analytics (acronym DSDA as defined in EDSF or short DSA), Data Science Engineering (DSENG or DSE), Data Management (DSDM or DM), Research Methods and Project Management or Business Process Management (DSRMP or RMP), and Domain Knowledge (DSDK or DK). The benefit of this assumption is that the majority of current university curricula (in fields related to Data Science and Data Stewardship) already contain the above mentioned courses²⁴, and that it will be easy to further map identified competences and knowledge to typical academic courses and/or learning units.

A typical job vacancy has the following structure and contains the following information that can be mapped to different components of a competence definition (such as Competence, skills, knowledge, education, proficiency level):

- Job/position name, sometimes provided with the description of organisational roles and relations;
- Functions/responsibilities and abilities which can be mapped to competences, if competences are not explicitly defined (job vacancies usually use the term 'skills' instead of 'competences');)
- Skills and experiences, also including experience with tools and programming languages that all can be directly mapped to skills;
- Required knowledge or expected familiarity with named technologies or theories. This can be mapped to knowledge topics;
- Education, certification and proficiency level – can be mapped to a proficiency level that indicates mastering a certain level of a specific competence; but this information is rarely specified in the typical job vacancy.

It is also important to clarify the relations between competences, skills and knowledge as illustrated in Figure 3 and used in the EDSF (which itself was adopted from the European e-Competences Framework (e-CF) and the corresponding standard EN 16234-1: 2019):

- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results.
- Competence includes/is supported by the **knowledge** that is obtained from education or (self-)training and by **skills** that are acquired as a result of practical experience.
- Professional profiles suggest necessary competences, skills and knowledge and ensure the ability to perform organisational functions

²⁴ Tomasz Wiktorski, et al, Model Curricula for Data Science EDISON Data Science Framework, Proc. The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong. [online] <https://ieeexplore.ieee.org/document/8241134>

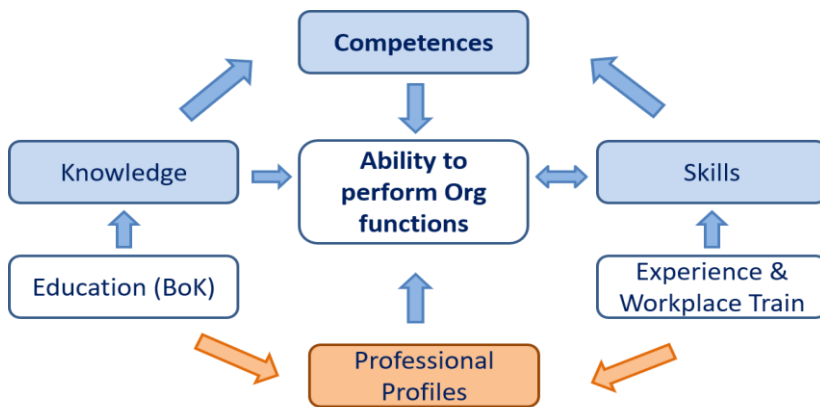


Figure 3. Relation between competences, skills and knowledge

7.4.2 EDISON methodology to collect and analyse job market and competence related data [1, 2]

To verify existing frameworks and potentially identify new competences, different sources of information have been investigated:

- First of all, job advertisements that represent the demand side for Data Stewards and data management specialists and based on practical tasks and functions that are identified by organisations for specific positions. This source of information provided factual data to define demanded competences and skills.
- Structured presentation of Data Steward related competences and skills produced by different studies as mentioned above, in particular EDSF definition of Data Science and Data Stewardship that identifies the following groups of competences, namely Data Analytics, Data Science Engineering, Data Management, Research Data, and Domain expertise. This information was used to correlate with information obtained from job advertisements.
- Blog articles and community forums discussions that represented valuable community opinion. This information was specifically important for defining practical skills and required tools.

The following approach has been used when analysing the job advertisement data

- 1) Collect data on required competences and skills
- 2) Extract information related to competences, skills, knowledge, qualification level, and education; translate and/or reformulate if necessary
- 3) Split extracted information on initial classification or taxonomy facets, first of all, on required competences, skills, knowledge; suggest mapping if necessary
- 4) Apply existing taxonomy or classification: for the purpose of this study, we used skills and knowledge groups as defined by the EDSF definition of Data Science and Data Stewardship (i.e. Data Analytics, Data Engineering, Data Management, Research Methods, Domain Knowledge)
- 5) Identify competences and skills groups that do not fit into the initial/existing taxonomy and create new competences and skills groups
- 6) Do clustering and aggregations of individual records/samples in each identified group
- 7) Verify the proposed competences groups definition by applying to originally collected and new data
- 8) Validate the proposed competence framework via community surveys and individual interviews.

The outlined above process has been applied to the collected information and all steps are tracked in the two Excel workbooks provided as supplementary material, which is available on the EDSF community GitHub.

7.4.3 Collected data

This section provides a summary of the information extracted from the Data Stewards vacancies analysis.

The following are the general characteristics of the collected data:

- Period data collected from: 30 August – 1 September 2020
- Sites: Indeed.com – NL, UK, DE, USA: monsterboard.nl - NL
- Days vacancy open: >50% more than 30 days
- Data Steward and related vacancies discovered:

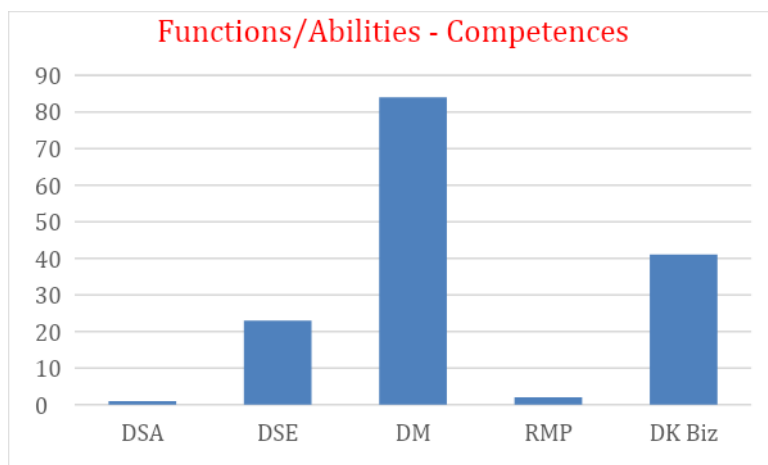
- NL – 51, UK – 30+, DE ~20, US – 300+
- Information collected/downloaded:
 - Key skills snapshot – for all or the first 200 for the USA
 - Full vacancy texts – approx. 40 in total
- Detailed analysis of sample vacancies
 - NL, UK – 20, US - 6
- Number of companies and organisations posted Data Steward related jobs – more than 50

Mapping of vacancies information to competences, skills and knowledge items was done manually using simple text extraction and content ordering in Excel. Appendix B provides details on the approach used for analysing data. The Excel workbooks are available in the DSP-CF git folder²⁵.

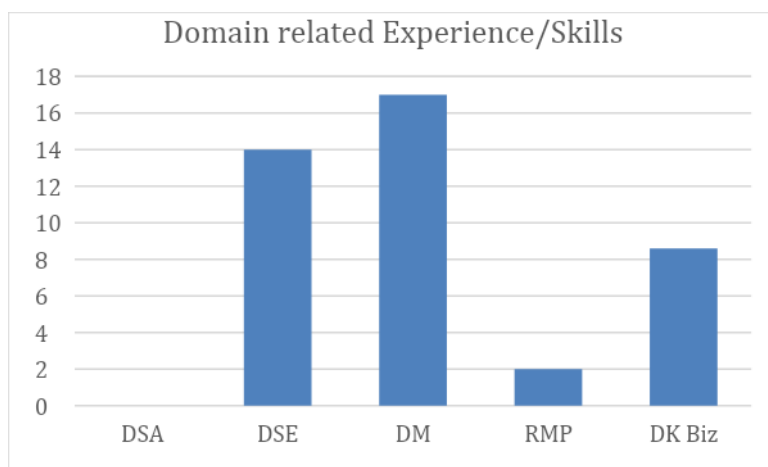
7.4.4 Identified competences, skills and knowledge and their mapping to CF-DS

One of the goals of the undertaken analysis was to identify what competence groups are demanded on the job market and if they can be mapped to competences defined in CF-DS/EDSF. If confirmed that the EDSF can be further used for the FAIR4HE and CF-DSP competences definition, this will bring the benefits of using the EDSF approach for linking competences to intended Learning Outcomes and model curricula.

The diagrams below illustrate what types/groups of competences are required in the Data Steward vacancies. Figure 4 (a) illustrates the mapping of functions/abilities to competence groups, Figure 4 (b) maps experiences and skills to skill groups.



(a) Competences present in Data Steward vacancies



(b) Skills present in Data Steward vacancies

²⁵ <https://github.com/EDISONcommunity/EDSF/tree/master/data-stewardship-professional-competence-framework>

Figure 4. Competence and skill groups present/required in the Data Steward vacancies. Legend: DSA - Data Science Analytics, DSE - Data Science Engineering, DM - Data Management, RMP - Research Methods and Project Management or Business Process management, DK - Domain Knowledge (such as Business Analytics).

Another valuable information obtained from the vacancies analysis is the spectrum of required knowledge topics presented in Table 2 which can also be classified into the same competence and knowledge groups as shown in Figure 5.

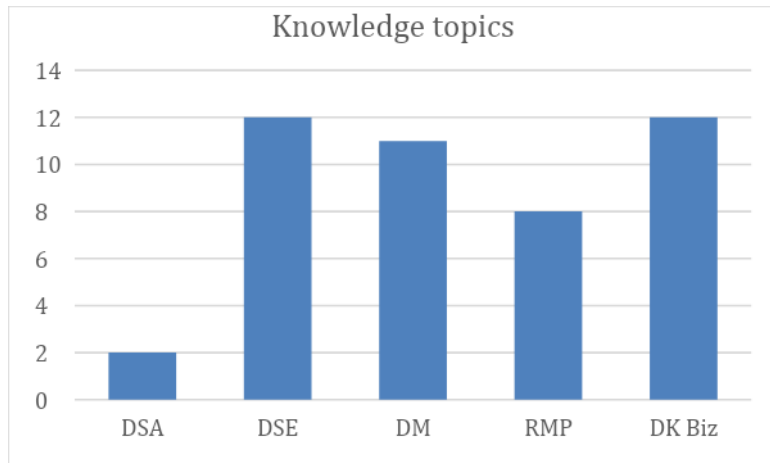


Figure 5 Knowledge topics present/required in the Data Steward vacancies. Legend: DSA - Data Science Analytics, DSE - Data Science Engineering, DM - Data Management, RMP - Research Methods and Project Management or Business Process management, DK - Domain Knowledge.

Table 2. Knowledge topics obtained from the Data Steward vacancies analysis

Competence/Knowledge Group	Extracted knowledge topics
Data Management	<ul style="list-style-type: none"> • Data Management techniques • FAIR data principles • Data Management and Data Governance principles • Data integrity • Metadata, PID and linked data • Ontology and Semantics • FAIR metrics and Maturity framework, FAIR certification • Data compliance regulations and standards • Data privacy law • GDPR • Ethics
Research Methods	<ul style="list-style-type: none"> • Research methods (general and domain related) • Project management
Data Analytics	<ul style="list-style-type: none"> • Data analysis and visualisation tools • Data lifecycle, lineage, provenance
Domain knowledge and Business processes	<ul style="list-style-type: none"> • Business process management • Marketing • Banking financial services and data management • Multilevel Bill of Materials • Data Warehouses • Version control system • Master Data Management (MDM) and Reference Data
Data Science Engineering	<ul style="list-style-type: none"> • Visual Basic for Applications (VBA) and interface design • WebAPI use for data access, collection and publishing • DevOps, Agile, Scrum methods and technologies • Data formats, standards

	<ul style="list-style-type: none"> • Data modeling (SQL and EDBMS, NoSQL) • Modern data infrastructure: Data registries, Data Factories, Semantic storage, SQL/NoSQL
--	--

A more detailed analysis of competences is done in section 3.4 and included the extraction of individual competences and their comparison to the current definition of the competences in EDSF and existing Data Stewardship and FAIR competence frameworks (refer to section 3.4).

7.4.5 Outcome of the job vacancies analysis and further steps

The following conclusions and assumptions can be done based on the initial vacancies analysis:

- The published Data Stewards vacancies demonstrated a variety of competences, skills and knowledge required from the candidates.
- The extracted competences can be successfully mapped to the competence groups defined for the Data Science professional family that includes Data Stewards.
- The presented analysis confirms the applicability of EDSF to the analysis and further structured development of the intended FAIR4HE and Data Stewardship Competence Framework.

The most populated competence group is Data Management what reflects the nature of the Data Steward profession and responsibilities. Two other well populated groups are Domain Knowledge and Data Science Engineering what reflects another side of the Data Steward profession to act as a bridge between ICT teams operating data facilities and domain specialists. This demonstrates the need for related knowledge at a level sufficient for coordination and communication. This fact is clearly reflected in the distribution of required knowledge topics.

The collected and extracted set of competences, skills and knowledge topics was used for detailed competences analysis and mapping to current definitions and vocabulary in EDSF and necessary updates and extensions/additions will be suggested. This information is presented in the next section.

7.4.6 Technological and organisational aspects of the FAIR data principles implementation

Besides collecting information from the job market, we can also look at the technological and organisational aspects of the implementation of the FAIR data principles in a typical research organisation as described in section 2.4 and 2.6. Table 3 correspondingly provides the mapping of the FAIR metadata requirements to required technological and management domains: standardisation, policy, infrastructure, and tools or platforms. Table 4 links the typical data management lifecycle stages to organisational roles related to organisational data management and governance.

Table 3. FAIR metadata requirements

FAIR metadata requirements and technology aspects	Standardisation	Policy	Infra structure	Tools
Findable				
F1. (meta)data are assigned a globally unique and persistent identifier	x		X	
F2. data are described with rich metadata		x		X
F3. metadata clearly and explicitly includes the identifier of the data it describes		x		x
F4. (meta)data are registered or indexed in a searchable resource		x	x	X
Accessible				
A1. (meta)data are retrievable by their identifier using a standardized communications protocol <ul style="list-style-type: none"> • A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary	x		x	X

A2 metadata are accessible, even when the data are no longer available		X	X	
Interoperable				
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	x	x		X
I2. (meta)data use vocabularies that follow FAIR principles	X	X		X
I3. (meta)data include qualified references to other (meta)data		X		X
Reusable				
R1. meta(data) are richly described with a plurality of accurate and relevant attributes <ul style="list-style-type: none"> ● R1.1 (meta)data are released with a clear and accessible data usage license ● R1.2 (meta)data are associated with detailed provenance ● R1.3 (meta)data meet domain-relevant community standards 		x		X

Table 4. Data Lifecycle Management and organisational roles

Data lifecycle stage	Data Steward key activities	Other roles involved
Data collection	Data model and metadata definition and implementation	Researchers Data engineers Data entry workers
Data preservation and curation	Data storage facility Data quality control Data integration	Data curators Data custodians/archivists
Data analysis	General contacts with data analytics team (not special tasks)	Data Scientists Data Architects Application developers
Data publication	Data publications in open repositories and archiving services, metadata Ensure data discoverability and findability	Data curators
Data governance and data management	Develop Data Governance Policy and Data Management Plan Ensure FAIR data principles implementation Coordinate and monitor data governance and management implementation Interact with ICT team and data infrastructure services Organise and conduct necessary training for data management policy	Chief Data Officer Data quality managers Data Controller (GDPR)
Data infrastructure and tools	Define and communicate requirements to data infrastructure and tools, coordinate their implementation Organise necessary training for tools and services	Infrastructure engineers Database managers/engineers Master data managers

Effect on FAIR and Data Stewardship competences

The implementation of the FAIR data principles in operation and practice of research infrastructures requires a strong technical base, infrastructure services and tools. This is a task for ICT and data infrastructure services/teams; but Data Stewards need to be aware of these technologies and tools and maintain a link between ICT and data policy, providing also the necessary training to researchers and data workers.

7.5 Defining a Competence Framework for Data Stewardship and FAIR Data Principles (CF-DSP)

As the basis for elaborating the Data Stewardship and FAIR data Competence Framework (CF-DSP), we used the Data Science Competence Framework (CF-DS) defined in EDSF. This allows us to benefit from other EDSF components, such as the Body of Knowledge and Model Curriculum. In this context, we treat the CF-DSP for Data Stewardship as a profile or subset of the more general CF-DS for the Data Science professional family.

The data collected and classified from the Data Stewards job vacancies are used for identifying the set of individual competences that match with the CF-DS competence groups. Based on this, original CF-DS competences are revised and/or extended, new competences are suggested to create a consistent Data Stewardship Competence Framework that reflects the current job market demand for Data Stewards and their essential competences. The final definition of the CF-DSP will be composed of the essential competences identified in this analysis.

It is also important to note that in the current, market-based definition of Data Steward competences and skills, the primary focus lies on data management skills (DSDM group), understanding of the required data management platforms and infrastructure (DSENG group) and domain-related or organisational competences (DSDK or DSBA group). A general understanding of research methods and project management competences is required, whereas Data Science and Analytics competences (DSDA group) may only be required at the level of general literacy. The following tables (Tables 5 to 8) list the original CF-DS competence groups together with the suggested changes and extensions to individual competences for the intended/proposed CF-DSP profile.

7.5.1 Data Management and Governance competence group (DSDM)

As a consequence of the wide recognition by organisations of the importance of quality data management, almost all individual competences have been updated (see Table 5). It is also important to mention that the growing adoption of the Data Steward profession as an important organisational role and wide adoption of the FAIR data principles motivate the addition of three competences into CF-DSP. These are:

- DSDM07: Manage Data Management/Data Stewards team, coordinate related activity between organisational departments, external stakeholder to fulfill Data Governance policy requirements
- DSDM08: Develop organisational policy and coordinate activities for sustainable implementation of the FAIR data principles
- DSDM09: Specify requirements in terms of and supervise the organisational infrastructure for data management (and archiving), maintain the pool of data management tools

Table 5. CF-DS competence group Data Management (DSDM) and suggested extensions for CF-DSP

Data Management (DSDM)	Relevance and proposed changes and extensions (posted as revised text and bulleted extensions)
DSDM Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	DSDM – extended, relevant Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing, <ul style="list-style-type: none"> • ensure compliance with FAIR data principles.
DSDM01 Develop and implement data strategy, in particular in the form of data management policy and Data Management Plan (DMP)	DSDM01 – extended, essential Develop and implement data management and governance strategy, in particular in the form of a Data Governance Policy and Data Management Plan (DMP) <ul style="list-style-type: none"> • Ensure compliance with standards and best practices in Data Governance and Data Management

<p>DSDM02 Develop and implement relevant data models, define metadata using common standards and practices for different data sources in a variety of scientific and industry domains</p>	<p>DSDM02 – extended, essential Develop and implement relevant data models, define metadata using common standards and practices for different data sources in a variety of scientific and industry domains.</p> <ul style="list-style-type: none"> • Ensure metadata compliance with FAIR requirements • Be familiar with the metadata management tools
<p>DSDM03 Integrate heterogeneous data from multiple sources and provide them for further analysis and use</p>	<p>DSDM03 – extended, essential Integrate heterogeneous data from multiple sources and provide them for further analysis and use</p> <ul style="list-style-type: none"> • Perform data preparation and cleaning • Match/transfer data models of individual datasets
<p>DSDM04 Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)</p>	<p>DSDM04 – extended, highly essential Maintain historical information on data handling, including reference to published data and corresponding data sources</p> <ul style="list-style-type: none"> • Publish data, metadata and related metrics • Perform and maintain data archiving • Develop necessary archiving policy, comply with Open Science and Open Access policies if applicable • Maintain data provenance and ensure continuity through the whole data lifecycle, ensure data provenance
<p>DSDM05 Ensure data quality, accessibility, interoperability, compliance to standards, and publication (data curation)</p>	<p>DSDM05 – extended, essential Develop policy and metrics for data quality management, maintain data quality and compliance to standards, perform data curation</p> <ul style="list-style-type: none"> • Interact/Collaborate with data providers and data owners to ensure data quality
<p>DSDM06 Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management</p>	<p>DSDM06 – extended, essential Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management, address legal issues if necessary.</p> <ul style="list-style-type: none"> • Ensure GDPR compliance in data management and access • Develop data access policies and coordinate their implementation and monitoring, including security breaches handling
<p>None</p>	<p>DSDM07* - added new, essential Manage Data Management/Data Stewards team, coordinate related activity between organisational departments, external stakeholder to fulfill Data Governance policy requirements, provide advice and training to staff. Define domain/organisation specific data management requirements, communicate to all departments and supervise/coordinate their implementation. Coordinate/supervise data acquisition.</p>
<p>None</p>	<p>DSDM08* - added new, essential Develop organisational policy and coordinate activities for sustainable implementation of the FAIR data principles and Open Science, define corresponding requirements to data infrastructure and tools, ensure organisational awareness.</p>

None	DSDM09* - added new, essential Specify requirements to and supervise the organisational infrastructure for data management and (and archiving), maintain the park for data management tools, provide support to staff (researchers or business developers), coordinate solving problems.
------	---

7.5.2 Data Engineering competence group (DSENG)

Table 6 describes the relevance of, and proposed changes to DSENG competences to align them with the changes applied to corresponding Data Stewardship competences. Updates/extensions were added to the competences DSENG03-DSENG06 to reflect FAIR related requirements to data infrastructure, data management tools and metadata management during the whole data lifecycle.

Table 6. CF-DS competence group Data Science Engineering (DSENG) and suggested extensions for CF-DSP

Data Science Engineering (DSENG)	Relevance and proposed changes and extensions (posted as revised text and bulleted extensions)
DSENG Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle.	DSENG – no changes, generally relevant Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle.
DSENG01 Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis and visualisation	DSENG01 – no changes, low relevance Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis and visualisation
DSENG02 Develop and apply computational and data driven solutions to domain related problems using a wide range of data analytics platforms, with a special focus on Big Data technologies for large datasets and cloud based data analytics platforms	DSENG02 – no changes, low relevance Develop and apply computational and data driven solutions to domain related problems using a wide range of data analytics platforms, with a special focus on Big Data technologies for large datasets and cloud based data analytics platforms
DSENG03 Develop and prototype specialised data analysis applications, tools and supporting infrastructures for data driven scientific, business or organisational workflow; use distributed, parallel, batch and streaming processing platforms, including online and cloud based solutions for on-demand provisioned and scalable services	DSENG03 – extended, relevant Develop and prototype specialised data analysis applications, tools and supporting infrastructures for data driven scientific, business or organisational workflow; use distributed, parallel, batch and streaming processing platforms, including online and cloud based solutions for on-demand provisioned and scalable services <ul style="list-style-type: none"> Develop new tools and applications, ensure support of the data FAIRness requirements by existing and new tools and applications

<p>DSENG04 Develop, deploy and operate large scale data storage and processing solutions using different distributed and cloud based platforms for storing data (e.g. Data Lakes, Hadoop, HBase, Cassandra, MongoDB, Accumulo, DynamoDB, others)</p>	<p>DSENG04– extended, essential Develop, deploy and operate data infrastructure, including data storage and processing facilities, using different distributed and cloud based platforms.</p> <ul style="list-style-type: none"> Implement requirements for data storage facilities to comply with the data management policies and FAIR data principles in particular.
<p>DSENG05 Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.</p>	<p>DSENG05– extended, relevant Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection, ensure standards and corresponding data protection regulation compliance, in particular GDPR.</p> <ul style="list-style-type: none"> Define and implement (coordinate) data access policies for different stakeholders and organisational roles
<p>DSENG06 Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets</p>	<p>DSENG06– extended, essential Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform), OLTP, OLAP processes for large datasets</p> <ul style="list-style-type: none"> Define, implement and maintain data model, reference data, master data definitions, implement consistent metadata

7.5.3 Research Methods and Project Management competence group (DSRMP)

The Research Methods and Project Management competences are important for Data Stewards in supporting research projects in an organisation, to work effectively with the domain related researchers and to serve as a link between the researchers and other roles during the whole cycle of the research process and corresponding data lifecycle. Minor extensions were added to DSRMP03 and DSRMP05.

Table 7. CF-DS competence group Research Methods and Project Management (DSRMP) and suggested extensions for CF-DSP

Research Methods and Project Management (DSRMP)	Relevance and proposed changes and extensions (posted as revised text and bulleted extensions)
<p>DSRMP Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals</p>	<p>DSRMP – revised, generally relevant Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals</p> <ul style="list-style-type: none"> Base research on collected scientific facts and collected data
<p>DSRMP01 Create new understandings by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods</p>	<p>DSRMP01 – no changes, generally relevant Create new understandings, discover new relations by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods</p>

DSRMP02 Direct systematic study toward the understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSRMP02 – no changes, generally relevant Direct systematic study toward the understanding of the observable facts, and discovers new approaches to achieve research or organisational goals
DSRMP03 Analyse domain related research process model, identify and analyse available data to identify research questions and/or organisational objectives and formulate a sound hypothesis	DSRMP03- extended, essential Analyse domain related research process model, identify and analyse available data to identify research questions and/or organisational objectives and formulate a sound hypothesis <ul style="list-style-type: none"> • Link domain-related concepts and models to general/abstract Data Science concepts and models,
DSRMP04 Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications, contribute to the development of organizational objectives	DSRMP04 – no changes, generally relevant Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and use this knowledge to devise new (data-driven) applications, contribute to the development of organizational or project objectives
DSRMP05 Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	DSRMP05 – extended, essential Design experiments which include data collection (passive and active) for hypothesis testing and problem solving <ul style="list-style-type: none"> • Work with Data Science, Data Stewardship and data infrastructure teams to develop project/research goals.
DSRMP06 Develop and guide data driven projects, including project planning, experiment design, data collection and handling	DSRMP06 – no changes, essential Develop and guide data driven projects, including project planning, experiment design, data collection and handling

7.5.4 Domain related competence (DSDK/DSBA)

Domain-related knowledge and competences are important for Data Stewards as one of their roles is to support organisational (and project) data management during the whole data lifecycle and correspondingly through all business process or research process stages. Our job vacancies analysis indicated the importance for Data Stewards to understand and know the main organisational and business processes with a focus on data management, provenance and quality.

Analysis of the Data Steward positions in the context of the organisational needs, both for the research and the business domain, identified necessary extensions that can be applied to the initial definitions in EDSF CF-DS and also a need for specific activities related to the coordinating role of Data Steward in data management and governance:

- DSBA07: Coordinate intra-organisational activities related to data analytics, data management and data provenance/lineage along all data flow stages.

We use the business related domain competence group DSDA as it is well represented in the business related Data Steward positions and has a well-defined focus on organisational needs. Table 8 summarises the proposed extensions and defines a new competence DSBA07.

Table 8. CF-DS competence group Domain Knowledge (Organisational specific and Business related, DSBA) and suggested extensions for CF-DSP

Domain related Competences (DSDK): Applied to Business Analytics (DSBA)	Relevance and proposed changes and extensions (posted as revised text and bulleted extensions)
DSDK Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations	DSDK – no changes, generally relevant Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
DSBA01 Analyse information needs, assess existing data and suggest/identify new data required for specific business context to achieve organizational goal, including using social network and open data sources	DSBA01 – extended, relevant for organisation processes and data Analyse information needs, assess existing data and suggest/identify new data required for specific business context to achieve organizational goal, including using social network and open data sources <ul style="list-style-type: none"> • Data management and Quality Assurance of organisational data assets
DSBA02 Operationalise fuzzy concepts to enable key performance indicators measurement to validate the business analysis, identify and assess potential challenges	DSBA02 – extended, relevant for organisation processes and data Operationalise fuzzy concepts to enable key performance indicators measurement to validate the business analysis, identify and assess potential challenges <ul style="list-style-type: none"> • Specify requirements/develop data models for organisational data
DSBA03 Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make a business case as a result of organisational data analysis and identified trends	DSBA03 – extended, generally relevant Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make a business case as a result of organisational data analysis and identified trends <ul style="list-style-type: none"> • Ensure data availability and quality for BA/BI needs
DSBA04 Analyse opportunities and suggest the use of historical data available at organisation for organizational processes optimization	DSBA04 – extended, relevant for organisation processes and data Analyse opportunities and suggest the use of historical data available at organisation for organizational processes optimization <ul style="list-style-type: none"> • Coordinate implementation of FAIR data principles for collected data, ensure proper lineage and provenance of collected data
DSBA05 Analyse customer relations data to optimise/improve interaction with specific user groups or in the specific business sectors	DSBA05 – no changes, relevant for organisation processes and data Analyse customer relations data to optimise/improve interaction with specific user groups or in the specific business sectors
DSBA06 Analyse multiple data sources for marketing purposes; identify effective marketing actions	DSBA06 – no changes, relevant for organisation processes and data Analyse multiple data sources for marketing purposes; identify effective marketing actions

none	DSBA07 – added, essential Coordinate intra organisational activities related to data analytics, data management and data provenance/lineage along all data flow stages, ensure data FAIRness
------	---

7.6 Data Steward professional and transversal skills

It is evident that the new profession of Data Stewards and the emerging FAIR data management culture will create new types of professional transversal skills (often referred to as soft skills), which can be defined using such concepts as attitude or aptitude (referring to such concepts introduced in the FAIR4S competence framework).

It is important to compile such skills related to Data Stewardship and FAIR principles. The workplace skills for Data Scientists defined in EDSF can provide an example and a basis for the definition of such skills for Data Stewards.

Although transversal or soft skills are not usually included directly in academic curricula, they can be a part of Professional and Academic skills training that is established at many universities.

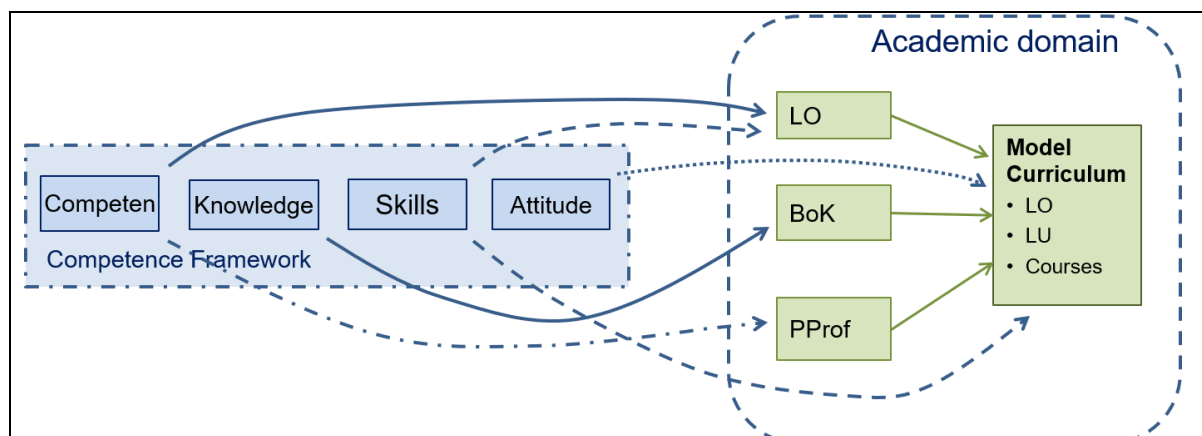
For reference purposes, Appendix C provides an example of how such skills are defined in the EDSF for Data Science Professionals.

7.7 Comparing/Mapping CF-DSP to other Competence Frameworks

This section compares the proposed Data Steward Professional Competence Framework (CF-DSP) with the existing frameworks discussed in chapter 2 with the goal to provide alignment between the proposed CF-DSP and other frameworks. This will simplify educational and training courses exchange, re-use and blending. This is especially important when designing academic curricula for universities and vocational education where existing training materials and courses can be included as self-study and practical study materials.

The comparison was made by mapping the CF-DSP components to similar components in other frameworks, such as competence groups, individual competences, responsibilities, capabilities, skills and knowledge topics. In fact, the mapping presented is the result of an iterative process during which an initial mapping has been done for the initial set of DSP competences to clarify the initial set, discover necessary extensions and incorporate these into the current CF-DSP.

The mapping and alignment presented below have been done for the following frameworks FAIR4S, ELIXIR Data Stewardship Competency Framework (DSP4LS), DeIC Data Stewardship Curriculum recommendations, and Foster Learning Objectives for Open Science. Figure 6 below illustrates (a) the general relation between different components of the entire professional framework for Data Stewardship and (b) the link between different components in existing frameworks and CF-DSP.



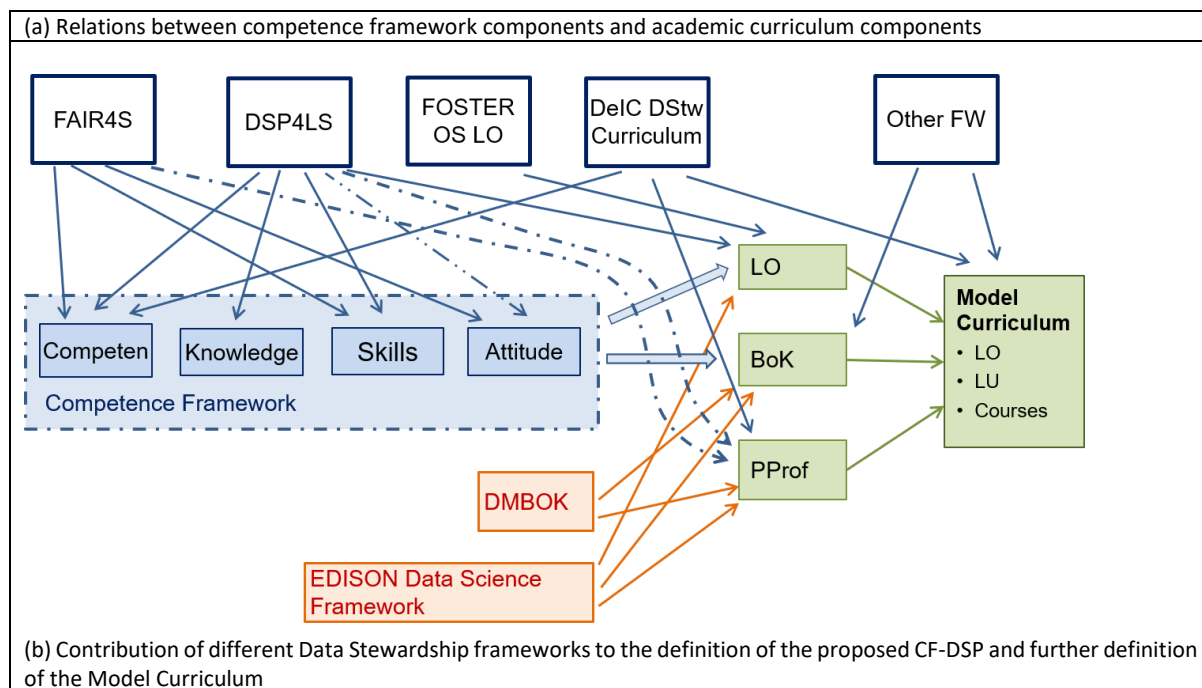


Figure 6. Relations and mapping between CF-DSP and other frameworks
 Legend: LO – Learning Outcomes; LU – Learning Units; BoK – Body of Knowledge; PProf – Professional Profiles. Dotted lines reflect indirect/implied relations via Data Steward organisational roles described in FAIR4S and DSP4LS.

To allow constructive comparison and mapping between frameworks, the competences definitions in individual frameworks were decomposed into individual components (statements) and enumerated for better comparison. Individual components of competences, skills, knowledge topics, learning objectives – whatever relevant to different frameworks, were enumerated for better components grouping and mapping to the corresponding CF-DSP competence groups. A number of individual components in each group was used to assess the relevance of the proposed competences. This was further used for improving the definition of individual competences in CF-DSP.

Figures 7-9 below illustrate the mapping between CF-DSP competences and related definitions in other competence frameworks, where the vertical axis presents the count of the relevant individual competences, skills or knowledge in referred frameworks corresponding to CF-DSP competences. The following general approach has been used when mapping existing frameworks to the proposed CF-DSP:

- Most of the frameworks include only definitions of competences, which are also profiled for different organisational roles (such as Policy, Research and Infrastructure). However, when applying for education and training purposes, the competences should be defined in a more general form and preferably be linked to the Body of Knowledge and well established academic disciplines.
- In many cases, the defined competences and Learning Objectives (LO)²⁶ can be linked to education or training curricula, but others can rather be achieved through career and work experience. It is also understood that future graduates may possess good knowledge of the subject and have a set of necessary competences for the junior role but time is needed to gain workplace experience and corresponding knowledge.
- Most of the existing frameworks do not provide direct links between responsibilities and competences, such as are illustrated in Figure 3 and Figure 6 and implemented in EDSF. This kind of linking was applied during the mapping exercise.
- When analysing the ELIXIR Data Stewardship Competence Framework for Life Sciences (DSP4LS) the defined combined set of Skills, Knowledge and Attitude (SKA) was broken down into individual

²⁶ This analysis uses original term Learning Objectives as it is used in referred frameworks DSP4LS and FOSTER, while in EDSF, ACM/IEEE Curricula guidelines the term Learning Objectives is used.

elements related to competences and knowledge. Similar mapping was done for the Learning Objectives defined in DSP4LS.

Competences defined in both FAIR4S by EOSCpilot and the DeIC Data Stewardship curriculum correspond with the main competence groups defined in CF-DSP, as shown in Figure 7 and Figure 8 respectively. Both frameworks show the majority of required competences in the groups DSDM – Data Management and DSENG – Data infrastructure, services and tools. Moreover, FAIR4S also shows the importance of the Data Science and Analytics competences. However, it is motivated by the need for data quality assurance and tools development. This is usually done by analytics and engineering teams in coordination with the Data Steward who defines user needs and requirements. FAIR4S also identifies the importance of general research methods competences and domain related competences, as well as the importance of professional skills or attitudes. The DeIC Data Stewardship framework similarly confirms the importance of general research methods and project management, including different levels (basic, intermediate, expert) to work effectively with scientific and data publications.

The FOSTER Learning Objectives for Open Science²⁷ provide an important view of the competences that are required for Open Science, which are closely related to the intended application of the FAIR data principles. Figure 9 illustrates the mapping of the FOSTER LOs to extended competences in the DSDM group: DSDM01 – DSDM09, with more stress on the organisational policy and compliance in DSDM06, DSDM08, DSDM09. It also indicates the importance of research methods in general.

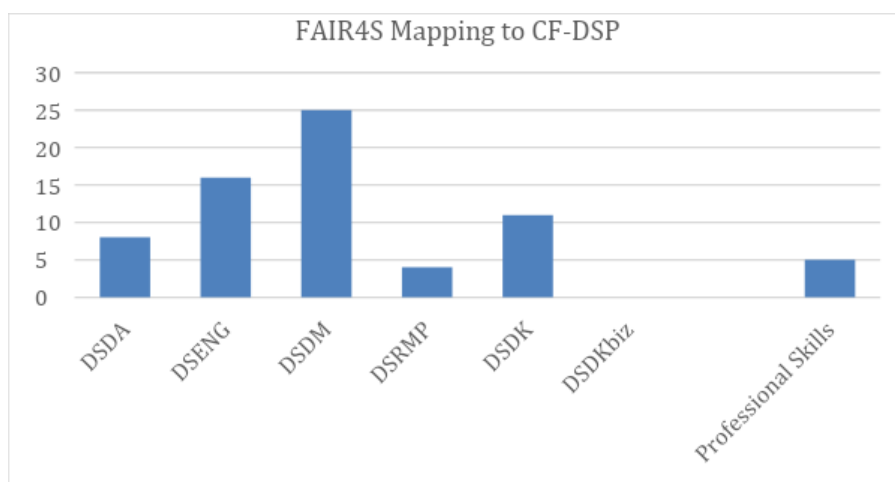


Figure 7. Mapping FAIR4S competences to CF-DSP competence groups (refer to Table 5 for extended DSDM competences definition)

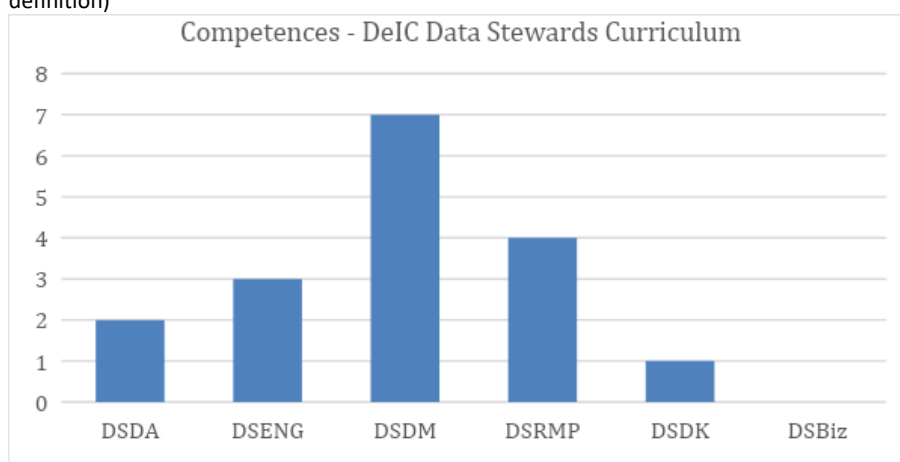


Figure 8. Mapping DeIC Data Stewardship Curriculum competences to CF-DSP competences groups

²⁷ The FOSTER Learning Objectives for Open Science, 23 February 2015 [online] <https://zenodo.org/record/15603#.YA8rD-hKhPY>

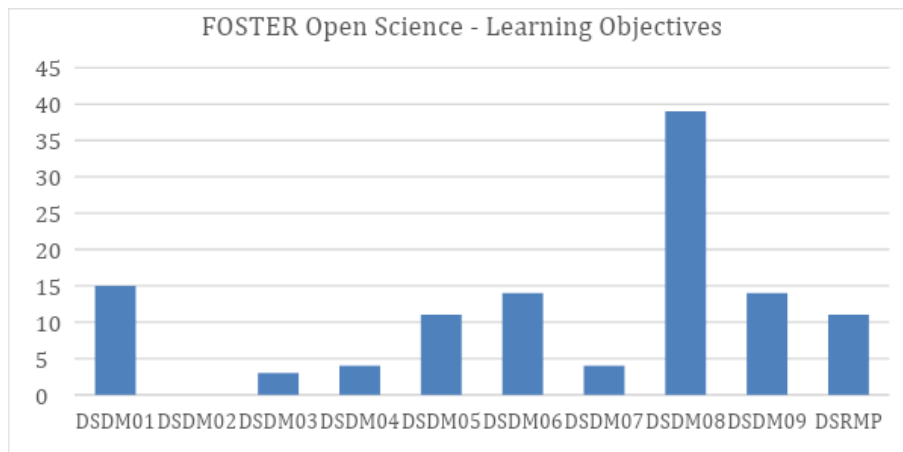
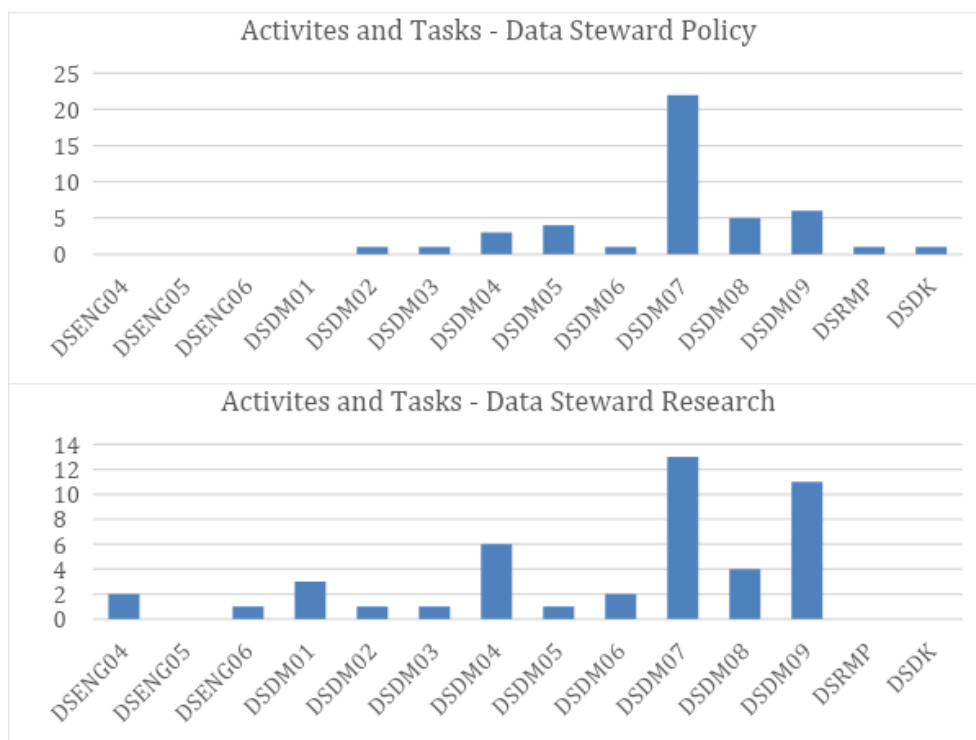


Figure 9. Mapping FOSTER Learning objectives to CF-DSP DSDM Competences (refer to Table 5 for extended DSDM competences definition for CF-DSP and to for FOSTER Learning Objectives for Open Science)

The ELIXIR Data Stewardship Competence Framework for Life Sciences (DSP4LS) provides a detailed inventory of the Activities, Tasks, and Knowledge, Skills, and Abilities for the three organisational roles Research, Infrastructure and Policy; the DSP4LS framework also defines the extended list of Learning Objectives which are aimed for graduate level Data Stewards to develop necessary competences that are grouped in the following competence areas: Policy/strategy; Compliance; Alignment with FAIR data principles; Services; Infrastructure; Knowledge management; Network; Data archiving.

Figures 10 and 11 illustrate the mapping of the DSP4LS Activities, Tasks, and Knowledge, Skills, and Abilities to the proposed CF-DSP competences that include all competences DSDM01-DSDM09 in the data management group, DSENG04-DSENG06 of the engineering group, and also to DSRMP and DSDK that indicate expected general competences and knowledge in these groups. Similarly to the previous diagrams, the vertical axis presents the count of the relevant individual activities and tasks, and knowledge, skills, abilities in DSP4LS corresponding to CF-DSP competences.



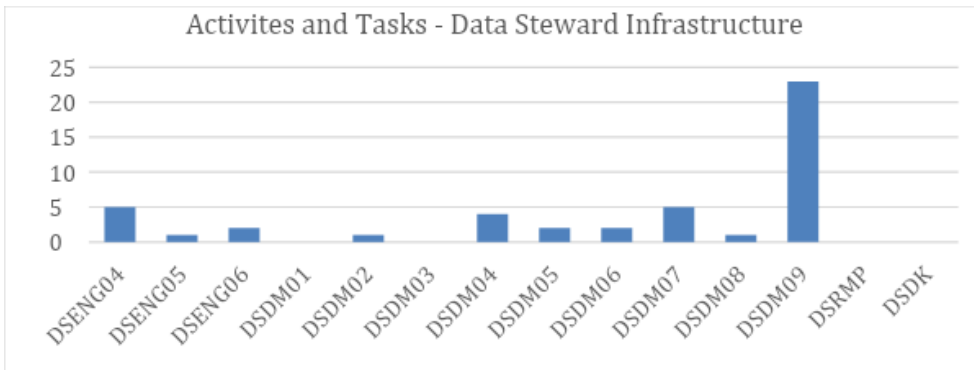


Figure 10. Mapping ELIXIR DSP4LS Activities and Tasks to selected CF-DSP competences

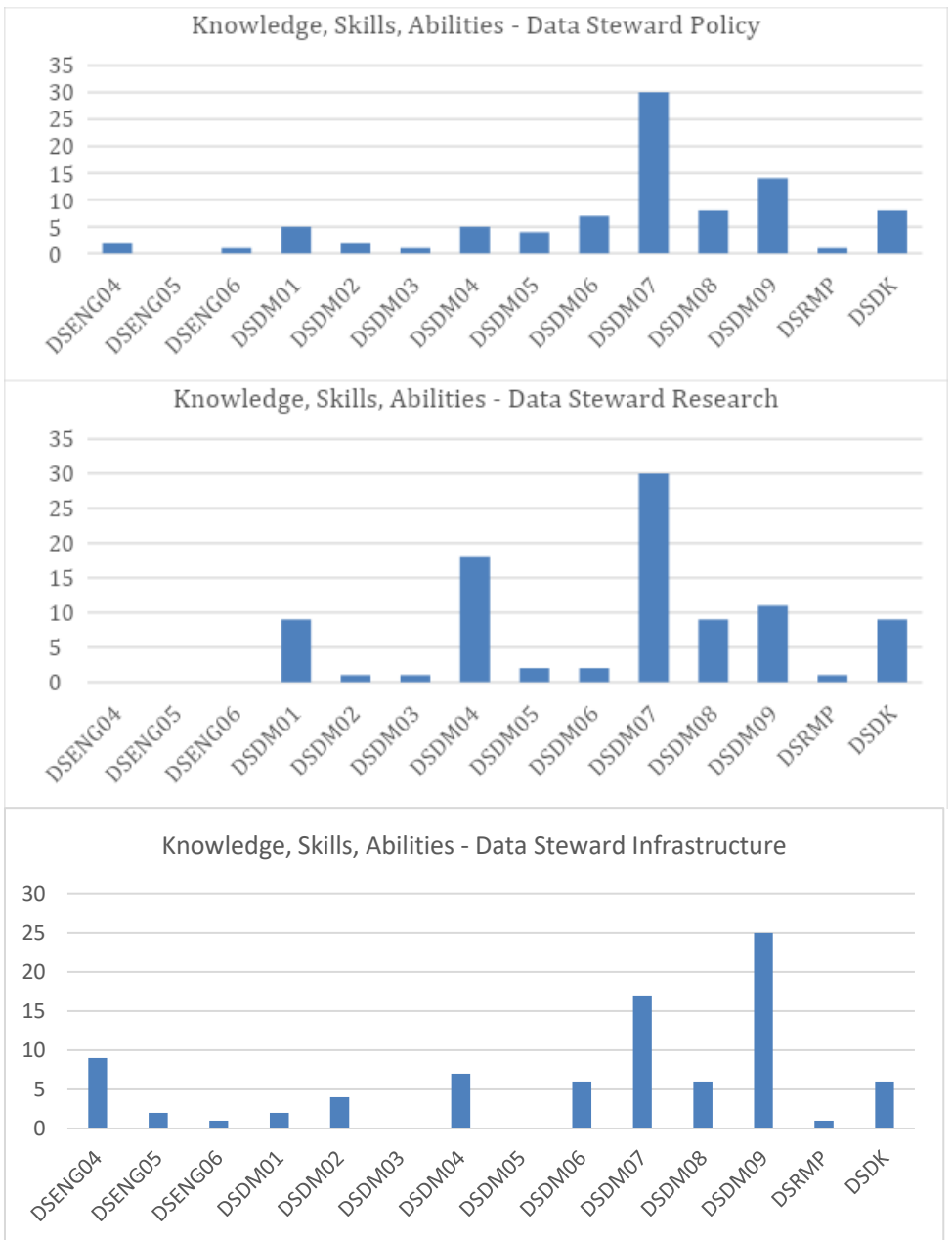


Figure 11. Mapping ELIXIR DSP4LS KSA (Knowledge, Skills, Abilities) group to selected CF-DSP competences

The analysis of the presented mapping allows us to conclude that in general DSP4LS includes competences from the DSDM group and selected competences from the DSENG group of the proposed CF-DSP. Both frameworks

can be effectively used for defining Data Stewardship curricula and to address FAIR data competences. At the same time, the FAIR4HE and CF-DSP competence frameworks can benefit from the advanced development of the DSP4LS framework and its ongoing implementation in the ELIXIR Data Stewardship training programme.

7.7.1 Summary on defining the Data Stewardship and FAIR competence framework for Higher Education

The detailed analysis and mapping of existing competences allowed us to identify important competences and knowledge topics that are required for the successful work of Data Stewards in different roles and organisations. The consolidated definition of the proposed CF-DSP competences is available in the form of an Excel workbook in the EDSF GitHub working directory.

The definition of the proposed CF-DSP is based on the general Data Science Competence Framework created in the EDISON project. Necessary changes and extensions have been added to the initial CF-DS definition to create the proposed Data Stewardship competence profile.

Further analysis and mapping should allow more precise linking between the proposed and existing frameworks, also creating a basis for future exchange of the practical implementation experience. To do this effectively, an ontology for Data Stewardship competences should be developed.

7.8 Defining Data Stewardship and FAIR Body of Knowledge

The Body of Knowledge is an important element linking competence framework with academic curricula. A Body of Knowledge defines a set of Knowledge Areas (KA) and Knowledge Units (KU) that need to be included in a curriculum to achieve the intended Learning Outcome (LO) and defines a set of academic disciplines that can be taught in a curriculum. The definition of the Body of Knowledge is typically based on the classification of the scientific disciplines, such as Classification Computer Science (CCS).

This section provides information about the Data Stewardship Body of Knowledge (hereafter referred to as DSP-BoK), which is defined as a subset or a profile of the general Data Science Body of Knowledge (DS-BoK) defined in EDSF. The DSP-BoK is extended with the FAIR data related knowledge topics as well as with knowledge topics supporting the coordination role of the Data Steward in the organisational data governance and management. The DSP-BoK inherits the benefits of the DS-BoK definition that is based on an overview and analysis of existing bodies of knowledge that are relevant to Data Science and required to fulfill the competences and skills identified in CF-DS. DS-BoK adopts essential knowledge elements from multiple BoKs, such as DMBOK, BABOK, CS-BOK (see Table 9), and introduces a number of new Knowledge Units that reflect the practice in academic and professional training courses by universities and professional training organisations.

The DS-BoK can be used as a basis for defining Data Science related curricula, courses, instructional methods, educational/course materials, and necessary practices for university post- and undergraduate programs and professional training courses. The DS-BoK is also intended to be used for defining certification programs and certification exam questions. While CF-DS (comprising of competences, skills and knowledge) can be used for defining job profiles (and correspondingly content of job advertisements), the DS-BoK can provide a basis for interview questions and evaluation of the candidate's knowledge and related skills, as well as for professional certification exams and training.

7.8.1 Data Science Body of Knowledge Areas and Knowledge Units

The Data Science Body of Knowledge realized in EDSF is structured by knowledge area groups (KAG) corresponding with CF-DS competence groups:

- KAG1-DSDA: Data Analytics group including Data Analytics methods, Machine Learning, statistical methods, and data visualisation
- KAG2-DSENG: Data Science Engineering group including software engineering, database and Big Data technologies
- KAG3-DSDM: *Data Management group including data curation, preservation and data modeling*
- KAG4-DSRMP: *Research Methods and Project Management*
- KAG5-DSBA: Business Analytics (strongly based on KAG1-DSDA)

- KAG*-DSDK: Placeholder for the Data Science Domain Knowledge groups to include domain specific knowledge

The Data Management and Governance knowledge area group (KAG3 DSDM) is a key and distinguishing KAG for DSP-BoK. It includes general principles and concepts in data management and stewardship, data management and governance policies and procedures, data storage systems, data modeling and data warehouses, data libraries and archives. It is extended with the FAIR data principles and other knowledge topics related to the Data Steward role in organisations.

The KAG3-DSDM group includes most of the KAs from the DAMA DMBOK however extends it with KAs related to RDA recommendations, community data management models (Open Access, Open Science, Open Data, etc.) and general Data Lifecycle Management, which is used as a central concept in many data management related education and training courses. For the DSP-BOK, FAIR data related knowledge area and knowledge units must be included as an additional knowledge areas.

The following are the commonly defined Data Management and Governance Knowledge Areas:

- KA03.01 (DSDM.01/DMORG) General principles and concepts in Data Management and organisation
- KA03.02 (DSDM.02/DMS) Data management systems
- KA03.03 (DSDM.03/EDMI) Data Management and Enterprise data infrastructure
- KA03.04 (DSDM.04/DGOV) Data Governance
- KA03.05 (DSDM.05/BDSTOR) Big Data storage (large scale)
- KA03.06 (DSDM.05/DLIB) Data libraries and archives

Other knowledge areas are sufficiently defined in the original EDSF DS-BoK such as KAG1-DSDA, KAG2-DSENG, KAG4-DSRMP and KAG5-DSDK for domain related knowledge areas.

7.8.2 Defining a DSP BoK profile

Table 9 provides the general structure of KAG3-DSDM and relevant Knowledge Areas from other KAG. Extensions of the original DS-BoK are proposed based on the recent Data Stewards job market analysis in chapter 2. It contains the definition of the Knowledge Areas and Knowledge Units that need to be added to properly address Data Stewardship and FAIR data principles. Knowledge Units (KU) corresponding to suggested KAs are defined from different sources: existing BoK, CCS2012, and from practices in designing academic curricula and corresponding courses by universities and professional training organisations²⁸.

Further work on the DSP-BoK will include a mapping of identified knowledge topics to the corresponding Knowledge Units defined in the original DS-BoK. The DSP-BoK defined at this stage will undergo further development and will be updated based on feedback on the model curricula and courses that will be designed in the context of the FAIRsFAIR Task T7.4 activity.

Table 9. DS-BoK Knowledge Area Groups (KAG) and Knowledge Areas (KA) related to the Data Stewardship DSP-BoK

KA Groups	Suggested additional Knowledge Areas (KA)	Knowledge Areas from existing BoK, CCS2012 scientific subject groups and exiting DS&FAIR frameworks
KAG2-DSENG: Data Science Engineering	KA02.01 (DSENG.01/BDIT) Big Data Infrastructure and Technologies KA02.04 (DSENG.04/SEC) Data and Applications security KA02.07 (DSENG.07/IS) Information systems (to support data driven decision making)	ACM CS-BoK selected KAs: IM - Information Management Data and Information systems related scientific subjects from CCS2012: CCS2012: Information systems CCS2012: Software and its engineering

²⁸ KAs and KUs defined in such a way are not exclusive (as mentioned above) but have the benefit of being close to academic practice and allowing easier and faster implementation.

KAG3-DSDM: Data Management	KA03.01 (DSDM.01/DMORG) General principles and concepts in Data Management and organisation KA03.02 (DSDM.02/DMS) Data management systems KA03.03 (DSDM.03/EDMI) Data Management and Enterprise data infrastructure KA03.04 (DSDM.04/DGOV) Data Governance KA03.05 (DSDM.05/BDSTOR) Big Data storage (large scale) KA03.06 (DSDM.05/DLIB) Digital libraries and archives	DM-BoK selected KAs (1) Data Governance, (2) Data Architecture, (3) Data Modelling and Design, (4) Data Storage and Operations, (5) Data Security, (6) Data Integration and Interoperability, (7) Documents and Content, (8) Reference and Master Data, (9) Data Warehousing and Business Intelligence, (10) Metadata, and (11) Data Quality. RDA recommendations on FAIR Data Principles
KAG4-DSRMP: Research Methods and Project Management	KA04.01 (DSRMP.01/RM) Research Methods KA04.02 (DSRMP.02/PM) Project Management	There are no formally defined BoK for research methods PMI-BoK selected KAs <ul style="list-style-type: none"> ● Project Integration Management ● Project Scope Management ● Project Quality ● Project Risk Management
KAG5-DSBPM: Business Analytics	KA05.01 (DSBA.01/BAF) Business Analytics Foundation KA05.02 (DSBA.02/BAEM) Business Analytics organisation and enterprise management	BABOK selected KAs *) <ul style="list-style-type: none"> ● Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts. ● Requirements Analysis and Design Definition. ● Requirements Life Cycle Management (from inception to retirement). ● Solution Evaluation and improvements recommendation.

*) BABOK KA are more business focused and related to KAG5-DSBA. However, its specific topics related to data management can be reflected in the KAG1-DSDA

Referred bodies of knowledge:

ACM/IEEE CS-BoK - ACM and IEEE Computer Science Curricula 2013 (CS2013) [online] <http://dx.doi.org/10.1145/2534860>

DMBOK – Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [online] <http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf>

BABOK - Business Analytics Body of Knowledge (BABOK) [online] <http://www.iiba.org/babok-guide.aspx>

PM-BoK - Project Management Professional Body of Knowledge (PM-BoK) [online] <http://www.pmi.org/PMBOK-Guide-and-Standards/pmbok-guide.aspx>

7.8.3 Using CF-DSP and DSP-BoK for Data Stewardship curriculum definition

The DSP-BoK, together with the CF-DSP, can be used to define Data Stewardship university curricula and courses that respond to the needs of a given community or target stakeholders. In this case, the required competences are expressed in the form of intended learning outcomes that, together with the knowledge topics, define the knowledge units from the BoK that need to be included in the curricula.

References

- [1] Final results of the European Data Market study measuring the size and trends of the EU data economy, ECIDC, March 2017 [online] <https://ec.europa.eu/digital-singlemarket/en/news/final-results-european-data-market-studymeasuring-size-and-trends-eu-data-economy>
- [2] Realising the European Open Science Cloud: First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud, European Commission, 2016 [online] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf
- [3] Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data, 2018 [online] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf
- [4] D7.2 “Briefing on FAIR Competences and Synergies”, FAIRsFAIR Project Deliverable, September 2021
- [5] FAIR4HE Design Workshop, 8-9 Oct 2020, Program [online] <https://docs.google.com/document/d/11VLSJL41wtZa40Y7rHoYQahjNGXQ02dsk45goniR9sA/edit#heading=h.89nh0hqdf91>
- [6] Workshop “Data Stewardship and FAIR Competences in Academic Curricula,” part of CODATA FAIR Convergence Symposium [online] <https://conference.codata.org/FAIRconvergence2020/sessions/222/>
- [7] RDA IG on Professionalising Data Stewardship [online] <https://www.rd-alliance.org/groups/professionalising-data-stewardship-ig>
- [8] RDA IG on Education and Training on handling of research data (IG ETHRD) [online] <https://www.rd-alliance.org/groups/education-and-training-handling-research-data.html>
- [9] Towards FAIR Data Steward as profession for the Life Sciences, Final report ZonMw & ELIXIR-NL projects (Oct 3, 2019) [online] <https://doi.org/10.5281/zenodo.3471707>
- [10] The Danish e-Infrastructure Cooperation (DeIC) and Danish National Forum for Research Data Management (DM Forum) Report on National Coordination of Data Steward Education in Demark [online] https://www.deic.dk/sites/default/files/Data%20Steward%20Education%20in%20Denmark_0.pdf
- [11] GO FAIR Initiative [online] <https://www.go-fair.org/go-fair-initiative/>
- [12] EOSCpilot deliverable “D7.5: Strategy for Sustainable Development of Skills and Capabilities” The FAIR Guiding Principles for scientific data management and stewardship, March 2016, Scientific Data 3(160018 (2016)) DOI: 10.1038/sdata.2016.18
- [13] FAIR Data Maturity Model [online] https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v0.90.pdf
- [14] RDA Data maturity model Working Group [online] <https://www.rd-alliance.org/groups/fair-data-maturity-model-wg>
- [15] Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMA) [online] <http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf>
- [16] EDISON Data Science Framework (EDSF). [online] Available at <https://github.com/EDISONcommunity/EDSF>
- [17] The Data Science Framework, A View from the EDISON Project, Editors Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer Nature Switzerland AG 2020, ISBN 978-3-030-51022-0, ISBN 978-3-030-51023-7
- [18] CCS, 2012 The 2012 ACM Computing Classification System. Available at <http://www.acm.org/about/class/class/2012>
- [19] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at <https://ec.europa.eu/esco/portal/#modal-one>
- [20] The FOSTER Learning Objectives for Open Science, 23 February 2015 [online] <https://zenodo.org/record/15603#.YA8rD-hKhPY>
- [21] Terms4FAIRskills Initiative [online] <https://terms4fairskills.github.io/Announcement.html>
- [22] P21's Framework for 21st Century Learning [online] http://www.p21.org/storage/documents/P21_framework_0515.pdf
- [23] Tomasz Wiktorski, Yuri Demchenko, Adam Belloum, Model Curricula for Data Science EDISON Data Science Framework, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong. <https://ieeexplore.ieee.org/document/8241134>
- [24] Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, ISBN978-1-4503-6186-6/19/03.
- [25] 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/wpcontent/uploads/2014/02/European-e-CompetenceFramework-3.0_CEN_CWA_16234-1_2014.pdf

8 Data Management and Data Stewardship for Industry, Research and Academia

Establishing effective Data Management and Data Governance (DMG) in organisation is considered as a first step in digital transformation. Best practices in DMG are well defined by the DAMA (Data Management Association) and published as Data Management Body of Knowledge (DMBOK) [25] that defines a set of knowledge, competences and responsibilities of the main organisational roles and actors in the Data Management and Governance. The DNV GL Data Quality Assessment Framework [26] provides a set of industry best practices recommendations on how to achieve the best use of company's data resource, converting them into assets and bringing competitive benefits. Widely adopted in the research community, the FAIR data principles [27, 28] and Data Stewardship competence framework [29] provide a good contribution to building practically oriented DMG curricula. Below are outlines of the two distinctive courses that are related to DMG: Enterprise Data Management and Governance and Research Data Management and Stewardship.

8.1 Research Data Management and Stewardship (RDMS)

The research data management has numerous implementations and is well supported with training materials but in most cases, this is focused on the specific scientific domain. The courses also include growing popular FAIR principles and Data Stewardship related topics.

The following RDMS course example is structured along with practical aspects of research data management.

A. Use cases for data management and stewardship

- Preserving the Scientific Record

B. Data Management elements (organisational and individual)

- Goals and motivation for managing your data
- Data formats, Metadata, related standards
- Creating documentation and metadata, metadata for discovery
- Using data portals and metadata registries
- Tracking Data Usage, data provenance, linked data
- Handling sensitive data
- Backing up data, backup tools and services
- Data Management Plan (DMP)

C. Responsible Data Use (Citation, Copyright, Data Restrictions)

- Data privacy and GDPR compliance

D. FAIR principles in Research Data Management, supporting tools, maturity model and compliance

E. Data Stewardship and organisational data management

- Responsibilities and competences
- DMP management and data quality assurance

F. Open Science and Open Data (Definition, Standards, Open Data use and reuse, open government data)

- Research data and open access
- Repository and self- archiving services
- RDA products and recommendations: PID, data types, data type registries, others
- ORCID identifier for data and authors
- Stakeholders and roles: engineer, librarian, researcher
- Open Data services: ORCID.org, Altmetric Doughnut, Zenodo

G. Hands on practice includes the following topics:

- a) Data Management Plan design
- b) Metadata and tools
- c) Selection of licenses for open data and contents (e.g. Creative Common, and Open Database)

8.2 FAIR Teaching Handbook: Curricula Topic 15: Research Data Management: Overview and Best Practices

Audience/Target learners group:

This lesson is intended to deliver a concise overview of the Research Data Management (RDM) principles and practices for master students or professional audiences of vocational education and training.

Learning outcomes:

1. Understanding RDM process and main use cases
2. Understanding Open Research and Open Data (Definition, Standards, Open Data use and reuse, open government data, European policies and initiatives)
3. Understanding FAIR principles in Research Data Management, maturity model and compliance
4. Working with sensitive, personal or private data (General Data Protection Regulation [GDPR] and its requirements, Ethics approval process and form)
5. Understand what a Data Management Plan is, its purpose and benefits for a project or organisation
6. Know tools, guides, templates to support RDM, metadata management, DMP creation
7. Apply the acquired knowledge in practice, namely, be able to create a DMP, create and publish data and metadata
8. Understand the key roles in RDM: Data Steward, Chief Data Officer, Data Protection Officer and other employees of the institution who can support the creation of DMP;

Delivery format:

This lesson can be delivered in the form of tutorial, webinar or self-paced self-study course
Required time: 2 sessions of lecture (1.5 hrs) and 1 session of practice (approx 1.5 hrs)

Prerequisites:

Basic knowledge of computer software and applications
Understanding of organisational and/or research process and data used or produced

Lesson topics (Summary of Tasks / Actions):

- A. Use cases for research data management and stewardship
 - Preserving the Scientific Record
- B. Data Management elements (organisational and individual)
 - Goals and motivation for managing your data
 - Data formats, Metadata, related standards
 - Creating documentation and metadata, metadata for discovery
 - Using data portals and metadata registries
 - Tracking Data Usage, data provenance, linked data
 - Handling sensitive data
 - Backing up data, backup tools and services
- C. Responsible Data Use (Citation, Copyright, Data Restrictions)
 - Data privacy and GDPR compliance
- D. FAIR principles in Research Data Management, supporting tools, maturity model and compliance
- E. Data Management Plan (DMP)
- F. Data Stewardship and organisational data management
 - Responsibilities and competences
 - DMP management and data quality assurance
- G. Open Research and Open Data (Definition, Standards, Open Data use and reuse, open government data)
 - Research data and open access
 - Repository and self-archiving services

- Research Data Alliance (RDA) products and recommendations: Persistent Identifiers (PID), data types, data type registries, others
- ORCID identifier for data and authors
- Stakeholders and roles: engineer, librarian, researcher
- Open Data services: ORCID.org, Altmetric Doughnut, Zenodo

Practice:

Hands on practice including the following topics:

- a) Data Management Plan design, templates and tools
- b) Metadata and tools, metadata registries
- c) Selection of licences for open data and contents (e.g. Creative Common, and Open Database)

Materials / Equipment

1. Collection of DMP templates
2. Example metadata for research data and publications
3. Collection of links to RDM tools, metadata registries,

References

- General Data Protection regulation - <https://gdpr-info.eu/>
- License selector - <https://ufal.github.io/public-license-selector/>
- DMP Online - <https://dmponline.dcc.ac.uk/>
- DMP Templates - <https://guides.lib.umich.edu/c.php?g=283277&p=2138498>
- Towards FAIR principles for research software - <https://content.iospress.com/articles/data-science/ds190026>
- FAIR Cookbook, developed by Life Sciences academics and pharmas, 2021 <https://fairplus.github.io/the-fair-cookbook/content/home.html>
- [FAIRsharing](#) for (meta)data standards and interlinked repositories

Take Home Tasks

Organisational Data Management Plan creation (using provided template and/or using online tools)

8.3 Data Management and Governance (enterprise scope)

The DMG course uses DMBOK as general framework covering majority of topics, extending them with the Data Science and Big Data Analytics platforms and enriching with the FAIR and industry best practices. The following are the main topics that can be included in the course:

- Introduction. Big Data Infrastructure and Data Management and Governance.
- Data Management concepts. Data management frameworks: DAMA Data Management framework, the Amsterdam Information Model. Extensions for Big Data and Data Science.
- Enterprise Data Architecture. Data Lifecycle Management and Service Delivery Model. Data management and data governance activities and roles.
- Data Science Professional profiles and organisational roles, Skills management and capacity building.
- Data Architecture, Data Modelling and Design. Data types and data models. Data modeling. Metadata. Relational and not relational (SQL and NoSQL) databases overview. Distributed systems: CAP theorem, ACID and BASE properties.
- Enterprise Big Data infrastructure and integration with enterprise IT infrastructure. Data Warehouses. Distributed file systems and data storage.
- Big Data storage and platforms. Cloud based data storage services: data object storage, data blob storage, Data Lakes (services by AWS, Azure, GCP).
- Trusted storage, blockchain enabled data provenance.

- FAIR data principle and Data Stewardship, Data Quality assessment and maturity model. Data repositories, Open Data services, public services.
- Maturity: DNV-GL Data Quality Framework, DCC RISE, CIMM, etc
- Big Data Security and Compliance. Data security and data protection. Security of outsourced data storage. Cloud security and compliance standards and cloud provider services assessment.

The outlined topics above can be included in the practical courses for different target groups and at the different competence levels from Data literacy courses to professional training and academic curricula.

8.4 FAIR Teaching Handbook: Curricula Topic 16: Data Management and Governance in Industry and Research

Audience/Target learners group:

This lesson is targeted to deliver a concise overview of the Data Management and Governance (DMG) practices in research and industry for master students or professional audiences of vocational education and training, primarily with Computer or information science background.

Learning outcomes:

1. Understanding the Enterprise Data Management and Governance process and main use cases. DAMA (Data Management Association) Data Management Body of Knowledge (DMBOK)
2. Understanding European Data Spaces concept and initiatives, European policies and regulations, GDPR (General Data Protection Regulation)
3. Understanding elements of the enterprise data management infrastructure and services: Data Warehouses, cloud based storage, data lakes
4. Understanding data modelling process, data models, data structures. Master data management
5. Understanding FAIR principles in Research Data Management and their applicability to industrial use cases
6. Understanding data management maturity frameworks and best practices
7. Understand what a Data Management Plan is, its purpose and benefits for a project or organisation
8. Apply the acquired knowledge in practice, namely be able to create a DMP, assess organisational data security and compliance
9. Understand the key organisational roles in DMG: Chief Data Officer, Data Steward, Data Protection Officer and other roles

Delivery format:

This lesson can be delivered in the form of lecture+practice, tutorial or self-paced self-study course.

Suggested time: 2 sessions of lecture (1.5 hrs) and 1 session of practice (approx 1.5 hrs)

Prerequisites:

Basic knowledge of computer software and applications.

Understanding of organisational processes (HR/staff, customers, products, shipments, orders, etc.) and data used or produced.

Basic understanding of SQL for Advanced course

Lesson topics (Summary of Tasks / Actions):

The DMG course uses DAMA DMBOK as a general framework covering the majority of topics, extending them with Data Science and Big Data Analytics platforms and enriching them with FAIR and industry best practices. The following are the main topics that can be included in the course:

- Introduction. Big Data Infrastructure and Data Management and Governance. European Data Spaces: Definitions, Use cases. European policy on Data Governance, Data Protection, GDPR

- Data Management concepts. Data management frameworks: DAMA Data Management framework, the Amsterdam Information Model. Extensions for Big Data and Data Science.
- Enterprise Data Architecture. Data Lifecycle Management and Service Delivery Model. Data management and data governance activities and roles.
- Data Science Professional profiles and organisational roles, Skills management and capacity building.
- Data Architecture, Data Modelling and Design. Data types and data models. Metadata. Relational and not relational (SQL and NoSQL) databases overview. Distributed systems: CAP theorem, ACID and BASE properties.
- Enterprise Big Data infrastructure and integration with enterprise IT infrastructure. Data Warehouses. Distributed file systems and data storage.
- Big Data storage and platforms. Cloud based data storage services: data object storage, data blob storage, Data Lakes (services by AWS, Azure, GCP).
- Trusted storage, blockchain enabled data provenance.
- FAIR data principles and Data Stewardship, FAIR Digital Object and Persistent Identifier (PID).
- Data repositories, Open Data services, public services.
- Data Quality assessment. Data Management maturity frameworks: DNV-GL Data Quality Framework, DCC RISE, CIMM, etc
- Big Data Security and Compliance. Data security and data protection. Security of outsourced data storage. Cloud security and compliance standards and cloud provider services assessment.

Practice:

Hands-on practice including the following topics:

- a. Data Management Plan design, templates and tools
- b. Metadata and tools, metadata registries
- c. Assessing organisation's data security and compliance requirements
- d. Advanced: Data Modelling, Relational data model creation

Materials / Equipment

1. Collection of DMP templates
2. Example metadata for research data and publications
3. Collection of links to enterprise Data Management and Governance practices and recommendations,

References

1. DAMA Data Management Body of Knowledge (DMBOK), DAMA International, 2017
2. GO FAIR Initiative [online] <https://www.go-fair.org/gofair-initiative/>
3. General Data Protection Regulation - <https://gdpr-info.eu/>
4. [DMP Templates](https://guides.lib.umich.edu/c.php?g=283277&p=2138498) - <https://guides.lib.umich.edu/c.php?g=283277&p=2138498>
5. Towards FAIR principles for research software - <https://content.iospress.com/articles/data-science/ds190026>
6. A European strategy for data COM(2020) 66 final, 19.02.2020 <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066>
7. European Data Governance <https://ec.europa.eu/digital-single-market/en/european-data-governance>
8. EU/Parlament Regulation on European data governance (Data Governance Act) SEC(2020) 405 final, Nov 2020, <https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A52020PC0767>
9. GAIA-X – A Federated Data Infrastructure for Europe - <https://www.gaia-x.eu/>
10. FAIR Cookbook, developed by Life Sciences academics and pharmas, 2021 <https://fairplus.github.io/the-fair-cookbook/content/home.html>

Take Home Task

Organisational Data Management Plan creation (using provided template and/or using online tools)

9 EDISON Data Science Framework (EDSF): Addressing Demand for Data Science and Analytics Competences for the Data Driven Digital Economy²⁹

Emerging data driven economy including industry, research and business, requires 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. Data Science is becoming a newly recognised field of science that leverages Data Analytics methods with the power of Big Data technologies and Cloud Computing that both provide a basis for the effective use of the data driven research and economy models. Data Science research and education require a multi-disciplinary approach and data driven/centric paradigm shift. Besides core professional competences and knowledge in Data Science, increasing digitalisation of Science and Industry also requires new type of workplace and professional skills that raise the importance of critical thinking, problem solving and creativity required to work in a highly automated and dynamic environment. The education and training of the data related professions must reflect all multi-disciplinary knowledge and competences that are required from the Data Science and handling practitioners in modern, data driven research and the digital economy. In modern conditions with the fast technology change and strong skills demand, Data Science education and training should be customizable and delivered in multiple forms, also providing sufficient lab facilities for practical training. This paper discusses aspects of building customizable and interoperable Data Science curricula for different types of learners and target application domains. The proposed approach is based on using the EDISON Data Science Framework (EDSF) initially developed in the EU funded Project EDISON and currently being maintained by the EDISON Community Initiative.

9.1 Introduction

Emerging data economy, as a part of the more general The Fourth Industrial Revolution (referred to as Industry 4.0) is powered by the convergence of previously disconnected fields such as Cloud Computing, Big Data, Data Science and Analytics (DSA), Artificial Intelligence (AI), robotics, mobile technologies, 3D printing, nanotechnology and biotechnologies, that all are based on automation and digitalisation of organisational, industrial and business processes. Industry 4.0 will be characterized by fast development, a high level of technologies convergence and increased role of knowledge, skills and human factors to enable continuous and sustainable science and technology development. Such type of economy requires new type of data driven and Data Science and Analytics enabled competences and workplace skills.

Sustainable development of the modern data driven economy requires re-thinking and re-design of both traditional educational models and existing courses reflecting multi-disciplinary nature of Data Science and its application domains. However, at the 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 required by research and industry. In conditions of continuous technology development and shortened technology change cycle, Data Science education requires an effective combination of theoretical, practical and workplace skills. The importance of effective use of existing data analytics and data management platforms and tools and corresponding hands-on experience is growing and their elements need to be generically incorporated into modern curriculum design. The EDISON Data Science Framework (EDSF) [1, 2-5], which is the product of the EDISON Project, provides a basis for building an effective educational environment combining educational or training components and practical hands-on experience with virtual and data labs. The future educational model and approach should also solve different aspects of the future professionals that include both theoretical knowledge and practical skills that must be supported by corresponding education infrastructure and educational labs environment.

The paper refers to the previous authors' works that researched new approaches to building effective curricula in Cloud Computing, Big Data and Data Science [6, 7, 8, 9] and provides examples of curricula that are important to enable the digital transformation of organisations.

²⁹ Based on paper Yuri Demchenko, Tomasz Wiktorski, Steve Brewer, Juan José Cuadrado Gallego, EDISON Data Science Framework (EDSF): Addressing Demand for Data Science and Analytics Competences for the Data Driven Digital Economy, In Proc. Data Science Education (DSE), Special Session, EDUCON2021 – IEEE Global Engineering Education Conference, 21-23 April 2021, Vienna, Austria

9.2 Demand for Data Science and Data Skills

Growing demand for Data Science and Analytics enabled and general data driven professions is confirmed by multiple European and global market studies. Demand for data related professions will grow even more with the emerging Industry 4.0 [10] that will bring tremendous changes both to business models and the labour market with the big change in the skills set needed to thrive in the new economy landscape.

The key Industry 4.0 elements that both empower new data economy and will be facilitated by the new business and consumer models:

- Cyber-physical systems
- Internet of things
- Internet of services
- Smart factory
- Mobile technologies
- Cloud computing
- Big data

The World Economic Forum (WEF) published the report “The Future of Jobs” (2016) [11] that is focused on the employment, skills, and workforce strategy for the future economy. The report summarised the vision of the leading high-tech companies on future skills demand. The following 10 top skills are identified as critical for 2020 (reflecting a shift from currently required skills in the direction of independent critical thinking, creativity and cognitive flexibility) [12, 13]:

1. Complex problem solving
2. Critical thinking
3. Creativity
4. People management
5. Coordinating with others
6. Emotional Intelligence
7. Judgment and decision making
8. Services orientation
9. Cognitive flexibility

The IDG report 2017 [14] provided a deep analysis of the European data market and growing demand for data workers, the value of the data market, the number of data user enterprises, the number of data companies and their revenues, and the overall value of the impact of the data economy on EU GDP. The EU data market was estimated as EUR 60 Bln with growth to EUR 106 Bln in 2020. With the total number of data workers to grow 6.1 mln (2016) 10.4 million in 2020 the data worker skill gap is estimated as 769,000 or 9.8% (2020). Addressing this demand and gap is becoming critical for European economy and a challenge for universities. The report stresses that not satisfied demand in data workers will lead to an under-performing economy, industry, research and loss of competitiveness.

9.3 EDISON Data Science Framework (EDSF)

Designing a 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 EDISON Data Science Framework (EDSF) that is a core product of the EDISON Project. EDSF provides a basis for the definition of the Data Science profession and other components related to Data Science education, training, organisational roles definition and skills management, as well as professional certification and career transferability. Figure 1 below illustrates the main EDSF components and their inter-relations:

- CF-DS – Data Science Competence Framework [2]
- DS-BoK – Data Science Body of Knowledge [3]
- MC-DS – Data Science Model Curriculum [4]
- DSPP - Data Science Professional profiles and occupations taxonomy [5]
- Data Science Taxonomy and Scientific Disciplines Classification.

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

- Data Science Education Environment (DSEE) intended to be cloud based, customizable and aligned with the new workplace practices and skills
- Education and Training Directory connected to Marketplace and Virtual Data Labs
- Data Science Community Portal (CP) that provides information and community support services. It also provides a gateway to DSEE, Marketplace and Virtual Data Labs. CP is intended to include tools for individual competences benchmarking and personalized educational path building

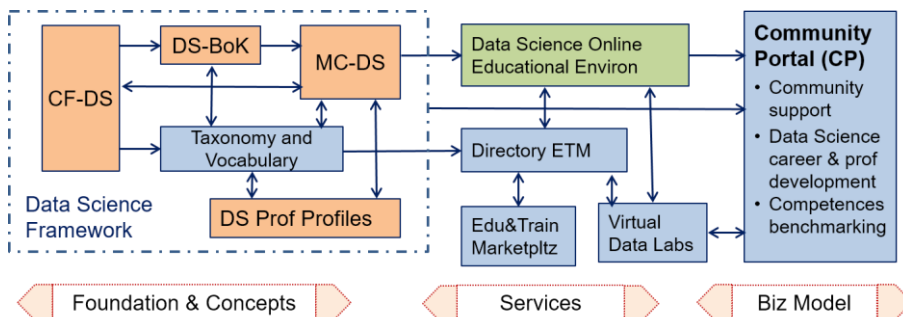


Figure 1. EDISON Data Science Framework components and Data Science Educational environment.

9.3.1 Data Science Competence Framework (CF-DS)

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 has competence structured according to the major identified functional groups (as explained below).

The CF-DS is structured along four dimensions (similar to European e-Competence Framework e-CFv3.0 [15]) that include (1) competence groups, (2) individual competences definition, (3) proficiency levels, and (4) corresponding knowledge and skills. In this context, each individual competence includes a set of required knowledge topics and a set of skills type A and skills type B. Such CF-DS structure allows for competence based curriculum design where competences can be defined based on the professional profile (see DSPP [5] for mapping between professional profiles and competences) or target learners group when designing a full curriculum, or based on competence benchmarking for tailored training to address identified competences and knowledge gaps.

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 the 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:

- Skills Type A which are related to professional experience and major competences, and
- Skills Type B that are related to a wide range of practical computational skills including using programming languages, development environment and cloud based platforms (refer to CF-DS [2] for a full definition of the identified knowledge and skills groups).

9.3.2 Workplace skills

Workplace skills, also referred to as “soft” skills or professional attitude skills, are becoming increasingly important in the modern data driven and future Industry 4.0 economy.

The CF-DS defined two groups of skills that are demanded by employers and required for Data Scientist to efficiently work in modern data driven agile companies:

- Data Science Professional and Attitude skills (Thinking and acting like Data Scientist) that define a special mindset that is developed by a practicing Data Scientist along with their career progression
- 21st Century skills that comprise a set of workplace skills that include critical thinking, communication, collaboration, organizational awareness, ethics, and others.

Universities should pay attention to developing such skills and include them in curricula or extra-curricula activities. Refer to CF-DS for a detailed skills definition.

9.3.3 Data Science Professional Profiles (DSPP)

The proposed Data Science professional profiles definition is based on the analysis of the research and industry demand in data related professions. The identified professional profiles are classified using ESCO taxonomy [16], and necessary extensions are proposed to support the following hierarchy of data handling related occupations:

- Managers: Chief Data Officer (CDO), Data Science (group/department) manager, Data Science infrastructure manager, Research Infrastructure manager
- Professionals: Data Scientist, Data Science Researcher, Data Science Architect, Data Science (applications) programmer/engineer, Data Analyst, Business Analyst, etc.
- Professional (database): Large scale (cloud) database designers and administrators, scientific database designers and administrators
- Professional (data handling/management): Data Stewards, Digital Data Curator, Digital Librarians, Data Archivists
- Technicians and associate professionals: Big Data facilities operators, scientific database/infrastructure operators
- Support and clerical workers: Support and data entry workers.

The individual profiles are defined in accordance with the CWA 16458 (2012): European ICT Professional Profiles [17] standard (and its revision 2018)

The DSPP document also defines mastery levels and corresponding learning outcomes for the targeted education or training. The following mastery levels are defined (using workplace terminology that can be easily mapped to mastery levels defined in MC-DS):

A - Awareness

- 1) Understand Terminology
- 2) Understand the Principles
- 3) Apply principles
- 4) Understand the Methods

U - Use/Application

- 5) Apply basics
- 6) Supervised use
- 7) Unsupervised Use

P - Professional/Expert

- 8) Development of applications using a wide range of technologies
 - 9) Supervise project development, a team of professionals,
- where borderline mastery levels 4 and 7 actually belong to both higher level and lower level groups.

9.3.4 Data Science Body of Knowledge and Model Curriculum

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) [18], incorporates best practices in defining domain specific BoK's and provides a reference to existing related BoK's. It also includes the 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. The practical curriculum should be supported by a corresponding educational environment for hands-on labs and educational projects development.

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

9.4 Example Curricula to Facilitate Digital Transformation

This section provides example curricula “Data Science Accelerator” developed and taught by the EDISON project university partners specifically oriented on supporting digital transformation of organisations to adopt the Agile Data Driven Enterprise model (ADDE).

9.4.1 Data Management and Data Governance

Establishing effective Data Management and Data Governance (DM&DG) in an organisation is considered the first step in digital transformation. Best practices in DM&DG are well defined by the DAMA (Data Management Association) and published as Data Management Body of Knowledge (DMBOK) [22] and corresponding guidelines.

The DM&DG course uses DMBOK as a general framework covering the majority of topics and extending them with the Data Science and Big Data Analytics platforms. The following are the main topics included in the course:

- Introduction. Big Data Infrastructure and Data Management and Governance.
- Data Management concepts. Data management frameworks: DAMA Data Management framework, the Amsterdam Information Model. Extensions for Big Data and Data Science.
- Enterprise Data Architecture. Data Lifecycle Management and Service Delivery Model. Data management and data governance activities and roles. Data Science Professional profiles family. Skills management and capacity building.
- Data Architecture, Data Modelling and Design. Data types and data models. Data modeling. Metadata. SQL and NoSQL databases overview. Distributed systems: CAP theorem, ACID and BASE properties.
- Enterprise Big Data infrastructure and integration with enterprise IT infrastructure. Data Warehouses. Distributed file systems and data storage. Cloud based data storage services: data object storage, data blob storage, Data Lakes (services by AWS, Azure, GCP).
- Big Data storage and platforms. Big Data Security and Compliance. Data security and data protection. Security of outsourced data storage. Cloud security and compliance standards and cloud provider services assessment.

9.4.2 Data Science and Analytics Foundation (DSAF)

The goal of this course is to introduce the students to the whole spectrum of Data Science and Analytics technologies and, at the same time, provide strong statistical background for future mastering core data analytics methods and Machine Learning technologies. This defines the main stress in DSAF course on statistical methods, probability theory, hypothesis testing, data preparations, methods of qualitative and quantitative analytics. The primary analytics for this course is recommended to have a low programming threshold to enable fast learning. The RapidMiner visual data analytics environment (<https://rapidminer.com/>) was identified as a preferable choice against R or python tools to introduce the trainees to the key data analytics methods and enable active experimentation.

The following topics are included in DSAF curriculum:

- Introduction and course overview: Data Science and Big Data technologies, Data Science competence and skills, Research Methods in Data Science, Machine Learning and Data Mining overview.
- Statistical methods and Probability theory
- Data description and Statistical Data Analysis
- Data preparation: data entry, data cleaning, data pre-processing
- Qualitative and Quantitative data analysis
- Classification: methods and algorithms
- Cluster analysis basics and algorithms
- Performance of data analytics algorithms and tools

9.4.3 Professional Issues in Data Science

The goal of this course is to equip the students and practitioners with the knowledge and skills for further focused study of more specific Data Science and Analytics areas and courses.

- Data Science Competences and Skills management and capacity building, EDISON Data Science Framework.
- Data Science professional skills (“Act and think as Data Scientist”) and 21st Century Skills.
- Data Science and Analytics methods and technologies overview.
- Research Methods; Business processes management.
- Data Management in research, industry and personal: standards and best practices.
- FAIR (Findable, Accessible, Interoperable, and Re-usable) principles in Open Data and enterprise data management.
- Ethical and legal principles and regulations. Privacy enabling technologies.

It is also beneficial to supply this course with the guided/tutored groups and/or individual training on essential professional skills such as complex problem solving, critical thinking, creativity, etc. defined as critical for Industry 4.0 workforce.

9.4.4 Cloud based DSEE and Virtual Data Labs

The educational Data Science labs and project development environment can benefit from using clouds and available data analytics and data handling applications and services that can be made available on demand for specific time periods when the education or training takes place. Using cloud resources to build an effective and up-to-date professional Data Science education environment is inevitable with current fast technology development and required computational performance that can be requested on-demand.

Major Cloud Service Providers (CSP) provides a wide range of data analytics and business analytics services and platforms that can be equally used by big, small and medium companies and individuals on a pay-per-use basis. In addition to the possibility to use the same resources for education and training purposes, the major CSPs also provide designated education and self-training resources that are in many cases also supported educational grants for students and teachers.

An important component of Data Science education is educational datasets that often need to be provided with their specific applications. While many educational datasets are available from mentioned above cloud platforms, from the community run Kaggle (<https://www.kaggle.com/>) and UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/index.php>), the use of cloud based VDLabs allows instantiating the whole experimental setup or environment together with used data sets in case of specific domain focused education or training.

9.5 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 the fully enumerated definition of EDSF components and individual units. Besides defining academic components of the effective and consistent curriculum, EDSF also provides advice on the required Data Science Education Environment to facilitate fast practical knowledge and skills acquisition by students and learners.

Business Higher Education Forum (BHEF) has published two important reports in cooperation with PriceWaterhouseCoopers (PwC), IBM and Burning Glass Technologies (BGT) [19, 20] that studied the 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.

The authors’ experience of developing a pilot project for re-/up-skilling employees of one of the Dutch governmental organisations confirmed a trend that organisations, in a way to become data driven and agile, will intend to make the existing organisational roles DSA enabled and require corresponding DSA training in a customizable and flexible form.

An effective professional education needs to 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. The wide use of available online resources and platforms for so demanded Data Science and other digital and data skills will facilitate the adoption of FAIR principles in the future Open Education to become Findable, Accessible, Interoperable, and Re-usable that were initially proposed for Open Data [21]. The universities can contribute to building FAIR life-long educational space that can serve both organisational and individual needs of students and learners, including support for widely apprised citizen scientists.

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.

References

- [1] EDISON Data Science Framework (EDSF). Available at <http://edison-project.eu/edison/edison-data-science-framework-edsf>
- [2] Data Science Competence Framework. Available at <http://edison-project.eu/data-science-competence-framework-cf-ds>
- [3] Data Science Body of Knowledge. Available at <http://edison-project.eu/data-science-body-knowledge-ds-bok>
- [4] Data Science Model Curriculum. Available at <http://edison-project.eu/data-science-model-curriculum-mc-ds>
- [5] Data Science Professional Profiles. Available at <http://edison-project.eu/data-science-professional-profiles>
- [6] Demchenko, Yuri, Emanuel Gruengard, Sander Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. 1st IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science, in Proc.The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 December 2014, Singapore.
- [7] Manieri, Andrea 2015, 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
- [8] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Steve Brewer, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, 3rd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2016), in Proc.The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 December 2016, Luxembourg.
- [9] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong.
- [10] The Fourth Industrial Revolution: what it means, how to respond. [online] <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond>
- [11] The Future of Jobs, World Economic Forum Report, 18 January 2016 [online] http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf
- [12] The 10 skills you need to thrive in the Fourth Industrial Revolution [online] <https://www.weforum.org/agenda/2016/01/the-10-skills-you-need-to-thrive-in-the-fourth-industrial-revolution/>
- [13] Are you ready for Industry 4.0. [online] <http://www.delivered.dhl.com/en/articles/2017/02/skills-for-industry-4-0.html>
- [14] 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>
- [15] 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
- [16] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at <https://ec.europa.eu/esco/portal/#modal-one>
- [17] European ICT Professional Profiles CWA 16458 (2012) (Updated by e-CF3.0) [online] http://relaunch.ecompetences.eu/wp-content/uploads/2013/12/EU_ICT_Professional_Profiles_CWA_updated_by_e_CF_3.0.pdf
- [18] CCS, 2012 The 2012 ACM Computing Classification System. Available at <http://www.acm.org/about/class/class/2012>
- [19] 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>
- [20] 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/IML14576usen/IML14576USEN.PDF>
- [21] Barend Mons, et al, The FAIR Guiding Principles for scientific data management and stewardship [online] <https://www.nature.com/articles/sdata201618>
- [22] DAMA Data Management Body of Knowledge (DMBOK2), DAMA International, 2017

10 Transversal Skills required by Emerging Industry 4.0 Transformation

The emerging data-driven economy (also defined as Industry 4.0 or simply 4IR), encompassing industry, research and business, requires new types of specialists that are able 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 collectively defined as the Data Science family of professions. Data Science as a research and academic discipline provides a basis for Data Analytics and ML/AI applications. The education and training of the data related professions must reflect all multi-disciplinary knowledge and competences that are required from the Data Science and handling practitioners in modern, data-driven research and the digital economy. In the modern era, with ever faster technology changes, matched by strong skills demand, the Data Science education and training programme should be customizable and deliverable in multiple forms, tailored for different categories of professional roles and profiles. Referring to other publications by the authors on building customizable and interoperable Data Science curricula for different types of learners and target application domains, this paper is focused on defining a set of transversal competences and skills that are required from modern and future Data Science professions. These include workplace and professional skills that cover critical thinking, problem solving, and creativity required to work in a highly automated and dynamic environment. The proposed approach is based on the EDISON Data Science Framework (EDSF) initially developed within the EU funded Project EDISON and currently being further developed in the EU funded MATES project and also the FAIRsFAIR projects.

10.1 Introduction

The emerging data-driven economy, as a part of a more general Fourth Industrial Revolution (also referred to as Industry 4.0 or simply 4IR) is powered by the convergence of previously disconnected fields such as Mathematics, Cloud Computing, Big Data, Data Science and Analytics (DSA), Artificial Intelligence (AI), robotics, mobile technologies, 3D printing, internet of things (IoT), nanotechnology and biotechnologies, that are all based on automation and the digitalisation of organisational, industrial and business processes. Industry 4.0 is characterized by fast development, high levels of technology convergence, and an increased role for knowledge, skills and human factors to enable continuous and sustainable science and technology development. Such a type of economy requires new types of data-driven approaches, and Data Science and Data Analytics enabled competences and related workplace skills.

In conditions of continuous technological development and shortened technology change cycles, Data Science education requires an effective combination of theoretical, practical and workplace skills. The importance of effective application of existing data analytics and data management platforms and tools, and corresponding hands on experience is growing, and their elements need to be generically incorporated into modern curriculum design.

The EDISON Data Science Framework (EDSF) [1, 2-5], which is the product of the EDISON Project, provides a basis for building such effective education and training curricula. This paper discusses ongoing developments to extend EDSF with the set of transversal competences and skills that are required for modern and future Data Science professions. These include workplace and professional skills such as critical thinking, problem solving, and creativity, and the ability to thrive in a highly automated and dynamic environment.

10.2 Digital Competences and Data Literacy

In the context of digital transformation and growing AI-based automation, all professional profiles should possess sufficient levels of digital competences and skills to operate successfully in Industry 4.0. The EC study and report on “Digital Competences for Citizen” (DigComp) published in 2018 [20] provides good advice for addressing both digital skills and data skills in professional workplace training and vocational education. An important part of digital competences is understanding the role of data and processes related to data handling in modern applications, social media, industrial processes, research, and specifically how data are used in AI based decision making and control. Special training and skills development must be focused on data processing and management issues.

- Digital Competences and Data Literacy
- The Digital Competences for Citizens (DigComp 2.1, 2017)

- An important part of digital competences is understanding the role of data and processes related to data handling
 - In modern data centric applications
 - Social media
 - Industrial processes
 - Research
 - Specifically, how data are used in AI based decision making and control
 - Special training and skills development must be focused on data processing and management issues
- [ref] <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/digcomp-21-digital-competence-framework-citizens-eight-proficiency-levels-and-examples-use>

Table 1 Digital Competences defined in DigComp 2.1

Competence area	Competence
Competence area 1: Information and data literacy	1.1 Browsing, searching and filtering data, information and digital content
	1.2 Evaluating data, information and digital content
	1.3 Managing data, information and digital content
Competence area 2: Communication and collaboration	2.1. Interacting through digital technologies
	2.2. Sharing through digital technologies
	2.3. Engaging in citizenship through digital technologies
	2.4. Netiquette
	2.4 Collaborating through digital technologies
	2.5 Netiquette
Competence area 3: Digital content creation	3.1 Developing content
	3.2 Integrating and re-elaborating digital content
	3.3 Copyright and licenses
	3.4 Programming
Competence area 4. Safety	4.1. Protecting devices
	4.2. Protecting personal data and privacy
	4.3. Protecting health and well-being
	4.4. Protecting the environment
Competence area 5. Problem Solving	5.1. Solving technical problems
	5.2. Identifying needs and technological responses
	5.3. Creatively using digital technologies
	5.4. Identifying digital competence gaps

10.3 Suggested Knowledge and curriculum topics

10.3.1 Digital and Data Literacy Topics

A. Data – B. Cloud – C. Digital Content – D. Data Science & Analytics

Example Digital competences curriculum for training

A. Data related competences and technologies

- A.1. Big Data definition and technologies: 6V of Big Data and challenges for companies. Big Data examples from research and industry
- A.2. Data collection, access and sharing
- A.3. Data formats, data models, metadata
- A.4. Data Storage and databases, SQL scripting and simple commands
- A.5. Data inspection, Data protection, data backup and archiving
- A.6. Cloud based services and tools for data storage, sharing and management
- A.7. Open Data repositories, test datasets, developer communities
- A.8. Organisational and private Data Management, FAIR Data Principles, organisational roles, Data Stewards

B. Cloud services and cloud economics

- B.1. Cloud service models: IaaS, PaaS, SaaS, Apps. Use examples and Cloud Service Providers. Cost model of cloud services.
- B.2. Company IT infrastructure migrating to cloud: benefits and problems
- B.3. Cloud and Big Data, Cloud based Big Data platform and services
- B.4. Data storing, backing up, sharing and processing in clouds (for organisational and private data)
- B.5. Practical exercises with cloud services: Cloud management console and cloud services deployment and access.

C. Digital content creation, access and management

This group of competences and skills is expected to be acquired as self-study or known to trainees modern practitioners. Additionally, online tutorials can be recommended.

D. Data Science and Big Data Analytics

This course is provided as a general overview of the listed below topics. More in-depth training and learning will require more time commitment and pre-requisite knowledge.

- D.1. Statistical methods and Probability theory
- D.2. Data description and Statistical Data Analysis
- D.3. Data preparation: data loading, data cleaning, data pre-processing, parsing, transforming, merging, and storing data
- D.4. Qualitative and Quantitative data analysis
- D.5. Classification: methods and algorithms
- D.6. Cluster analysis basics and algorithms
- D.7. Performance of data analytics algorithms and tools
- D.8. Visualizations of data analysis and dashboards
- D.9. Organizing data analytics process following CRISP-DM and Data Science Process

11 Big Data Value Data Science Badges

Authors

Ernestina Menasalvas, Ana M. Moreno, and Nik Swoboda
Universidad Politécnica de Madrid, Madrid, Spain

11.1 Introduction

With the development of new technology and the digital transformation of our economy, the labor market has also evolved. Nowadays, applicants for a job are no longer asked to submit a traditional paper resume, this information is presented digitally; recruiters and headhunters search the Internet (on an international level) for candidates who have the required skills for their needs; and some assessment of candidates can be done on-line. Moreover, the demands for labor are 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 chapter, we will first discuss mechanisms for recognizing skills in the EU with a focus on the internationalization, digitalization, and flexibility of those credentials. Then we will consider their application to data science with the goal of proposing a new framework for the recognition skills in data science. We begin with a brief review of the main challenges we hope to address.

How can we standardize credentials throughout Europe?

Although political institutions in the EU have strived to coordinate and standardize 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 recognize 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 evaluate the capabilities of potential employees.

Currently, there is no automatic system for the EU-wide recognition of academic diplomas; students can only request a "statement of comparability" for their university degree. This statement of comparability details how the student's diploma compares to the diplomas of another EU country. Something similar happens with the recognition of professional qualifications: 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. 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 harmonized minimum training conditions, i.e., nurses, midwives, doctors, dentists, 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 professional activities such as carpenters, upholsterers, beauticians etc.;

Additionally, since January 18, 2016 the European professional card (EPC) has been available for five professions (general care nurses, physiotherapists, pharmacists, real estate agents and mountain guides). It is an electronic certificate which can be applied for online.

Unfortunately, these existing mechanisms do not easily accommodate many professions, including that of data science.

How can data science credentials be: digital, verifiable, granular, and quickly evolving?

Traditionally skills and credentials were conveyed via a resume on paper and other paper-based credentials. Nowadays, this information can be shared via the Internet in web pages, social media, and in many other forms. The digitalization 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, and
- 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.

Overview

This chapter begins with a summary of current trends regarding education and skills in Europe. From these trends, we extract a series of desirable properties for a data science skills recognition scheme. Then, a survey of different mechanisms for the recognition of skills is presented. Next, a comparison and critical analysis of these recognitions is provided while taking into account the previously identified desirable properties. Lastly, a recommendation for a data science skills recognition process is proposed.

11.2 The Strategic Framework for Education and Training

Even though the educational systems of each country in the EU are managed separately at the national level, a common EU policy exists to support those systems and to help address common challenges faced by the EU. The Strategic Framework for Education & Training 2020 (ET-2020) contains the current policies of the EC for cooperation on education and training. These policies were initially adopted in 2009 and contained four common objectives that should be met by 2020 in the EU:

- Making lifelong learning and mobility a reality.
- Improving the quality and efficiency of education and training.
- Promoting equity, social cohesion, and active citizenship.
- Enhancing creativity and innovation, including entrepreneurship, at all levels of education and training.

In reviewing the progress of the ET-2020, the 2015 Joint Report of the Council and the Commission on the implementation of the strategic framework for European cooperation in education and training (2015-JR-SFECT) included a new set of “priority areas for European cooperation in education and training”:

1. Relevant and high-quality skills and competences for employability, innovation, active citizenship
2. Inclusive education, equality, non-discrimination, civic competences
3. Open and innovative education and training, including by fully embracing the digital era
4. Strong support for educators
5. Transparency and recognition of skills and qualifications
6. Sustainable investment, performance and efficiency of education and training systems

As the emphasis on promoting the “transparency and recognition of skills and qualifications” is particularly relevant to the task of recognizing data science skills we will focus further on that priority area. To explain this priority area, the report identifies these concrete needs:

- Fostering transparency, quality assurance, validation and recognition of skills and/or qualifications, including those acquired through digital, online and open learning and the validation of informal and non-formal learning
- Simplifying and rationalizing the transparency, documentation, validation and recognition tools that involve direct outreach to learners, workers and employers and further implementing the EQF [European Qualifications Framework for lifelong learning] and NQFs [National Qualifications Frameworks]
- Supporting the mobility of pupils, apprentices, students, teachers, members of educational staff and researchers
- Developing strategic partnerships and joint courses, in particular through increasing internationalization of higher education and vocational education and training

To help reach these goals the EU has promoted several programs:

- *The European Qualifications Framework for Lifelong Learning (EQF)* supports the process of validating qualifications by providing a common reference for qualification levels throughout Europe and the linking of member state validation systems with formal qualifications systems.
- *Validation of non-formal and informal learning* is the process of recognizing an individual’s knowledge, skills and competences gained outside formal educational systems. To help to achieve this recognition,

the CEDEFOP (European Centre for the Development of Vocational Training) in cooperation with the EC have defined some guidelines for validating non-formal and informal learning. They have also proposed an up-to-date European Inventory that provides an overview for each country and good practices for the design and implementation of validation initiatives.

- The *Europass portfolio* also relates to validation systems because it documents learning and enables users to display their skills, qualifications and experiences in a uniform way across Europe.
- The *European Credit Transfer and Accumulation* system for higher education and the *European Credit system for Vocational Education and Training* (ECVET) standardize the quantification of formal learning throughout Europe, making it easier to compare the time spent in educational programs.
- *Quality assurance* arrangements in higher education and vocational training to make education systems easier to understand for students and employers, by improving transparency tools.

We will now elaborate more upon the most relevant points of a number of key initiatives within this framework.

11.2.1 New Skills Agenda for Europe

The New Skills Agenda for Europe (NSAE) is one of the initiatives of the EU to help meet the targets of the ET-2020. On June 10, 2016, the European Commission presented this agenda, which sets out different guidelines to guarantee that the most adequate training, skills and career guidance is accessible to everyone in the EU.

The New Skills agenda emphasizes “the strategic importance of skills for sustaining jobs, growth and competitiveness,” and is focused on three key points:

- Improving the quality and relevance of skills formation
- Making skills and qualifications more visible and comparable
- Improving skills intelligence and information for better career choices

The most relevant concerns and recommendations mentioned in the New Skills Agenda regarding data science skills in Europe include:

- Future credentials should easily allow the comparison of students’ skills throughout the EU.
- Both the employed and the unemployed need adequate ways to present their skills and qualifications. Employers need ways to identify and recruit new employees with the skills that they need.
- Once skills qualifications are easily accessible, current and future demands for skills could be identified with data science analysis (skills intelligence).
- Current qualification systems focus on the learning outcomes of formal education programs, but do not validate non-formal and informal learning. Ongoing learning, including learning at the workplace, needs to be encouraged.
- Skills acquisition should not only be in formal education and training (literacy, numeracy, science, foreign languages) but also transversal skills (teamwork, creative analysis, problem solving, entrepreneurship, etc.).

11.2.2 New Europass framework

In February 2005, the Europass was launched, following the decision of the European Parliament and the Council to create a single framework whose goal was to make individuals skills and qualifications more comprehensible in Europe. This was done in the interest of facilitating the mobility of students and workers. The Europass consisted of a portfolio of five documents: the Europass Curriculum Vitae (CV); the Europass Language Passport; the Europass Mobility; the Europass Certificate Supplement; and the Europass Diploma Supplement.

Despite the fact that the European CV has undergone significant improvements to adapt to the changes brought by the technological revolution, the initial Europass did not address the changing educational, training and labor market conditions:

- It focused on documents and templates that are not compatible with the use of social media, mobile devices, Big Data analysis and job matching tools;
- It did not face the growing relevance of modern learning, that needed an easy way to record skills and qualifications acquired through non-formal or informal learning, including on-line learning;
- It did not take into account the use of tools such as ‘open badges.’

On October 4, 2016 the Commission decided to revise the Europass Decision, by building a new Europass framework that contributes to the display of people's skills and qualifications in a unified manner for all EU countries.

The proposal addresses challenges regarding the way that information technology has changed the labor market and new educational possibilities:

- The publication of employment offers, job applications, candidate's evaluation and recruiting are increasingly done online through tools that use social media, Big Data and other technologies, making it easier to find information on skills and qualifications.
- Education and training is increasingly offered on-line using digital platforms; at the same time, skills, experiences and learning achievements (formal and non-formal) are recognized in different forms, such as open badges.

With the new framework users can display their skills and qualifications in new formats, as the revised Europass uses open standards to facilitate the exchange of electronic data and defines authentication measures to ensure the validity of the digital content.

To achieve this goal, the new infrastructure includes several new tools, which allow users to give evidence of their skills and qualifications in all EU languages; these tools are:

- An online tool to create persona profiles, including both the traditional CV, with work experience and training/education and the skills recognition.
- Applications to help to evaluate the users' skills
- Information on learning opportunities across Europe
- Assistance on how to get a user's skills recognized
- Labor market intelligence, to learn which skills are more valuable

The new Europass is connected to other EU tools and services related to work, education and training systems, to encourage the exchange of information and to help users in their education and career path decisions.

11.2.3 European Qualifications Framework for lifelong learning

In April 2008, the European Parliament and Council resolved to establish the European Qualifications Framework for Lifelong Learning. The process was voluntary, and countries were invited to implement the framework in two stages: the first, that was to be completed by 2010, relating national qualification levels to the EQF; and the second, by 2012, ensuring that all new qualifications issued in Europe include references to the appropriate EQF level.

The aim of this framework is to provide a way to compare and interpret the levels of different qualification in the EU and thereby make those qualifications more transparent. This also facilitates mobility, having positive effects for learners and workers, who can have their level of competence recognized using a standard description all across Europe. This proposal is also beneficial for recruiters and education providers, who will be able to understand the applicants' qualifications. The adoption of a common reference framework eases the comparison and recognition of traditional qualifications issued by national authorities and those awarded by third parties (e.g., multinational companies). This allows the comparison of formal and non-formal education by increasing the transparency of qualifications awarded outside the formal education system.

The EQF can be applied to any kind of education, training or qualification including required basic education to advanced academic and professional and training. It consists of eight qualification levels, given in terms of learning outcomes: knowledge, skills and competences. These levels take into consideration theoretical knowledge, practical skills, technical skills, and social competences.

However, the adoption process of the EQF has proved difficult: differences have appeared when comparing general education certificates in different national systems with the EQF levels. For example, for a similar school certificate, some countries assign a level 2 or 3 (for secondary education) and others a level 4 or 5 (for higher education). This same problem has occurred with vocational education.

11.2.4 European Skills, Competences, Qualifications and Occupations

The European Skills, Competence, Qualifications and Occupations (ESCO) framework is a multilingual classification system that aims to bridge the communications gap between industry and those offering training through common reference terminology. The ESCO initiative was launched in 2010 by the EU and the first version of the ESCO framework was published in July 2017.

The ESCO benefits those in the labor market and in the education and training sector in a variety of ways:

- By providing a better matching of people to jobs by employment services or electronic tools:
 - Helping employers define the set of skills, competences and qualifications for a vacant job.
 - Helping job seekers build professional profiles in a terminology that suits job vacancies
 - Enable mobility through Europe
- By supporting education and training systems in the move to learning outcomes that better meet labor market needs:
 - Supporting the provision of information to education and training institutions that can help them in the development of new curricula.
 - Helping to provide more transparent information to students on learning outcomes and the relevance of qualifications to the labor market before they commence education or training.

By supporting evidence-based policy making:

- Enhancing the “collection, comparison and dissemination of data in skills intelligence and statistics tools, among others, in the European Skills Panorama.”

The ESCO is built upon three pillars: Occupations, Skills/Competences and Qualifications.

Regarding qualifications, the ESCO is based on the EQF framework and the national databases that most Member States developed or are developing, in which they assign an EQF level to each qualification and describe the expected outcome.

The most relevant ESCO guiding principles are:

- Useful, ESCO aims to become the de facto standard of the identification of occupations, skills competence and qualifications.
- Accepted, ESCO aims to be voluntarily adopted by stakeholders.
- Updated, ESCO will be continuously updated and adapted.
- Flexible, ESCO does not aim to standardize the scope of occupations but to provide standard terminology.
- High-quality, different stakeholders carefully ensured the quality of the ESCO.
- Transparent and open development, results were shared with interested parties and it was open to all stakeholders.
- Machine readable and compatible with existing IT systems and standards.

11.2.5 Highlights and common threads in these initiatives

After reviewing these political trends in Europe, we can now extract some of the key properties that a data science skills recognition program should contain in order to be in agreement with these existing efforts.

A data science skills recognition system should:

- P1 (2015-JR-SFECT, NSAE, Europass, EQF, ESCO) be transparent, accessible and allow the easy comparison of students’ skills throughout the EU
- P2 (2015-JR-SFECT, ESCO) include an assurance of quality
- P3 (2015-JR-SFECT) provide tools for their verification and validation
- P4 (2015-JR-SFECT, NSAE, Europass, EQF) include skills acquired through traditional, digital, online, and open learning, as well as the validation of informal and non-formal learning
- P5 (2015-JR-SFECT, EQF, ESCO) be compatible with the EQF
- P6 (NSAE) influence the relevance of the skills being acquired
- P7 (NSAE, Europass, ESCO) allow the digital analysis of both the demand for and the availability of skills.

- P8 (Europass) allow their use online: in platforms like the Europass, social media and on mobile devices
- P9 (EQF, ESCO) focus on learning outcomes and not on traditional measures such as hours of study

With these goals in mind, we will now look at the most common and popular methods of recognizing skills.

11.3 Education and Training Recognitions

11.3.1 A Survey of Recognitions

Accreditations

"Accreditation [is] the formal recognition by an independent body, generally known as an accreditation body, that a certification body operates according to international standards."

In the context of higher education in Europe, accreditation is the process by which an educational program acquires the right to grant degrees. In Europe, most accreditation agencies are endorsed by national governments and accredit all of that country's degrees. Sometimes these agencies recognize European or international accreditations and simplify the national accreditation process for programs already accredited at the European or International Level.

Requirements for accreditation vary depending on the accreditation agency. Some examples of accreditation agencies include: the Agencia Nacional de Evaluación de la Calidad y Acreditación in Spain, the UK Accreditation Service in the United Kingdom, the Accreditation Board for Engineering and Technology or the World Association of Conformity Assessment Accreditation Bodies.

University/Academic degrees

Of the collection of recognitions we will review, this is most certainly the one with the longest history. For example, the notion of a doctorate was established in medieval Europe and was considered to be a license to teach at the university level. Nowadays, the European system of university degrees (bachelor degree, master degree, and the doctorate) are used worldwide.

Accredited College and University programs have the right to award academic degrees. It should be noted that in some parts of the world, unaccredited programs (sometimes referred to as degree mills) can legally confer degrees but these degrees are often considered to be of little worth.

A great variety of requirements exist but typically four years of university study must be completed to be awarded a bachelor degree, two additional years of study for a master degree, and a significant research contribution is required for the awarding of a doctorate.

In many private and public sector jobs, both pay scales and position prerequisites are directly related to the degrees held by a candidate. In some parts of the world, holding a degree results in a change of title (Doctor for example) and in others it results in the right to use post-nominal letters (BA for example).

Noteworthy examples in Big Data/Data Science include: M.Sc. Big Data & Business Analytics, University of Amsterdam, M.Sc. Applied Informatics, Vytautas Magnus University, M.Sc. Data Science, Sapienza Università di Roma or M.Sc. Computer Science, National University of Ireland.

Certificates

A certificate is simply a document that attests to the fact that a certain individual has "received specific education or has passed a test or series of tests." Though in reality they are very varied in use, in the context of the computing industry certificates are most commonly used to recognize knowledge regarding a specific set of skills. There are both academic and professional certificates with the former being awarded by higher education providers while the latter are awarded by professional organizations or individual companies related to their own products.

Certificates typically require less effort to obtain than an academic degree. The examples given below have a duration of between 6 and 18 months. Many certificate programs are specifically designed for 'continuing education' students who are already employed full-time but are trying to advance their careers. Many academic certificates are associated with traditional coursework and are basically equivalent to having passed with a certain grade a set of courses. Most professional certificates require passing a test, some also require a certain

amount of professional experience to be eligible for certification. Many professional certificates expire after a certain period of time and require renewal by completing continuing education courses and/or exams.

Corporations often require that service providers have staff with certain certifications before they can provide specific services. Often job advertisements specifically require certain certifications. For example, many Network Engineer positions require that applicants have a Cisco Certified Network Associate (CCNA) certificate. Relevant examples of academic certificates in Big Data/Data Science include:

- Harvard Data Science Certificate: Cost 11,360 USD, Duration: 1.5 years, all courses can be completed online
- University of California, Irvine Data Science Certificate Program: Cost: free for UCI graduate students, Duration: 32 hours of courses/workshops
- Georgetown University Certificate in Data Science: Cost: 7,496.00 USD, Duration: 6 months, courses held on Friday evenings and Saturdays
- UC Berkeley Certificate Program in Data Science: Cost 5,100 USD (+ materials + registration), Duration 150 hours of class

Labels

Labels are a distinction awarded to an existing degree program. The use of the term label appears to be almost exclusively European. Like accreditations, but unlike the other recognitions mentioned here, it is applied to the program itself and is not normally thought of as being applied to the program's participants (except through association). A label publicly recognizes that the program in question meets the requirement of the label issuer. There is no clear consensus regarding who can confer a label, but at the moment, the most noteworthy labels are conferred by programs funded by the European Commission. Each organization is free to design any requirements, which it sees fit.

Noteworthy examples are Erasmus Mundus Master of Excellence, Erasmus+, EIT Digital, Eur-ace, Euro-Inf, and The European Language Label (ELL).

Badges

Badges are a very recent arrival in the skills recognition landscape. A badge is a graphical representation of any kind of achievement, goal or milestone based on a digital file that integrates the criteria and evidence used to obtain the badge. In an industrial setting, badges seem to be a lesser and more accessible form of recognition when compared with certificates. One of the principal motivations behind the "Badge Movement" was to provide a new mechanism for the recognition of skills which is better adapted to recent changes in learning:

- Nowadays, learning happens in many different contexts and while using various kinds of media. Instead of only recognizing credentials from students enrolled in established learning institutions, employers should have a mechanism for recognizing skills (and experiences) acquired by anyone through: professional training programs, participating in competitions, volunteer programs, MOOC's etc.
- Learning is not something which should only be 'recognized' during the 'student' phase of one's career but should rather be an ongoing process.
- Traditional skills recognitions do not capture elements that cannot be easily evaluated by test scores and short-term projects.
- Academic diplomas are for the most part monolithic documents. Recognitions of lesser granularity and much more flexibility are needed. Also, these recognitions should contain both the criteria used to evaluate the skills and evidence of the acquisition.
- Recognitions need to be digital, and capable of being displayed online

By design, badges can be awarded by any organization. Each organization is free to design any set of requirements, which it sees fit.

Noteworthy examples are IBM Digital Badges , Digital Badges at Purdue University or Stackoverflow's badge program.

The next section compares these recognition strategies according to the main characteristics identified from the EU political trends discussed in Section 2.

11.3.2 A Comparison of Recognitions

We began by reviewing current political trends in Europe regarding skills recognition and then we gave a brief overview of different popular skills recognition tools. The goal of this section is to evaluate the suitability of those recognition tools based upon those previously identified European political trends along with a short list of properties specific to data science. Lastly, based on that evaluation we will recommend a scheme for the recognition of data science skills in Europe.

Properties specific to data science

Before proceeding to the comparison, we would like to include a few additional properties not previously mentioned in Section 2.5 and which are specific to the rapidly evolving area of data science. Additionally, skills recognition in data science should:

- P10 require renewal after a set period of time
- P11 provide a framework which can quickly adapt to changes in skill requirement
- P12 measure skills on a highly granular and an individual by individual basis

The comparison

It should be noted that many of the previously recognition tools are very flexible, thus in certain cases, they could or could not satisfy a certain property depending upon the implementation of the tool. When this occurs, we will mark that fact with a “v?” in the table.

Desired Property	AC	UD	CE	LA	Badge
P1 - transparent, accessible and allow the easy comparison of students' skills throughout the EU			✓	✓?	✓
P2- include an assurance of quality	✓	✓	✓	✓	✓?
P3 - provide tools for their verification and validation	✓?	✓?	✓?	✓?	✓
P4 - include skills acquired through traditional, digital, online, and open learning, as well as the validation of informal and non-formal learning		✓?	✓	✓?	✓
P5 - compatible with the EQF		✓?	✓	✓	✓
P6 - influence the relevance of the skills being acquired		✓?	✓	✓	✓
P7 - allow the digital analysis of both the demand for and the availability of skills			✓?		✓
P8 - allow their use online: in platforms like the Europass, social media and on mobile devices		✓?	✓?	✓?	✓
P9 - focus on learning outcomes and not on traditional measures such as hours of study		✓?	✓	✓	✓
P10 - require renewal after a set period of time			✓		✓
P11 - provide a framework which can quickly adapt to changes in skill requirement			✓	✓?	✓
P12 - measure skills on a highly granular and an individual by individual basis			✓		✓

Discussion

In the previous comparison, the two tools which showed the worst results were accreditations and labels. Simply put, accreditation is a tool used to ensure that an educational program meets some minimum requirements in order for it to issue degrees. As such, it is not suited for the recognition of data science skills. Labels have more promise but again are a tool used to recognize traditional educational programs as a whole and not an individual's learning outcomes. University degrees have many strong points but fall short in that they:

- are not very transparent or individual,
- do not traditionally recognize informal and non-formal learning,
- are not very granular,
- are not traditionally digital, and lastly
- are not well suited to recognizing quickly changing skills.

Badges and certificates both offer all of the desired properties and thus our recommendation will be a mix of both of these tools.

11.3.3 Recommendations for Data Science skills recognition

Given that both certificates and badges manifest the properties which we found desirable in a skills recognition tool, we propose a hybrid approach drawing from the strengths of both badges and certifications.

We begin with a summary of the needs of all stakeholders in the data science ecosystem.

Data scientists need:

- (DS-N1) Credentials, which are widely recognized
- (DS-N2) Credentials, which can be easily verified online
- (DS-N3) A simple way to digitally display their skills online and in social networks
- (DS-N4) Mechanisms to formally recognize skills acquired through informal and non-formal training

Employers who hire data scientists need:

- (EM-N1) Tools to verify the authenticity of credentials
- (EM-N2) A skills recognition framework, which facilitates the comparison of candidate skills throughout the EU
- (EM-N3) Influence in the process of designing the types of training data scientists receive
- (EM-N4) A scheme for recognizing skills in data science, which can quickly adapt to changes in the data science ecosystem

Educators who train data scientists need:

- (ED-N1) Publicity for their programs and the added value that an externally branded recognition of their training can provide
- (ED-N2) Recognitions for the partial completion of their programs to assist students who are seeking employment while studying or students who abandon their studies
- (ED-N3) Contact with employers, a mechanism to clarify the changing needs of industry and clear recommendations regarding how to adapt to those needs

Accordingly, to meet these needs, our recognition approach will be based on the following elements:

- OB. The use of Open Badges. (DS-N2, DS-N3, EM-N1, EM-N4, ED-N2)
- E. Experts from industry and academia will contribute to both defining and maintaining the badge scheme. (DS-N1, EM-N2, EM-N3, EM-N4, ED-N1, ED-N3)

An initial proposal regarding both the types and the requirements of the badges will be based on the work of the EDISON and EDSA projects. These projects have focused a great deal of time and effort on establishing a consensus regarding frameworks for data science education in both industry and academia.

- TP. Badges will only be issued by trusted third parties. By carefully vetting badge issuers and their practices the reputation of the credentials will increase. (DS-N1, ED-N1)

- Pr. Supposing that the badges become popular, their requirements will influence the training of data scientists, give prestige to those who issue the badges and give publicity to their branding. (EM-N3, ED-N1)
- InF. Include the recognition of skills acquired through informal and non-formal training. (DS-N4)

Table 3 shows that our proposal covers all of the previously identified needs

	DS-N1	DS-N2	DS-N3	DS-N4	EM-N1	EM-N2	EM-N3	EM-N4	ED-N1	ED-N2	ED-N3
OB	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
E	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
TP	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
Pr	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
InF	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪

Though what we are recommending is really a hybrid approach inspired by aspects of certificates and badges, given the novelty and interest by large corporations such as IBM and Microsoft in the use badges, we believe that the term “badge” should be used for the data science skills recognition system.

The only disadvantage to the use of the term “badge” that we see is their current strong association with informal and non-formal learning. But in light of the growing interest, both in academia and in industry, to use badges we believe that in the medium-term that association will decrease.

11.3.4 The Logistics of the BDV Data Science Badge Program

In this section, we will briefly describe the logistics of the BDV Data Science Badge program.

Preliminaries:

- A collection of experts, including representatives from industry and academia, establish both the types and the requirements of the badges included in the program. They also define the process for applying to issue badges.
- This group of experts also approves the members of a group of reviewers whose mission is to evaluate applications received to issue badges.

Applying to issue badges:

- 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.
- Reviewers are assigned applications to assess and decide whether the applicant program meets the established standards to issue badges. Applications can be rejected, accepted conditionally for one year or accepted for four years.

Issuing badges:

- Students in a data science program authorized to issue BDV Badges acquire and demonstrate their data science skills through their studies. Students in the program can submit an application to their program to receive a badge.
- The program reviews badge applications and if the applicant has met the requirements of the badge then it issues the student that badge. Badges are individualized and contain metadata including: the requirements to earn the badge and evidence of the student’s achievements.

Displaying and viewing badges:

- Students can display their badges online: in their CV, in 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.

Revision

- The group of experts will periodically meet to review the program. Based upon progress reports, they can propose changes and improvements in the program.

A graphical representation of the process for applying to issue and issuing badges is given in Figure 27.

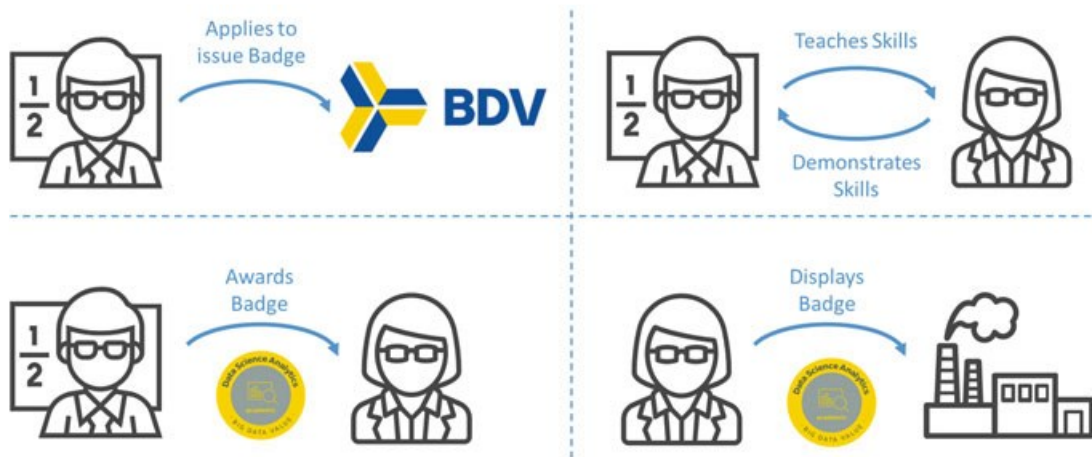


Figure 1 BDV Badges - Application and Issuing Process

11.4 BDV Data Science Badges Types and Requirements

In order to define both the types and requirements of the BDV Data Science Badge program we focused on EDISON’s Data Science Competence Framework (CF-DS) and Data Science Model Curriculum (MC-DS). We selected these documents as the basis of our proposal as they directly relate competences with professional profiles. So from an industrial perspective, it should be easy to understand the value of the badges through their relationships to the corresponding competences. Furthermore, the professional profiles and their required competency levels as given by EDISON can be used as an example of how the badges can be used in each organization. These examples can then be adapted to the structure and needs of an individual organization to provide a mapping between the badges and its hiring needs.

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

We also considered the possibility of creating one badge for each of the competences in each competence group. But this would have resulted in an excessively large number of badges that we thought would be unmanageable.

Though the BDV Badges are based on the work produced by the EDISON project, it should be noted that they can also be related to EDSA’s curriculum. Each badge has several required skills (or in EDISON’s terminology “learning outcomes”), which relate to one or several of EDSA’s topics. In this sense, the description given by EDSA for each

topic constitutes one possible learning resources source. In fact, EDSA based their four learning pathways (Data Analytics, Data Science Engineering, Data Management, and Business Process Management) upon four of the five competence groups in EDISON's framework.

11.4.1 Refining and Evaluating this Initial Proposal

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 this initial stage we focused only on the first badge, the Data Science Analytics Badge.

Twelve companies were contacted to participate in different stages of the assessment. The aim was 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.

Additionally, fifteen universities were also 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 badges applicants), and about the requirements of the badge.

Finally, the members of the BDVA Skills and Education Task Force were also requested to provide their opinions on the initial version of the badges as well as on the comments gathered from the industry and academy throughout the entire process.

As result of the assessment process several changes were proposed in the initial version of the requirements of the Data Science Analytics Badge. One of the most relevant ones is the replacement of the three levels of proficiency (basic, intermediate and expert) with three different sets of required skills with 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.

Furthermore, based on comments received the descriptions of some of the requirements were modified. Some of these changes were, for example, to highlight the role of descriptive models and to separate requirements related to data preparation, data visualization and data analytics tasks. The resulting requirements for the BDV Data Science Analytics Badge are shown below and the images of both the academic and professional badges are shown in Figure 28.

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 visualization techniques to improve the presentation of the results of a data science project in any of its phases.
- Table 4. BDV Data Science Analytics Badge Skills

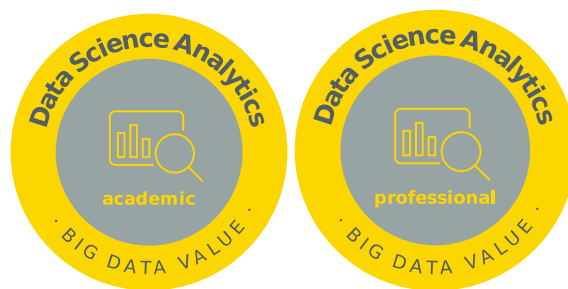


Figure 2. Data Science Analytics Badges with academic and professional levels (v1.0)

A pilot of the entire application process to issue the Academic Level of the Data Science Analytics Badge was conducted. The goal of the pilot was to check the workflow to be followed by universities applying to issue the badge and reviewers of the applications to issue badges. Three applications were received and reviewed as part of the pilot. No major problems in the process were found and the pilot was considered to be a success. Two of the applications received as part of the pilot were accepted:

- Master in Big Data Analytics, Universitat Politècnica de València, València, Spain

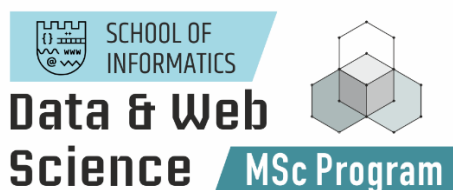
Máster **Big Data Analytics**



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



- Data and Web Science MSc Program, Aristotle University of Thessaloniki, Thessaloniki, Greece



11.5 Conclusions

After a survey of current trends in education in the EU was conducted, a collection of desirable properties for a data science skills recognition system was established. Those properties were compared to the benefits of existing tools for recognizing skills and a hybrid between certifications and badges was selected as the appropriate tool for recognizing skills in data science. Lastly, based upon the needs of stakeholders in the data science ecosystem, the details of a recognition system for data science skills were defined and the system was successfully piloted. At the moment of writing, the first open call for application to issue the Academic Level of the BDV Data Science Analytics Badge is open and can be found online at:

<https://www.big-data-value.eu/skills/skills-recognition-program/call-for-academic-level-data-science-analytics-badge-issuers/>.

Portions of this document were taken from D4.6: A Framework for the Recognition of Data Science Skills in Europe and D4.2: Skills, Education, and Centers of Excellence Period I Report which were produced as part of the Big Data Value Ecosystem project.

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

12 Part 5 Conclusion and further document extension

The proposed EDSF Release 4 Part 5 is composed based on recommendations from the EDSF Release 4 Design Workshop on 20 November 2019. It collects information about known use cases and application of the EDSF in different areas of the Data Science profession management and ecosystem. or organisational activities on Data Science and data related capacity building.

The presented applications and use cases strongly rely on the overall EDSF definition that is described in Parts 1-4 and continuously maintained by the EDISON Initiative community.

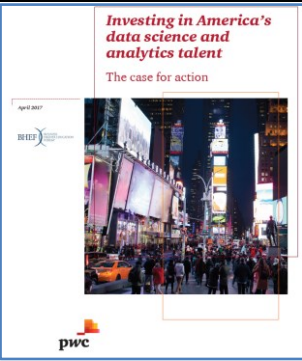
This document will continue collecting EDSF based or EDSF inspired developments and will be regularly updated.

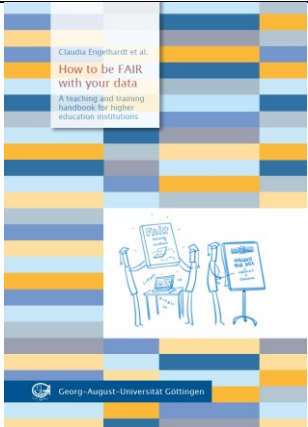
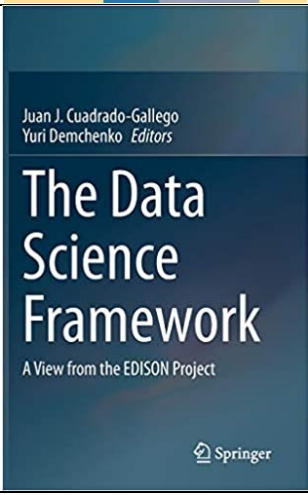
Acronyms

Acronym	Explanation
ACM	Association for Computer Machinery
BABOK	Business Analysis Body of Knowledge
CCS	Classification Computer Science by ACM
CF-DS	Data Science Competence Framework
CODATA	International Council for Science: Committee on Data for Science and Technology
CRISP-DM	Cross Industry Standard Process for Data Mining
CS	Computer Science
DigComp	Digital Competences for citizens (EU report 2017)
DM-BoK	Data Management Body of Knowledge by DAMAI
DS-BoK	Data Science Body of Knowledge
EDSA	European Data Science Academy
EOEE	EDISON Online E-Learning Environment
ETM-DS	Data Science Education and Training Model
EUDAT	http://eudat.eu/what-eudat
EGI	European Grid Initiative
ELG	EDISON Liaison Group
EOSC	European Open Science Cloud
ERA	European Research Area
ESCO	European Skills, Competences, Qualifications and Occupations
EUA	European Association for Data Science
HPCS	High Performance Computing and Simulation Conference
ICT	Information and Communication Technologies
IEEE	Institute of Electrical and Electronics Engineers
IPR	Intellectual Property Rights
LERU	League of European Research Universities
LIBER	Association of European Research Libraries
MC-DS	Data Science Model Curriculum
NIST	National Institute of Standards and Technologies of USA
P21	21st Century Skills Framework
PID	Persistent Identifier
PM-BoK	Project Management Body of Knowledge
PRACE	Partnership for Advanced Computing in Europe
RDA	Research Data Alliance
SWEBOK	Software Engineering Body of Knowledge

Appendix A. EDSF reviews, citation and references in other research and projects

A.1. EDSF reviews and references

Report/Study	Image	
<p>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 2.35 mln postings, 23% Data Scientist, 67% DSA enabled jobs DSA enabled jobs growing at higher rate than main Data Science jobs</p>	 <p>Citing EDISON and EDSF</p>	
<p>Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017) - Edited https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF DSA enabled jobs takes 45-58 days to fill: 5 days longer than average Commonly required work experience 3-5 yrs</p>		
<p>Evaluation of EDISON’s Data Science Competency Framework through a Comparative Literature Analysis (2021) https://par.nsf.gov/servlets/purl/10314377 https://www.aimsocieties.org/article/doi/10.3934/fods.2021031</p>		
<p>Data Science in Perspective, Paper 2021 by Rogério Rossi https://arxiv.org/ftp/arxiv/papers/2201/2201.05852.pdf</p>		
<p>Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards, by Usama Fayyad and Hamit Hamutcu, 30 June 2020 https://hdsr.mitpress.mit.edu/pub/6wx0qmk1/release/4 Institute of Data Science Management: Certification Standard: The Credentialing Framework (2020) https://www.datascienceinstitute.net/ https://www.datascienceinstitute.net/about-the-data-science-institute</p>		
<p>Computing Competencies for Undergraduate Data Science Curricula, Initial Draft, January 2019 ACM Data Science Task Force - https://dstf.acm.org/DSReportInitialFull.pdf</p>		
<p>EDISON at CERN eLearning wiki (Maria Dimou, 2017) - https://twiki.cern.ch/ELearning/Edison</p>		
<p>FAIRsFAIR Deliverable D7.3 Data Stewardship and FAIR Competence Framework (2021) https://zenodo.org/record/5361917</p>		

<p>How to be FAIR with your data A teaching and training handbook for higher education institutions by Claudia Engelhardt et al. Published: 2022 http://resolver.sub.uni-goettingen.de/purl?univerlag-isbn-978-3-86395-539-7</p>		
<p>The Data Science Framework, A View from the EDISON Project, Editors Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer Nature Switzerland AG 2020, ISBN 978-3-030-51022-0, ISBN 978-3-030-51023-7 (eBook, printed book)</p>		

A.2. Projects Using or contributing to the EDSF development

FAIRsFAIR - Fostering FAIR Data Practices in Europe, Grant Agreement 831558 (2019-2022) - <https://fairsfair.eu/>

ERASMUS+ MATES - Maritime Alliance for fostering the European Blue Economy through a Marine Technology Skilling Strategy, Grant Agreement 591889 (2018-2022) - <https://www.projectmates.eu/>

A.3. Publications by Authors

- [1] How to be FAIR with your data: A teaching and training handbook for higher education institutions, by Claudia Engelhardt et al. Published 2022, DOI: <https://doi.org/10.17875/gup2022-1915>, (Free online version) A modular and community-driven FAIR teaching and training handbook for higher education institutions - To be submitted 2022
- [2] The Data Science Framework, A View from the EDISON Project, Editors Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer Nature Switzerland AG 2020, ISBN 978-3-030-51022-0, ISBN 978-3-030-51023-7 (eBook, printed book) Publisher page
- [3] Demchenko, Yuri, Cuadrado-Gallego, Juan J., Data Science among other Data Driven Technology Domains Revisited, SciDataCon2022, 20-23 June 2022, Seoul, Korea. Published online at <https://www.scidatacon.org/IDW-2022/sessions/469/poster/253/> (PDF)
Demchenko, Yuri, Data Stewardship Competence Framework: Instrumental in Organisational Skills Management, Career Path building and Training, SciDataCon2022, 20-23 June 2022, Seoul, Korea. Published online at <https://www.scidatacon.org/IDW-2022/sessions/469/poster/255/> (PDF)
- [4] Demchenko, Yuri, Digital and Data Skills Training to Enable the Digital Transformation of the Maritime Industry, SciDataCon2022, 20-23 June 2022, Seoul, Korea. Published online at <https://www.scidatacon.org/IDW-2022/sessions/469/poster/254/> (PDF) 2021

- [5] Yuri Demchenko, Mathijs Maijer, Luca Comminiello, Data Scientist Professional Revisited: Competences Definition and Assessment, Professional Development and Education Path Design, International Conference on Big Data and Education (ICBDE2021), February 3-5, 2021, London, United Kingdom (PDF)
Yuri Demchenko, Lennart Stoy, Research Data Management and Data Stewardship Competences in University Curriculum, In Proc. Data Science Education (DSE), Special Session, EDUCON2021 – IEEE Global Engineering Education Conference, 21-23 April 2021, Vienna, Austria (PDF)
- [6] Yuri Demchenko, Tomasz Wiktorski, Steve Brewer, Juan José Cuadrado Gallego, EDISON Data Science Framework (EDSF): Addressing Demand for Data Science and Analytics Competences for the Data Driven Digital Economy, In Proc. Data Science Education (DSE), Special Session, EDUCON2021 – IEEE Global Engineering Education Conference, 21-23 April 2021, Vienna, Austria (PDF)
- [7] Cuadrado-Gallego, Juan J., Losada, Miguel, Demchenko, Yuri, Ormandjieva, Olga, Classification and Analysis of Techniques and Tools for Data Visualization Teaching, In Proc. Data Science Education (DSE), Special Session, EDUCON2021 – IEEE Global Engineering Education Conference, 21-23 April 2021, Vienna, Austria (PDF) 2020
- [8] Tomasz Wiktorski, Yuri Demchenko and Oleg Chertov, Data Science Model Curriculum Implementation for Various Types of Big Data Infrastructure Courses, Proc. 5th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2019), part of the eScience 2019 Conference, September 24 – 27, 2019, San Diego, California, USA (PDF)
- [9] Yuri Demchenko, Tomasz Wiktorski, Steve Brewer, Juan José Cuadrado Gallego, EDISON Data Science Framework (EDSF) Extension to Address Transversal Skills required by Emerging Industry 4.0 Transformation, Proc. 5th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2019), part of the eScience 2019 Conference, September 24 – 27, 2019, San Diego, California, USA (PDF)
- [10] Yuri Demchenko, Big Data Platforms and Tools for Data Analytics in the Data Science Engineering Curriculum, Proc 2019 3rd International conference on Cloud and Big Data (ICCBDC 2019), August 28-30, 2019, Oxford, UK (PDF)
- [11] Yuri Demchenko, Cloud based Big Data Platforms and Tools for Data Analytics in the Big Data Engineering Curriculum, Proc 15th Int. Conference on Data Science (ICDATA'19), July 29 - August 1, 2019, Las Vegas (PDF)
- [12] Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, ISBN978-1-4503-6186-6/19/03. (PDF).
- [13] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong. (PDF)
- [14] Tomasz Wiktorski, Yuri Demchenko, Adam Belloum, Model Curricula for Data Science EDISON Data Science Framework, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong. (PDF)
- [15] Demchenko, Yuri, Adam Belloum, Wouter Los, Cees de Laat, EDISON Data Science Framework: A Foundation for Data Science Competence Management and Curricula Development, Proc Conference ICT.OPEN2017, Section "Computing and Imaging (by ASCII)", 22-23 March 2017, Amersfoort [online] (PDF)
Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Steve Brewer, EDISON Data Science Framework: Defining the Data Science Professions Family for Research and Industry, Data Science Journal, 2017. Submitted paper (PDF)
- [16] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Steve Brewer, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, 3rd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2016), in Proc. The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 December 2016, Luxembourg. (PDF)
- [17] Tomasz Wiktorski, Yuri Demchenko, Adam Belloum and Anoosheh Shirazi, Quantitative and Qualitative Analysis of Current Data Science Programs from Perspective of Data Science Competence Groups and

- Framework, 3rd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2016), in Proc.The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 December 2016, Luxembourg. (PDF)
- [18] Yuri Demchenko, Adam Belloum, Wouter Los, Steve Brewer, Andrea Manieri, EDISON Data Science Framework for defining the Data Science Profession, SciDataCon2016, 11-13 September 2016, Denver, Colorado, USA. Published online at <http://www.scidatacon.org/2016/sessions/98/poster/75/> (PDF)
- [19] Yuri Demchenko, Adam Belloum, Wouter Los, Steve Brewer, Defining Customisable Model Curriculum for Research Data Management Training, SciDataCon2016, 11-13 September 2016, Denver, Colorado, USA. Published online <http://www.scidatacon.org/2016/sessions/57/paper/167/> (PDF)
- [20] Yuri Demchenko, Adam Belloum, Wouter Los, Steve Brewer, Defining Data Science Professions Family, SciDataCon2016, 11-13 September 2016, Denver, Colorado, USA. Published online <http://www.scidatacon.org/2016/sessions/98/poster/20/> (PDF)
- [21] Demchenko, Yuri, Adam Belloum, Wouter Los, Spiros Koulouzis, Cees de Laat, EDISON Project: Building Data Science Profession for European Research and Industry. Proc Conference ICT.OPEN2016, Section "Computing and Imaging (by ASCII)", 22-23 March 2016, Amersfoort ISBN/EAN 978-90-73461-932 [online] <http://www.ictopen.nl/binaries/content/assets/bestanden/ict-open-2016/proceedings/demenchenko-yuri.pdf>
- [22] Andrea Manieri, Yuri Demchenko, Tomasz Wiktorski, Steve Brewer, Matthias Hemmje, Ruben Riestra, Tiziana Ferrari, Jeremy Frey, Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists, 2nd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science, in Proc.The 7th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2015), 30 November - 3 December 2015, Vancouver, Canada (PDF)