EDISON
building the data science profession

EDISON Data Science Framework:

Part 1. Data Science Competence Framework (CF-DS)

Release 4 (EDSF04 or EDSF2022)

EDISON Community Initiative
(Maintaining the H2020 EDISON project outcome)

Release Date	31 December 2022
Document Editor/s	Yuri Demchenko
Version	Release 4, v11
Status	Working document, request for comments

Document Version Control

Version	Version	Date	Change Made (and if appropriate reason for change)	Contributors and Editors (initials)
Release 1	03	10/10/2016	Release 1 after ELG03 meeting discussion	YD, AB, AM, TW, WL, SB
Release 2	07	03/07/2017	Release 2 documents published. Updated after multiple discussions and comments, DARE Project alignment, mapping CRISP-DM, ELG04 comments	YD, AB, AM, TW, WL, SB, ES
Pre-Release 3	09	07/09/2018	Pre-release 3. Definition of competences revised and extended.	YD
Release 3	10	31/12/2018	Release 3. Pre-release document updated based on received comments and feedback from practical implementations.	YD
Release 4	11	31/12/2022	Release 4. Extended with the Data Stewardship Professional Competence Framework (CF-DSP). Document updated based on feedback from practical implementations; Skills groups revised related to technology and tools.	YD, JJCG, SB

Contributors

Document Editors: Yuri Demchenko		
Author Initials	Name of Contributor	Institution
YD	Yuri Demchenko (editor)	University of Amsterdam
AB	Adam Belloum	University of Amsterdam
AM	Andrea Manieri	Engineering
TW	Tomasz Wiktorski	University of Stavanger
WL	Wouter Los	University of Amsterdam
SB	Steve Brewer	University of Southampton
ES	Erwin Spekschoor	Independent expert
JJCG	Cuadrado Gallego Juan José	Alcala University

Acknowledgment

Terry Hook, Clock-IT-Skills
Gábor Kismihók, Center of Job Knowledge Research, University of Amsterdam
Oleg Chertov, Professor, Sikorski National Technical University of Ukraine “Kyiv Polytechnic Institute”
Luca Comminiello, Erasmus+ Master, University of Perugia
Mathijs Maijer, University of Amsterdam
Thom Hartog, University of Amsterdam

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 four main components: Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), Data Science Professional Profiles (DSPP), which are extended with new Part 5. Use cases and guidelines. 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 Data Science Competence Framework (CF-DS) is a cornerstone component of the whole EDISON framework. CF-DS provides a basis for the Data Science Body of Knowledge (DS-BoK) and Model Curriculum (MC-DC) definitions, and further for the Data Science Professional Profiles definition and certification. The CF-DS incorporates many of the underpinning principles of the European e-Competence Framework (e-CF3.0) that have been used for the Data Science competences definition; in its own turn, this allowed to provide extensions of the new e-CF4.0 version (published as CEN EN 16234-1, 2019) with the Data Science competences. The CF-DS and DSPP have also adopted the classification structure of the European Skills, Competences, Occupations (ESCO) Framework. Corresponding information is provided in the corresponding documents CF-DS and DSPP.

This presented Data Science Competence Framework definition is based on the analysis of existing frameworks for Data Science and ICT competences and skills and is supported by the analysis of the demand side (job market) for Data Science professionals in industry and research. The presented CF-DS Release 4 is extended with the skills and knowledge subjects/units related to all competences groups. The document also refined the Data Science workplace) skills definition that includes the Data Science professional skills (Acting and thinking like Data Scientist) and the definition of the general "soft" skills often referred to as 21st Century skills.

Since its publication in 2017 with the EDSF Release 2, the EDSF methodology has been referred to and used by many projects worldwide and in particular such European projects as ELIXIR/RITrain, MATES, FAIRsFAIR, and CEN Workshop 14568 to develop domain related and sectoral competence frameworks and body of knowledge, in particular, digital and data related competences and skills

The current EDSF Part 1 document defines the Data Science Competence Framework and includes the following components:

- The CF-DS defines five groups of competences for Data Science that include Data Analytics, Data Science Engineering, Domain Knowledge, *Data Management* and Governance, *Research Methods* and Project Management for research related occupations, or Business Process Management for business related occupations.
- The document provides examples of the individual competences mapping to identified skills and knowledge topics for the Data Science Analytics competence group.
- The identified competences, skills, and knowledge subjects are provided as enumerated lists to allows easy use in applications and provide a basis for developing compatible APIs.
- The presented CF-DS definition is supported by the corresponding Excel documents and ontology definition that contain a full list of enumerated EDSF attributes.

The proposed EDSF, and CF-DS in particular, are intended to provide guidance and a basis for universities and education 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.

It is also intended that the proposed CF-DS can provide a basis for building interactive/web based tools for individual or organizational Data Science competences benchmarking, Data Science team building, and creating customized Data Science education and training programs.

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) ... 10
3 Existing frameworks for ICT and Data Science competences and skills definition ... 12
3.1 NIST definition of Data Science ... 12
3.2 European e-Competence Framework (e-CF) ... 13
3.3 ACM Information Technology Competencies Model ... 16
4 EDISON Data Science Competence Framework (CF-DS) ... 18
4.1 Relation to and use of existing framework and studies ... 18
4.2 Identified Data Science Competence Groups ... 19
4.3 Identified Data Science Skills and their mapping to Competences ... 24
4.3.1 Data Science skills related to the main competence groups – Skills Type A ... 24
4.3.2 Data Science skills related to the Data Analytics languages, tools, platforms and Big Data infrastructure – Skills Type B ... 27
4.4 Knowledge required to support identified competences ... 30
4.5 Proficiency levels ... 32
4.6 Data Scientist Workplace skills (aka "soft" or transversal skills) ... 32
4.6.1 Data Science Professional or Attitude skills (Thinking and acting like a Data Scientist) ... 32
4.6.2 21st Century skills ... 34
4.6.3 Design Thinking and System Thinking ... 34
4.7 Data Scientist in the modern agile data driven organisation ... 35
4.8 Data Science Literacy: Commonly required competences and skills for Data Science related and enabled occupations/roles ... 36
4.9 Relation between Data Scientist and Subject Domain specialist ... 36
5 Example of Data Science and Analytics Competences definition ... 38
6 Alignment with other competence frameworks and suggested extensions ... 45
6.1 Mapping Data Science competences to e-Competence Framework (e-CF) ... 45
6.1.1 Proposed e-CF3.0 extension with the Data Science related competences ... 45
6.2 Mapping Data Science competences to CRISP-DM model ... 46
6.3 Process Groups in Data Management and their mapping to CF-DS competences ... 49
7 Practical uses of CF-DS ... 50
7.1 Usage example: Competences assessment [31] ... 50
7.2 Data Stewardship Professional Competence Framework (CF-DSP) definition ... 50
7.2.1 Data Management and Governance competence group (DSDM) ... 51
7.2.2 Data Engineering competence group (DSENG) ... 53
7.2.3 Research Methods and Project Management competence group (DSRMP) ... 55
7.2.4 Domain related competence (DSDK/DSBA) ... 56
8 Conclusion and further developments ... 58
8.1 Summary of the recent developments ... 58
8.2 Further developments to formalize CF-DS ... 59
9 References ... 60
10 Acronyms ... 62
11 Appendix A. Data used in the study of demanded Data Science competences and skills ... 63
A.1. Selecting sources of information ... 63
A.2. EDISON approach to analysis of collected information ... 64
A.3. Regular Job Market analysis ... 64
Appendix B. Overview: Studies, reports and publications related to Data Science competences and skills ... 65
B.1. O'Reilly Strata Survey (2013) [18] ... 65
B.2. Skills and Human Resources for e-Infrastructures within Horizon 2020 [19] ... 66
B.3. UK Study on demand for Big Data Analytics Skills (2014) [36] ... 67
B.4. IWA Data Science profile [37] ... 67
B.5. Other studies, reports and projects on defining Data Science competences, skills and profiles ... 68
12 Appendix C. Mapping between CF-DS and e-CF ... 69
C.1. e-CF3.0 competences relevant to CF-DS: Essential and additional ... 69
C.2. Proposed e-CF3.0 extension with the Data Science related competences ... 70
C.3. Data Science related Competences included in the EN 16234-1 (2018) ... 71

13 Appendix D. Mapping between e-CF proficiency levels and EQF qualification levels 72
14 Appendix D. Concepts and models related to CF-DS definition 74
D.1. Scientific Data Lifecycle Management Model 74
D.2. Scientific methods and data driven research cycle 75
D.3. Business Process Management lifecycle 76

1 Introduction

Data Science Competence Framework (CF-DS) 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], Data Science Professional Profiles (DSPP) [4], and Use cases and Guidelines [5].

The CF-DS definition is a cornerstone component of the whole EDISON Data Science Framework. CF-DS provides a basis for the Data Science Body of Knowledge that defines a set of knowledge required from the Data Scientist, or related Data Science enabled roles to support required competences and effectively operate in their organisational roles, which are in their own turn defined based in functions, responsibilities, and competences. Competences defined in CF-DS are used for defining learning outcomes when defining the Data Model Curriculum. The CF-DS incorporates many of the underpinning principles of the European e-Competence Framework (e-CF3.0) and provides suggestions for e-CF3.0 extension with the Data Science related competences and skills. Furthermore, the CF-DS and DSPP have also adopted and intend to comply with the structure of European ICT Professional Profiles and European Skills, Competences, Occupations (ESCO) Framework. Corresponding information is provided in both documents, CF-DS and DSPP.

The current EDSF Part 1 document defines the Data Science Competence Framework that includes the following components:

- The presented CF-DS defines five groups of competences for Data Science that include Data Analytics, Data Science Engineering, Domain Knowledge, *Data Management* and Governance, *Research Methods* and Project Management for research related occupations, or Business Process Management for business related occupations.
- The document provides the individual competences mapping to identified skills and knowledge topics for the Data Science Analytics competence group.
- The identified competences, skills, and knowledge subjects are provided as enumerated lists to allow easy use in applications and developing compatible APIs.
- The presented CF-DS definition is supported by the corresponding Excel documents that contain a full list of enumerated EDSF attributes.

The Research Methods competences are essential for the Data Scientist to discover new relations and provide actionable insight into available data, and have the ability to formulate good research questions, hypotheses and evaluate them based on collected data.

This presented Data Science Competence Framework definition is based on the analysis of existing frameworks for Data Science and ICT competences and skills, and supported by the analysis of the demand side for the Data Scientist profession in industry and research. The CF-DS has been widely discussed at numerous workshops, conferences and meetings with wide community contribution. The core CF-DS competences have been reviewed by experts and validated in numerous practical implementations.

The presented CF-DS Release 4 is extended with the skills and knowledge topics related to all competence groups. The document also refined the Data Science workplace skills definition that includes Data Science professional skills (Acting and thinking like a Data Scientist) and the definition of the general “soft” skills often referred to as 21st Century skills that are increasingly demanded by modern data driven companies.

The proposed EDSF comprising of the mentioned above components intends to provide guidance and a basis for universities and education 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 proposed CF-DS can be used for building interactive/web based tools or applications for knowledge and skills (self-) assessment, job vacancy design, and assessment of the candidate’s profile for a specific profile/role or job vacancy. All individual competences, knowledge topics and skills are enumerated to allow easier design of API for applications that may use CF-DS.

The document has the following structure. Section 2 provides an overview of the EDISON Data Science Framework and related components of the Data Science professional ecosystem. Section 3 provides an overview of existing frameworks for ICT and Data Science competences and skills definition, including NIST Special

Publication 1500-1, e-CF3.0, ACM Computing Classification System (2012). Section 4 presents the full CF-DS definition that includes identified competence groups, identified skills, and knowledge that all together should enable the Data Scientist to effectively work with a variety of Data Analytics methods and Big Data platforms to deliver insight and value to organisations. Section 4 also provides a description of the Data Science professional (workplace) and general attitude skills demanded from the modern specialists/professionals intended to work in modern agile data driven companies. Section 5 provides examples of the individual competences definition together with their linking to knowledge and skills. Section 6 provides a mapping of the CF-DS competences to other frameworks, such as e-CF3.0 and competences related to CRISP-DM processes. Section 7 provides suggestions for the practical use of CF-DS, in particular for other ESDF components definition and possible uses by organisations for competences assessment and skills management.

Appendices to this document contain important supplementary information: information about the approach and data sets used for deriving the proposed CF-DS competences groups; overview of known studies, reports and publications related to Data Science competences and skills; concepts and models related to the Data Science competences definition, such as data lifecycle management models, scientific methods, and business process management lifecycle models.

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 ecosystem[1] , 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

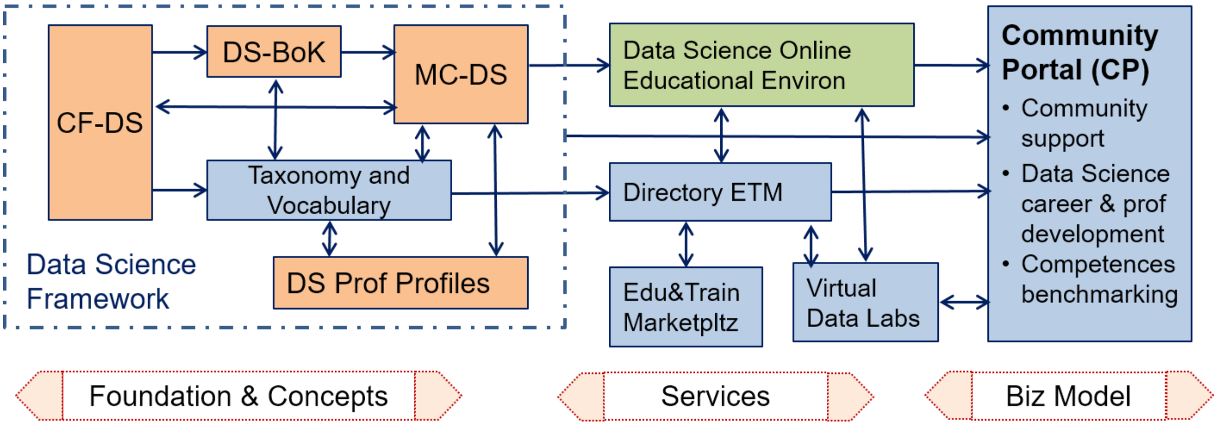


Figure 2.1 EDISON Data Science Framework components and Data Science professional ecosystem.

The EDSF Release 4 includes Part 5 EDSF Use cases and Guidelines [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 profiles[2].

[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 Amsterdam[3].

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.

[3] EDISON Community Initiative website https://edisoncommunity.github.io/EDSF/

3 Existing frameworks for ICT and Data Science competences and skills definition

This section provides a brief overview of existing standard and commonly accepted frameworks that have been used for defining Data Science and general Computer Science and ICT competences, skills, and subject domain classifications that can be, with some alignment, built upon and re-used for better acceptance from research and industrial communities... The information in this section is also complemented with the overview of other works and publications to define required Data Science competences and skills, which are placed in Appendix B.

3.1 NIST definition of Data Science

NIST Big Data Working Group (NBD-WG) published its first release of the Big Data Interoperability Framework (NBDIF) in September 2015, consisting of 7 volumes. Final Version 3 is published in 2019 as NIST Standard SP 1500 comprising of 9 volumes [9]. Volume 1. Definitions provides a number of definitions in particular Data Science, Data Scientist and Data Life Cycle, which we will use as a starting point for our analysis:

> **Data science** is the extraction of actionable knowledge directly from data through a process of discovery or hypothesis formulation and hypothesis testing. Data science can be understood as the activities happening in the processing layer of the system architecture, against data stored in the data layer in order to extract knowledge from the raw data.
>
> Data science across the entire data life cycle incorporates principles, techniques, and methods from many disciplines and domains, including data cleansing, data management, analytics, visualization, engineering, and in the context of Big Data, now also includes Big Data Engineering. Data science applications implement data transformation processes from the data life cycle in the context of Big Data Engineering.
>
> A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle.
> Data scientists and data science teams solve complex data problems by employing deep expertise in one or more of these disciplines, in the context of business strategy, and under the guidance of domain knowledge. Personal skills in communication, presentation, and inquisitiveness are also very important given the complexity of interactions within Big Data systems.
>
> The **data life cycle** is the set of processes in an application that transform raw data into actionable knowledge.

The term analytics refers to the discovery of meaningful patterns in data and is one of the steps in the data life cycle of collection of raw data, preparation of information, analysis of patterns to synthesize knowledge, and action to produce value. Analytics is used to refer to the methods, their implementations in tools, and the results of the use of the tools as interpreted by the practitioner. The analytics process is the synthesis of knowledge from information.

The NBDIF Volume 1 also provides an overview of other definitions of Big Data and Data Science from IDG, McKinsey, O'Reilly reports and popular blogs published by experts in new technology.

Figure 3.1 from the BDIF publication provides a graphical presentation of the multi-factor/multi-domain Data Science definition.

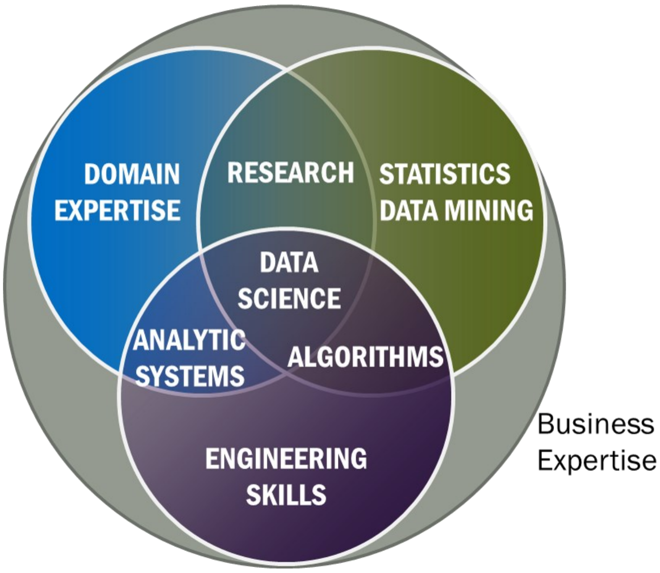


Figure 3.1. Data Science definition by NIST BD-WG [9].

3.2 European e-Competence Framework (e-CF)

The EDISON CF-DS development follows the European e-Competences Framework (e-CF)[4] approach and guiding principles:

- CF-DS adopts a holistic e-CF definition: "Competence is a demonstrated ability to apply knowledge, skills and **attributes** for achieving desirable results" in organisational or role context.
- Competence is a durable concept and although technology, jobs, marketing terminology and promotional concepts within the ICT environment change rapidly, the e-CF remains durable, requiring maintenance approximately every three years to maintain relevance.
- CF-DS should work as an enabler for multiple applications that can be used by different types of users, from individual to organisational; it should support common understanding and not mandate specific implementation.
- Competence can be a part of a job definition but cannot be used to substitute a similarly named job definition; one single competence can be assigned to multiple job definitions.

The European e-Competence Framework (e-CF) [10,11, 12] was established as a tool to support mutual understanding and provide transparency of language through the articulation of competences required and deployed by ICT professionals (including both practitioners and managers).

The e-CF is structured by four dimensions:

Dimension 1:
5 e-Competence areas, derived from the ICT business processes PLAN – BUILD – RUN – ENABLE – MANAGE

Dimension 2:
A set of reference e-Competences for each area, with a generic description for each competence. 40 competences identified in total provide the European generic reference definitions of the e-CF 3.0.

Dimension 3:
Proficiency levels of each e-Competence provide European reference level specifications on e-Competence levels e-1 to e-5, which are related to the EQF levels 3 to 8 [13].

[4] e-CF3.0 and e-CF4.0 version use the same methodology but different in details of the competences definition. e-CF4.0 e-CF3.0 was available at the time EDSF Release 2 was published. e-CF4.0 is currently standardised as CEN CEN EN 16234-1:2019 e-Competence Framework (e-CF) - A common European Framework for ICT Professionals in all sectors - Part 1: Framework [11]

Dimension 4:
Samples of knowledge and skills relate to e-Competences in dimension 2. They are provided to add value and context and are not intended to be exhaustive.

Whilst competence definitions are explicitly assigned to dimension 2 and 3 and knowledge and skills samples appear in dimension 4 of the framework, attitude is embedded in all three dimensions.

Dimension 1. Competence Area defined by ICT Business Process stages from organisational perspective:
A. Plan: Defines activities related to planning services or infrastructure, may also include elements of design and trends monitoring.
B. Build: Includes activities related to applications development, deployment, engineering, and monitoring
C. Run: Includes activities to run/operate applications or infrastructure, including user support, change support, and problems management
D. Enable: Includes numerous activities related to supporting production and business processes in organisations that include sales support, channels management, knowledge management, personnel development and education and training.
E. Manage: Includes activities related to ICT/projects and business processes management including management of risk, customer relations, and information security.

e-competences in Dimensions 1 and 2 are presented from the organisational perspective as opposed to an individual's perspective. Figure 3.2 illustrates the ICT process stages as they are defined in the e-CF3.0 document. Dimension 3, which defines e-competence levels related to the European Qualifications Framework (EQF) [13], is a bridge between organisational and individual competences. Refer to Appendix C which provides a mapping between e-CF proficiency levels and EQF qualification levels.

Table 3.1 below contains competences defined for areas A-E. For a more detailed definition of e-CF3.0 dimensions 1-3 and dimension 4 refer to the original e-CF3.0 definition [10].

Table 3.1. e-CF3.0 competences defined for areas A-E

Dimension 1: 5 e-CF areas (A – E)	Dimension 2: 40 e-Competences identified	Dimension 1: 5 e-CF areas (A – E)	Dimension 2: 40 e-Competences identified
A. PLAN	A.1. IS and Business Strategy Alignment	D. ENABLE	D.1. Information Security Strategy Development
	A.2. Service Level Management		D.2. ICT Quality Strategy Development
	A.3. Business Plan Development		D.3. Education and Training Provision
	A.4. Product / Service Planning		D.4. Purchasing
	A.5. Architecture Design		D.5. Sales Proposal Development
	A.6. Application Design		D.6. Channel Management
	A.7. Technology Trend Monitoring		D.7. Sales Management
	A.8. Sustainable Development		D.8. Contract Management
	A.9. Innovating		D.9. Personnel Development
			D.10. Information and Knowledge Management
B. BUILD	B.1. Application Development		D.11. Needs Identification
	B.2. Component Integration		D.12. Digital Marketing
	B.3. Testing		
	B.4. Solution Deployment	E. MANAGE	E.1. Forecast Development
	B.5. Documentation Production		E.2. Project and Portfolio Management
	B.6. Systems Engineering		E.3. Risk Management
			E.4. Relationship Management
C. RUN	C.1. User Support		E.5. Process Improvement

	C.2. Change Support		E.6. ICT Quality Management
	C.3. Service Delivery		E.7. Business Change Management
	C.4. Problem Management		E.8. Information Security Management
			E.9. IS Governance

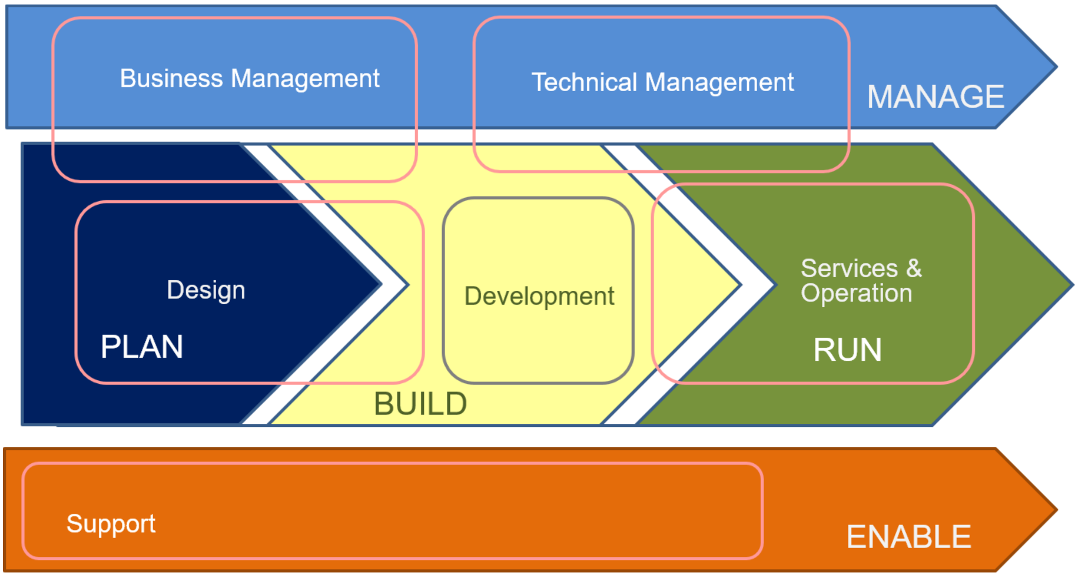


Figure 3.2. ICT process stages aligned with the organisational production workflow (as used in e-CF3.0)

Figure 3.3 illustrates the multi-purpose use of the European e-Competence Framework within ICT organisations. The e-CF has a multidimensional structure and is flexible in using for different purposes, it can be easily adopted for organisation specific models and roles. The e-CF3.0 is used for job-profiles definition in CWA 16458 (see [14] and EDSF DSPP document [4]) that are linked to the organisational processes aligned with the product or services lifecycle which are often aligned with the organisational structure. However, this may creates limitations for cross-organisational professional profiles and roles matching, in particular for the organisational role of the Data Scientist that works with data and serves data analytics tasks in cross-department way. Combining competences from different competence areas and using them as building blocks can allow flexible job-profiles definition. This enables the derived job-profiles to be easily updated by changing the set of competences related to profiles without the need to restructure the entire profile.

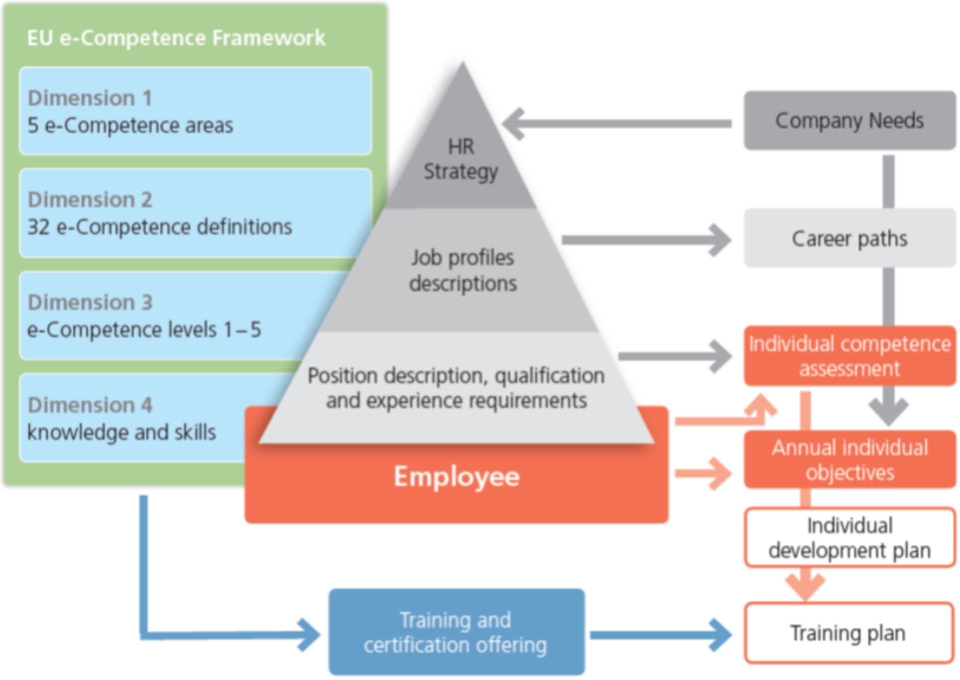

EU e-Competence Framework
Dimension 1
5 e-Competence areas
Dimension 2
32 e-Competence definitions
Dimension 3
e-Competence levels 1 – 5
Dimension 4
knowledge and skills
HR Strategy
Job profiles descriptions
Position description, qualification and experience requirements
Employee
Company Needs
Career paths
Individual competence assessment
Annual individual objectives
Individual development plan
Training and certification offering
Training plan

Figure 3.3. e-CF3.0 structure and use for definition of the job profile definition and training needs [12].

3.3 ACM Information Technology Competencies Model[5]

The ACM Information Technology Competency Model (IT-CM) of Core Learning Outcomes and Assessment for Associate-Degree Curriculum (2014) has been developed by ACM Committee for Computing Education in Community Colleges (ACM CCECC) [7, 15].

ACM currently categorizes the overarching discipline of computing into five defined sub- disciplines (ACM, 2005): computer science, computer engineering, software engineering, information systems and information technology. This report specifically focuses on information technology defined by the ACM CCECC as follows:

> *Information Technology involves the design, implementation and maintenance of technology solutions and support for users of such systems. Associated curricula focus on crafting hardware and software solutions as applied to networks, security, client- server and mobile computing, web applications, multimedia resources, communications systems, and the planning and management of the technology lifecycle (ACM CCECC, 2009).*

The document refers to the U.S. Department of Labour Information Technology Competency Model [16] which was one of the sources that provided a foundation for the curricular guidance outlined in IT-CM report

Competencies are used to define the learning outcome. In formulating assessment rubrics, the ACM CCECC uses a structured template comprised of three tiers: "emerging", "developed", and "highly developed", that can actually be mapped to the level of Bloom's verbs from the lower order thinking skills (LOTS) to the higher order thinking skills (HOTS), including "analysing" and "evaluating."

The ACM Competencies Model provides a basis for the Competency -based learning that is Instead of focusing on how much time students spend learning a particular topic or concept (Carnegie unit credit hour), the outcomes-based model assesses whether students have mastered the given competencies, namely the skills, abilities, and knowledge.

[5] Note: The EDSF adopted "competence" spelling as it is used in e-CF, while ACM/IEEE documents use common for US spelling "competency" Both forms can be used and they don't have grammatical or historical preference.

The document defines 50 learning outcomes (that also define the Body of Knowledge) that represent core or foundational competencies that a student in any IT-related program must demonstrate. Curricula for specific IT programs (e.g., networking, programming, digital media, and user support) will necessarily include additional coursework in one or more defined areas of study. The core IT learning outcomes are grouped into technical competency areas and workplace skills.

The ACM CCECC classification is supported by the web portal http://ccecc.acm.org/. The portal provides related information, linking and mapping between different classification systems, in particular:

- ACM Computing Classification System 2012 [7]
- U.S. Dept. of Labor IT 2012 Competency Model [16]
- European e-Competence Framework 3.0 (Proficiency Levels 1 & 2) [10, 11]
- European Qualifications Framework (EQF) [13]
- Bloom's Taxonomy Revised [17]

4 EDISON Data Science Competence Framework (CF-DS)

This section describes the proposed Data Science Competence Framework (CF-DS) that serves as a foundation for the definition of the other EDSF components. The presented CF-DS provides a full and comprehensive view of the demanded Data Science competences, skills and knowledge what provides a more consistent view comparing to the existing Data Science definitions that primarily cover data analytics and software engineering competences while modern data driven enterprises and processes require advanced skills for heterogeneous data management and use of research methods to uncover full data value. In the current version, the proposed Data Science Competence Framework has evolved from the initially proposed as a result of the job market study supported by extensive desk research covering professional blogs, community discussions, and existing standards and best practices overview, to the mature framework reviewed by expert groups and individual experts, and feedback received from multiple practical implementations by universities, professional training organisations, and different projects dealing with data related skills management.

4.1 Relation to and use of existing framework and studies

The following describes what existing frameworks and documents were used for defining the proposed set of Data Science competences and skills.

a) NIST NBDIF Data Science and Data Scientist definition [9]

It provided the general approach to the Data Science competences and skills definition, in particular, as having 3 groups: Data Analytics, Data Science Engineering, and Domain expertise, that may define possible specialisation of actual Data Science curricula or individual Data Scientists competences profile.

b) European e-Competence Framework (e-CFv3.0/e-CF4.0) [10, 11, 12]

e-CF provided a general framework for ICT competences definition and possible mapping to Data Science competences. However, it appeared that the current e-CF doesn’t contain competences that reflect specific Data Scientist roles in organisation. Furthermore, e-CF is built around organisational workflow while the anticipated Data Scientist’s role is cross-organisational bridging different organisational roles and departments in providing data centric view or organisational processes.

c) European ICT profiles CWA 16458 (2012) [13]

European ICT profiles and its mapping to e-CF3.0 provided a good illustration of how individual ICT profiles can be mapped to e-CF3.0 competences and areas. Similarly, the additional ICT profiles are proposed to reflect Data Scientist’s role in the organisation.

d) European Skills, Competences, Qualifications, and Occupations (ESCO) [8]

ESCO provides a good example of a standardised competences and skills taxonomy. The presented study will provide a contribution to the definition of the Data Scientist as a new profession or occupation with related competences, skills and qualifications definition. The CF-DS definition will re-use, extend and map the ESCO taxonomy to the identified Data Science competences and skills.

e) ACM Computing Classification System (ACM CCS2012) [7]

ACM Computing Classification System will be used as a basis to define the proposed Data Science Body of Knowledge, and an extension to ACM CCS2012 will be provided to cover the identified knowledge and required academic subjects. Necessary contacts will be made with the ACM CCS body and corresponding ACM curriculum defining committees.

f) O'Reilly Strata Survey (2013) [18]

It was one of the first extensive studies on Data Scientist organisational roles, profiles and skills. Although skills are defined as very technically and technologically specific, the proposed definition of profiles is important for defining required competence groups; in particular, identification of Data Science Creative and Data Science

Researcher profiles indicates an important role of scientific approach and need for research method training in Data Scientist professional education. This group of competences is included in the proposed CF-DS.

g) EC Report on the Consultation Workshop (May 2012) "Skills and Human Resources for e-Infrastructures within Horizon 2020" [19].

This report provided important information about EC and European research community vision of the needs for Data Science skills for e-Infrastructure, in particular, to support e-Infrastructure development, operation and scientific use. The identified nine skills gap areas provide additional motivation for specific competences and skills training for future Data Scientists who will work in e-Infrastructure that in particular include data management, curation and preservation.

In the course of defining, validating and refining the proposed CF-DS, the EDISON project interacted with the following external projects and studies contributing to them and influencing their approach and alignment with EDSF:

- PwC and BHEF report "Investing in America's data science and analytics talent: The case for action" (April 2017) [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) [21]
- DARE project [6] where the EDISON contributed to the definition of recommended Data Science Analytics competences [22] and their alignment with the EDSF

4.2 Identified Data Science Competence Groups

The results of the job market study and analysis for Data Science and Data Science enabled vacancies, conducted at the initial stage of the project, provided a basis and justification for defining the main competence groups that are commonly required by companies, including identification of such skills as Data Management and Research methods that were not required formerly required for data analytics jobs.

The following CF-DS competence and skills groups have been identified:

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

DSDA, DSENG and DSDM competence groups constitute the core Data Science competences that actually define the main Data Science professional profiles and roles, including those related to different application domains[7]. DSDM and DSRMP competence groups are considered as commonly required for all Data Science professional profiles to ensure effective work with modern data driven technologies and in modern data driven organisations.

Data management, curation and preservation competences are already attributed to the existing (research) data related professions such as data steward, data manager, data librarian, data archivist, and others. Data management is an important component of the European Research Area and Open Data and Open Access policies. It is a basis for consistent implementation of the FAIR (Findable, Accessible, Interoperable, Reusable)[8]

[6] DARE (Data Analytics Rising Employment) project is commissioned by Asia Pacific Economic Cooperation (APEC) council and is focused on defining the Recommended Data Science Analytics competences. The DARE project recommendation is to include the basic competences or literacy in the overall Data Science competences definition.

[7] See section 4.8 discussion about relation between core Data Science concepts and competences and those related to knowledge and technology domains.

[8] FAIR data principles [28] widely accepted by the research community and a core concept in Open Science and Open Data.

data principles widely adopted by the European research community [23]. It is extensively addressed by the Research Data Alliance (RDA) and supported by numerous projects, initiatives and training programmes[9].

Knowledge of the research methods and techniques is something that makes the Data Scientist profession different from all previous professions. Data analysis, data mining, and data exploration are the main processes in Data Science that actually use research methods and hypothesis testing. It should also be coupled with basic project management competences and skills.

From the education and training point of view, the identified competences can be treated or linked to expected learning or training outcomes. This aspect is discussed in detail in relation to the definition of the Data Science Body of Knowledge and Data Science Model Curriculum.

The identified five Data Science related competence groups provide a basis for defining consistent and balanced education and training programmes for Data Science related jobs, re-skilling and professional certification.

Figures 5 (a) and (b) provide a graphical presentation of the relations between identified competence groups for research and business oriented professional profiles that correspondingly include Research Methods competences or Business Process Management competences. The figure illustrates the importance of the Data Management competences and skills and Research Methods or Business Process Management knowledge for all categories and profiles of Data Scientists.

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

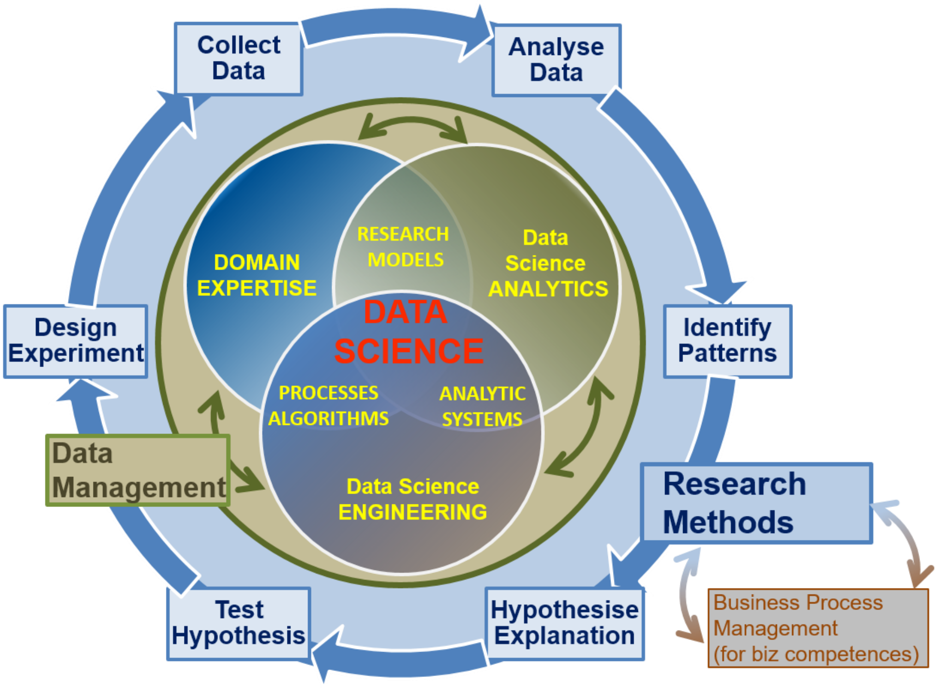
An important part of the research process is theory building, but this activity is attributed to the domain or subject matter researcher. The Data Scientist (or related role) should be aware of domain related research methods and theory as a part of their domain related knowledge and team or workplace communications. See the example of Data Science team building in the Data Science Professional Profiles definition provided as a separate document [4].

There is a number of Business Process Operations models depending on their purpose, but typically they contain the following stages that are generally similar to those for Scientific methods, in particular in collecting and processing data (see reference to exiting definitions (see Appendix C for reference to existing Business Process Management stages definitions):

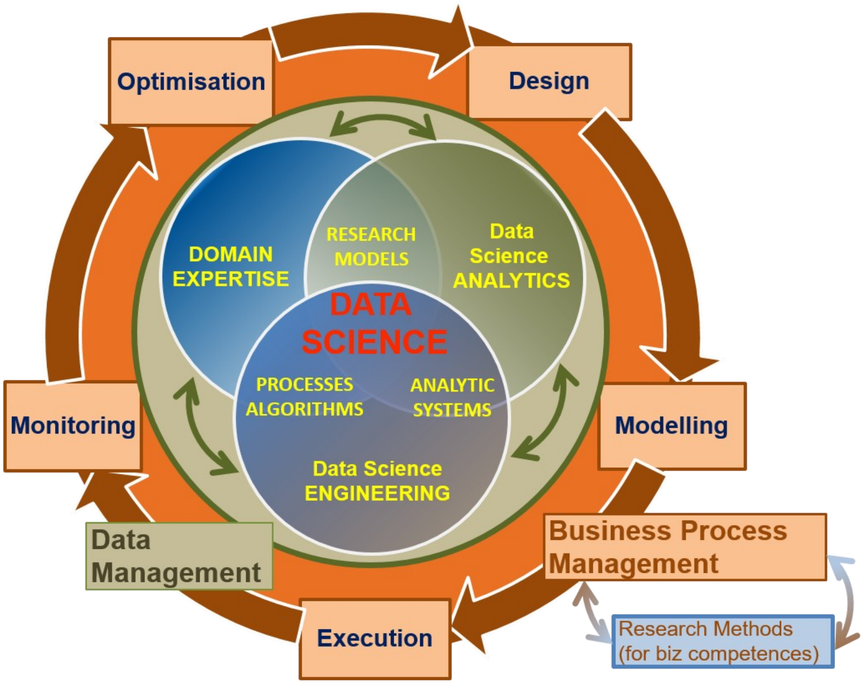
- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

Table 4.1 provides the proposed Data Science competences definition for different groups supported by the data extracted from the collected information. The presented competences definition has been reviewed by a number of expert groups and individual experts as a part of the project EDISON engagement and network activities. The presented competences are required for different professional profiles, organisational roles and throughout the whole data lifecycle, but not necessarily to be provided by a single role or individuum. The presented competences are enumerated to allow easy use and linking between all EDSF documents.

[9] Research Data Alliance Europe https://europe.rd-alliance.org/



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



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

Figures 5. Relations between identified Data Science competence groups for (a) general or research oriented and (b) business oriented professions/profiles: Data Management and Research Methods or Business Processes Management competences and knowledge are important for all Data Science profiles.

Table 4.1. Competences definition for different Data Science competence groups

Data Analytics (DSDA)	Data Science Engineering (DSENG)	Data Management (DSDM)	Research Methods and Project Management (DSRMP)	Domain related Competences (DSDK): Applied to Business Analytics (DSBA)
DSDA Use appropriate data analytics and statistical techniques on available data to discover new relations and deliver insights into research problems or organizational processes and support decision-making.	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.	DSDM Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	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	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
DSDA01 Effectively use a variety of data analytics techniques, such as Machine Learning (including supervised, unsupervised, semi-supervised learning), Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through 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	DSDM01 Develop and implement data strategy, in particular, in the form of data management policy and Data Management Plan (DMP)	DSRMP01 Create new understandings by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods	DSBA01 Analyse information needs, assess existing data and suggest/identify new data required for specific business context to achieve organizational goal, including using social networks and open data sources
DSDA02 Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation, to deploy appropriate models for analysis and prediction	DSENG02 Develop and apply computational and data driven solutions to domain related problems using wide range of data analytics platforms, with the special focus on Big Data technologies for large datasets and cloud based data analytics platforms	DSDM02 Develop and implement relevant data models, define metadata using common standards and practices, for different data sources in variety of scientific and industry domains	DSRMP02 Direct systematic study toward understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSBA02 Operationalise fuzzy concepts to enable key performance indicators measurement to validate the business analysis, identify and assess potential challenges
DSDA03 Identify, extract, and pull together available and pertinent heterogeneous data, including modern data sources such as social media data, open data, governmental data, verify data quality	DSENG03 Develop and prototype specialised data analysis applicaions, 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	DSDM03 Integrate heterogeneous data from multiple sources and provide them for further analysis and use	DSRMP03 Analyse domain related research process model, identify and analyse available data to identify research questions and/or organisational objectives and formulate sound hypothesis	DSBA03 Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends

DSDA04 Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	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)	DSDM04 Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)	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	DSBA04 Analyse opportunity and suggest use of historical data available at organisation for organizational processes optimization
DSDA05 Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making, apply different Data Science process model (CRISP-DM, ASUM, TDSP)	DSENG05 Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	DSDM05 Ensure data quality, accessibility, interoperability, compliance to standards, and publication (data curation)	DSRMP05 Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	DSBA05 Analyse customer relations data to optimise/improve interacting with the specific user groups or in the specific business sectors
DSDA06 Visualise results of data analysis, design dashboard and use storytelling methods	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	DSDM06 Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management	DSRMP06 Develop and guide data driven projects, including project planning, experiment design, data collection and handling	DSBA06 Analyse multiple data sources for marketing purposes; identify effective marketing actions

The identified demand for general competences and knowledge on Data Management and Research Methods needs to be implemented in the future Data Science education and training programs, as well as to be included in re-skilling training programmes. It is important to mention that knowledge of Research Methods does not mean that all Data Scientists must be talented scientists; however, they need to know the general research methods such as formulating hypothesis, applying research methods, producing artefacts, and evaluating hypothesis (so called 4 steps model). Research Methods training are already included in master programs and graduate students at many universities.

The presented CF-DS can be used as a basis for defining competences for other data related professional groups, such as Data Stewards, Artificial Intelligence, and Machine Learning that require an extended set of competences in some of the CF-DS competence groups. Such an approach was used for defining the Data Stewardship Professional Competence Framework (CF-DSP) in the FAIRsFAIR project [6]; short information about CF-DSP extension is provided in section 7.

4.3 Identified Data Science Skills and their mapping to Competences

Required Data Science skills are defined based on the job market study of the current analysis of the Data Science job market, extended with the numerous blog articles analysis[10] published by Data Science practitioners, which provide valuable information in the such new emerging area as Data Science.

The identified skills can be organised in the following groups:

- Data Science skills related to the main competence groups that cover knowledge and experience related to effectively realise defined competences and related organisational functions;
- Data analytics and data handling languages, tools, platforms, and applications, including SQL based applications and data management tools;
- Knowledge and experience with the Big Data infrastructure platforms and tools.

Separately defined are personal and attitude skills, also referred to as the 21st century skills and Data Science professional skills those that define specific (personal) skills that the Data Scientist need to develop to successfully work as a Data Scientist in different organisational roles and along their career. The analysis and identified Data Science soft skills are described in section 4.6.

4.3.1 Data Science skills related to the main competence groups – Skills Type A

Table 4.2 lists identified skill groups:
Group 1. Data Science skills related to the main competence groups

- Data Science Analytics covering extensive skills related to using different Machine Learning, Data Mining, statistical methods and algorithms;
- Data Science Engineering skills related to design, implementation and operation of the Data Science (or Big Data) infrastructure, platforms and applications
- Data Management and governance (including both general data management and research data management)
- Research Methods and Project Management
- Business Analytics as an example of domain related skills

The Data Science Analytics group is the most populated what reflects wide spectrum of required skills in this group as a core for Data Science. It is followed by the Data Science Engineering skills that are important for the Data Scientist to have the ability to implement effective data analytics solutions and applications.

[10] It was anticipated that for such new technology domain as Data Science the blog articles provided a valuable source of information. Information extracted from them was correlated with other sources and in many cases provided valuable expert opinion. Opinion based research is one of the basic research methods and can produce valid results.

Table 4.2. Identified Data Science skills related to the main Data Science competence groups (hereafter referred to as Skills Type A Group 1)

-

SDSDA **Data Science Analytics**	**SDSENG** **Data Science Engineering**	**SDSDM** **Data Management**	**SDSRMP** **Research Methods and Project Management**	**SDSBA** **Business Analytics**
SDSDA01 Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)	SDSENG01 Use systems and software engineering principles to organisations information system design and development, including requirements design	SDSDM01 Specify, develop and implement enterprise data management and data governance strategy and architecture, including Data Management Plan (DMP)	SDSRMP01 Use research methods principles in developing data driven applications and implementing the whole cycle of data handling	SDSBA01 and Business Intelligence (BI) methods for data analysis; apply cognitive technologies and relevant services
SDSDA02 Use Data Mining techniques	SDSENG02 Use Cloud Computing technologies and cloud powered services design for data infrastructure and data handling services	SDSDM02 Data storage systems, data archive services, digital libraries, and their operational models	SDSRMP02 Design experiment, develop and implement data collection process	SDSBA02 Apply Business Processes Management (BPM), general business processes and operations for organisational processes analysis/modelling
SDSDA03 Use Text Data Mining techniques	SDSENG03 Use cloud based Big Data technologies for large datasets processing systems and applications	SDSDM03 Define requirements to and supervise the implementation of the hybrid data management infrastructure, including enterprise private and public cloud resources and services	SDSRMP03 Apply data lifecycle management model to data collection and data quality evaluation	SDSBA03 Apply Agile Data Driven methodologies, processes and enterprises
SDSDA04 General statistical analysis methods and techniques, Descriptive analytics	SDSENG04 Use agile development technologies, such as DevOps and continuous improvement cycle, for data driven applications	SDSDM04 Develop and implement data architecture, data types and data formats, data modeling and design, including related technologies (ETL, OLAP, OLTP, etc.)	SDSRMP04 Apply a structured approach to use cases analysis	SDSBA04 Use Econometrics for data analysis and applications
SDSDA05 Use Quantitative Analytics methods	SDSENG05’ Develop and implement systems and data security, data access, including data anonymisation, federated access control systems	SDSDM05 Implement data lifecycle support in organisational workflow, support data provenance and linked data	SDSRMP05 Develop and implement Research Data Management Plan (DMP), apply data stewardship procedures	SDSBA05 Develop data driven Customer Relations Management (CRP), User Experience (UX) requirements and design
SDSDA06 Use Quantitative Analytics methods	SDSENG06 Apply compliance based security models, in particular for privacy and IPR protection	SDSDM06 Consistently implement data curation and data quality controls, ensure data integration and interoperability	SDSRMP06 Consistently apply project management workflow: scope, planning, assessment, quality and risk management, team management	SDSBA06 Apply a structured approach to use cases analysis in business and industry

SDSDA07 Apply Predictive analytics methods	SDSENG07 Use relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for structured and unstructured data	SDSDM07 Implement data protection, backup, privacy, mechanisms/ services, comply with IPR, ethics and responsible data use		SDSBA07 Use Data Warehouses technologies for data integration and analytics, including use open data and social media data
SDSDA08 Apply Prescriptive Analytics methods	SDSENG08 Effectively use Big Data infrastructures, high-performance networks, infrastructure and services management and operation	SDSDM08 Use and implement metadata, PID, data registries, data factories, standards and compliance		SDSBA08 Use data driven marketing technologies
SDSDA09 Use Graph Data Analytics for organisational network analysis, customer relations, other tasks	SDSENG09 Use and apply modeling and simulation technologies and systems	SDSDM09 Adhere to the principles of Open Data, Open Science, Open Access, use ORCID based services		SDSBA09 Mechanism Design and/or Latent Dirichlet Allocation
SDSDA10 Apply analytics and statistics methods for data preparation and pre-processing	SDSENG10 Use and integrate with the organisational Information systems, collaborative system			
SDSDA11 Use performance and accuracy metrics for data analytics assessment and validation	SDSENG11 Design efficient algorithms for accessing and analysing large amounts of data, including API to different databases and data sets			
SDSDA12 Use effective visualiation and storytelling methods to create dashboards and data analytics reports	SDSENG12 Use of Recommender or Ranking system			
KDSDA13 Apply Data Science Process Models (CRISP-DM, ASUM, TDSP)				
SDSDA14 Use Natural Language Processing methods				
SDSDA15 Operations Research methods for practical tasks				
SDSDA16 Optimisation methods for practical tasks				
SDSDA17 Simulation methods for practical tasks				

It is important to mention that the whole complex of Data Science related competences, skills and knowledge are strongly based on the mathematical foundation that should include knowledge of mathematics (including linear algebra, calculus, etc), statistics and probability theory.

4.3.2 Data Science skills related to the Data Analytics languages, tools, platforms and Big Data infrastructure – Skills Type B

Table 4.3 lists identified skills related to the Data Analytics languages, tools, platforms and Big Data infrastructure that are split into the following groups:

Group 3. Big Science and Big Data platforms and tools

- DSDALANG - Data Analytics and Statistical languages and tools
- DSADB - Databases and query languages
- DSVIZAPI - Data Collection and Visualization, WebAPI
- DSADM - Data Management and Curation platform
- DSBDA - Big Data Analytics platforms (cloud based)

Group 4. Data Science development and project management platforms and tools

- DSDEV - Development and project management frameworks, platforms and tools

It is important for Data Scientist to be familiar with multiple data analytics languages and demonstrate proficiency in one or a few of the most popular languages (what should be supported by several years of practical experience)[11], such as

- Python and related data analytics libraries
- R including extensive data analysis libraries
- KNIME
- SPSS
- Additionally (but might be legacy): Julia, SAS, WEKA, Orange, others

Data Science practitioner must be familiar and have experience with the general programming languages, software versioning and projects management environments such as

- Java, JavaScript and/or C/C++ as general applications programming languages
- Git versioning system as a general platform for software development
- Scrum agile software development and management methodology and platform, recently developed to the DataOps and MLOps frameworks

It is essential to mention that all modern Big Data platforms and general data storage and management platforms are cloud based. The knowledge of Cloud Computing and related platforms for application deployment and data management are included in the table. The use of cloud based data analytics tools is growing, and most big cloud services providers provide whole suites of platforms and tools for enterprise data management from Enterprise Data Warehouses, data backup and archiving to business data analytics, data visualization and content streaming

[11] Consider proposed here lists as examples and refer to other more focused and extended research and discussions such as for example blog article "Data Scientist Core Skills", Blog article by Mitchell Sanders, posted on August 27, 2013 [online] http://www.datasciencecentral.com/profiles/blogs/data-scientist-core-skills

Table 4.3. Required skills related to analytics languages, tools, platforms, data management and Big Data infrastructure (Skills Type B Group 3 and Group 4) [12]

DSDALANG **Data Analytics and Statistical languages and tools**	**DSADB** **Databases and query languages**	**DSVIZAPI** **Data collection and visualization**	**DSADM** **Data Management and Curation platform**	**DSBDA** **Big Data Analytics platforms**	**DSDEV** **Development and project management frameworks, platforms and tool**
DSDALANG01 Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, pytorch, scikit-learn, seaborn, etc.)	DSADB01 SQL and relational databases, incl. open source: PostgreSQL, mySQL, etc.)	DSVIZAPI01 Data visualization Libraries (mathpoltlib, seaborn, D3.js, FusionCharts, Chart.js, other)	DSADM01 Data modelling and related technologies (ETL, OLAP, OLTP, etc.)	DSBDA01 Big Data and distributed computation models and tools (Hadoop, Spark, MapReduce, Mahout, Lucene, NLTK, Pregel, etc.)	DSDEV01 Frameworks: Python, Java, C/C++, GO, D3.js (Data-Driven Documents), jQuery, others
DSDALANG02 R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)	DSADB02 SQL and relational databases (proprietary: Oracle, MS SQL Server, others)	DSVIZAPI02 Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)	DSADM02 Data Warehouse platform and related tools	DSBDA02 Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)	DSDEV02 Python, R, Java or C/C++ Development platforms/IDE (Anaconda/Jupyter Notebook, R Studio, Eclipse, Visual Studio Code, Atom, others)
DSDALANG03 SAS	DSADB03 Data Warehouses, Master Data Management, Complex Events Processing	DSVIZAPI03 Online visualization tools (Datawrapper, Google Visualisation API, Google Charts, Flare, etc)	DSADM03 Data curation platform, metadata management (Curator's Workbench, DataUp, etc)	DSBDA03 Real time and streaming analytics systems (Flume, Kafka, Storm)	DSDEV03 Git versioning system as a general platform for software development
DSDALANG04 IBM SPSS	DSADB04 NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)	DSVIZAPI04 WebAPI and access to external data/datasets access	DSADM04 Backup and storage management (iRODS, XArch, Nesstar, others)	DSBDA04 Hadoop Ecosystem/platform (Apache, Cloudera, etc)	DSDEV04 Scrum agile software development and management methodology and platform
DSDALANG05 Julia	DSADB 05 Hive (query language for Hadoop)	DSVIZAPI05 Web Scraping and data collection from the web	DSADM05 Big Data and cloud based storage platforms and services	DSBDA05 Azure Data Analytics platforms (Machine Learning Studio, HDInsight, Data Lake Analytics, Power BI, etc)	DSDEV05 Data Science oriented DevOps platforms (DataOps and MLOps)
DSDALANG06 RapidMiner	DSADB 06 Data Modeling (UML, ERWin, DDL, etc)		DSADM06 FAIR Data Management and metadata tools	DSBDA06 Amazon Data Analytics platform (SageMaker, EMR, Kinesis, Data	

[12] The presented here Big Data platforms and tools are examples of the most popular platforms and tools and are not exhaustive. This list also doesn’t include possible proprietary companies’ platforms and tools. Please search for general and domain specific other general and domain specific reviews and inventories, for example: Data Science Knowledge Repo https://datajobs.com/data-science-repo/

				Pipeline, Machine Learning, etc)	
DSDALANG07 Other analytics, statistical and programming languages (WEKA, KNIME, Scala, Stata, Orange, etc)				DSBDA07 Google Analytics Platform (Google Data Studio, Machine Learning, TensorFlow, others)	
DSDALANG08 Scripting language, e.g. JavaScript, PHP, Octave, Pig, others				DSBDA08 IBM Watson Analytics	
DSDALANG09 Matlab Data Analytics				DSBDA09 Other cloud based Data Analytics platforms (Cloudera Data Science Workbench Vertica, LexisNexis HPCC System, etc)	
DSDALANG10 Analytics tools (R/R Studio, Python/Anaconda, SPSS, Matlab, etc)				DSBDA10 Cognitive platforms (such as IBM Watson, Microsoft Cortana, others)	
DSDALANG11 Data Mining tools: RapidMiner, Orange, R, WEKA, NLTK, others				DSBDA11 Kaggle competition, resources and community platform	
DSDALANG12 Excel Data Analytics (Analysis ToolPack, PivotTables, etc)					

*) The majority of the listed Big Data and analytic platforms, online data management platforms are cloud based. They are becoming increasingly popular for enterprise and business applications and provide important features for modern businesses, such as scalability and on-demand resources allocation. The cloud based services and applications are typically well supported by the providers' deployment and monitoring

4.4 Knowledge required to support identified competences

Table 4.4 provides an enumerated list of knowledge topics[13] that are required to support corresponding competence groups. There is no direct mapping between individual competences and knowledge units, singe competence may be mapped to multiple knowledge units.

Table 4.4. Knowledge required to support identified competences

KDSDA **Data Science Analytics**	**KDSENG** **Data Science Engineering**	**KDSDM** **Data Management**	**KDSRMP** **Research Methods and Project Management**	**KDSBA** **Business Analytics**
KDSDA01 Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others	KDSENG01 Systems Engineering and Software Engineering principles, methods and models, distributed systems design and organisation	KDSDM01 Data management and enterprise data infrastructure, private and public data storage systems and services	KDSRMP01 Research methods, research cycle, hypothesis definition and testing	KDSBA01 Business Analytics (BA) and Business Intelligence (BI); methods and data analysis; cognitive technologies
KDSDA02 Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)	KDSENG02 Cloud Computing, cloud based services and cloud powered services design	KDSDM02 Research Data Management, FAIR data principles, Data Stewardship Data storage systems, data archive services, data libraries, and their operational models	KDSRMP02 Experiment design, modelling and planning, Experimental research reproducibility	KDSBA02 Business Processes Management (BPM), general business processes and operations, organisational processes analysis/modelling
KDSDA03 Machine Learning (reinforced value based, policy based, model based): Q-Learning, TD-Learning, Genetic Algorithms	KDSENG03 Big Data technologies for large datasets processing: batch, parallel, streaming systems, in particular cloud based	KDSDM03 Data governance, data governance strategy, Data Management Plan (DMP)	KDSRMP03 Open Access, Open Science, Open Data, research data archives/ repositories, ORCID, FAIR data principles	KDSBA03 Agile Data Driven methodologies, processes and enterprises
KDSDA04 Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)	KDSENG04 Applications software requirements engineering and design, agile development technologies, DevOps and continuous improvement cycle	KDSDM04 Data Architecture, data types and data formats, data modeling and design, including related technologies (ETL, ELT, etc.)	KDSRMP04 Research Data Management, Data Management Plan (DMP), data stewardship, FAIR data principles	KDSBA04 Econometrics: data analysis and applications
KDSDA05 Text Data Mining: statistical methods, NLP, feature selection, apriori algorithm, etc.	KDSENG05' Systems and data security, data access, incl data anonymisation, federated access control systems	KDSDM05 Data lifecycle and organisational workflow, data provenance and linked data	KDSRMP05 Data lifecycle and data collection, data quality assurance	KDSBA05 Data driven Customer Relations Management (CRM), User Experience (UX) design

[13] Note, EDSF uses terminology convention when referring to knowledge aspects: (1) competences definition includes required knowledge topics; (2) Body of Knowledge (BoK) uses hierarchy Knowledge Area Groups (KAG) corresponding to competence groups, Knowledge Areas (KA), Knowledge Units (KU) as a lowest level. Knowledge topics can be mapped to KAs or KUs in the BoK.

KDSDA05 General statistical analysis methods and techniques, Descriptive analytics	KDSENG06 Compliance based security models, privacy and IPR protection	KDSDM06 Data curation and data quality, data integration and interoperability	KDSRMP06 Use cases analysis: research infrastructure and projects	KDSBA06 Use cases analysis: business and industry
KDSDA07 Quantitative Analytics	KDSENG07 Relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL/ELT, OLTP/OLAP processes	KDSDM07 Data protection, backup, privacy, IPR, ethics and responsible data use	KDSRMP07 Project management: scope, planning, assessment, quality and risk management, team management	KDSBA07 Data Warehouses technologies, data integration and analytics
KDSDA08 Qualitative Analytics	KDSENG08 Big Data infrastructure services, high-performance networks	KDSDM08 Metadata, PID, FDO (FAIR Digital Object), data registries, data factories, standards and compliance		KDSBA08 Data driven marketing technologies
KDSDA09 Predictive Analytics	KDSENG09 Modeling and simulation, theory and systems			
KDSDA10 Prescriptive Analytics	KDSENG10 Information systems, collaborative systems			
KDSDA11 Data Science Process Models (CRISP-DM, ASUM, TDS)P				
KDSDA12 Data Analytics Model Formats (PMML, PFA, ONNX, TensorFlow, etc)				
KDSDA11 Graph Data Analytics: (path analysis, connectivity analysis, community analysis, etc.				
KDSDA12 Natural language processing				
KDSDA13 Data preparation and pre-processing				
KDSDA14 Performance and accuracy metrics				
KDSDA15 Markov Models, Conditional Random Fields				
KDSDA16 Operations Research				
KDSDA17 Optimisation				
KDSDA18 Simulation				

4.5 Proficiency levels

It is essential to mention that for the such complex professional domain as Data Science, the practical experience of working with data analytics languages, tools and platforms are essential and typically required from minimum 1 to 3 years to be able to develop the minimum required experiences with related analytics methods and applications required to solve critical organisational needs[14]. Although many companies explicitly require experience up to 5 years, the current shortage of skilled Data Scientists will demand novel approaches to targeted competences and skills development that should combine individual competences assessment, design of tailored training for deficient skills development and personalised workplace (self-)training.

The definition of the proficiency levels of individual competes is an important dimension in the CF-DS definition. The CF-DS follows the e-CF3.0 approach in defining the proficiency levels of individual competences. The 5 proficiency levels defined in e-CF are mapped to levels 4-8 of the EQF (European Qualification Framework) [17] (refer to Appendix C for mapping between e-CF proficiency levels and EQF qualification levels). For easier linking to the Model Curricula, the CF-DS defines 3 levels of the Data Science competences what is considered sufficient for practical purposes of job description, certification and education and training courses development (however future development or some practical considerations may intend to define 5 levels similar to e-CF):

- Associate (entry level): basic or entry level that defines minimum competences and skills to be able to work in a Data Science team under the supervision
- Professional that indicates the ability to solve major tasks independently, use multiple languages, tools and platforms and develop specialised applications
- Expert (lead professional) that requires wide knowledge and experience with multiple Data Analytics, engineering and data management areas, and related tools. Platforms and Big Data infrastructure services. Expert level is typically required from the lead Data Scientist, manager of the Data Science team, or similar.

Examples of the proficiency levels definition for Data Science Analytics competences are provided in section 5. It is essential that all Data Science competences are strongly based on the common/foundational competences and skills that include necessary knowledge and practical experience in mathematics, statistics, statistical and analytics languages, general computation skills, visualisation as defined in the previous sections (see section 4.7 for additional considerations).

4.6 Data Scientist Workplace skills (aka “soft” or transversal skills)

Although it is commonly agreed on the importance of soft skills for Data Scientists, the job market analysis additionally confirmed the importance of personal skills and identified a number of specific Data Science professional skills (what can be defined as “Thinking and acting like a Data Scientist”) that are required for the Data Scientist to effectively work in the modern agile data driven organisations and project teams. These should also be complemented with the general personal skills referred to as 21st Century skills (P21C skills) [24]. The importance of such skills for Data Scientist is defined by their cross-organisational functions and responsibilities in collecting and analysing organisational data to provide insight for decision making. In such a role, the Data Scientist often reports to the executive level or to other departments and teams. These skills extend beyond traditionally required communication or team skills. In addition, the ideal Data Scientist is expected to bring and spread new (data analytics) knowledge to organisation and ensure that the results of their work contribute to the consistency of the data collection, analysis, exploitation, and decision making processes.

4.6.1 Data Science Professional or Attitude skills (Thinking and acting like a Data Scientist)

Data Science is growing as a distinct profession and consequently will need professional identification via definition of the specific professional skills and code of conduct that can be defined as “Thinking and acting like Data Scientist”. Understanding, recognising and acquiring such skills is essential for Data Scientists to progress successfully in their careers. It is also important for team leaders to correctly build relations in the team or a project group.

[14] Whilst the authors recognise a common for employers to state in the job description the experience requirements in terms of years, timeframes do not have direct correlation with competences. Competence is based upon capability derived from knowledge, skills and attitude gained from a variety of inputs, for example, education and training. In fact, time does not have a linear relationship with the competence (level) but may depend on overall employee’s experience and education.

Table 4.5 lists the Data Science professional (or attitude) skills which are identified by the Data Science practitioners and educators. Although some of the skills are common in the 21st Century skills, it is important to provide the whole list of skills that can provide guidance for future Data Scientists on what skills are expected from them and need to be developed in their careers.

Table 4.5. Data Science Professional skills (Thinking and acting like Data Scientist)

Skill ID	Skill definition
DSPS	General group definition: Thinking and acting like a Data Scientist
DSPS01	Accept/be ready for iterative development, know when to stop, comfortable with failure, accept the symmetry of outcome (both positive and negative results are valuable)
DSPS02	Ask the right questions
DSPS03	Recognise what things are important and what things are not important
DSPS04	Respect domain/subject matter knowledge in the area of data science
DSPS05	Data driven problem solver and impact-driven mindset
DSPS06	Recognise value of data, work with raw data, exercise good data intuition
DSPS07	Good sense of metrics, understand importance of the results validation, never stop looking at individual examples
DSPS08	Be aware of the power and limitations of the main machine learning and data analytics algorithms and tools
DSPS09	Understand that most of data analytics algorithms are statistics and probability based, so any answer or solution has some degree of probability and represents an optimal solution for a number of variables and factors
DSPS10	Working in an agile environment and coordinating with other roles and team members
DSPS11	Work in multi-disciplinary teams, ability to communicate with the domain and subject matter experts
DSPS12	Embrace online learning, continuously improve your knowledge, use professional networks and communities
DSPS13	Story Telling: Deliver actionable results of your analysis
DSPS14	Attitude: Creativity, curiosity (willingness to challenge status quo), commitment to finding new knowledge and progress to completion
DSPS15	Ethics and responsible use of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

4.6.2 21st Century skills

21st Century skills comprise a set of general workplace skills that include critical thinking, creativity, communication, collaboration, organizational awareness, ethics, and others. The importance of this kind of skills is motivated by the fast technologies development and the ongoing digital transformation of the modern economy and Industry 4.0 in particular.

Table 4.6 lists the 21st Century skills defined based on the recommendations of the DARE Project [22], P21's Framework for 21st Century Learning [25], and the OECD Report on industry digitalisation [26].

Table 4.6. The 21st Century workplace skills

Skill ID	**Skill definition**
SK21C	General group definition: Critical thinking, communication, collaboration, organizational awareness, attitude, etc.
SK21C01	1. Critical Thinking: Demonstrating the ability to apply critical thinking skills to solve problems and make effective decisions
SK21C02	2. Communication: Understanding and communicating ideas
SK21C03	3. System thinking and Design thinking: Use system thinking and design thinking cycle stages in delivering intended solution
SK21C04	3. Collaboration: Working with others, appreciation of multicultural difference
SK21C05	4. Creativity and Attitude: Deliver high quality work and focus on the final result, initiative, and intellectual risk
SK21C06	5. Planning & Organizing: Planning and prioritizing work to manage time effectively and accomplish assigned tasks
SK21C07	6. Business Fundamentals: Having fundamental knowledge of the organization and the industry
SK21C08	7. Customer Focus: Actively look for ways to identify market demands and meet customer or client needs
SK21C09	8. Working with Tools & Technology: Selecting, using, and maintaining tools and technology to facilitate work activity
SK21C10	9. Dynamic (self-) re-skilling: Continuously monitor individual knowledge and skills as shared responsibility between employer and employee, ability to adapt to changes
SK21C11	10. Professional network: Involvement and contribution to professional network activities
SK21C12	11. Ethics: Adhere to high ethical and professional norms, responsible use of power data driven technologies, avoid and disregard un-ethical use of technologies and biased data collection and presentation

4.6.3 Design Thinking and System Thinking

Although Design Thinking and System Thinking are often attributed to the future 21st Century skills, we see the benefit of defining them in details. Table 4.7 below provides an overview pf the design thinking cycle and stages together with the system thinking cycle that allows a complex/system approach to problem solving and aligning problem space, solution space, and user/consumer needs. The main difference of the two thinking models is the focus on the system in system thinking and the focus on the users and needs of people in design thinking.

Table 4.7. Design and System thinking cycle stages

Design Thinking	**System Thinking**
• Understand	• Problem definition

• Observe • Define point of view • Ideate • Prototype • Test • Reflect	• Mapping of reality • Situation analysis • Goal formulation (decision making criteria) • Search for a solution • Evaluation • Decision

Design thinking is best realized in inter-disciplinary team work and is closely related to Agile technologies which are focused on fast product development and continuous improvement based on user feedback. In its own turn, design thinking empowers agile teams for consistent problem solving. Design thinking should continuously switch and iterate between problem space and solution space. The Design Thinking mindset includes the following aspects:

- Driven by the problem solving curiosity
- Focused on people as a target for products or services
- Accept complexity
- Develop process awareness and the whole lifecycle
- Visualise and show relations
- Prototype, experiment and iterate
- Co-create, grow, and scale with varying perspectives and frameworks
- Collaborate in networks
- Reflect on actions

System approach and system thinking are considered as an important set of skills for the future economy reflecting a complex and interconnected environment and a problem space in which future workers will work. The system approach has been developed and widely used in the technical domain; however, its benefits recognized in other domains. The mindset of the system thinker should:

- Always look at the big picture
- Think positively about system improvement, and don’t complain if the system doesn’t work
- Check the results and improve results with each iteration
- Reflect on the way of thinking because it may affect what will happen
- Take time to penetrate even complex interconnections
- Search for the key to the system
- Consider facts from different perspectives
- Accept the change takes place gradually and interconnections also trigger changes
- Identify an effect that was triggered by an action

Design and system thinking can be implemented by adopting Agile and DevOps methodology that actually defines a culture of project/goal/mission based collaboration and team management.

4.7 Data Scientist in the modern agile data driven organisation

Companies intending to implement data driven business methods and benefit from available data or data that can be collected expect that Data Scientist will provide the necessary expertise and insight to achieve the company’s goals. In these cases, Data Scientist will face and will need to cope with the expectations to his or her role in organisation which are, in some cases, far beyond ordinary analyst, engineer or programmer. The following list of expected Data Scientist’s contribution is compiled from the collected information and other studies:

- Optimise, improve what related to organizational mission, goals, performance
- Support, advise what related to organizational processes, roles
- Develop, implement and operate data driven services
- Prepare insightful report, targeted analysis
- Monitor processes and services with smart data
- Discover new relations and realise new possibilities
- Use scientific/research methods to discover new relations and solve problems
- Translate business/organizational needs to computational tasks

- Manage data: collect, aggregate, curate, search, visualize

Reflecting on the observation that organisations expect that Data Scientist will bring general Big Data and Data Science knowledge to organisation, we can see a need for the general Data Science literacy and data driven thinking in organisation that would ensure a necessary level of understanding of the result of the Data Science processes and data analytics results. This means that management and all workers would need to obtain general knowledge of data analytics methods, visualisation, data management, data presentation and structures, and understand data analytics and other tools.

This should motivate general Data Science literacy training in organisations what should be the responsibility of the management. Such training should also focus on a general data presentation and visualisation to enable effective communication of the results and a clear definition of the data analytics tasks.

4.8 Data Science Literacy: Commonly required competences and skills for Data Science related and enabled occupations/roles

Data Scientist can do data analytics work and provide important insight into organisational, process, or events related data. However, for the Data Scientists or data analytic team to work effectively, there is a need for common knowledge and understanding of the data analysis process and its place in the whole data lifecycle and organisational data driven workflow. This can be achieved by defining a common required knowledge and skills in data handling and data analytics. The goal of this is to enable all workers and roles correctly handle data, collect and present them to analysis, understand the outcome the analysis and provide possible feedback from the domain expertise point of view.

Following the outcome and recommendations of the DARE project, we define the basic Data Science Analytics competences and skills (also can be referred to as data literacy) that must be required for all roles working in the Data Science team or interacting with the data analytics teams.

The following competences and skills are defined as basic or common literacy level:

- **Mathematics:** functions, equations, algebra, linear algebra, calculus, vector analysis, other.
- **Statistical methods and techniques:** General statistical analysis techniques and their use for data inspection, exploration, analysis and visualisation (as supporting activity for more complex data analysis).
- **Computational thinking and programming with data:** Apply information technology, computational thinking, and utilize programming languages and software and hardware solutions for data analysis.
- **Programming languages and tools for data analysis:** Use general and specialised statistical and data analysis programming languages and tools to develop specialised data analysis processes and applications
- **Data visualization languages and tools:** Create and communicate compelling and actionable insights from data using visualization and presentation tools and technologies.
- **Data Management:** Data collection, data entry and annotation, data preparation, data and files versioning, Data Management Plan (DMP), FAIR data principles, metadata, Open Data, data repositories

A further definition of basic or Data and Data Science literacy can be done by incorporating definitions from the recently published DigComp 2.2 The Digital Competence Framework for Citizens [26] and EntreComp: The Entrepreneurship Competence Framework [27].

4.9 Relation between Data Scientist and Subject Domain specialist

Data Scientist by definition is playing assistant role to the main organisational management (decision making) role or a subject domain scientific/researcher role to help them with data processing and organizing data management to achieve their specific management or research role. However, Data Scientist has also an opportunity to play a leading role in some data driven projects or functions because of their potentially wider vision of the organisational processes or influencing factors.

To understand this, we need to look closer at the relation between Data Scientist and the subject domain specialist. The subject domain is generally defined by the following components:

- Model (and data types)
- Methods (and additionally theory)

- Processes
- Domain specific data types and presentation, including visualization methods
- Organisational roles and relations

Data Scientist as an assistant to the subject domain specialist will do the following work that should bring benefits to organisation or facilitate scientific discovery:

- Translate subject domain Model, Methods, Processes into abstract data driven form
- Implement computational models in software, build required infrastructure and tools
- Do (computational) analytic work and present it in a form understandable to subject domain
- Discover new relations originated from data analysis and advice subject domain specialist
- Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Figure 6 illustrates relations between subject domain components and those mapped to the Data Science domain which is abstract, formalised and data driven.

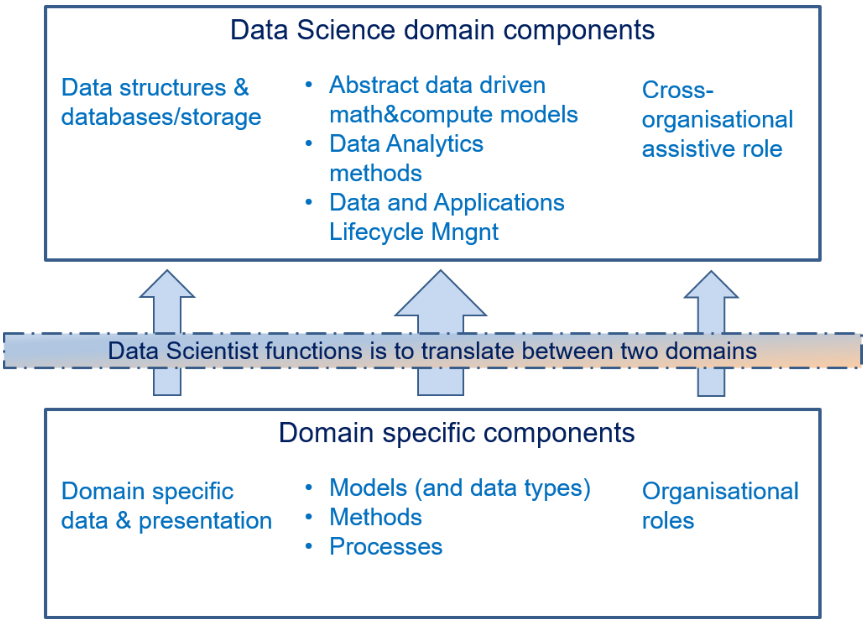

Data Science domain components
Data structures & databases/storage
• Abstract data driven math&compute models
• Data Analytics methods
• Data and Applications Lifecycle Mngnt
Cross-organisational assistive role
Data Scientist functions is to translate between two domains
Domain specific components
Domain specific data & presentation
• Models (and data types)
• Methods
• Processes
Organisational roles

Figure 6. Relations between subject domain and Data Science domain and role of Data Scientist.

Formalisation of the relations between the components and work activities of the subject domain specialist/scientist and Data Science domain provides additional arguments to the discussion about the Data Scientist contribution to the scientific research and discovery that has been recently disputed in many forums: Should Data Scientist be treated as an author of the potential scientific discovery, or just be acknowledged for contribution as assistant role.

5 Example of Data Science and Analytics Competences definition

This section provides examples of the detailed competences definition for the Data Science competence group in a format similar to e-CF3.0/e-CF4.0. This includes the definition of the proficiency levels, mapping to identified skills and knowledge subjects.

Note: The authors acknowledge that the presented here competences are provided as examples, for the full definition of CF-DS competences and other EDSF components[15] refer to the EDSF development github project[16]. It should also be noted that it is not necessarily the case that competences can be performed at every proficiency level. In many cases, competence descriptions are applicable at only one or two levels. (see e-CF framework overview, this demonstrates this principle in a matrix overview)

[15] Full and up-to-date competences definition and other components of EDSF is maintained via controlled vocabulary and attributes collection in the form of Excel workbook. It is untended that it will be migrated to ontology and other API ready format as the EDSF based applications development progresses.

[16] https://github.com/EDISONcommunity/EDSF /

Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA01	Effectively use variety of data analytics techniques, such as Machine Learning (including supervised, unsupervised, semi-supervised learning), Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through the whole data lifecycle

Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 2 (Professional)	Level 3 (Expert)
	Understand and be able to select an approach to analysing selected datasets. Demonstrate understanding and perform statistical hypothesis testing, explain statistical significance.	Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction	Develop and plan required data analytics for organizational tasks, including: evaluating requirements and specifications of problems to recommend possible analytics-based solutions

Dimension 4	Knowledge ID	Knowledge topics definition
Knowledge	KDSDA01	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others
	KDSDA02	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)
	KDSDA03	Machine Learning (reinforced): Q-Learning, TD-Learning, Genetic Algorithms)
	KDSDA04	Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)
	KDSDA06	Predictive Analytics
	KDSDA07	Prescriptive Analytics
	KDSDA11	Data preparation and pre-processing
	KDSDA12	Performance and accuracy metrics

	Skill ID	Skills definition
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)
	SDSDA02	Use Data Mining techniques
	SDSDA04	Apply Predictive Analytics methods
	SDSDA05	Apply Prescriptive Analytics methods
	SDSDA06	Use Graph Data Analytics for organisational network analysis, customer relations, other tasks
Skills Data Analytics languages, tools and platforms	DSALANG01	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, pytorch, scikit-learn, seaborn, etc.)
	DSALANG02	R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)
	DSADB01	SQL and relational databases (open source: PostgreSQL, mySQL, etc.)
	DSADB03	NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)
	DSAVIZ02	Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
	DSABDA03	Real time and streaming analytics systems (Flume, Kafka, Storm)
	DSABDA09	Kaggle competition, resources and community platform
	DSADEV03	Git versioning system as a general platform for software development

•

Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA02	Apply designated quantitative techniques, including statistics, time series analysis, optimization, and simulation to deploy appropriate models for analysis and prediction

Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 2 (Professional)	Level 3 (Expert)
	Be familiar and use related methods and tools. Work under supervision or guidance	Independent work and development. Knowledge and experience with multiple techniques ad tools. Full applications development and deployment	Expert knowledge and experience with multiple data analytics techniques, tools and platforms, Architecture level development, assessment and selection of appropriate solution. Suggestions for new approaches and applications, including relevant data collection.

Dimension 4	Knowledge ID	Knowledge topics definition
Knowledge	KDSDA01	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others
	KDSDA02	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)
	KDSDA04	Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)
	KDSDA06	Predictive Analytics
	KDSDA14	Optimisation
	KDSDA15	Simulation
	Skill ID	Skills definition
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)
	SDSDA02	Use Data Mining techniques
	SDSDA04	Apply Predictive Analytics methods
	SDSDA13	Apply oprtimisation methods
	SDSDA14	Use computer simulation methods
Skills Data Analytics languages, tools and platforms	DSALANG01	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)
	DSADB01	SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)
	DSADB03	NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
	DSABDA03	Real time and streaming analytics systems (Flume, Kafka, Storm)
	DSADEV01	Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others
	DSADEV03	Git versioning system as a general platform for software development

•

<table>
<tr><td>Dimension 1 Competence Group</td><td>DSDA</td><td>Data Science Analytics</td></tr>
<tr><td>Dimension 2 Competence</td><td>DSDA03</td><td>Identify, extract, and pull together available and pertinent heterogeneous data, including modern data sources such as social media data, open data, governmental data</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>Dimension 3 Proficiency level</td><td colspan="2"><table><tr><th>Level 1 (Entry/Associate)</th><th>Level 2 (Professional)</th><th>Level 3 (Expert)</th></tr><tr><td>Collect data from multiple sources, apply data quality check, use corresponding APIs to access different data sources. Be able to write SQL and ETL scripts.</td><td>Collect and integrate necessary data sources. Define necessary transformations and data preparation procedures, write necessary pipelines.</td><td>Identify existing and suggest new data required for organisational analytics tasks to deliver maximum insight. Verify data quality and veracity. Define policy and manage IPR issues.</td></tr></table></td></tr>
<tr><td>Dimension 4</td><td>Knowledge ID</td><td>Knowledge topics definition</td></tr>
<tr><td>Knowledge</td><td>KDSDA10</td><td>Natural language processing</td></tr>
<tr><td></td><td>KDSDA11</td><td>Data preparation and pre-processing</td></tr>
<tr><td></td><td>KDSDM04</td><td>Data Architecture, data types and data formats, data modeling and design, including related technologies (ETL, OLAP, OLTP, etc.)</td></tr>
<tr><td></td><td>KDSDM05</td><td>Data lifecycle and organisational workflow, data provenance and linked data</td></tr>
<tr><td></td><td>KDSDM06</td><td>Data curation and data quality, data integration and interoperability</td></tr>
<tr><td></td><td>KDSDM08</td><td>Metadata, PID, data registries, data factories, standards and com0liance</td></tr>
<tr><td></td><td>KDSDM09</td><td>Open Data, Open Science, research data archives/repositories, Open Access, ORCID</td></tr>
<tr><td></td><td>Skill ID</td><td>Skills definition</td></tr>
<tr><td>Skills Data Analytics methods and algorithms</td><td>SDSDA08</td><td>Apply analytics and statistics methods for data preparation and pre-processing</td></tr>
<tr><td></td><td>SDSENG03</td><td>Use cloud based Big Data technologies for large datasets processing systems and applications</td></tr>
<tr><td></td><td>SDSENG07</td><td>Use relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for structured and unstructured data</td></tr>
<tr><td></td><td>SDSDM06</td><td>Consistently implement data curation and data quality controls, ensure data integration and interoperability</td></tr>
<tr><td>Skills Data Analytics languages, tools and platforms</td><td>DSALANG01</td><td>Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)</td></tr>
<tr><td></td><td>DSADB01</td><td>SQL and relational databases (open source: PostgreSQL, mySQL, Nettezza, etc.)</td></tr>
<tr><td></td><td>DSADB03</td><td>NoSQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, etc.)</td></tr>
<tr><td></td><td>DSADM02</td><td>Data Warehouse platform and related tools</td></tr>
<tr><td></td><td>DSADM05</td><td>Big Data and cloud based storage platforms and services</td></tr>
<tr><td></td><td>DSABDA02</td><td>Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)</td></tr>
</table>

-

<table>
<tr><td>Dimension 1 Competence Group</td><td>DSDA</td><td colspan="3">Data Science Analytics</td></tr>
<tr><td>Dimension 2 Competence</td><td>DSDA04</td><td colspan="3">Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval</td></tr>
<tr><td></td><td></td><td colspan="3"></td></tr>
<tr><td rowspan="2">Dimension 3 Proficiency level</td><td>Level 1 (Entry/Associate)</td><td colspan="2">Level 2 (Professional)</td><td>Level 3 (Expert)</td></tr>
<tr><td>Be familiar and be able to use different performance and accuracy metrics as part of used data analytics platforms</td><td colspan="2">Select appropriate performance metrics and apply them for specific analytics applications. Develop new metrics and use it for fine tuning the used analytics solutions.</td><td>Not specifically defined. Advanced knowledge and experience.</td></tr>
<tr><td>Dimension 4</td><td>Knowledge ID</td><td colspan="3">Knowledge topics definition</td></tr>
<tr><td>Knowledge</td><td>KDSDA01
KDSDA02
KDSDA06
KDSDA11
KDSDA14</td><td colspan="3">Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others
Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)
Predictive Analytics
Performance and accuracy metrics
Optimisation</td></tr>
<tr><td></td><td>Skill ID</td><td colspan="3">Skills definition</td></tr>
<tr><td>Skills Data Analytics methods and algorithms</td><td>SDSDA01
SDSDA04
SDSDA09</td><td colspan="3">Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)
Apply Predictive Analytics methods
Be able to use performance and accuracy metrics for data analytics assessment and validation</td></tr>
<tr><td>Skills Data Analytics languages, tools and platforms</td><td>DSALANG01
DSALANG02
DSABDA02
DSABDA09</td><td colspan="3">Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, pytorch, scikit-learn, seaborn, etc.)
R and data analytics libraries (cran, ggplot2, dplyr, reshap2, etc.)
Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
Kaggle competition, resources and community platform</td></tr>
</table>

•

Dimension 1 Competence Group	DSDA	Data Science Analytics
Dimension 2 Competence	DSDA05	Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making

Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 2 (Professional)	Level 3 (Expert)
	Develop analytics solutions for specific tasks and pre-defined data sets. Ensure correct interaction with other components of the application.	Develop organisational analytics applications that support the whole organisational data lifecycle. Integrate and deploy all components. Integrate analytics application with the enterprise information system.	Plan, design, develop, implement analytics for organizational tasks. Develop the whole data processing workflow and integrate it with organisational workflow. Use research methods principles in developing data driven applications and implementing the whole cycle of data handling

Dimension 4	Knowledge ID	Knowledge topics definition
Knowledge	KDSDA01	Machine Learning (supervised): Decision trees, Naïve Bayes classification, Ordinary least square regression, Logistic regression, Neural Networks, SVM (Support Vector Machine), Ensemble methods, others
	KDSDA02	Machine Learning (unsupervised): clustering algorithms, Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Independent Components Analysis (ICA)
	KDSDA04	Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)
	KDSDA06	Predictive Analytics
	KDSDA07	Prescriptive Analytics
	KDSENG04	Applications software requirements and design, agile development technologies, DevOps and continuous improvement cycle
	KDSENG07	Relational, non-relational databases (SQL and NoSQL), Data Warehouse solutions, ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets
	Skill ID	Skills definition
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)
	SDSDA04	Apply Predictive Analytics methods
	SDSDA05	Apply Prescriptive Analytics methods
	SDSENG01	Use systems and software engineering principles to organisations information system design and development, including requirements design
	SDSENG02	Use Cloud Computing technologies and cloud powered services design for data infrastructure and data handling services
	SDSRM01	Use research methods principles in developing data driven applications and implementing the whole cycle of data handling
Skills Data Analytics languages, tools and platforms	DSALANG01	Python and data analytics libraries
	DSALANG02	R and data analytics libraries
	DSADB01	SQL and relational databases (open source: PostgreSQL, mySQL, etc.)
	DSADB03	Data Warehouses, Master Data Management, Complex Events Processing
	DSAVIZ02	Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
	DSADEV01	Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others
	DSADEV03	Git versioning system as a general platform for software development

•

Dimension 1 Competence Group	DSDA Data Science Analytics		
Dimension 2 Competence	DSDA06	Visualise results of data analysis, design dashboard and use storytelling method	
Dimension 3 Proficiency level	Level 1 (Entry/Associate)	Level 2 (Professional)	Level 3 (Expert)
	Use visualisation techniques and tools for existing data set and applications. Develop simple dashboards	Use multiple visualisation techniques, languages for existing and new analytics applications and processes. Develop new visualisation solutions and advanced dashboards.	Define best visualisation approach and solutions for specific business bases. Use multiple techniques to create interactive dashboards.
Dimension 4	Knowledge ID	Knowledge topics definition	
Knowledge	KDSDA04	Data Mining (Text mining, Anomaly detection, regression, time series, classification, feature selection, association, clustering)	
	KDSDA06	Predictive Analytics	
	KDSDA07	Prescriptive Analytics	
	Skill ID	Skills definition	
Skills Data Analytics methods and algorithms	SDSDA01	Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)	
	SDSDA02	Use Data Mining techniques	
	SDSDA10	Use effective visualiation and storytelling methods to create dashboards and data analytics reports	
Skills Data Analytics languages, tools and platforms	DSALANG01	R and data analytics libraries for data visualisation	
	DSALANG02	Python and data analytics libraries (pandas, numpy, mathplotlib, scipy, scikit-learn, seaborn, etc.)	
	DSAVIZ01	Data visualization Libraries (mathpoltlib, seaborn, D3.js, FusionCharts, Chart.js, other)	
	DSAVIZ02	Visualisation software (D3.js, Processing, Tableau, Raphael, Gephi, etc.)	
	DSAVIZ03	Online visualization tools (Datawrapper, Google Visualisation API, Google Charts, Flare, etc)	
	DSABDA02	Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)	
	DSABDA05	Azure Data Analytics platforms (HDInsight, APS and PDW, etc)	

6 Alignment with other competence frameworks and suggested extensions

6.1 Mapping Data Science competences to e-Competence Framework (e-CF)

6.1.1 Proposed e-CF3.0 extension with the Data Science related competences

The proposed new competence groups provide a basis for defining new competences related to Data Science that can be added to the existing e-CF3.0. In particular, this report suggests the following additional e-competences related to Data Scientist functions as listed in Table 3.4 (assigned numbers are the continuation of the current e-CF3.0 numbering). When defining an individual professional profile or role the presented competences can be combined with those generic listed in original e-CF3.0 because normally Data Scientist need to have basic or advanced knowledge and skills in the general ICT domain.

Table 6.1. Proposed e-CF3.0 extension with the Data Science related Competences[17]

Competence group	Competences related to Data Science	Corresponding CF-DS competence groups
A. PLAN (and Design)	A.10* Organisational workflow/processes model definition/formalization A.11* Data models and data structures	DSDA DSENG
B. BUILD (Develop and Deploy/ Implement)	B.7* Apply data analytics methods (to organizational processes/data) B.8* Data analytics application development B.9* Data management applications and tools B.10* Data Science infrastructure deployment (including computing, storage and network facilities)	DSDA DSENG DSDM
C. RUN (Operate)	C.5* User/Usage data/statistics analysis C.6* Service delivery/quality data monitoring	DSDM DSENG
D. ENABLE (Use/Utilise)	D10. Information and Knowledge Management (powered by Data Science Analytics) - *refactored* D.13* Data analysis, insight or actionable information extraction, visualisation D.14* Support business processes/roles with data analytics, visualisation and reporting (support to D.5, D.6, D.7, D.12) D.15* Data management, curation, preservation, provenance	DSDA DSDK/DSBA
E. MANAGE	E.10* Support Management and Business Improvement with data and insight (data driven organisational processes management) (support to E.5, E.6) E.11* Data analytics for (business) Risk Analysis/Management (support to E.3) E.12* ICT and Information security monitoring and analysis (support to E.8)	DSDA DSENG DSDM

Analysis of the demanded Data Scientist functions and responsibilities in relation to typical organisational workflow revealed that Data Scientist roles and functions can be treated as rather cross-organisational and crossing-multiple competence area (as defined by e-CF3.0); they are rather linked to research or business process management lifecycle than to organisational structure.

[17] The current version 3 of the e-CF (2018) has limited Data and Data Science focused content and will benefit from the perspective of the Data Science competence framework. However, the broader scope and granularity of the e-CF will determine the transferability of these suggestions in part or entirety.

6.2 Mapping Data Science competences to CRISP-DM model

Although initially proposed in 1990s, CRISP-DM (Cross Industry Standard Process for Data Mining) [28] model is still used in defining Data Mining and Data Analytics workflows and processes. It is also used for defining common Data Mining and Data Analytics processes and stages, however not limited to data analytics or data management. Figure 7 illustrates CRISP-DM stages. It is important to mention that modern agile technologies and agile business technologies engage the main data handling and data analytics processes into continuous development and continuous improvement cycle.

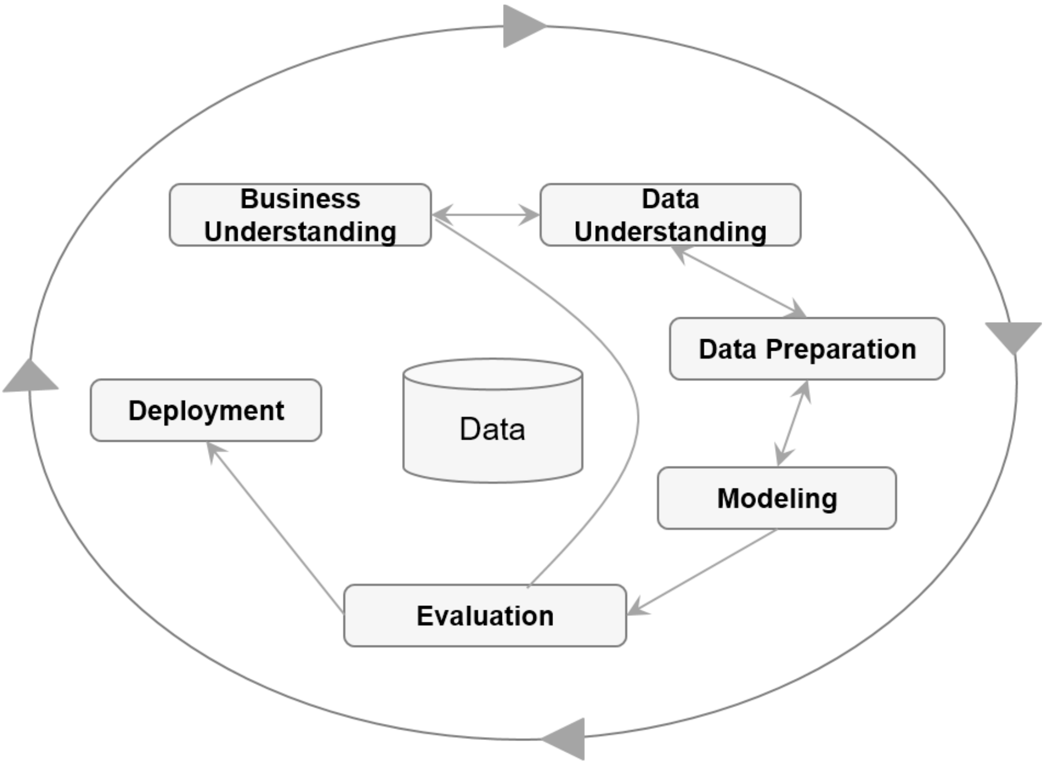


Figure 7. CRISP-DM (Cross Industry Standard Process for Data Mining)

Table 6.2. provides an example of initial mapping CF-DS competences to the CRISP-DM processes and stages that will ned to undergo cross-checking with the corresponding knowledge subjects in DS-BoK.

Table 6.2. Mapping CF-DS competences to the CRISP-DM processes and stages

CRISP-DM Processes and Stages	Description	Mapping to CF-DS
Business Understanding	General Business understanding, role of data and required actionable information	DSBAxx DSRMPxx
Determine Business Objectives	Business Objectives (SMART approach). Specific, Measurable, Attainable (in principle), Relevant and Timely. This is performed by Business Stakeholders!	DSBA01
	Business Success Criteria (or benchmark or threshold values)	
Assess Situation	Inventory of Resources, Requirements, Assumptions and Constraints	DSBA01
	Risks and Contingencies	
	Costs and Benefits	
Determine Data Mining Goals	Data Mining Goals	DSRMP05, DSRMP06

	Data Mining Success Criteria	
Produce Project Plan	Project Plan	DSRMP05, DSRMP06
	Initial Assessment of Tools and Techniques	
Data Acquisition and Understanding	Collect data, assign metadata, explore data, run ETL processes	DSDAxx DSDMxx
Collect Initial Data	Acquire access to data from internal and external sources (API, webscrapping). In a steady state, data extraction and transfer routines would be in place.	DSDA03
Describe Data	Describe data, add metadata	DSDA03 DSDM04
Explore Data	Checking on definitions and meaning of data acquired. This requires Business Knowledge (Business Analyst/ or business stakeholder)	DSDA03 DSBA01
	Examine the ´surface´ properties of the acquired data	
	Understand distribution of Key attributes, perform initial visualisation, understand initial relationships between a small number of attributes, perform simple aggregations	
Verify Data Quality	Checking if data is up-to-date	DSDA03
	Checking if data is complete, correct, error-free	
Data Preparation (select and cleanse)	Data preprocessing, cleaning, reduction, sampling	DSDAxx
Select Data	Decide on data to be used for analysis	DSDA03
Clean Data	Increase data quality by substitution, imputation (estimating of missing data) / insertion of suitable defaults. Identification of outliers, anomalies and patterns	DSDA03, DSDM05
Construct Data	Transform data set, produce derived values, produce new (composed) records	DSDA03
Integrate Data	Merging of tables (joins) or aggregations of data	DSDA03, DSDM03
Format data	Syntactic modifications (not changing meaning but producing format required by modeling tool (e.g. Convert dataset to JSON)	DSDA03, DSDM03
Hypothesis and Modelling		DSDAxx
Select Modelling Techniques	Decide on techniques to be used, depending on the type of problem (ML, Decision Tree, Neural Nets, etc.)	DSDA01
Generate Test Design	Generate a procedure or mechanism to test model quality and validity. Separate data set in train and validation sets and test sets	DSDA01
Build Model	Run the modeling tool on the prepared dataset to create one or more models. Perform parameter selection (e.g. hyper param)	DSDA01, DSDA02
Assess Model & Revise Parameters	Judge the success of the application of modeling and discovery technically: contact business analysts and domain experts in order to discuss the model.	DSDA04

	Summarize qualities of generated models (i.e. Accuracy, etc).	
	Revise parameter settings and tune them for the next run in the Build Model task	
Evaluate Results	Summarize assessment results in terms of business success criteria. This involves Business Stakeholders. This is not to evaluate the model's accuracy/ generalization: this is already done in the previous step	DSDA04
Review Process and determine improvement	Formally assess the data analytics process.	DSDA04, DSRM01
Deployment, Operations & Maintenance	Deploy application or process, maintain, prepare	DSRM06
Plan deployment	Determine a strategy for deployment, determine how the information will be propagated to users, decide how the use of results will be monitored and benefits measured	DSRM06
	Plan potential re-coding (e.g. from python to Java for production environment)	
Plan Monitoring & Maintenance	Check for dynamic aspects, decide how accuracy will be monitored, determine the threshold below which result cannot be used anymore or should be updated/recalibrated. Monitor and measure the performance of model.	DSDA05, DSRMP06
	Plan for DEVOPS or Continuous delivery/ Agile development	DSENGxx
Produce Final Report	Produce the final report, create the dashboard for proper visualisation.	DSDA06, DSRM06

6.3 Process Groups in Data Management and their mapping to CF-DS competences

Data handling includes multiple stages and processes that can be defined as the data lifecycle that can be related to data management processes or to more general project management processes.

The following Process Groups can be identified based on existing Data Lifecycle Management models (as reviewed in Appendix C) and corresponding processes definitions in Data Management BoK (DMBOK) [29] and Project Management BoK (PMBOK) [30]:

1. **Data Identification and Creation**: how to obtain digital information from in-silico experiments and instrumentations, how to collect and store in digital form, any techniques, models, standards and tools needed to perform these activities, depending on the specific discipline.
2. **Data Access and Retrieval**: tools, techniques and standards used to access any type of data from any type of media, and retrieve it in compliance with IPRs and established legislations.
3. **Data Curation and Preservation:** includes activities related to data cleansing, normalisation, validation and storage.
4. **Data Fusion (or Data integration)**: the integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and useful representation.
5. **Data Organisation and Management**: how to organise the storage of data for various purposes required by each discipline, tools, techniques, standards and best practices (including IPRs management and compliance to laws and regulations, and metadata definition and completion) to set up ICT solutions in order to achieve the required Services Level Agreement for data conservation.
6. **Data Storage and Stewardship**: how to enhance the use of data by using metadata and other techniques to establish long term access and extended use to that data also by scientists and researchers from other disciplines and after very long time from the data production time.
7. **Data Processing**: tools, techniques and standards to analyse different and heterogeneous data coming from various sources, different scientific domains and of a variety of sizes (up to Exabytes) – it includes notion of programming paradigms.
8. **Data Visualisation and Communication**: techniques, models and best practices to merge and join various data sets, techniques and tools for data analytics and visualisation, depending on the data significance and the discipline.

The majority of the Data Management processes can be mapped to the DSDM competence group, but Data Processing and Data Visualisations will also require DSDA competences, while development, deployment and operation corresponding tools may require DSENG competences.

Note, the defined Data Management processes are linked to but don’t substitute the research or business processes management lifecycle, which are focused on the delivery of value to scientific research or business.

7 Practical uses of CF-DS

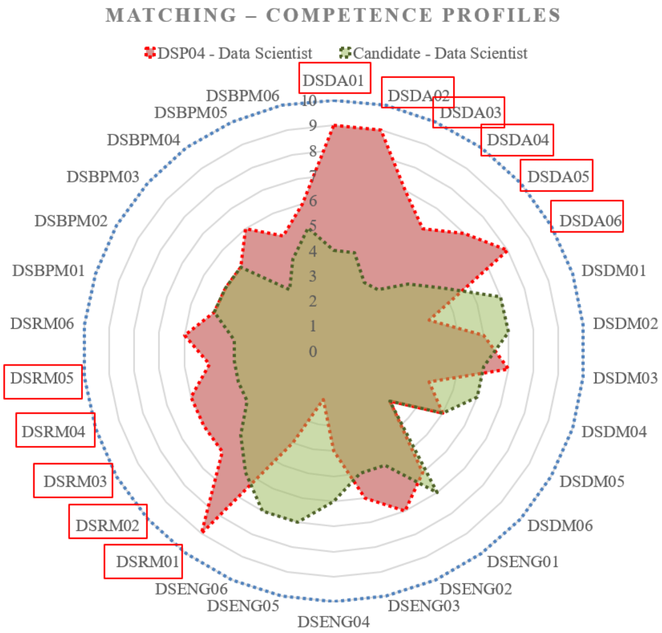
The presented CF-DS provides the basis for the definition of all other EDSF components: Data Science Body of Knowledge, Model Curriculum and Data Science Professional Profiles. Competences are used to define required knowledge and learning outcomes are applied to the curriculum design. Data Science professional Profiles are defined based on the set of competences required for each professional profile or group of profiles.

Other practical uses include but are not limited to:

- Assessment of individual and team competences, as well as balanced Data Science team composition
- Developing tailored curriculum for academic education or professional training, in particular, to bridge the skills gap and staff up/re-skilling
- Professional certification and self-training.

7.1 Usage example: Competences assessment [31]

Figure 7.1 illustrates an example of the individual competences assessment that may be used for one of the general use cases: the Data Science practitioner competences assessment against the target/desirable competence profile or role; or competences matching between the job vacancy and the candidate's competence profile.



Individual Education/Training Path based on Competence benchmarking

- Red polygon indicates the chosen professional profile: Data Scientist (general)
- Green polygon indicates the candidate or practitioner competences/skills profile
- Insufficient competences (gaps) are highlighted in *red*
  - *DSDA01 – DSDA06 Data Science Analytics*
  - *DSRM01 – DSRM05 Data Science Research Methods*
- Can be use for team skills match marking and organisational skills management

Figure 7.1. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

The intended professional profile or job vacancy are defined in the radial coordinates based on CF-DS competences required for the profiles or vacancy. The candidate's profiles can be defined based on a self-assessment or using a simple test. The illustrated competences mismatch can be used either for deciding on the suitability of the candidate or suggesting the necessary training program.

Using the enumerated set of competences, skills and knowledge units can be used for different applications dealing with competences assessment, knowledge assessment, job vacancy design and candidate assessment.

7.2 Data Stewardship Professional Competence Framework (CF-DSP) definition

D7.3 [6]
EDUCON paper [32]

Book hwo to FAIR [33]

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.

7.2.1 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 7.1). It is also important to mention that the growing adoption of the Data Steward profession as an important organisational role and the 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 7.1. 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 a data management strategy for data collection, storage, preservation, and availability for further processing, • 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) • Ensure compliance with standards and best practices in Data Governance and Data Management
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	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. • Ensure metadata compliance with FAIR requirements • Be familiar with the metadata management tools

DSDM03 Integrate heterogeneous data from multiple sources and provide them for further analysis and use	DSDM03 – extended, essential Integrate heterogeneous data from multiple sources and provide them for further analysis and use • Perform data preparation and cleaning • Match/transfer data models of individual datasets
DSDM04 Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)	DSDM04 – extended, highly essential Maintain historical information on data handling, including reference to published data and corresponding data sources • 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
DSDM05 Ensure data quality, accessibility, interoperability, compliance to standards, and publication (data curation)	DSDM05 – extended, essential Develop policy and metrics for data quality management, maintain data quality and compliance to standards, perform data curation • Interact/Collaborate with data providers and data owners to ensure data quality
DSDM06 Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management	DSDM06 – extended, essential Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management, address legal issues if necessary. • Ensure GDPR compliance in data management and access • Develop data access policies and coordinate their implementation and monitoring, including security breaches handling
None	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.
None	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.
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.

7.2.2 Data Engineering competence group (DSENG)

Table 7.2 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 7.2. 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 • Develop new tools and applications, ensure support of the data FAIRness requirements by existing and new tools and applications
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)	DSENG04– extended, essential Develop, deploy and operate data infrastructure, including data storage and processing facilities, using different distributed and cloud based platforms. • Implement requirements for data storage facilities to comply with the data management policies and FAIR data principles in particular.
DSENG05 Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	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. • Define and implement (coordinate) data access policies for different stakeholders and organisational roles

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	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 • Define, implement and maintain data model, reference data, master data definitions, implement consistent metadata

7.2.3 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 7.3. 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)
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	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 • Base research on collected scientific facts and collected data
DSRMP01 Create new understandings by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods	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
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 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 sound hypothesis ● 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 • 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

7.2.4 Domain related competence (DSDK/DSBA)

Domain-related knowledge and competences are important for Data Stewards as one of their roles are 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 7.4. 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 • 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 • 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 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 business case as a result of organisational data analysis and identified trends • Ensure data availability and quality for BA/BI needs
DSBA04 Analyse opportunity and suggest the use of historical data available at organisation for organizational processes optimization	DSBA04 – extended, relevant for organisation processes and data Analyse opportunity and suggest the use of historical data available at organisation for organizational processes optimization • 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 the 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 the 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

8 Conclusion and further developments

The presented Data Science Competence Framework Release 4 summarises the framework development since the published Release 3 in December 2018. The initial CF-DS versions have been created based on extensive analysis of available information that includes Data Science job market study (primarily demand side, i.e., job advertisement), existing standards, best practices, academic publications and blog articles that are posted by experts, practitioners and enthusiasts of the new technology domain and profession of Data Scientist.

The focused work on defining all the foundational components of the whole EDISON Data Science Framework for consistent Data Science profession definition have been done during the EDISON project duration (2015-2017) with wide consultation and engagement of different stakeholders, primarily from the research community and Research Infrastructures, but also involving industry via standardisation bodies, professional communities and directly via the project network. The work was also regularly reviewed by the EDISON Liaison Groups (ELG) and practically implemented by the champion universities as described in the final project deliverables [34].

The proposed EDSF Release 4 documents have been updated and extended based on review and contributions from individual experts, discussions at a number of workshops and conferences where the EDSF related developments and use cases have been presented, it also used feedback from practitioners that assessed usability and practically used EDSF for curriculum design or review, as well as by organisations used EDSF for defining their skills management and training needs. The new release incorporated results from the two projects: FAIRsFAIR that defined the Data Stewardship Professional Competence Framework; and MATES that developed the training programme on the digital and data skills for maritime industry.

8.1 Summary of the recent developments

This document presents ongoing results of the Data Science Competence Framework definition based on the analysis of existing frameworks for Data Science and ICT competences and skills, and supported by the analysis of the demand side for the Data Scientist profession in industry and research. For consistency, the recent developments are listed for EDSF Release 2 (2017) and current Release 3 (2018).

The following developments were included in CF-DS Release 2:

- Definition of skills and knowledge topics related to the competences groups significantly extended.
- Added the Data Science workplace skills definition that includes Data Science professional skills ("Thinking and acting like Data Scientist") and general "soft" or attitude skills often referred to as 21st Century skills.
- The document provides an example of the individual competences mapping to identified skills and knowledge for the Data Science and Analytics competence group.
- The identified competences, skills and knowledge subjects are provided as enumerated lists to provide the basis for future applications development and API definitions to ensure applications interoperability.
- The report suggests possible extensions to e-CF3.0 on the Data Science related competences.

The following developments were included in CF-DS Release 3:

- Reviewed, clarified and extended the definition of individual competences, skills and knowledge topics
- Mapping between CF-DS competence proficiency levels, e-CF proficiency levels and EQF qualification levels
- Reference to P21's Framework for 21st Century Learning and DigComp 2.1 Digital competences for citizens added, corresponding competences incorporated into the Data Science workplace skills definition and used to motivate necessary data literacy aspects.

The following developments were included in CF-DS Release 4:

- Reviewed, clarified and extended the definition of individual competences, skills and knowledge topics, in particular those that revised and extended based on contributions from the FAIRsFAIR and MATES projects and those related to new technologies, platforms and tools.
- Competences, skills and knowledge topics extended with the aspects related to FAIR data principles and Open Science in general.
- Provided reference to and a short summary of the Data Stewardship Professional Competence Framework as extension to CF-DS, what is an outcome of the FAIRsFAIR project.

- Extended section on workplace or transversal skills: added references to the existing initiatives for 21st Century Learning and recently published DigComp 2.2 Digital competences for citizens, added section on the design and system thinking.
- Book published that documents the EDSF Release 3 extended with use cases and implementation examples of using EDSF and CF-DS in particular: 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) [50]

8.2 Further developments to formalize CF-DS

The CF-DS Release 4 presents an already mature framework for the Data Science competences definition that has been validated via numerous practical uses during the EDISON project duration, in many reviews and implementations by different projects worldwide.

Further development will depend on the support and contribution by future projects and initiatives that will have interest and resources to use the EDSF for their purposes and occasionally contribute to the EDSF develop. In particular, the following CF-DS related developments are seen as beneficial for the community:

- Define the EDSF ontology for all EDSF components: CF-DS, DS-BoK, MC-DS, DSPP, what should improve its use and interoperability with existing ontology based frameworks, such as ESCO.e-CF4.0, and ACM CCS2013.
- Provide extended guidelines for intended CF-DS and EDSF usage in general: job profile and vacancy description generation; individual competences and skills assessment, certification profiles, and others.
- Specify EDSF API that would facilitate applications development and ensure applications compatibility.

To ensure successful acceptance of the proposed EDSF and its core components, an essential role belongs to the standardisation in the related technology and educational domains. This work has been done in the EDISON project. Necessary contacts with European and international standardisation bodies and professional organisations have been established and need to be maintained.

The EDSF maintenance and ongoing development is supported and coordinated by the EDISON Community Initiative via the EDISON Community github project that contains the EDSF working documents as Open Source under CC BY 4.0 License https://github.com/EDISONcommunity/EDSF/

9 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 guidelines, EDSF Part 5 [online] https://github.com/EDISONcommunity/EDSF/tree/master/data-science-edsf-use-cases-guidelines

[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#.Y6uctnbMK38

[7] The 2012 ACM Computing Classification System [online] http://www.acm.org/about/class/class/2012

[8] European Skills, Competences, Qualifications and Occupations (ESCO) [online] https://ec.europa.eu/esco/portal/home

[9] NIST SP 1500-1 NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions, Version 2, 2018 [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

[10] European e-Competences Framework [online] https://itprofessionalism.org/about-it-professionalism/competences/the-e-competence-framework/

[11] CEN EN 16234-1:2019 e-Competence Framework (e-CF) - A common European Framework for ICT Professionals in all sectors - Part 1: Framework [online] https://www.en-standard.eu/csn-en-16234-1-e-competence-framework-e-cf-a-common-european-framework-for-ict-professionals-in-all-sectors-part-1-framework/

[12] User guide for the application of the European e-Competence Framework 3.0. CWA 16234:2014 Part 2. [online] https://itprofessionalism.org/app/uploads/2019/11/User-guide-for-the-application-of-the-e-CF-3.0_CEN_CWA_16234-2_2014.pdf

[13] European Qualifications Framework (EQF) [online] https://ec.europa.eu/ploteus/content/descriptors-page

[14] European ICT Professional Profiles CWA 16458 (2012) (Updated by e-CF3.0) [online] https://www.cencenelec.eu/media/CEN-CENELEC/AreasOfWork/CEN%20sectors/Digital%20Society/CWA%20Download%20Area/ICT_SkillsWS/16458-1.pdf

[15] Information Technology Competency Model of Core Learning Outcomes and Assessment for Associate-Degree Curriculum (2014) http://www.capspace.org/uploads/ACMITCompetencyModel14October2014.pdf

[16] The U.S. Department of Labor IT Competency Model is available at www.careeronestop.org/COMPETENCYMODEL/pyramid.aspx?IT=Y

[17] Bloom's taxonomy: the 21st century version. [online] http://www.educatorstechnology.com/2011/09/blooms-taxonomy-21stcentury-version.html

[18] Harris, Murphy, Vaisman, Analysing the Analysers. O'Reilly Strata Survey, 2013 [online] http://cdn.oreillystatic.com/oreilly/radarreport/0636920029014/Analyzing_the_Analyzers.pdf

[19] Skills and Human Resources for e-Infrastructures within Horizon 2020, The Report on the Consultation Workshop, May 2012. [online] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/report_human_skills.pdf

[20] 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

[21] Burning Glass Technology, IBM, and BHEF report "The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market" (April 2017) http://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market

[22] DARE Project Recommended Data Science and Analytics Skills, Working report, APEC, 2017 [online] https://www.apec.org/Press/Features/2017/0620_DSA

[23] Barend Mons, et al, The FAIR Guiding Principles for scientific data management and stewardship [online] https://www.nature.com/articles/sdata201618

[24] P21 Partnership for 21st Century Learning [online] https://www.battelleforkids.org/networks/p21/frameworks-resources

[25] P21's Framework for 21st Century Learning [online] http://www.p21.org/storage/documents/P21_framework_0515.pdf Going Digital in a Multilateral World, Meeting of the OECD Council at Ministerial Level, Paris, 30-31 May 2018, OECD Report, 2018 [online] https://www.oecd.org/going-digital/C-MIN-2018-6-EN.pdf

[26] DigComp 2.2: The Digital Competence Framework for Citizens, Joint Research Center, 2022 [online] https://eliant.eu/fileadmin/user_upload/pdf/DigComp_2.2_-_The_Digital_Competence_Framework_for_Citizens.pdf

[27] EntreComp: The Entrepreneurship Competence Framework, by Margherita Bacigalupo, Panagiotis Kampylis, Yves Punie, Godelieve Van den Brande, JRC Science for Policy Report, 2016, EUR 27939 EN [online] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/entrecomp-entrepreneurship-competence-framework

[28] Cross Industry Standard Process for Data Mining (CRISP-DM) Reference Model [online] http://crisp-dm.eu/reference-model/

[29] 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

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

[31] 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 Kigndom

[32] 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

[33] 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

[34] Deliverable D3.3 Final Report on the Use Cases support design, EDISON Project Deliverable [online] http://edison-project.eu/sites/edison-project.eu/files/filefield_paths/ EDISON_D3.3_Usecases_v1.0-final.v1.pdf

[35] Auckland, M. (2012). Re-skilling for research. London: RLUK. [online] http://www.rluk.ac.uk/files/RLUK%20Re-skilling.pdf

[36] Big Data Analytics: Assessment of demand for Labour and Skills 2013-2020. Tech Partnership publication, SAS UK & Ireland, November 2014 [online] https://www.e-skills.com/Documents/Research/General/BigData_report_Nov14.pdf

[37] Italian Web Association (IWA) WSP-G3-024. Date Scientist [online] http://www.iwa.it/attivita/definizione-profili-professionali-per-il-web/wsp-g3-024-data-scientist/

[38] LERU Roadmap for Research Data, LERU Research Data Working Group, December 2013 [online] http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

[39] ELIXIR community projects RITrain and CORBEL dealing with competences and skills definition for bioinformaticians as an example of Data Science enabled professions

[40] Data Life Cycle Models and Concepts, CEOS Version 1.2. Doc. Ref.: CEOS.WGISS.DSIG, 19 April 2012

[41] NIST SP 1500-6 NIST Big Data interoperability Framework (NBDIF): Volume 6: Reference Architetcure, September 2015 [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6.pdf

[42] European Union. A Study on Authentication and Authorisation Platforms For Scientific Resources in Europe. Brussels : European Commission, 2012. Final Report. Contributing author. Internal identification SMART-Nr 2011/0056. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/aaa-study-final-report.pdf

[43] Demchenko, Yuri, Peter Membrey, Paola Grosso, Cees de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA. ISBN: 978-1-4673-6402-7; IEEE Catalog Number: CFP1316A-CDR.

[44] E. Bright Wilson Jr., An Introduction to Scientific Research, Dover Publications; Rev Sub edition, January 1, 1991

[45] Scientific Methods, Wikipedia [online] https://en.wikipedia.org/wiki/Scientific_method

[46] Research Methodology [online] https://explorable.com/research-methodology

[47] The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4 [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[48] Business process management, Wikipedia [online] https://en.wikipedia.org/wiki/Business_process_management

[49] Theodore Panagacos, The Ultimate Guide to Business Process Management: Everything you need to know and how to apply it to your organization Paperback, CreateSpace Independent Publishing Platform (September 25, 2012)

[50] 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)

10 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 Universities Association
FAIR	Findable, Accessible, Interoperable, Reusable data principles in Research Data Management
FAIRsFAIR	EU funded H2020 project (EOSC cluster)
FDO	FAIR Digital Object
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
MATES	ERASMUS+ project on Maritime Industry Skills Alliance
MC-DS	Data Science Model Curriculum
NIST	National Institute of Standards and Technologies of USA
P21C	21st Century Skills (also 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
RDM	Research Data Management
SLICES	European Research Infrastructure for Experimental Research
SWEBOK	Software Engineering Body of Knowledge

11 Appendix A. Data used in the study of demanded Data Science competences and skills

The presented study and the proposed Data Science competences and skills definition is based on data collected from job advertisements on such popular job search and employment portals as IEEE Jobs portal and LinkedIn Jobs advertised that provided rich information for defining Data Science competences, skills and required knowledge of Big Data tools and data analytics software. The IEEE Jobs portal posts job advertisements predominantly from US companies and universities. LinkedIn posts vacancies related to the region or country from where the request originated and many job ads are posted in national language. In particular case of this study, the job advertisements were collected for positions available in Netherlands that appeared to be quite extensive and representing the whole spectrum of required competences and skills.

The initial study used a set of Data Science job openings from IEEE Jobs portal (around 120) and LinkedIn Netherlands (around 140) collected in the period of mid-September to the beginning of October 2015, taking into account that the Netherlands is one of leading countries in relation to Big Data and Data Science technologies acceptance and development. A number of Data Science related key words were used, like Data Science, Big Data, Data Intensive technologies, data analytics, machine learning. Initial analysis of collected information allowed us to make the assumption that collected information from more than 250 samples was sufficiently representative for the initial study.

The following are general characteristics of the collected data.

- Total number of advertisements collected: IEEE Jobs – 120; Linkedin Jobs – 140
- Number of advertisements selected for analysis IEEE Jobs – 28; Linkedin Jobs – 30
- Number of companies posted Data Science related jobs – more than 50
- The most active recruiting companies: Booking.com, Scandia, etc.

The collected and working data are available in the EDSF github project repository. The data can be used for training purposes and new applications development.

A.1. Selecting sources of information

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 Scientist specialists and are 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 Science related competences and skills produced by different studies as mentioned above, in particular NIST definition of Data Science that provided a basis for definition of initial 3 groups of skills, namely Data Analytics, Data Science Engineering, 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.

It appeared that the richest information can be collected from job advertisements on popular job search and employment portals such as IEEE Jobs portal and LinkedIn Jobs advertised. Important to admit that although IEEE Jobs designed to post international job openings, the advertisements are mostly from US companies and universities. LinkedIn posts vacancies related to the region or country from where the request is originated and many job ads are posted in the national language. In particular case of this report, it was possible to collect information from LinkedIn for Netherlands, however it was quite representative due to a large number of advertisements. This means that at the following stage, the information needs to be collected by EDISON partners in their own countries. The same relates to collecting information from different scientific, technology and industry domains that should take place at the next stage of this study.

- If referred to the category of job openings such as academic positions or industry and business related positions, the academic positions didn't provide valuable information as they don't specify detailed competences and skills but rather search for candidates who are capable to teach, create or support new academic courses on Data Science.

- In this initial stage, we used a set of Data Science job openings from IEEE Jobs portal (around 120) and LinkedIn Netherlands (around 140) collected in the period of mid-September to the beginning of October 2015. A number of Data Science related key words were used like Data Science, Big Data, Data Intensive technologies, data analytics, machine learning. Initial analysis of collected information allowed us to make the assumption that collected information from more than 250 samples was sufficiently representative for the initial study, taking into account that the Netherlands is one of the leading countries in relation to Big Data and Data Science technologies acceptance and development. See Appendix B for more details about the collected data.

A.2. EDISON approach to analysis of collected information

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 NIST Data Science definition (i.e. Data Analytics, Domain Knowledge, and Engineering)
5) Identify competences and skills groups that don’t 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 CF-DS via community surveys and individual interviews[18].

The Data Science competences and skills defined in this way will be used to provide input to existing professional competence frameworks and profiles:

- Map to e-CF3.0 if possible, suggest new competences
- Map to CWA ICT profiles where possible suggest new profiles if needed
- Identify inconsistencies in using current e-CF3.0 and CWA ICT profiles and explore alternative frameworks if necessary.

The outlined above process has been applied to the collected information and all steps are tracked in the two Excel workbooks provided as supplementary materials to this report that are available on the project shared storage and later to be available via the project wiki

A.3. Regular Job Market analysis

The EDSF maintenance team (i.e. currently coordinated by UvA) makes regular, at least two times a year, job market analysis to review the relevance of currently defined competences, identify new demands and necessary adjustments, also update and extend demanded skills on Data Science and Analytics platforms, tools and technologies.

[18] This activity has being done at a regular basis (approximately on yearly basis) by the EDSF maintenance team, also in occasions of new competence domains investigated such as Data Stewardship, Artificial Intelligence, or general data and digital skills that have been done in the recent projects.

Appendix B. Overview: Studies, reports and publications related to Data Science competences and skills

B.1. O'Reilly Strata Survey (2013) [18]

O'Reilly Strata industry research [15] defines the four Data Scientist profession profiles and their mapping to the basic set of technology domains and competencies as shown in Figure A.1. The four profiles are defined based on the Data Science practitioners self-identification:

- Data Businessperson
- Data Creative
- Data Developer
- Data Researcher

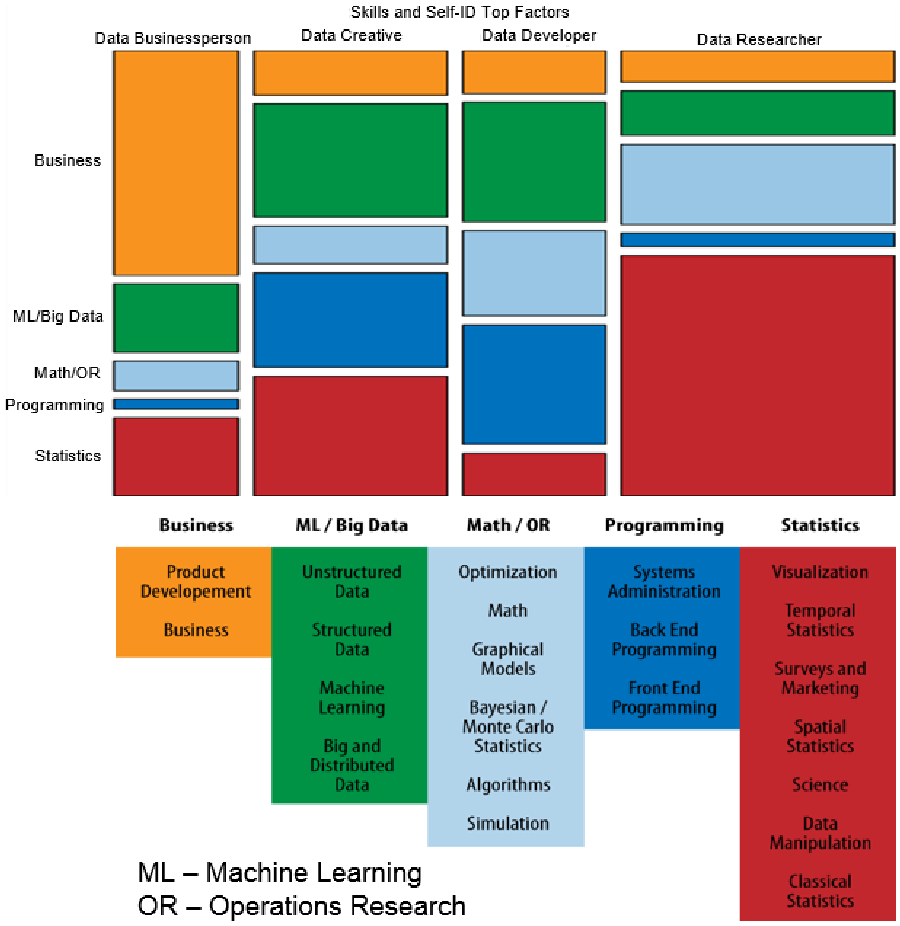


Figure A.1. Data Scientist skills and profiles according to O'Reilly Strata survey [18]

Table A.1 below lists skills for Data Science that are identified in the study. They are very specific in a technical sense but provide useful information when mapped to the mentioned above Data Science profiles. We will refer to this study in our analysis of CF-DS and related competence groups.

Table A.1. Data Scientist skills identified in the O'Reilly Strata study (2013)

Data Science Skills	Examples -> Knowledge and skills
Algorithms	computational complexity, CS theory
Back-End Programming	JAVA/Rails/Objective C

Bayesian/Monte-Carlo Statistics	MCMC, BUGS
Big and Distributed Data	Hadoop, Map/Reduce
Business	management, business development, budgeting
Classical Statistics	general linear model, ANOVA
Data Manipulation	regexes, R, SAS, web scraping
Front-End Programming	JavaScript, HTML, CSS
Graphical Models	social networks, Bayes networks
Machine Learning	decision trees, neural nets, SVM, clustering
Math	linear algebra, real analysis, calculus
Optimization	linear, integer, convex, global
Product Development	design, project management
Science	experimental design, technical writing/publishing
Simulation	discrete, agent-based, continuous)
Spatial Statistics	geographic covariates, GIS
Structured Data	SQL, JSON, XML
Surveys and Marketing	multinomial modeling
Systems Administration	*nix, DBA, cloud tech.
Temporal Statistics	forecasting, time-series analysis
Unstructured Data	NoSQL, text mining
Visualization	statistical graphics, mapping, web-based data visualisation

B.2. Skills and Human Resources for e-Infrastructures within Horizon 2020 [19]

The Report on the Consultation Workshop (May 2012) “Skills and Human Resources for e-Infrastructures within Horizon 2020” [17] summarises the outcomes of a consultation workshop that was organised by DG INFSO "GÉANT and e-Infrastructures" unit to consult the stakeholders on their views of approaching these challenges. The workshop discussions highlighted cross-cutting challenges of

i) new and changed skills needs which combine technical and scientific skills and require interdisciplinary thinking and communication;

ii) recognizing new job profiles and tasks rising from the emergence of computing intensive and data-driven science with integral role of e-infrastructures;

iii) need for effective European level collaboration and coordination to avoid duplication of efforts and join the forces for developing high quality human capital for e-infrastructures

Several concrete recommendations for supporting the suggested development aspects with e-Infrastructures activities under Horizon 2020 were devised. It was considered important to have both specific and integrated activities to support skills and human resources aspects within the e-Infrastructures projects.

The report defined three perspectives for e-infrastructure related skills needs:

- Development
  - Create new tools, further develop e-Infrastructure
  - Technological innovations
- Operation
  - Support users, maintain and operate services
  - Process/service innovations
- Scientific use
  - Use ICT tools, apply e-science methods
  - Scientific innovations

The nine areas identified as having potentially the most significant skills gap according to RLUK report “Re-skilling for Research” (2012) [35]

- Ability to advise on preserving research outputs (49% essential in 2-5 years; 10% now)
- Knowledge to advise on data management and curation, including ingest, discovery, access, dissemination, preservation, and portability (48% essential in 2X-5 years; 16% now)
- Knowledge to support researchers in complying with the various mandates of funders, including open access requirements (40% essential in 2-5 years; 16% now)
- Knowledge to advise on potential data manipulation tools used in the discipline/subject (34% essential in 2-5 years; 7% now)
- Knowledge to advise on data mining (33% essential in 2-5 years; 3% now)
- Knowledge to advocate, and advise on, the use of metadata (29% essential in 2-5 years; 10% now)
- Ability to advise on the preservation of project records e.g. correspondence (24% essential in 2-5 years; 3% now)
- Knowledge of sources of research funding to assist researchers to identify potential funders (21% essential in 2-5 years; 8% now)
- Skills to develop metadata schema, and advise on discipline/subject standards and practices, for individual research projects (16% essential in 2-5 years; 2% now)

B.3. UK Study on demand for Big Data Analytics Skills (2014) [36]

The study “Big Data Analytics: Assessment of demand for Labour and Skills 2013-2020” [19] provided extensive analysis of the demand side for Big Data specialists in UK in forthcoming year. Although majority of roles are identified as related to Big Data skills, it is obvious that all these roles can be related to a more general definition of the Data Scientist as an organisational role working with Big Data and Data Intensive Technologies.

The report lists the following Big Data roles:

- Big Data Developer
- Big Data Architect
- Big Data Analyst
- Big Data Administrator
- Big Data Consultant
- Big Data Project Manager
- Big Data Designer
- Data Scientist

B.4. IWA Data Science profile [37]

Italian Web Association (IWA) published the WSP-G3-024. Date Scientist Profile for web related projects [20]. It provides a good example of domain specific definition of the Data Science competences, skills and organisational responsibilities, it also suggests mapping to e-CF3.0 competences.

The Data Scientist is defined as “Professional that owns the collection, analysis, processing, interpretation, dissemination and display of quantitative data or quantifiable organization for analytical, predictive or strategic.”

The profile contains the following sections:

- Concise definition
- Mission
- Documentation produced
- Main tasks
- Mapping to e-CF competences
- Skills and knowledge
- Application area of KPI
- Qualifications and certifications (informational)
- Personal attitudes (informational)

- Reports and reporting lines (informational)

For reference purposes, it is worth mentioning that IWA Data Scientist profile maps its competences and skills to the following e-CF3.0 competences:

A.6. Application design: Level e-3
A.7. Monitoring of technological Bertrand: Level e-4
B.1. Development of applications: Level e-2
B.3. Testing: Level e-3
B.5. Production of documentation: Level e-3
C.1. User assistance: Level e-3
C.3. Service Delivery: Level e-3
C.4. Management Problem: Levels e-3, e-4.

B.5. Other studies, reports and projects on defining Data Science competences, skills and profiles

The following reports and studies and ongoing works to define the Data Science skills profiles and needs for European Research Area and industry are considered relevant to the current study and will be used to finalise the DS-CF definition:

- LERU Roadmap for Research Data (2013) [38]
- ELIXIR community projects RITrain and CORBEL dealing with competences and skills definition for bioinformaticians as an example of Data Science enabled professions [39]
- PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017) [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) [21]

12 Appendix C. Mapping between CF-DS and e-CF

This section provides example of mapping between CF-DS and e-CF3.0 which is relevant to the new e-CF4.0 published as CEN standard CEN EN 16234-1 in 2019.

C.1. e-CF3.0 competences relevant to CF-DS: Essential and additional

Initial results suggested refactoring of some e-CF3.0 competence areas and individual competences and adding new competences as shown in Table C.1. The competences are ordered by relevance and essential competences are highlighted in bold.

Table C.1. e-CF3.0 competences relevant to CF-DS: Essential and additional

A. PLAN and DESIGN	**Essential**
	A.2. Service Level Management
	A.3. Product / Service Planning
	A.5. Application Design
	A.4. Architecture Design
	Additional
	A.6. Sustainable Development
	A.7. Innovating and Technology Trend Monitoring
	A.8. Business/Research Plan Development and Grant application
	A.1. Research Infrastructure (RI) and Research Strategy Alignment
B. BUILD: DEVELOP and DEPLOY/IMPLEMENT	**Essential**
	B.1. Application Development (including Requirements Engineering, Function Specification, Application Programming Interfaces, Human Computer Interaction)
	B.2. Component Integration
	B.3. Testing (RI services and Scientific Applications)
	B.4. Solution/Application Deployment
	Additional
	B.5. Documentation Production
	B.6. Systems Engineering (DevOps)
C. OPERATE (RUN)	**Essential**
	C.1. User Support
	C.2. Service Delivery
	C.3. Problem Management
	Additional
	C.4. Change Support (Upgrade/Migration)
D. USE: UTILISE (ENABLE)	**Essential**
	D.1. Scientific Applications Integration (on running RI)
	D.5. Data collection and preservation
	D.4. New requirements and change Identification
	D.6. Education and Training Provision
	Additional
	D.2. Information Security Strategy Development
	D.3. RI/ICT Quality Strategy Development
	D.7. Purchasing/Procurement
	D.8. Contract Management
	D.9. Personnel Development
	D.10. Dissemination and outreach
E. MANAGE	**Essential**
	E.1. Overall RI management (by systems and components)
	E.5. Information/Data Security Management

	Additional
	E.6. Data Management (including planning and lifecycle management, curation)
	E.4. RI Security and Risk/Dependability Management
	E.2. Project and Portfolio Management
	E.3. ICT Quality Management and Compliance
	E.7. RI/IS Governance

The selected subset of e-CF3.0 can match general requirements to ICT competences required to operate and manage computer facilities of Research Infrastructures, but it doesn't reflect sufficiently specific competences required for Research Infrastructure or research data management, which have been identified in this study. The rationale for introducing specific data management competences and skills into the profile of the Research Infrastructure administrators and technicians will be further investigated in the IG-ETRD where the project members are actively involved.

C.2. Proposed e-CF3.0 extension with the Data Science related competences

The proposed new competence groups provide a basis for defining new competences related to Data Science that can be added to the existing e-CF3.0. In particular, this report suggests the following additional e-competences related to Data Scientist functions as listed in Table 3.4 (assigned numbers are continuation of the current e-CF3.0 numbering). When defining an individual professional profile or role the presented competences can be combined with those generic listed in original e-CF3.0 because normally Data Scientist need to have basic or advanced knowledge and skills in general ICT domain.

Table 4.4. Proposed e-CF3.0 extension with the Data Science related Competences

Competence group	Competences related to Data Science
A. PLAN (and Design)	A.10* Organisational workflow/processes model definition/formalization A.11* Data models and data structures
B. BUILD (Develop and Deploy/ Implement)	B.7* Apply data analytics methods (to organizational processes/data) B.8* Data analytics application development B.9* Data management applications and tools B.10* Data Science infrastructure deployment
C. RUN (Operate)	C.5* User/Usage data/statistics analysis C.6* Service delivery/quality data monitoring
D. ENABLE (Use/Utilise)	D10. Information and Knowledge Management (powered by DS) - refactored D.13* Data presentation/visualisation, actionable data extraction D.14* Support business processes/roles with data and insight (support to D.5, D.6, D.7, D.12) D.15* Data management/preservation/curation with data and insight
E. MANAGE	E.10* Support Management and Business Improvement with data and insight (support to E.5, E.6) E.11* Data analytics for (business) Risk Analysis/Management (support to E.3) E.12* ICT and Information security monitoring and analysis (support to E.8)

Analysis of the demanded Data Scientist functions and responsibilities in relation to typical organisational workflow revealed that Data Scientist roles and functions can be treated as rather cross-organisational and crossing-multiple competence areas (as defined by e-CF3.0); they are rather linked to research or business process management lifecycle than to organisational structure.

C.3. Data Science related Competences included in the EN 16234-1 (2018)

EN 16234-1 "e-Competence Framework" (almost referred to as e-CFv4.0) is a European standard developed by CEN TC 428 ICT Professionalism and Digital competences Working Group. New revision of EN 16234-1 includes the following new competences roles and skills related to Data Science and defined in correspondence with CF-DS:

D7. Data Science and Analysis

Uses and applies data analytics techniques such as data mining, machine learning, prescriptive and predictive analytics to apply data insight to address organisation's challenges and opportunities.

Data Specialist Role Endures the implementation of the organisations data management policy

Data Scientist Leads the process of applying data analytics. Delivers insights from data by optimising the analytics process and presenting visual data representations.

Skill K11 FAIR data management principles (Findability, accessibility, interoperability, reusability)

Skill K6 Data governance, data governance strategy, data management plan (DMP)

13 Appendix D. Mapping between e-CF proficiency levels and EQF qualification levels

e-CF3.0 defines 5 proficiency levels applied to individual competences, however individual competence may not necessarily have all proficiency levels but can require lower or upper set of proficiency demanding on the complexity of performed tasks. The following table provides a guide to how this might be expressed. For the complete table go to Annex 2 of the e-CF 3.0 overview brochure available at the IT Professionalism website [10]

Table C.1. Mapping between e-CF proficiency levels and EQF qualification levels.

EQF Levels	EQF Levels description	e-CF Levels	e-CF Levels Description	Typical Tasks
8	Knowledge at the most advanced frontier, the most advanced and specialised skills and techniques to solve critical problems in research and/or innovation, demonstrating substantial authority, innovation, autonomy, scholarly or professional integrity.	e-5	**Principal** Overall accountability and responsibility; recognised inside and outside the organisation for innovative solutions and for shaping the future using outstanding leading edge thinking and knowledge.	IS strategy or programme management
7	Highly specialised knowledge, some of which is at the forefront of knowledge in a field of work or study, as the basis for original thinking, critical awareness of knowledge issues in a field and at the interface between different fields, specialised problem-solving skills in research and/or innovation to develop new knowledge and procedures and to integrate knowledge from different fields, managing and transforming work or study contexts that are complex, unpredictable and require new strategic approaches, taking responsibility for contributing to professional knowledge and practice and/or for reviewing the strategic performance of teams.	e-4	**Lead Professional/Senior Manager** Extensive scope of responsibilities deploying specialised integration capability in complex environments; full responsibility for strategic development of staff working in unfamiliar and unpredictable situations.	IS strategy/ holistic solutions
6	Advanced knowledge of a field of work or study, involving a critical understanding of theories and principles, advanced skills, demonstrating mastery and innovation in solving complex and unpredictable problems in a specialised field of work or study, management of complex technical or professional activities or projects, taking responsibility for decision-making in unpredictable work or study contexts, for continuing personal and group professional development.	e-3	**Senior Professional/Manager** Respected for innovative methods and use of initiative in specific technical or business areas; providing leadership and taking responsibility for team performances and development in unpredictable environments**.**	Consulting
5	Comprehensive, specialised, factual and theoretical knowledge within a field of work or study and an awareness of the boundaries of that knowledge, expertise in a comprehensive range of cognitive and practical skills in developing creative solutions to abstract problems, management and supervision in contexts where there is unpredictable change, reviewing and developing performance of self and others.	e-2	**Professional** Operates with capability and independence in specified boundaries and may supervise others in this environment; conceptual and abstract model building using creative thinking; uses theoretical knowledge and practical skills to solve complex problems within a predictable and sometimes unpredictable context.	Concepts / Basic principles
4	Factual and theoretical knowledge in broad contexts within a field of work or study,			

	expertise in a range of cognitive and practical skills in generating solutions to specific problems in a field of work or study, self-management within the guidelines of work or study contexts that are usually predictable, but are subject to change, supervising the routine work of others, taking some responsibility for the evaluation and improvement of work or study activities.			
3	Knowledge of facts, principles, processes and general concepts, in a field of work or study, a range of cognitive and practical skills in accomplishing tasks. Problem solving with basic methods, tools, materials and information, responsibility for completion of tasks in work or study, adapting own behaviour to circumstances in solving problems.	e-1	**Associate** Able to apply knowledge and skills to solve straight forward problems; responsible for own actions; operating in a stable environment.	Support / Service

Establishing a ‘level’ relationship between CF-DS competences and that of the e-CF will supports orientation and provides a general context for the positioning of job roles constructed from competence combinations. Although direct connections between the e-CF and EQF are not possible, as they represent different concepts, expressing a general relationship helps to clarify the complexity of tasks or activities incorporated within a competence.

The following table offers a potential connection between Data Science competence, e-CF competence and EQF qualification levels.

Table C.2. Suggested mapping between CF-DS, e-CF and EQF levels.

CF-DS Competence proficiency Levels	**e-CF proficiency levels**	**EQF qualification levels**
Associate/Entry	e-1	eqf-3, eqf-4
Professional	e-2, e-3	eqf-5, eqf-6
Expert/Lead	e-4, e-5	eqf-7, eqf-8

14 Appendix D. Concepts and models related to CF-DS definition

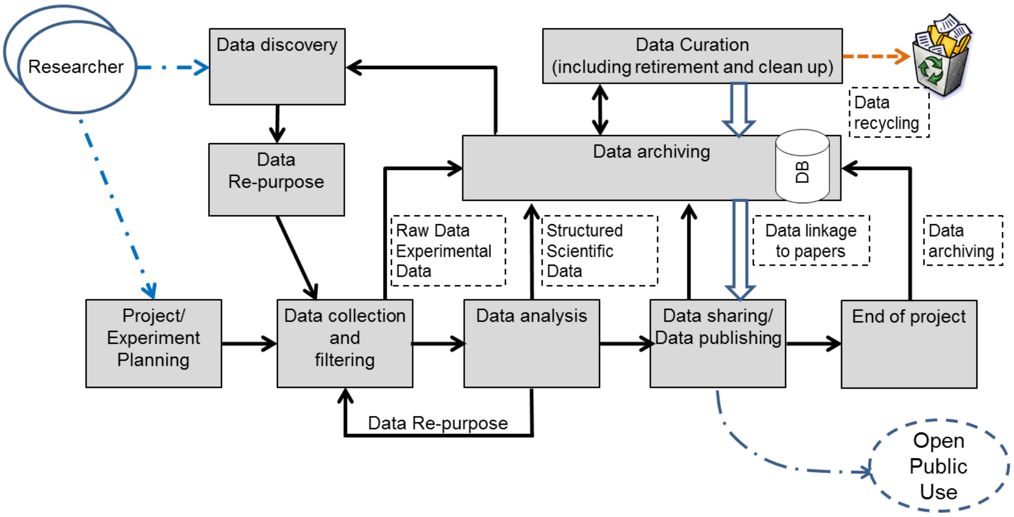
This section provides important definitions that are needed for consistent CF-DS definition in the context of organisational and business processes, e-Infrastructure and scientific research. First of all, this includes the definition of typical organisational processes and scientific workflow or research data lifecycle.

D.1. Scientific Data Lifecycle Management Model

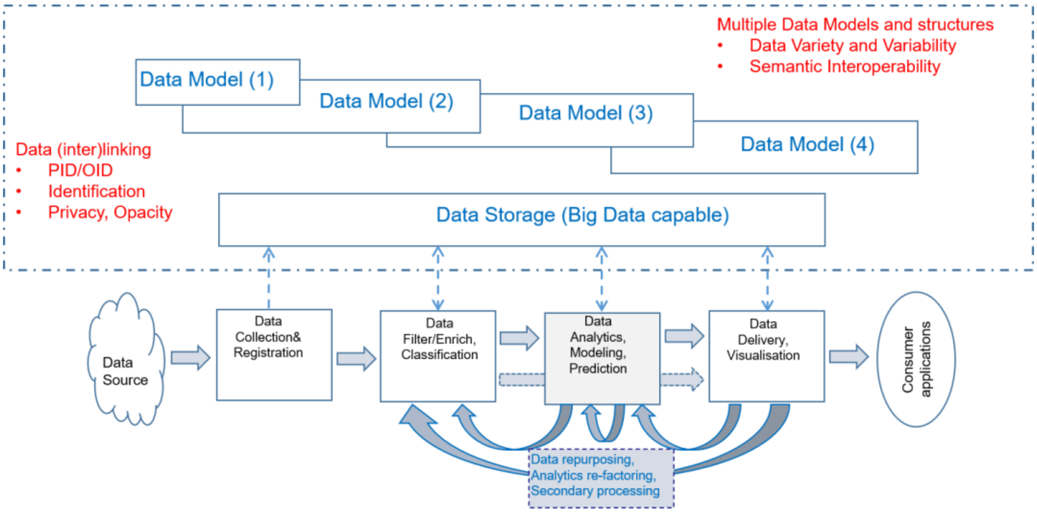
Data lifecycle is an important component of data centric applications, which Data Science and Big Data applications belong to. Data lifecycle analysis and definition is addressed in many domain specific projects and studies. Extensive compilation of the data life cycle models and concepts is provided in the CEOS.WGISS. DSIG document [39].

For the purpose of defining the major groups of competences required for Data Scientist working with scientific applications and data analysis we will use the Scientific Data Lifecycle Management (SDLM) model [30] shown in Figure 6 (a) defined as a result of analysis of the existing practices in different scientific communities. Figure D.1 (b) illustrates the more general Big Data Lifecycle Management model (BDLM) involving the main components of the Big Data Reference Architecture defined in NIST BDIF [7, 41, 42]. The proposed models are sufficiently generic and compliant with the data lifecycle study results presented in [29].

The generic scientific data lifecycle includes a number of consequent stages: research project or experiment planning; data collection; data processing; publishing research results; discussion, feedback; archiving (or discarding). SDLM reflects the complex and iterative process of scientific research that is also present in Data Science analytics applications.



(a) Scientific data lifecycle management - e-Science focused [42]


Multiple Data Models and structures
• Data Variety and Variability
• Semantic Interoperability
Data Model (1)
Data Model (2)
Data Model (3)
Data Model (4)
Data (inter)linking
• PID/OID
• Identification
• Privacy, Opacity
Data Storage (Big Data capable)
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer applications
Data repurposing, Analytics re-factoring, Secondary processing

(b) Big Data Lifecycle Management model (compatible with the NIST NBDIF definition) [43]

Figure D.1. Data Lifecycle Management in (a) e-Science and (b) generic Big Data Lifecycle Management model.

Both SDLM and BDLM require data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in Scientific Data Infrastructure (SDI). Data integrity, access control and accountability must be supported during the whole data lifecycle. Data curation is an important component of the discussed data lifecycle models and must also be done in a secure and trustworthy way. The research data management and handling issues are extensively addressed in the work of the Research Data Alliance[19].

D.2. Scientific methods and data driven research cycle

For a Data Scientist that is dealing with handling data obtained in the research investigation, understanding the scientific methods and the data driven research cycle is an essential part of knowledge that motivate necessary competences and skills for the Data Scientists for successfully perform their tasks and support or lead data driven research.

The scientific method is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge [44, 45, 46]. Traditional steps of scientific research were developed over time since the time of ancient Greek philosophers through modern theoretical and experimental research where experimental data or simulation results were used to validate the hypothesis formulated based on initial observation or domain knowledge study. The general research methods include: observational methods, opinion based methods, experimental and simulation methods.

The increased power of computational facilities and the advent of Big Data technologies created a new paradigm of the data driven research that enforced the ability of researchers to make an observation of the research phenomena based on bigger data sets and applying data analytics methods to discover hidden relations and processes not available to deterministic human thinking. The principles of the data driven research were formulated in the seminal work “The Fourth Paradigm: Data-Intensive Scientific Discovery” edited by Tony Hey [47].

The research process is iterative by its nature and allows scientific model improvement by using a continuous research cycle that typically includes the following basic stages:

- Define research questions
- Design experiment representing an initial model of research object or phenomena
- Collect Data
- Analyse Data
- Identify Patterns

[19] Research Data Alliance https://rd-alliance.org/

- Hypothesise Explanation
- Test Hypothesis
- Refine model and start new experiment cycle

The traditional research process may be concluded with the scientific publication and archiving of collected data. Data driven and data powered/driven research paradigm allows research data re-use and combining them with other linked data sets to reveal new relations between initially not linked processes and phenomena. As an example, biodiversity research when studying specific species populations can include additional data from weather and climate observation, solar activity, other species migration and technogenic factor.

The proposed CF-DS introduces research methods as an important component of the Data Science competences and knowledge and uses data lifecycle as an approach to defining the data management related competences group.

D.3. Business Process Management lifecycle

New generation Agile Data Driven Enterprises (ADDE) use Data Science methods to continuously monitor and improve their business processes and services. The data driven business management model allows combining different data sources to improve predictive business analytics what allows making more effective solutions, faster adaptation of services, and more specifically target different customer groups as well as doing optimal resources allocations depending on market demand and customer incentives.

Similarly, to the research domain the data driven methods and technologies change how the modern business operates attempting to benefit from the new insight that big data can give into business improvement including internal processes organisation and relation with customers and market processes. Understanding the Business Process Management lifecycle [48, 49] is important to identify necessary competences and knowledge for business oriented Data Science profiles.

The following are typical stages of the Business Process Management lifecycle:

- Define the business target: both services and customers
- Design the business process
- Model/Plan
- Deploy and Execute
- Monitor and Control
- Optimise and Re-design

The need for Business Process management competences and knowledge for business oriented Data Science profiles is reflected in the proposed CF-DS definition.