EDISON Data Science Framework:
Part 4. Data Science Professional Profiles (DSPP)
Release 2

Project acronym: EDISON
Project full title: Education for Data Intensive Science to Open New science frontiers
Grant agreement no.: 675419

Due Date	
Actual Date	03 July 2017
Document Author/s	Yuri Demchenko
Version	Release 2, v0.4
Dissemination level	PU
Status	Working document, request for comments
Document approved by	n/a

Document Version Control			
Version	Date	Change Made (and if appropriate reason for change)	Initials of Commentator(s) or Author(s)
0.0	11/07/2016	Initial internal draft	YD
0.1	09/09/2016	First draft (after Deliverable D2.2)	YD
0.2	03/10/2016	Updated after ELG discussion	YD
Release 1	10/10/2016	Release 1 after ELG03 meeting discussion	YD
0.3	30/01/2017	DSP Profiles definition is extend based on wide discussion with different communities	YD
0.4	03/07/2017	DSP profiles are updated after discussion with different communities; Data Steward is introduced as a distinct profile; new group of Data Science Enabled professional profiles suggested	YD

Document Editor
Yuri Demchenko

Document Contributors

Author Initials	Name of Author	Institution
YD	Yuri Demchenko	University of Amsterdam
AB	Adam Belloum	University of Amsterdam
TW	Tomasz Wiktorski	University of Stavanger

Executive summary

The EDISON project is designated to create a foundation for establishing a new profession of Data Scientist for European research and industry. The EDISON vision for building the Data Science profession will be enabled through the creation of a comprehensive framework for Data Science education and training that includes such components as Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), and Data Science Professional profiles definition.

The intended EDISON framework comprising of the mentioned above components will provide a guidance and a basis for universities to define their Data Science curricula and courses selection, on 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.

This document presents the results of the research and development in the EDISON project to define the Data Science Professional Profiles (DSPP) that is important for defining the Data Scientist roles in the organisation and their alignment with the organizational goals and mission. The Data Science Professional profiles definition is done in the context of the whole EDISON Data Science Framework.

- The proposed DSPP are defined as an extension to current ESCO (European Skills, Competences, Qualifications and Occupations) taxonomy and is intended to be proposed for formal inclusion of the new Data Science professions family into the future ESCO taxonomy edition.
- The proposed DSPP when adopted by the community will have multiple uses. First of all. they will help organisations to plan their staffing for data related functions when migrating to agile data driven organizational model. The Human Resource (HR) departments can effectively use DSPP for vacancy description construction and candidates assessment.
- The definition of the Data Science Professional profiles together with other EDSF components will provide a formal basis for Data Science professional certification, organizational and individual skills management and career transferability.

When used together with CF-DS, the DSPP can provide a basis for building interactive/web based tool for individual competences benchmarking against selected (or desirable) professional profiles as well as advising practitioners on the (up/re-) skilling path.

TABLE OF CONTENTS

1 Introduction

The revolutionary value of data in modern computer powered e-Science is recognized in earl works by technology visionaries. It is first described in the book by Tony Hey and others “The Fourth Paradigm” [5] and confirmed in the HLEG report “Riding the wave: How Europe can gain from the rising tide of scientific data” [6], that computational (and statistical) methods and data mining on large sets of scientific and experimental data will play a key role in discovering hidden and obscure relationships between processes and events that are necessary in order to make new scientific discoveries and support innovation in industry and the modern digital economy. Industry also recognises the benefits of Big Data technologies and the use of scientific methods in business/operational data analysis and in problem solving for managing enterprise operations, staying innovative and competitive, and being able to provide advanced customer-centric service delivery. Modern agile data driven companies are transforming their organizational to reflect the important role of data in optimizing business and operational processes. These changes have increased the demand for new types of specialists with strong technical background and deep knowledge of the data intensive technologies. This has been defined as a new profession of the Data Scientist.

This document presents the results of the research and development in the EDISON project to define the Data Science Professional profiles that important for defining the Data Scientist roles in the organisation and their alignment with organizational goals and mission. The Data Science Professional definition is done in the context of the whole EDISON Data Science Framework that includes such components as Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Model Curriculum (MC-DC).

The intended EDISON framework comprising of the mentioned above components will provide a guidance and a basis for universities to define their Data Science curricula and courses selection, on 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. Similar to e-CF3.0, the proposed CF-DS, will provide a basis for building interactive/web based tool for building custom Data Science profiles and (self-)evaluate candidate’s compliance with a created profile.

The document has the following structure. Section 2 describes the EDSF and its components. Section 3 provides an overview of existing profession profiles definition frameworks for ICT and Data Science competences and skills definition including e-CF3.0, CWA 16458 (2012) European ICT profiles, European Skills, European Competences, Qualifications and Occupations (ESCO) framework. Section 4 presents the definition of the proposed DSPP as an extension to ESCO taxonomy. It also provides example mapping different profiles to CF-DS competences what can be used for building curricula and training programs customised for specific professional profiles and roles. The document concludes with the suggested further development to finalise the DSPP definition.

2 EDISON Data Science Framework

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

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

- CF-DS – Data Science Competence Framework [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 basis for other components of the Data Science professional ecosystem 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

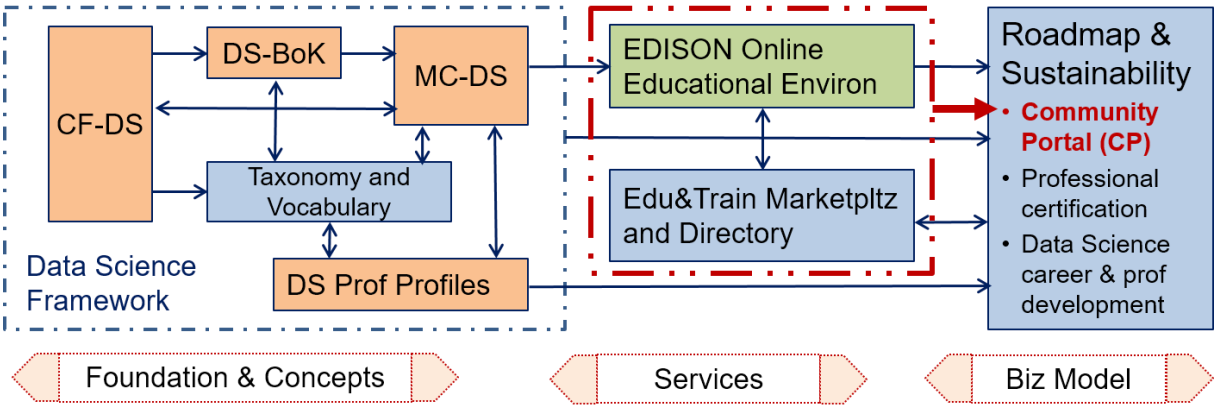


Figure 1 EDISON Data Science Framework components.

The CF-DS provides the overall basis for the whole framework, it first version has been published in November 2015 and was used as a foundation all following EDSF components developments. The CF-DS has been widely discussed at the numerous workshops, conferences and meetings, organised by the EDISON project and where the project partners contributed. The core CF-DS competences has been reviewed

The core CF-DS includes common competences required for successful work of Data Scientist in different work environments in industry and in research and through the whole career path. The future CF-DS development will include coverage of the domain specific competences and skills and will involve domain and subject matter experts.

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK follows the same approach to collect community feedback and contribution: Open Access CC-BY community discussion document is published on the project website. DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs defined based on the Classification Computer Science (CCS2012), components taken from other BoKs and proposed new KA 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 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 direct mapping to the enumerated competences in CF-DS. The preliminary version of MC-DS has been discussed at the first EDISON Champions Conference in June 2016 and collected feedback is incorporated in current version of MC-DS.

The DSPP are defined as an extension to European Skills, Competences, Qualifications and Occupations (ESCO) 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 be also used for building individual career path 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 subjects in CF-DS, knowledge groups, areas and units in DS-BoK, 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 ongoing EDSF dissemination and sustainability activity.

The EDISON Data Science professional ecosystem illustrated in Figure 1 uses core EDSF components to specify the potential services that can be offered for professional Data Science community and provide basis for the sustainable Data Science and related general data skills sustainability. In particular, CF-DS and DS-BoK can be used for individual competences and knowledge benchmarking and play instrumental role in constructing personalised learning paths and professional (up/re-) skilling programs based on MC-DS.

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 for defining professional profiles for general ICT occupations and currently defined data handling related professions. Appendix A provides additional overview of earlier works and publications that attempted to define required Data Science competences, skills and organisational roles.

3.1 CWA 16458 (2012): European ICT Professional Profiles

The European ICT Professional Profiles CWA 17458 (2012) was created to provide a basis for compatible ICT profiles definition by organisations and a basis for defining new profiles by European stakeholders [7].

The CWA defines 23 main ICT profiles the most widely used by organisations by defining organisational roles for ICT worker, that are grouped into the six ICT Profile families:

- Business Management
- Technical Management
- Design
- Development
- Service and Operation
- Support

The European ICT Profile descriptions are reduced to core components and constructed to clearly differentiate profiles from each other. Further context-specific elements can be added to the Profiles according to the specific environments in which the Profiles are to be integrated. Figure 2 illustrates six ICT profile families and related main profiles which are non-exhaustive.

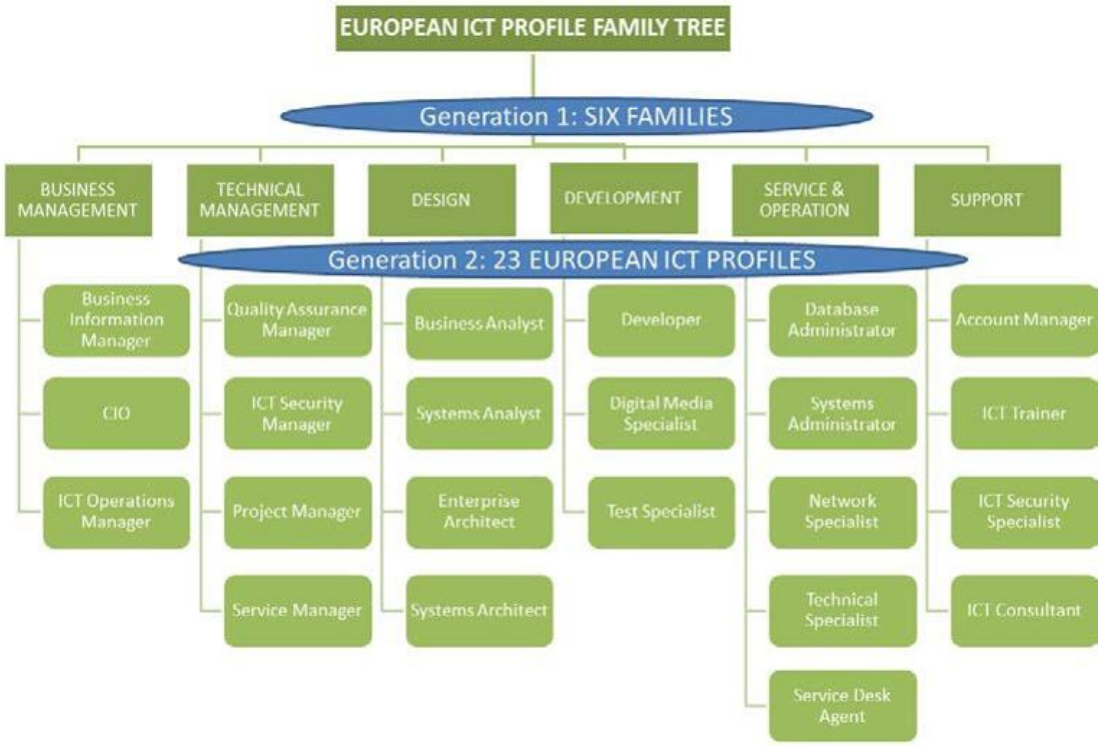


Figure 2. European ICT Profile Family Tree – Generation 1 and 2 as a shared European reference [7]

The 23 profiles constructed in CWA combined with e- competences from the e-CF3.0 [10], provide a pool for the development of tailored profiles that may be developed by European ICT sector players in specific contexts and with higher levels of granularity. The 23 Profiles cover the full ICT Business process; positioning them

into the e-CF Dimension 1 demonstrates this. Figure 4 below illustrates this together with the ICT Profiles family structure).

Figure 3 illustrates mapping between CWA families and e-CF3.0 competence areas and also CWA ICT profiles allocation to families and competence areas.

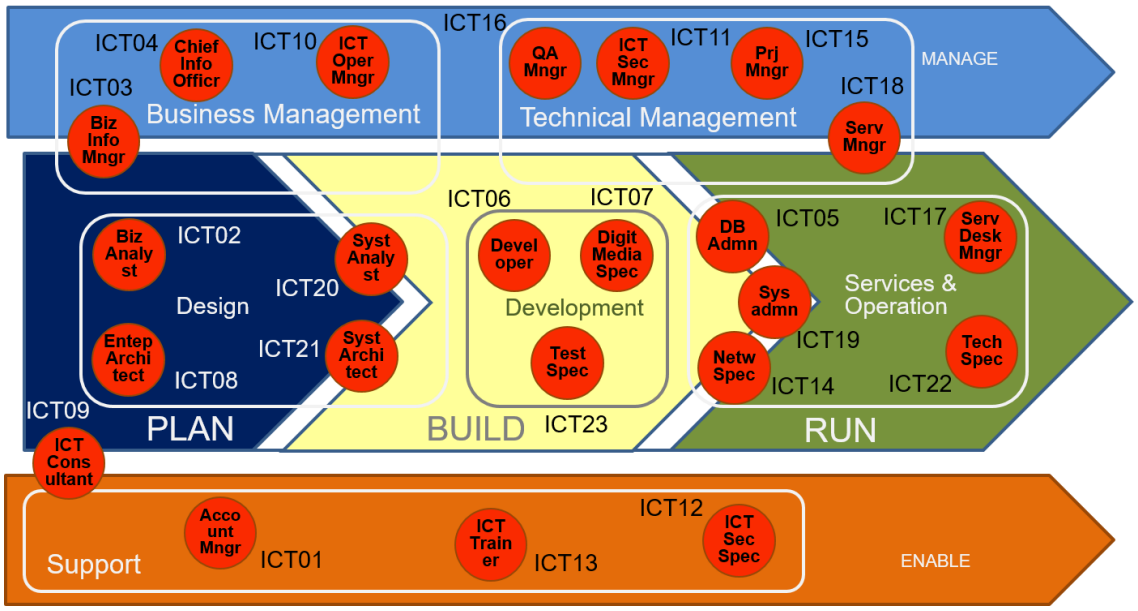


Figure 3. European ICT Professional Profiles structured by six families and positioned within the ICT Business Process (e-CF Dimension 1) (adopted from [8] and extended)

3.2 ESCO (European Skills, Competences, Qualifications and Occupations) framework and platform [9]

The Commission services launched the ESCO project in 2010 with an open stakeholder consultation. Currently, DG Employment, Social Affairs and Inclusion coordinates the development of ESCO with the support by the European Centre for the Development of Vocational Training Cedefop. Stakeholders are closely involved in the development and dissemination of ESCO.

The ESCO classification identifies and categorises skills, competences, qualifications and occupations relevant for the EU labour market and education and training. It systematically shows the relationships between the different concepts. ESCO has been developed in an open IT format, is available for use free of charge by everyone and can be accessed via the ESCO portal.

The first version of ESCO v0 was published on 23 October 2013. This version is based on the EURES classification but includes an enhanced semantic structure, cross-sector skills and competences and an initial small sample of qualifications. It includes the results of the Cross-Sector Reference Group, but not yet any sectoral updates.

ESCO v0 contains 4 761 occupations, around 5 000 skills and competences as well as some qualifications. As each concept in ESCO exists in all ESCO languages this amounts to more than 250 000 terms.

The qualifications pillar in ESCO v0 contains a small sample list of qualifications regulated at European level, international qualifications and certificates and licences linked to tasks, technologies, occupations or sectors. The list will be further developed in the next releases of ESCO. In addition, National Qualifications databases developed by the Member States and referenced to the European Qualifications Framework (EQF) [10] will, in the future, feed into the development of ESCO.

ESCO called for community review and contribution with the deadline of 31 December 2015[1]. Until end of 2016 the classification will be completely revised. The final product will be launched in 2017 as ESCO v1.

Table 1 contains data related occupations extracted from the ESCO classification together with related hierarchies. Table 2.3 is included for reference purposes to presents the ESCO top level occupations classification where data related occupations of different groups are highlighted in bold.

Table 1. Data related occupations in ESCO (2015) taxonomy

Occupations	Skills/Comp group	Hierarchy	Hierarchy	Top hierarchy
Security director (data processing/IT)	Database and network professionals not elsewhere classified	Database and network professionals	Information and communications technology professionals	Professionals
Security analyst (data processing/IT)				
Supervisor (data processing)				
Data processing investigator				
Data recorder	Database designers and administrators			
Operations manager (data processing)				
Data processing manager				
Data processing analyst				
Data processing supervisor				
Data processing consultant	Systems administrators			
Data processing strategist	Systems analysts	Software and applications developers and analysts		
Operations technician (data processing)	Information and communications technology operations technicians	Information and communications technology operations and user support technicians	Information and communications technicians	Technicians and associate professionals
Access supervisor, data processing/IT	Information and communications technology service managers	Production and specialised services managers		Managers

Table 2. ESCO top occupation hierarchy

ESCO Occupations top level hierarchy	
Armed forces occupations	
Clerical support workers *)	
	Numerical and material recording clerks
	Other clerical support workers
	Customer services clerks
	General and keyboard clerks
Craft and related trades workers	
Elementary occupations	
Managers	

[1] EDISON project provided a number of comments and suggestions to Data Science and education methods related terms and definitions.

	Administrative and commercial managers
	Chief executives, senior officials and legislators
	Hospitality, retail and other services managers
	Production and specialised services managers
Plant and machine operators and assemblers	
Professionals	
	Teaching professionals
	Science and engineering professionals
	Health professionals
	Legal, social and cultural professionals
	Business and administration professionals
	Information and communications technology professionals
Service and sales workers	
Skilled agricultural, forestry and fishery workers	
Technicians and associate professionals	
	Health associate professionals
	Information and communications technicians
	Legal, social, cultural and related associate professionals
	Science and engineering associate professionals
	Business and administration associate professionals

*) The highlighted bold font indicates which ESCO taxonomy groups are identified for proposed extension with the Data Science occupations.

4 Definitions of the Data Scientist

There is no well established definition of the Data Scientist due to a number of competences and skills expected from these specialists. The proposed Data Scientist definition is based on the definition provided in the NIST SP1500-1 document [11] and extended with the need to deliver value to the organisation or to the project:

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

The NIST document defines the following groups of skills required from the Data Scientists: domain experience, statistics and data mining, and engineering skills [11]. The EDSF has proposed structured definition of the Data Scientist via definition of the related competences, skills, knowledge and proficiency level.

Initial attempt to define the Data Scientist has been made by O'Reilly Strata Survey (2013) (see [13] and Appendix A) which recognised creativity as an important feature of Data Scientist.

Other definitions [12, 13] admit such desirable features as ability to solve variety of business problems, optimize performance and suggest new services for the organisation employing Data Scientist. Many practitioners admit a need for a successful Data Scientist to develop a special mindset, to be statistically minded, understand raw data and "appreciate data as a first class product" [14].

The qualified Data Scientist should be capable of working in different roles in different projects and organisations such as Data Engineer, Data Analyst or Data Architect, Data Steward, etc., and possess the necessary skills to effectively operate components of the complex data infrastructure and processing applications through all stages of the Data lifecycle till the delivery of expected scientific and business values to science and/or industry.

The Data Science Competence Framework defined the following main Data Science competence groups that must be possessed by Data Science practitioners to be able to work at different roles in the data driven organisations.

Core Data Science competences/skills groups defining profile of the Data Science related professional profiles

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Additional common competence groups demanded by organisations

- Data Management and Governance (including data stewardship, curation, and preservation)
- *Research Methods for research related professions and Business Process Management for business related professions*

Detailed definition of the CF-DS competences, skills and knowledge is provided in the Appendix B and in the CF-DS document [1].

5 Defining Data Science Professional profiles

This section presents initial results on defining the Data Science Professional profiles that can be also called Data relates occupations family. They are defined extension to the ESCO occupations taxonomy. The proposed new occupations are placed in four top classification groups: managers (for managerial roles); Professionals (for applications developers and for infrastructure engineers); technicians and associate professionals (for operators and technicians); and clerical support workers (for data curators and stewards).

5.1 Taxonomy of Data Science Occupations by extending ESCO Hierarchy

The presented here initial taxonomy of Data Science professional roles is based on the ESCO occupations classification and their competences and organisational roles are defined similar to CWA 16458 ICT profiles. Table 3 presents them in the context of the ESCO classification hierarchy, only Data Science related top level groups are presented (for overall top level ESCO occupations hierarchy refer to section 3.2 and Table 2).

The following suggestions were used when constructing the proposed taxonomy:

- Data Scientist occupations depending on organisational role can be placed in the following top level hierarchies:
    - Managers (for managerial roles);
    - Professionals (for analytics applications developers and for infrastructure and datacenter engineers);
    - Technicians and associate professionals (for operators and technicians)
- Correspondingly, new 3rd level occupation groups are proposed:
    - Data Science/Big Data Infrastructure Managers
    - Data Science Professionals
    - Data Science technology professionals
    - Data and information entry and access
- Group of occupations related to digital librarians, data archives management, data curations and support currently placed in the 3rd group "*Professionals > Information and communications technology professionals > Data Science technology professionals > Data handling professionals not elsewhere classified*", however potentially it can also put in a new 2nd level group "*Clerical support workers > Data handling support workers (alternative)*". Motivation for this is growing need for data support workers in all domain of human activities in the digital data driven economy.
- It is recognised that existing ESCO group "Database and network professionals" should extended with new occupations (or professions) related to Big Data and scientific data related profiles which examples are included in the table: Large scale (cloud) database administrator/operator and Scientific database administrator/operator, however further identification of such occupations need to be done.

Table 3. Data Science occupations extension to ESCO classification

Top level	Hierarchies existing and new		Occupations group (if any)	Occupations
Managers				
	Production and specialised services managers	Data Science/Big Data Infrastructure Managers		DSP01 Data Science (group) Manager
			Research Infrastructure Managers	DSP02 Data Science Infrastructure Manager
				DSP03 Research Infrastructure Manager
Professionals				
	Science and engineering professionals	Data Science Professionals	Data Science professionals not elsewhere classified	DSP04 Data Scientist
				DSP05 Data Science Researcher
				DSP06 Data Science Architect
				DSP07 Data Science (Application) Programmer/Engineer
				DSP08 (Big) Data Analyst
				DSP09 Business Analyst
	Information and communications technology professionals	Data Science technology professionals	Data handling professionals not elsewhere classified	DSP10 Data Steward
				DSP11 Digital Data Curator
				DSP12 Data Librarian
				DSP13 Data Archivist
	Science and engineering professionals	**Database and network professionals**	Large scale (cloud) data storage designers and administrators	DSP14 Large scale (cloud) database designer*)
			Database designers and administrators	DSP15 Large scale (cloud) database administrator*)
			Database and network professionals not elsewhere classified	DSP16 Scientific database administrator*)
Technicians and associate professionals				
	Science and engineering associate professionals	Data Science Technology Professionals	Data Infrastructure engineers and technicians	DSP17 Big Data facilities Operators
				DSP18 Large scale (cloud) data storage operators
			Database and network professionals not elsewhere classified	DSP19 Scientific database operator*)
Clerical support workers				
	General and keyboard clerks			
		Data handling and support workers	Data and information entry and access	DSP20 Data entry/access desk/terminal workers
				DSP21 Data entry field workers
				DSP22 User support data services

5.2 Definition of the Data Science Professional profiles

This section provides definition of the Data Science Professional profiles by defining their competences and organisational roles. The proposed definition can be instrumental in defining education and training profiles for students and for practitioners to acquire necessary competences and knowledge for specific professional profiles or occupations. It can be also used for defining certification profiles or career path building.

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

- Managers (for managerial roles);
- Professionals (for analytics applications developers and for infrastructure and datacenter engineers);
- Technicians and associate professionals (for operators and technicians)
- Optionally, some data management occupations can be also placed into the Clerical support workers group such as digital data archivist, digital librarians.

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

- Data Science/Big Data Infrastructure Managers
- Data Science Professionals
- Data Science technology professionals
- Data and information entry and access (this is a candidate group under Clerical support workers top level hierarchy)

It is proposed that the existing ESCO group "Database and network professionals" should be extended with new occupations (or professions) related to Big Data or cloud based databases: Large scale (cloud) database administrator/operator and Scientific database administrator/operator, however further identification of such occupations needs to be done.

A group of occupations related to digital librarians, data archives management, data stewardship and data curation are currently placed in the 3rd proposed group:

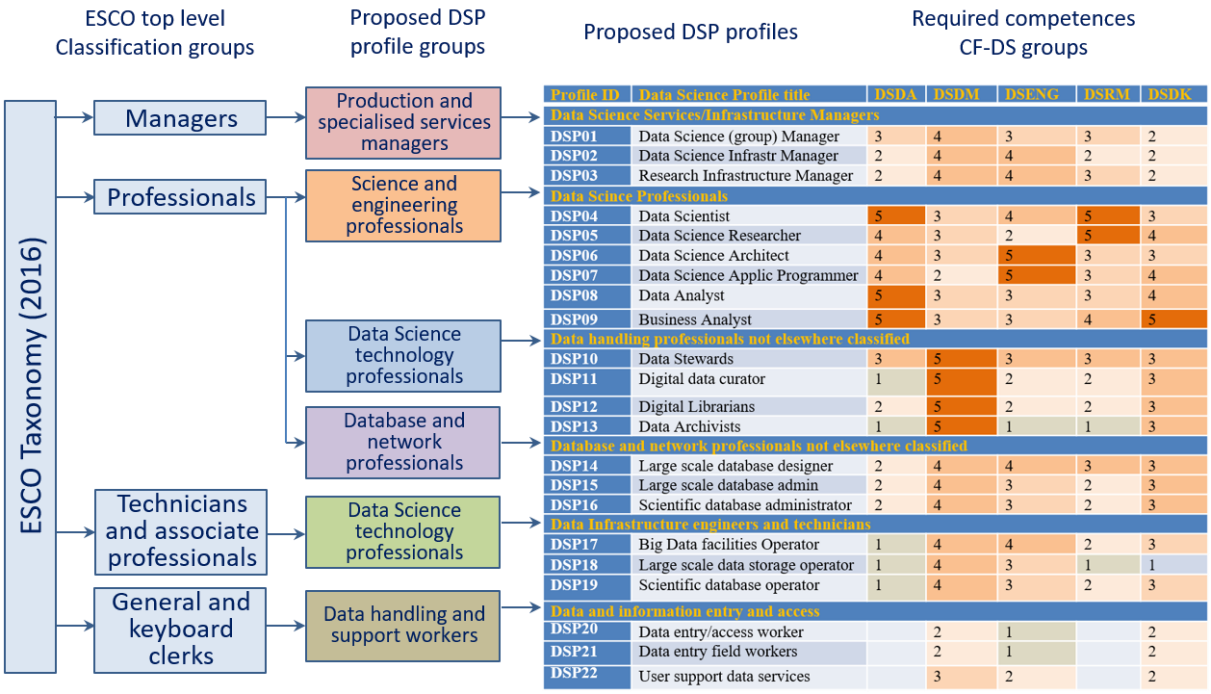
> *Professionals > Information and communications technology professionals > Data Science technology professionals > Data handling professionals not elsewhere classified,*

however potentially it can also be added in a new 2nd level group "*Clerical support workers > Data handling support workers (alternative)*". The motivation for this is a growing need for data support workers in all domains of human activities in the digital data driven economy.

To ensure a smooth Data Science professions acceptance by industry and employment bodies, the proposed profiles should be compatible with the relevant standards ESCO, CWA 16458 2012 ICT Profiles [7] , eCFv3.0 (future CEN standard EN 16324) [8].

Table 4 provides an ESCO compliant taxonomy and definition of the identified Data Science professional profiles collected from job advertisements, blogs and recent discussions at different forums, in particular, with the Research Data Alliance, and digital curation and data preservations communities.

Figure 3 graphically illustrates the existing ESCO hierarchy and the proposed new Data Science classification groups and corresponding new Data Science related profiles. The table in the figure also represent Table 4 in a compact way illustrating the CF-DS competences relevance to individual profiles. Figure 4 provides visual presentation of the identified DSPP and their grouping by the proposed high level classification groups.



Profile ID	Data Science Profile title	DSDA	DSDM	DSENG	DSRM	DSDK
Data Science Services/Infrastructure Managers						
DSP01	Data Science (group) Manager	3	4	3	3	2
DSP02	Data Science Infrastr Manager	2	4	4	2	2
DSP03	Research Infrastructure Manager	2	4	4	3	2
Data Scince Professionals						
DSP04	Data Scientist	5	3	4	5	3
DSP05	Data Science Researcher	4	3	2	5	4
DSP06	Data Science Architect	4	3	5	3	3
DSP07	Data Science Applic Programmer	4	2	5	3	4
DSP08	Data Analyst	5	3	3	3	4
DSP09	Business Analyst	5	3	3	4	5
Data handling professionals not elsewhere classified						
DSP10	Data Stewards	3	5	3	3	3
DSP11	Digital data curator	1	5	2	2	3
DSP12	Digital Librarians	2	5	2	2	3
DSP13	Data Archivists	1	5	1	1	3
Database and network professionals not elsewhere classified						
DSP14	Large scale database designer	2	4	4	3	3
DSP15	Large scale database admin	2	4	3	2	3
DSP16	Scientific database administrator	2	4	3	2	3
Data Infrastructure engineers and technicians						
DSP17	Big Data facilities Operator	1	4	4	2	3
DSP18	Large scale data storage operator	1	4	3	1	1
DSP19	Scientific database operator	1	4	3	2	3
Data and information entry and access						
DSP20	Data entry/access worker		2	1		2
DSP21	Data entry field workers		2	1		2
DSP22	User support data services		3	2		2

Figure 3. Proposed Data Science related extensions to the ESCO classification hierarchy and corresponding DSPP by classification groups.

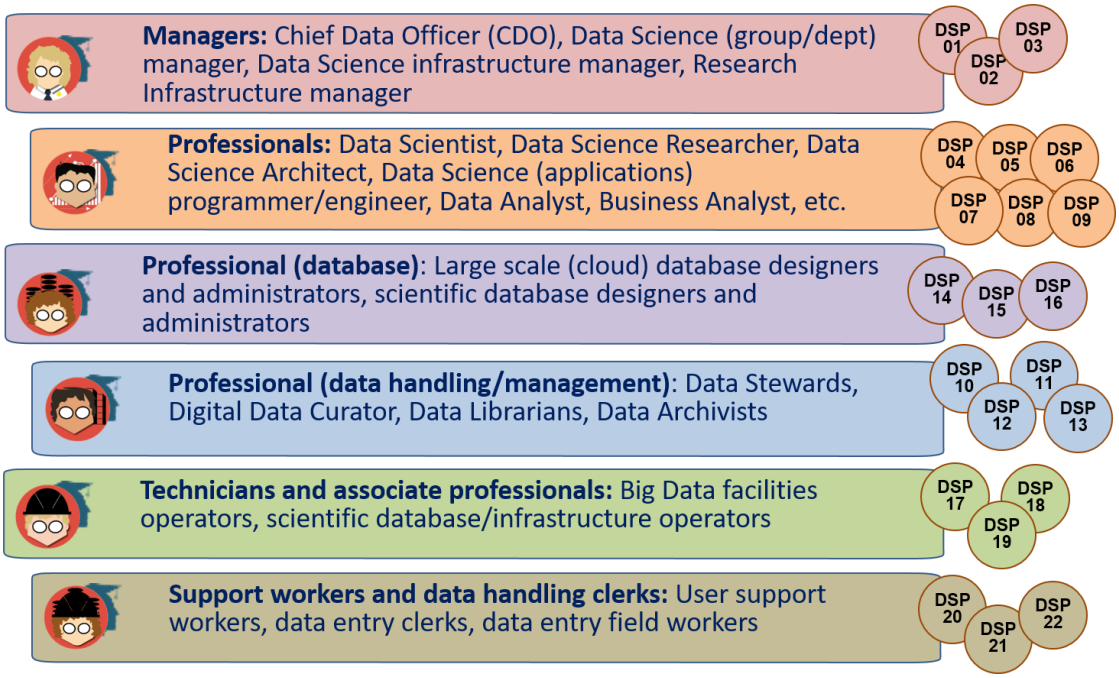


Figure 4. Data Science Professional profiles and their grouping by the proposed new professional groups compliant with the ESCO taxonomy.

Table 4 Data Science professional profiles definition

Profile ID	Data Science Profile title	Data Science Profile Summary statement	*Alternative titles and legacy titles*
Managers			
DSP01	Data Science (group) Manager	Proposes, plans and manages functional and technical evolutions of the data science operations within the relevant domain (technical, research, business).	Data analytics department manager
DSP02	Data Science Infrastructure Manager	Proposes plans and manages functional and technical evolutions of the big data infrastructure within the relevant domain (technical, research, business).	Big Data Infrastructure Manager
DSP03	Research Infrastructure Manager	Proposes plans and manages functional and technical evolutions of the research infrastructure within the relevant scientific domain.	Research Infrastructure data storage facilities manager
Professionals			
DSP04	Data Scientist	Data scientists find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of data-sets, and create visualisations to aid in understanding data. Build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data.	Data Analyst
DSP05	Data Science Researcher	Data Science Researcher applies scientific discovery research/process, including hypothesis and hypothesis testing, to obtain actionable knowledge related to scientific problem, business process, or reveal hidden relations between multiple processes.	Data Analyst
DSP06	Data Science Architect	Designs and maintains the architecture of Data Science applications and facilities. Creates relevant data models and processes workflows.	System Architect, Applications architect
DSP07	Data Science (Application) Programmer/Engineer	Designs/develops/codes large data (science) analytics applications to support scientific or enterprise/business processes.	Scientific Programmer
DSP08	Data Analyst	Analyses large variety of data to extract information about system, service or organisation performance and present them in usable/actionable form	
DSP09	Business Analyst	Analyses large variety of data Information System for improving business performance.	Business Development Manager (Data science role)
Professional (data handling/management) *)			
DSP10	Data Stewards	Plans, implements and manages (research) data input, storage, search, presentation; creates data model for domain specific data; support and advice domain scientists/ researchers. Creates data model for domain specific data, support and advice domain scientists/researchers during the whole research cycle and data management lifecycle.	

DSP11	Digital data curator	Finds, selects, organises, shares (exhibits) digital data collections, maintains their integrity, up-to-date status and freshness, discoverability	Digital curator, digital archivist, digital librarian
DSP12	Data Librarians	Data librarians perform or support one or more of the following: acquisition (collection development), organization (cataloguing and metadata), and the implementation of appropriate user services. Data librarians apply traditional librarianship principles and practices to data management, including data citation, digital object identifiers (DOIs), ethics and metadata.	Digital data curator
DSP13	Data Archivists	Maintain historically significant collections of datasets, documents and records, other electronic data, and seek out new items for archiving.	Digital Archivists
Professional (database)			
DSP14	Large scale (cloud) database designer	Designs/develops/codes large scale data bases and their use in domain/subject specific applications according to the customer needs.	Large scale (cloud) database developer
DSP15	Large scale (cloud) database administrator	Designs and implements, or monitors and maintains large scale cloud databases	
DSP16	Scientific database administrator	Designs and implements, or monitors and maintains large scale scientific databases	Large scale (cloud) database administrator
Technicians and associate professionals			
DSP17	Big Data facilities Operator	Manages daily operation of facilities, resources, and responds to customer requests. Includes all operations related to data management and data lifecycle	
DSP18	Large scale (cloud) data storage operator	Manages daily operation of cloud storage, including related to data lifecycle, and responds to requests from storage users	
DSP19	Scientific database operator	Manages daily operation of scientific databases, including related to data lifecycle, and responds to requests from database users	Large scale (cloud) data storage operators
Clerical and support workers (general and keyboard workers)			
DSP20	Data entry/access worker	Enter data into data management systems directly reading them from source, documents or obtained from people/users	Data entry desk/terminal worker
DSP21	Data entry field workers	The same work done on field when collecting data from disconnected sensors or doing direct counting or reading	
DSP22	User support data services	Provides support to users to entry their data into governmental service and user facing applications	

*) Note: The proposed Professional (data handling/management) taxonomy group doesn't include the occupation of the Digital Librarian as primarily related to digitising the library resources. The following is the commonly used definition of the Digital Librarian responsibilities and functions:
Selection, acquisition, organization, accessibility and preservation of digital information/library. Manages digital materials, takes a lead role in the creation, maintenance and stewardship of digital collections, including the digitization of special collections. Develops strategies for effective management and preservation of library digital assets.

5.3 Mapping Data Science related competences to professional profiles

Table 5 provides a mapping between professional profiles and Data Science competence groups, which are Defined in CF-DS [1] together with the suggested ranking the relevance of different competence groups to corresponding Data Science profiles (where 1 is less relevant and 5 is highly relevant).

The CF-DS competence groups are defined as follows (for full definition of the CF-DS competence see CF-FS document [1] and Appendix)

Data Analytics (DSDA)

Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations

Data Management (DSDM)

Develop and implement a data management strategy for data collection, storage, preservation, and availability for further processing.

Data Science Engineering (DSENG)

Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management

Scientific and Research Methods (DSRM) for research domain and Business Process Management (DSBP)

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 organizational goals

Data Science Domain Knowledge (DSDK)

Use domain knowledge (scientific or business) to develop relevant data analytics applications, and adopt general Data Science methods to domain specific data types and presentations, data and process models, organizational roles and relations

Table 5 Mapping Data Science competence groups to the proposed profiles

Profile ID	**Data Science Profile title**	Data Science Competences Groups (relevance 1 - low, 5 – high)				
		DSDA Data Analytics	DSDM Data Management	DSENG Data Science Engineering	DSRM Research Methods, Business methods	DSDK Subject Domain
Managers						
DSP01	**Data Science (group) Manager**	3	4	3	3	2
DSP02	**Data Science Infrastructure Manager**	2	4	4	2	2
DSP03	**Research Infrastructure Manager**	2	4	4	3	2
Professionals						
DSP04	**Data Scientist**	5	3	4	5	3
DSP05	**Data Science Researcher**	4	3	2	5	4
DSP06	**Data Science Architect**	4	3	5	3	3
DSP07	**Data Science (Application) Programmer/Engineer**	4	2	5	3	4
DSP08	**Data Analyst**	5	3	3	3	4
DSP09	**Business Analyst**	5	3	3	4	5
Professional (data handling/ management)						
DSP10	**Data Stewards**	3	5	3	3	3
DSP11	**Digital data curator**	1	5	2	2	3
DSP12	**Data Librarians**	2	5	2	2	3
DSP13	**Data Archivists**	1	5	1	1	3
Professional (database)						
DSP14	**Large scale (cloud) database designer**	2	4	4	3	3
DSP15	**Large scale (cloud) database administrator**	2	4	3	2	3
DSP16	**Scientific database administrator**	2	4	3	2	3
Technicians and associate professionals						
DSP17	**Big Data facilities Operator**	1	4	4	2	3
DSP18	**Large scale (cloud) data storage operator**	1	4	3	1	1
DSP19	**Scientific database operator**	1	4	3	2	3
Clerical support workers (general and keyboard workers)						
DSP20	**Data entry/access worker**		2	1		2
DSP21	**Data entry field workers**		2	1		2
DSP22	**User support data services**		3	2		2

6 Example DSP Profiles definition

6.1 Template CWA 16458 (2012) Profiles

The European ICT Professional Profiles CWA 16458 (2012) standard uses the following template for the individual professional profiles definition.

Table 8: The European ICT Profile description template and rules

Profile title	**Gives a commonly used name to a profile. TEMPLATE**		
Summary statement	**Indicates the main purpose of the profile.** The purpose is to present to stakeholders and users a brief, concise understanding of the specified ICT Profile. It should be understandable by ICT professionals, ICT managers and Human Resource personnel. It should provide a statement of the job's main activity.		
Mission	**Describes the rationale of the profile.** The purpose is to specify the designated job role defined in the ICT Profile.		
Deliverables	**Accountable (A)**	**Responsible (R)**	**Contributor (C)**
	Specifies the Profile by key deliverables. The purpose is to illuminate the ICT Profiles and to explain relevance including the perspective from a non-ICT point of view.		
Main task/s	**Provides a list of typical tasks to be performed by the profile.** A task is an action taken to achieve a result within a broadly defined context. Tasks may be associated with deadlines, resources, goals, specifications and/or the expected results.		
e-CF competences assigned	**Provides a list of necessary competences (from the e-CF) to carry out the mission.** Must include 1 up to 5 competences. Level assignment is important. Can be (usually) 1 or (maximum) 2 levels.		
KPI Area	**Based upon KPIs (Key Performance Indicators) KPI area is a more generic indicator, congruent with the overall profile granularity level. It is deployed to add depth to the mission.** Not prescriptive. Non-specific measurements. Use general examples. The principle is to provide KPI areas (which are stable, general and long lasting) providing users with an inspiration to enable development of specific KPI's for specific roles (such KPI measurements can be more short-term oriented). Must be related to the key deliverables in order to measure them.		

To ensure future compatibility and easier standardisation, the DSPP will use the same template however leaving some of the fields not filled in. Further DSPP development will include definition of all CWA defined components.

6.2 Example DSPP profiles in CWA 16458 (2012) format

Profile title	**DATA SCIENTIST (DSPP04)**		
Summary statement	Use data analytics to deliver data insight, optimise analytics process, present and visualise data		
Mission	Data scientists find and interpret rich data sources, manage large amounts of data, merge data sources, ensure consistency of data-sets, and create visualisations to aid in understanding data. Build mathematical models, present and communicate data insights and findings to specialists and scientists, and recommend ways to apply the data. Develop compelling visualisation applications, interactive dashboards.		
Deliverables	**Accountable**	**Responsible**	**Contributor**
	• Data collection and preparation • Data selection	• Data analytics applications • Data Analysis to support decision making	• Data Management • Data storage and processing infrastructure and tools
Main task/s	• Develop data analytics applications using Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning) • Apply Prescriptive Analytics methods to initial data insight and organisational workflow optimisation • Develop effective pipeline for data preparation and preprocessing • Define the whole data analysis workflow to support decision making • Identify, investigate and correct problems or inconsistencies related to data analysis • Develop effective visualiation and storytelling tools, create dashboards and data analytics reporting applications		
Competences *(from CF-DS)*	SDSDA01 Use Machine Learning technology, algorithms, tools (including supervised, unsupervised, or reinforced learning)		Level 3
	SDSDA05 Apply Prescriptive Analytics methods		Level 3
	SDSDA08 Apply analytics and statistics methods for data preparation and pre-processing		Level 2
	SDSDA10 Use effective visualiation and storytelling methods to create dashboards and data analytics reports		Level 2
	SDSRM01 Use research methods principles in developing data driven applications and implementing the whole cycle of data handling		Level 3
KPI area	Effective data analytics applications (measurable performance) Contribution to the organisational goals fulfilment, or scientific discovery by providing actionable data insight		

<table>
<tr><th>Profile title</th><th colspan="3">DATA STEWARD (DSPP10)</th></tr>
<tr><td>Summary statement</td><td colspan="3">Plans, implements and manages data collection, supports and advises domain scientist</td></tr>
<tr><td>Mission</td><td colspan="3">Plans, implements and manages (research) data input, storage, search, presentation; creates data model for domain specific data; supports and advices domain scientists/researchers. Interacts with the data analytics team. Do data preparation, inspection, visualisation; prepare for archiving and publication.</td></tr>
<tr><td>Deliverables</td><td>Accountable</td><td>Responsible</td><td>Contributor</td></tr>
<tr><td></td><td>• Data model
• Data Management Plan</td><td>• Data collection/in gest</td><td>• Domain related models
• Data analytics result inspection</td></tr>
<tr><td>Main task/s</td><td colspan="3">• Define/build/optimize data model and schemas
• Use existing or define new metadata framework
• Publish research data to existing scientific data archives
• Manage organisational or project related data
• Search and promote research data
• Assist main domain researcher/scientist in selecting right data analytics methods
• Interface between
• Monitor applying FAIR (Findable, Accessible, Interoperable, Reusable) and Open Data principles to data created by organisation or project</td></tr>
<tr><td rowspan="5">e-competences
(from e-CF)</td><td colspan="2">SDSDM02 Use data storage systems, data archive services, digital libraries, and their operational models</td><td>Level 1</td></tr>
<tr><td colspan="2">SDSDM05 Implement data lifecycle support in organisational workflow, support data provenance and linked data</td><td>Level 2</td></tr>
<tr><td colspan="2">SDSDM06 Consistently implement data curation and data quality controls, ensure data integration and interoperability</td><td>Level 2</td></tr>
<tr><td colspan="2">SDSDM08 Use and implement metadata, PID, data registries, data factories, standards and compliance</td><td>Level 3</td></tr>
<tr><td colspan="2">SDSDM09 Adhere to the FAIR principles of the Open Data, Open Science, Open Access, use ORCID based services</td><td>Level 3</td></tr>
<tr><td>KPI area</td><td colspan="3">Consistent data management workflow
Compliance with FAIR principles</td></tr>
</table>

6.3 Data Science Analytics enabled jobs and profiles

Recent studies by BHEF, PwC [20] and IBM, BGT and NHEF [21] identified strong growth of the Data Science and Analytics (DSA) enabled jobs that are not are not pure Data Scientists but require extensive DSA knowledge to work in the specific industry sectors. Figure 5 from PwC and BHEF study [20] provides illustration of currently highly demanded DSA enabled jobs in multiple industry and business sectors: Finance and Insurance; Healthcare and Social Assistance; Information; Manufacturing; Professional, Scientific and Technical Services; Retail Trade.

The study provides data that of 2.35 million job postings in the US in 2017 23% Data Scientist and 67% DSA enabled jobs. It is also strong demand for managers and decision makers with the Data Science (data analytics) skills/understanding. This creates a new challenge to deliver actionable knowledge and competences to CEO level managers

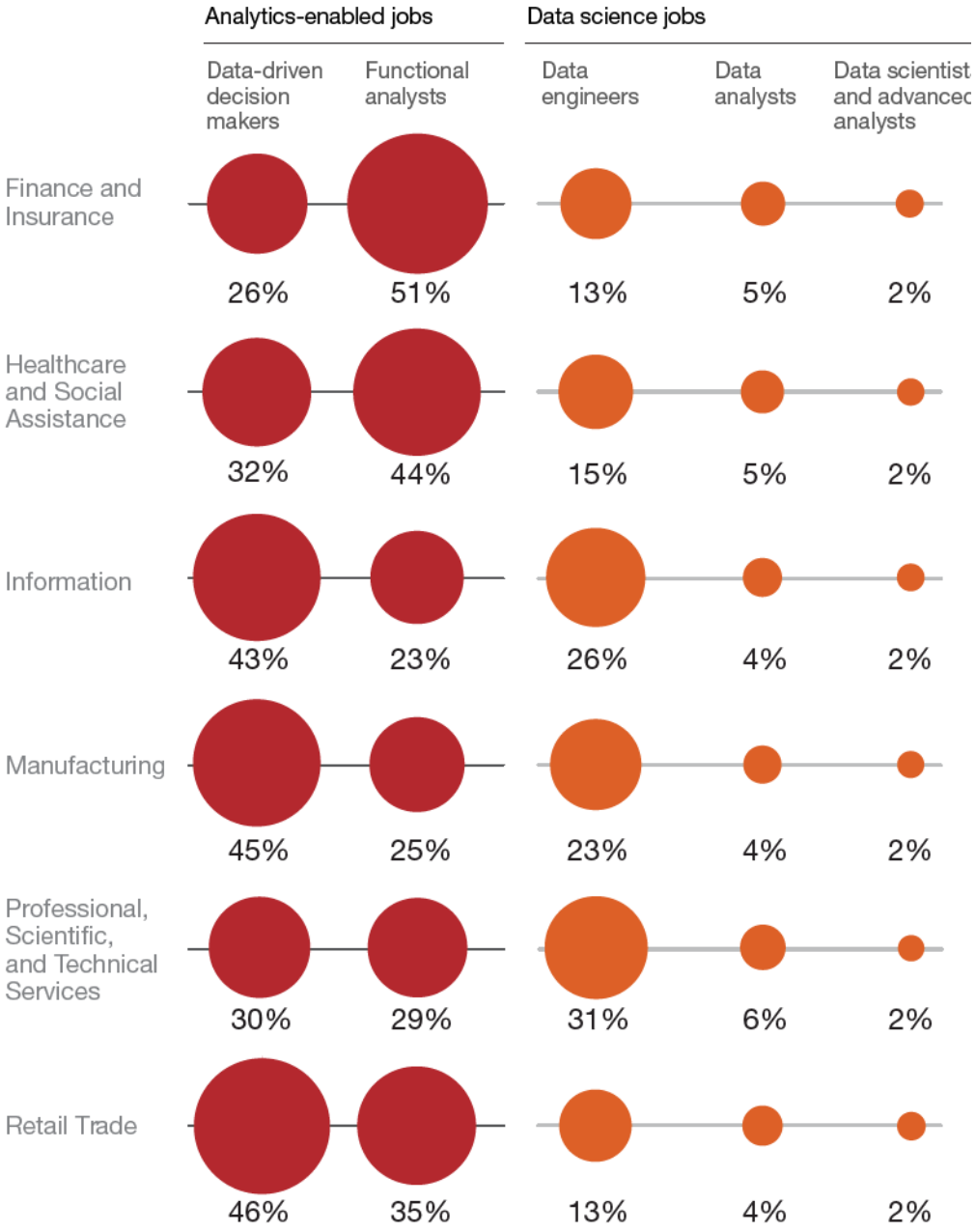


Figure 5. Strong demand for business people with analytics skills, not just data scientists in multiple industry sectors [2].

7 Practical use of the Data Science Professional profiles

The presented DSPP together with CF-DS and other EDSF documents provide a basis for multiple practical uses include but not limited to:

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

7.1 Usage example: Competences assessment

Figure 6 illustrates example of the individual competences assessment that maybe used for one of the general use cases: 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.

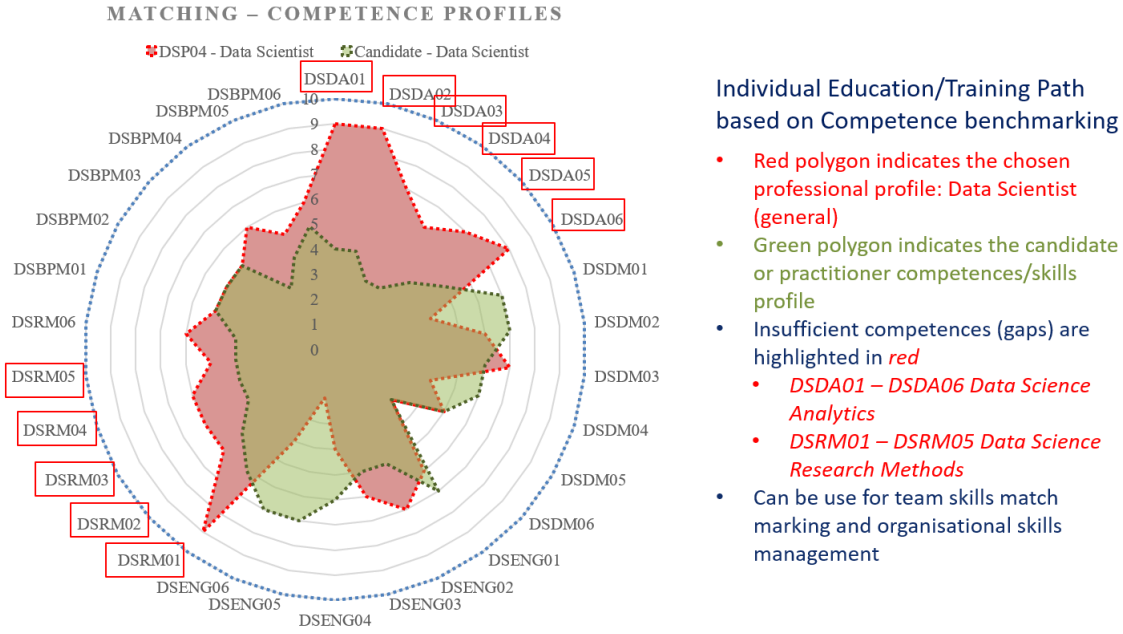


Figure 6. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

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 simple test. The illustrated competences mismatch can be used either for deciding on the suitability of the candidate or suggesting necessary training program.

Using 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 Science Team composition

Data Science team composition and competences matching is one of intended uses of the EDSF and DSPP in particular. Figure 7 illustrates a case of creating a Data Science team or group for an average size of the research organisation with affiliated number of researchers 200-300, what would require a Data Science team of 10-15 members whose responsibility would include supporting all main stages of data lifecycle: data collection, data input/ingest, data analysis, reporting, visualisation and storage. The figure also illustrates possible roles that may be assigned to perform different functions at different data workflow stages

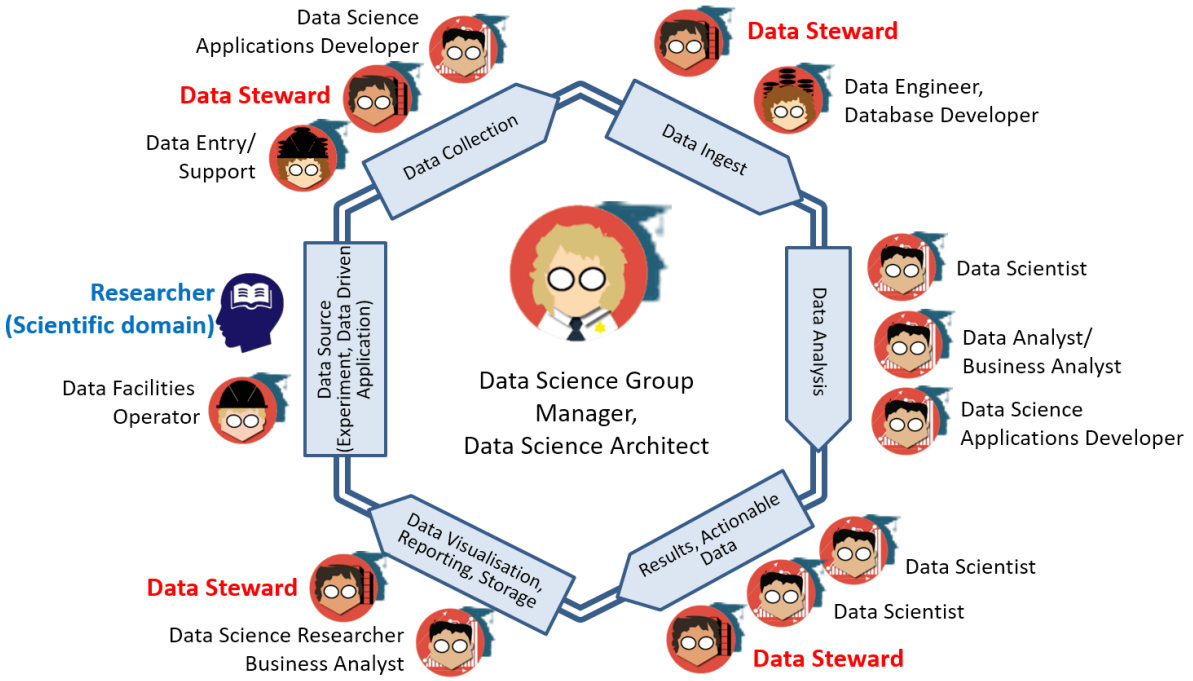


Figure 7. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

To support all data related research or production stages the following roles may be required (including suggested staffing for the team of 10-12 members):

- (Managing) Data Science Architect (1)
- Data Scientist (1), Data Analyst (1)
- Data Science Application architect/developer/programmer (2)
- Data Infrastructure/facilities administrator/operator: storage, cloud, computing (1)
- Data stewards, curators, archivists (3-5)

It is possible that some of the above roles can be re-defined and re-allocated to the Data Science team from the previous ICT and IT infrastructure groups or departments. In this case some basic Data Science training will be required for not initially data related professions.

It also suggested a distinct role of the Data Steward, a new emerging role for data driven research organisations and projects. Data Steward should play a bridging role between the subject domain researcher and the Data Science team or Data Scientist in particular cases to help to translate between subject domain and Data Science or data analytics domain. Data Stewards can have both backgrounds either ICT and computer or digital curation/librarian.

7.3 Data Steward Professional profile and organisational functions

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

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

Important role of the Data Steward is recognized in the HLEG report on European Open Science Cloud (October 2016) [19] identified critical need for core data experts and data stewards in particular.

8 Conclusion and further developments

This document provided information about the Data Science Professional profiles definition as a part of the overall EDISON Data Science Framework. The presented DSP profiles are defined based on and as an extension to ESCO taxonomy. They are enumerated and includes the following groups: Managers (DSP01-DSP03), Professionals (DSP04-DSP09), Professional Data Management/Handling (DSP10-DSP13), Professional (database) Technical (DSP14-DSP16), Professional Technicians (DSP17-DSP19), Support and clerical workers (DSP20 – DSP22).

The document also provides an example how the identified in CF-DS competences can be assigned to different profiles.

The presented information is a result of the EDISON project team discussion and a subject for further review and discussion by the research and industry community.

Further developments will be focused on the following activities:

- Finalise the Data Science Professional profiles definition by collecting feedback by consulting ESCO committee and practitioners from research and industry on their Human Resource management practices. Provide suggestion for ESCO extension with Data Science and data related occupations
- Cooperate with the CEN TC428 and define the proposed profiles in the format of the European ICT Professional Profiles according to CWA 16458
- Run the community survey and use a customisable questionnaire to run few key interviews, primarily with experts and top executives at universities and companies.

9 References

[1] Data Science Competence Framework [online] http://edison-project.eu/data-science-competence-framework-cf-ds

[2] Data Science Body of Knowledge [online] http://edison-project.eu/data-science-body-knowledge-ds-bok

[3] Data Science Model Curriculum [online] http://edison-project.eu/data-science-model-curriculum-mc-ds

[4] Data Science Professional Profiles [online] http://edison-project.eu/data-science-professional-profiles

[5] 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/

[6] Riding the wave: How Europe can gain from the rising tide of scientific data. *Final report of the High Level Expert Group on Scientific Data. October 2010*. [Online]. Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[7] European ICT Professional Profiles CWA 16458 (2012) (Updated by e-CF3.0) [online] http://relaunch.ecompetences.eu/wp-content/uploads/2013/12/EU_ICT_Professional_Profiles_CWA_updated_by_e_CF_3.0.pdf

[8] European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1 [online] http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf

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

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

[11] NIST SP 1500-1 NIST Big Data interoperability Framework (NBDIF): Volume 1: Definitions, September 2015 [online] http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

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

[13] What is a data scientist? 14 definitions of a data scientist! [online] http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

[14] LinkedIn's Daniel Tunkelang On "What Is a Data Scientist?" [online] http://www.forbes.com/sites/danwoods/2011/10/24/linkedins-daniel-tunkelang-on-what-is-a-data-scientist/

[15] 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

[16] 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/

[17] Computer Science 2013: Curriculum Guidelines for Undergraduate Programs in Computer Science http://www.acm.org/education/CS2013-final-report.pdf

[18] Cortnie Abercrombie, What CEOs want from CDOs and how to deliver on it [online] http://www.slideshare.net/IBMBDA/what-ceos-want-from-cdos-and-how-to-deliver-on-it

[19] Realising the European Open Science Cloud. First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud, October 2016 [online] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.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 – To be published 2017, currently in work.

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
CS	Computer Science
DM-BoK	Data Management Body of Knowledge by DAMAI
DS-BoK	Data Science Body of Knowledge
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
ICT	Information and Communication Technologies
IEEE	Institute of Electrical and Electronics Engineers
IPR	Intellectual Property Rights
LIBER	Association of European Research Libraries
MC-DS	Data Science Model Curriculum
NIST	National Institute of Standards and Technologies of USA
PM-BoK	Project Management Body of Knowledge
PRACE	Partnership for Advanced Computing in Europe
RDA	Research Data Alliance
SWEBOK	Software Engineering Body of Knowledge

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

A.1. O'Reilly Strata Survey (2013)

O'Reilly Strata industry research [25] 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 Scientists practitioners self-identification:

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

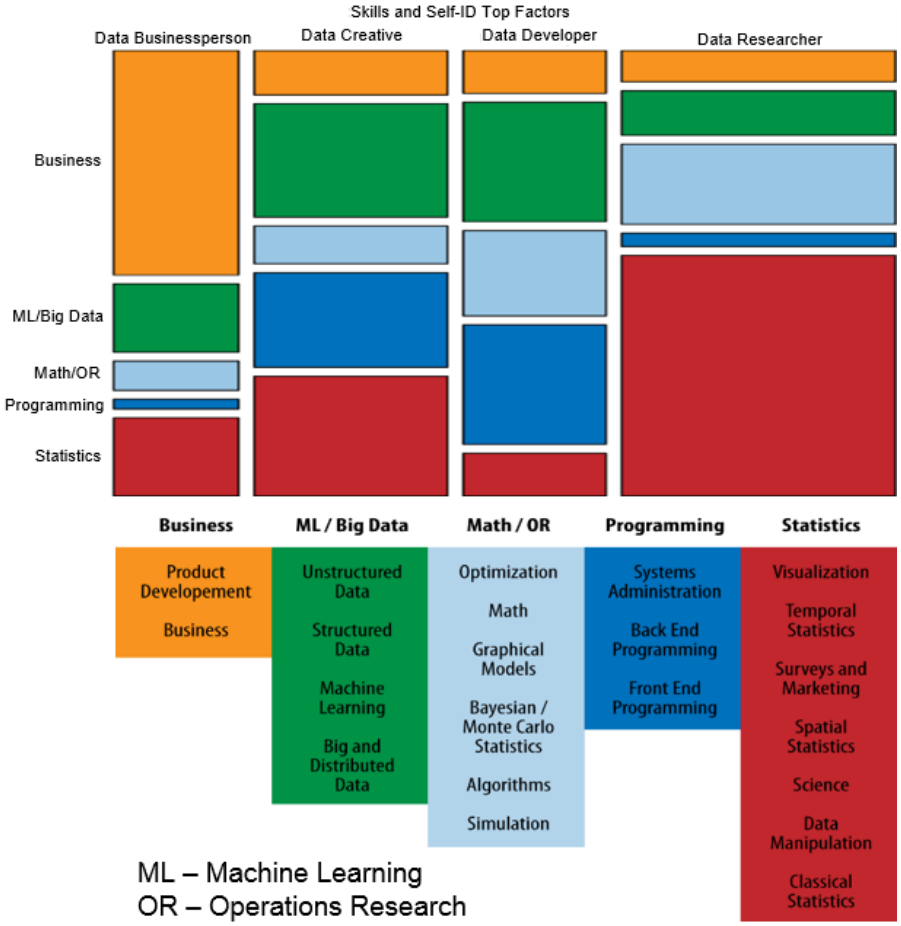


Figure A.1. Data Scientist skills and profiles according to O'Reilly Strata survey [25]

Table A.1 below lists skills for Data Science that are identified in the study. They are very specific in 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

A.3. UK Study on demand for Big Data Analytics Skills (2014)

The study "Big Data Analytics: Assessment of demand for Labour and Skills 2013-2020" [15] 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 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

A.4. IWA Data Science profile

Italian Web Association (IWA) published the WSP-G3-024. Date Scientist Profile for web related projects [16]. It provide a god example of domain specific definition of the Data Science competences, skills and organisational responsibilities, it suggests also 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 purpose, it is worth to mention 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.

Appendix B. Data Science Competence Framework (CF-DS) Excerption

This Appendix contains excerption from the original CF-DS document [1] that is required for understanding of the presented in this document the DSPP. The excerption includes the identified Data Science competences that are used for defining the DSPP. The full CF-DS definition including both competences and skills is available in the CF-DS document.

B.1. 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 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:

Core Data Science competences/skills groups defining profile of the Data Science related professional profiles

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Additional common competence groups demanded by organisations

- Data Management and Governance (including data stewardship, curation, and preservation)
- *Research Methods for research related professions and Business Process Management for business related professions*

Data management, curation and preservation competences are already attributed to the existing (research) data related professions such as data archivist, data manager, digital data librarian, and others. Data management is also important component of European Research Area and Open Data policies. It is extensively addressed by the Research Data Alliance and supported by numerous projects, initiatives and training programmes[2].

Knowledge of the scientific research methods and techniques is something that makes Data Scientist profession different from all previous professions.

From the education and training point of view, the identified competences can be treated or linked to expected learning or training outcome. 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 5 Data Science related competence groups provide a better basis for defining consistent and balanced education and training programmes for Data Science related jobs, re-skilling and professional certification.

Table B.1 provides the proposed Data Science competences definition for different groups supported by the data extracted for the collected information. The presented competences definition has been reviewed by a number of expert groups and individual experts (see Section 7 for details). The presented competences are required for different professional profiles, organisational roles and throughout the whole data lifecycle, but not necessary to be provided y a single role or individuum. The presented competences are enumerated to allow easy use and linking between all EDSF document.

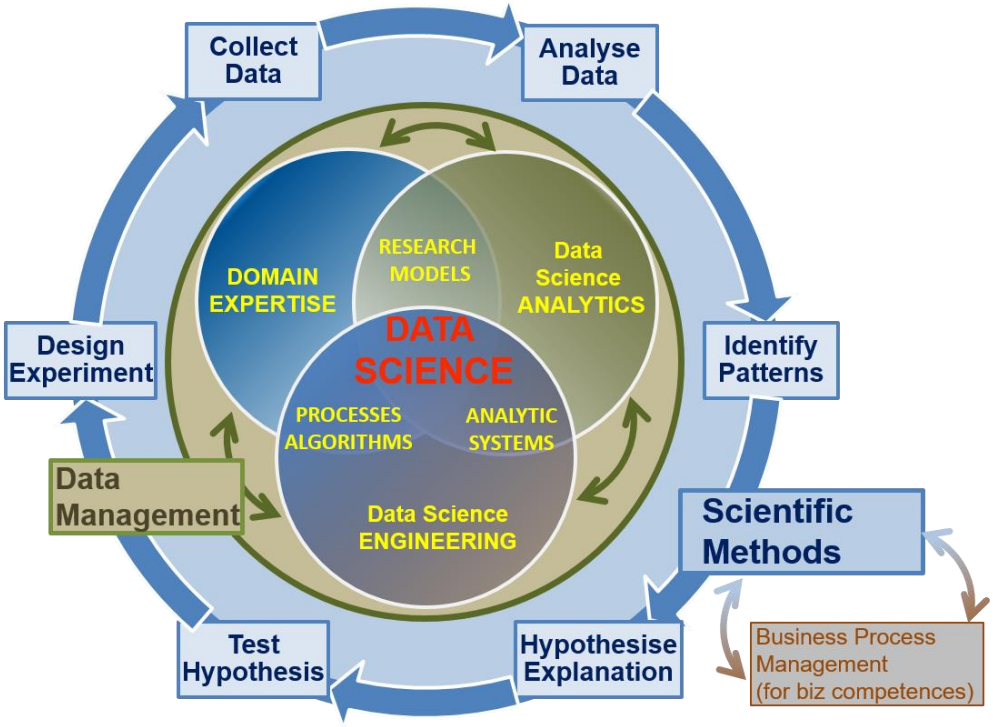
[2] Research Data Alliance Europe https://europe.rd-alliance.org/

Table B.1. Competences definition for different Data Science competence groups

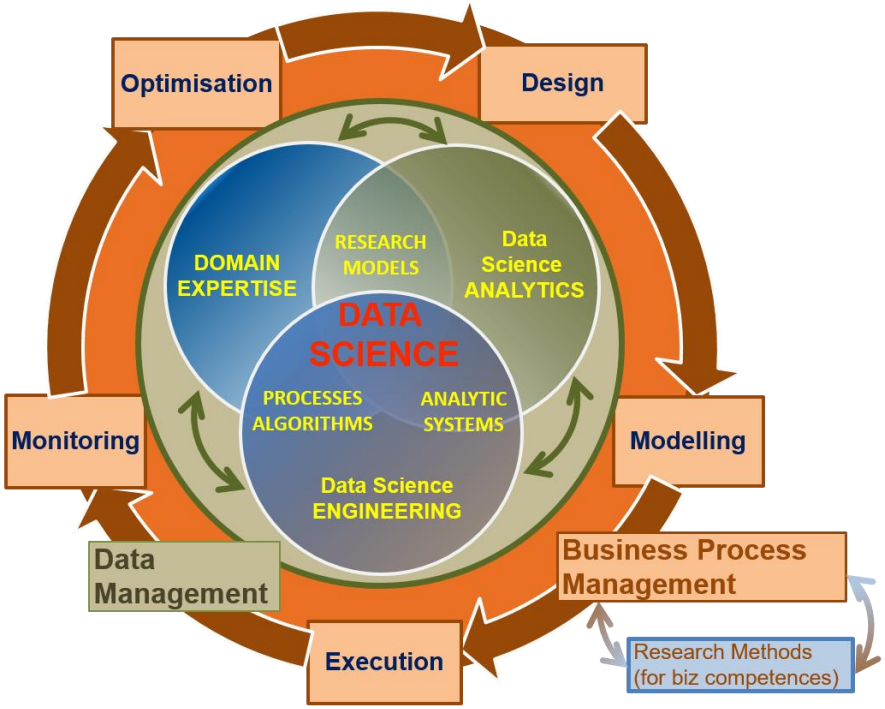
Data Analytics (DSDA)	Data Science Engineering (DSENG)	Data Management (DSDM)	Research Methods and Project Management (DSRM)	Domain related Competences (DSDK): Applied to Business Analytics (DSBA)
DSDA Use appropriate data analytics and statistical techniques on available data to discover new relations and deliver insights into research problem 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.	DSRM 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 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 a form of data management policy and Data Management Plan (DMP)	DSRM01 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 exisitng data and suggest/identify new data required for specific business context to achieve organizational goal, including using social network 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	DSRM02 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	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 source and provide them for further analysis and use	DSRM03 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)	DSRM04 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	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)	DSRM05 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	DSRM06 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

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



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



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

Figures B.1. Relations between identified Data Science competence groups for (a) general or research oriented and (b) business oriented professions/profiles: Data Management and Scientific/Research Methods or Business Processes Management competences and knowledge are important for all Data Science profiles.

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

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

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

There is a number of the 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

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 into 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 into master programs and graduate students of many master programs.

B.2. Identified Data Science Skills

For identified Data Science skills and technical platforms knowledge refer to the original CF-DS document [3].