# EDISON

building the data
science profession

EDISON Data Science Framework:
Part 4. Data Science Professional Profiles (DSPP)
Release 3

EDISON Community Initiative
(Maintaining the H2020 EDISON project outcome)

| Release Date | 7 September 2018 |
|---|---|
| Document Editor/s | Yuri Demchenko |
| Version | Release 3, v0.6 |
| Status | Working document, request for comments |

| Document Version Control | | | |
|---|---|---|---|
| Version | Date | Change Made (and if appropriate reason for change) | Editor(s) (Initials) |
| 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 |
| 0.6 | 07/09/2018 | Pre-release 3. Definition of professional profiles revised and updated. Comments incorporated. ESCO summary updated based on available ESCO documents. | YD |
| | | | |
| | | | |
| | | | |
| | | | |

Contributors

| Document Editors: Yuri Demchenko | | |
|---|---|---|
| Author Initials | Name of Contributor | Institution |
| YD | Yuri Demchenko | University of Amsterdam |
| TW | Tomasz Wiktorski | University of Stavanger |
| | | |
| | | |
| | | |
| | | |

## Executive summary

The initial definition of the EDISON Data Science Framework (EDSF) has been done in the Horizon2020 Project EDISON (Пκφте 675419) that produced Release 1 (2016 and Release 2 (2017). Currently, EDSF is maintained by the EDISON Community initiative that is coordinated by the University of Amsterdam.  The new EDSF Release 3 is the product of the wide community of academicians, researcher and practitioners that are practically involved into 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.

The EDISON Data Science framework (EDSF) includes such components as Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS), and Data Science Professional Profiles (DSPP). 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 proposed EDSF intends to 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 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 and is the result of project efforts and wide community contribution.

- The proposed Data Science professional profiles are defined as an intended extension to the 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 tools for individual competences benchmarking against selected (or desirable) professional profiles as well as advising practitioners on the (up/re-) skilling path.

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

The revolutionary value of data in modern computer powered digital economy the 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 (2015-2017) to define the Data Science Professional profiles that are important for defining the Data Scientist roles in the organisation and their alignment with organizational goals and mission. The Data Science Professional Profiles 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 professional profiles definition frameworks for ICT and Data Science competences and skills including CWA 16458 (2012) European ICT profiles, European Skills, Competences, and Occupations (ESCO) framework, International Standard Classification of Occupations (ISCO). Section 4 presents the definition of the proposed DSPP as an extension to ESCO taxonomy. It also provides example mapping of different profiles to CF-DS competences (using CWA 16458 template) 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.
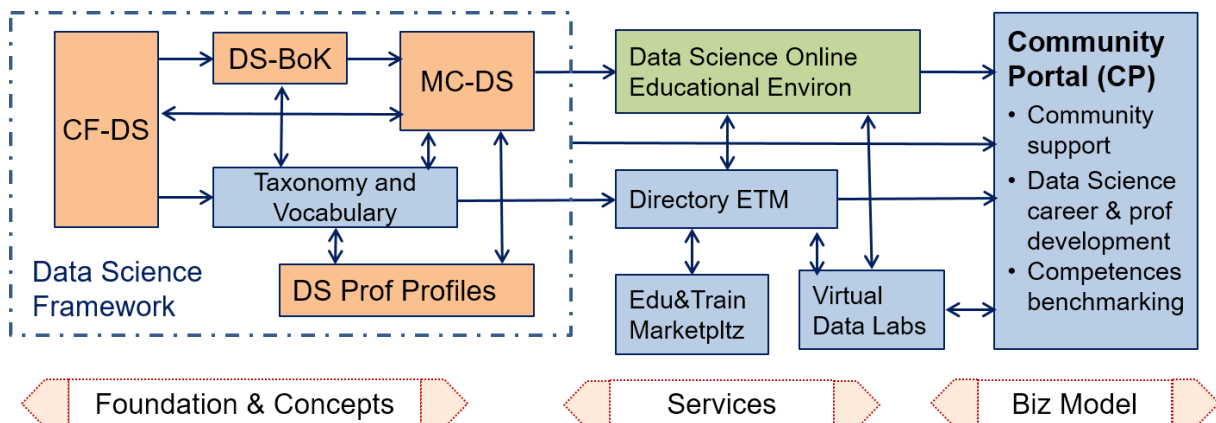
## 2    EDISON Data Science Framework

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[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 1 EDISON Data Science Framework components.**

The CF-DS provides the overall basis for the whole EDSF. 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.

---

[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.

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 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 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 professional profiles 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 [5].

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.



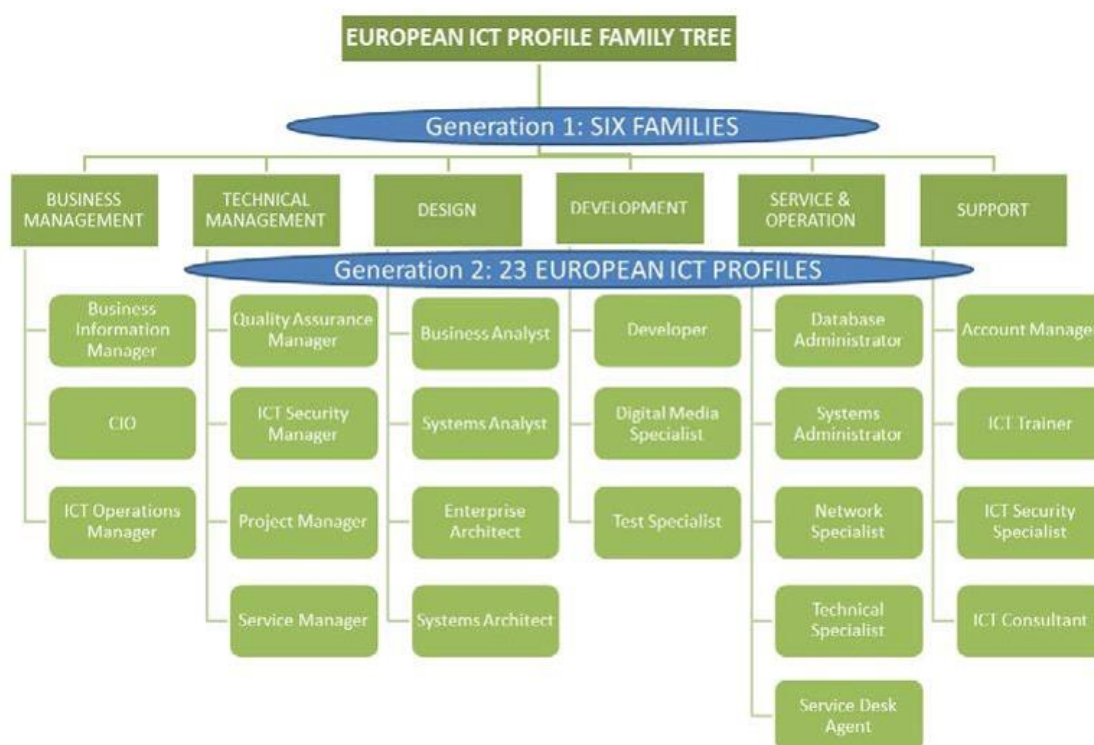Figure 2. European ICT Profile Family Tree – Generation 1 and 2 as a shared European reference [5]

The 23 profiles constructed in CWA combined with e- competences from the e-CF3.0 [6], 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.
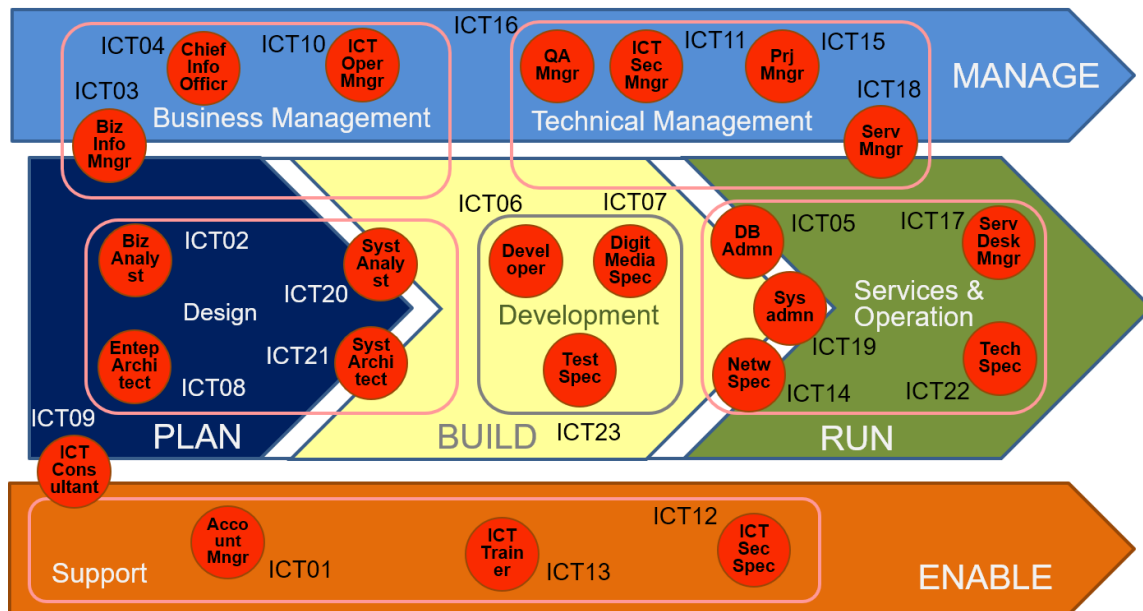


Figure 3. European ICT Professional Profiles structured by six families and positioned within the ICT Business Process (e-CF Dimension 1) (adopted from [6] and extended)

## 3.2 ESCO (European Skills, Competences, Qualifications and Occupations) framework and platform [7]

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.

The first full version of ESCO (ESCO v1) has been released on 28 July 2017. The release has been followed by the Conference "ESCO: Connecting people and jobs" whereby ESCO has been showcased in concrete applications to demonstrate its value in different use cases. ESCO v1 contains 2942 occupations, 13485 skills and competences as well as some qualifications. As each concept in ESCO exists in all 26 ESCO languages this amounts to more than +514 000 skills terms and +350 000 occupations terms [8].

ESCO is organised in three interrelated pillars:

- the occupations pillar that defines occupation profile which refers to essential and optional knowledge, skills and competences; higher level occupations classification in ESCO is compatible with ISCO as explained below.
- the knowledge, skills and competences pillar that besides the terms definition defines their re-usability as transversal, cross-sector, sector-specific, occupation specific;
- the qualifications pillar.

The following definitions are used:

**Knowledge:** The body of facts, principles, theories and practices that is related to a field of work or study. Knowledge is described as theoretical and/or factual, and is the outcome of the assimilation of information through learning.

**Skill:** The ability to apply knowledge and use know-how to complete tasks and solve problems. Skills are described as cognitive (involving the use of logical, intuitive and creative thinking) or practical (involving manual dexterity and the use of methods, materials, tools and instruments).

**Competence:** The proven ability to use knowledge, skills and personal, social and/or methodological abilities, in work or study situations, and in professional and personal development.

Transversal skills and competences are organised in a hierarchical structure with the following five headings:

- thinking
- language
- application of knowledge
- social interaction
- attitudes and values

The qualifications pillar aims to collect existing information on qualifications19. The final objective of the pillar is to provide a comprehensive list of qualifications relevant for the European labour market. The qualifications pillar in ESCO 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 qualifications pillar of ESCO is developed in full compliance and complementarity with the European Qualifications Framework (EQF) [9].

ESCO Strategic Framework [10] defines relation of ESCO with other European initiatives and standards, in particular ESCO will be used within the EURES network of employment services in order to exchange job vacancies and CVs between Member States and with the Commission.

Table 1 contains data related occupations extracted from the ESCO classification together with related hierarchies. Table 2 is included for reference purposes to present 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** | |
| | 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.

## 3.3 ISCO (International Standard Classification of Occupations) [9,10]

The International Standard Classification of Occupations (ISCO) is a tool for organizing jobs into a clearly defined set of groups according to the tasks and duties undertaken in the job. Its main aims are to provide:

- a basis for the international reporting, comparison and exchange of statistical and administrative data about occupations;
- a model for the development of national and regional classifications of occupations; and
- a system that can be used directly in countries that have not developed their own national classifications.

It is intended for use in statistical applications and in a variety of client oriented applications. Client oriented applications include the matching of job seekers with job vacancies, the management of short or long term migration of workers between countries and the development of vocational training programmes and guidance.

ISCO is a four-level classification of occupation groups managed by the International Labour Organisation (ILO). Its structure follows a grouping by education level. The two latest versions of ISCO are ISCO-88 (dating from 1988) and ISCO-08 (dating from 2008). Many current national occupational classifications are based on one of these ISCO versions.

ISCO 08 groups jobs together in occupations and more aggregate groups mainly on the basis of the similarity of skills required to fulfil the tasks and duties of the jobs. Two dimensions of the skill concept are used in the definition of ISCO 88 groups:

- Skill level, which is a function of the range and complexity of the tasks involved, where the complexity of tasks has priority over the range; and
- Skill-specialisation, which reflects type of knowledge applied, tools and equipment used, materials worked on, or with, and the nature of the goods and services produced. It should be emphasised that the focus in ISCO 88 is on the skills required to carry out the tasks and duties of an occupation and not on whether a worker in a particular occupation is more or less skilled than another worker in the same or other occupations.

In ESCO, each occupation is mapped to exactly one ISCO-08 code. ISCO-08 can therefore be used as a hierarchical structure for the occupations pillar in ESCO. ISCO-08 provides the top four levels for the occupations pillar. ESCO occupations are located at level 5 and lower. Figure 4 illustrates the role of ISCO 08 in the hierarchical structure of the ESCO occupations pillar.
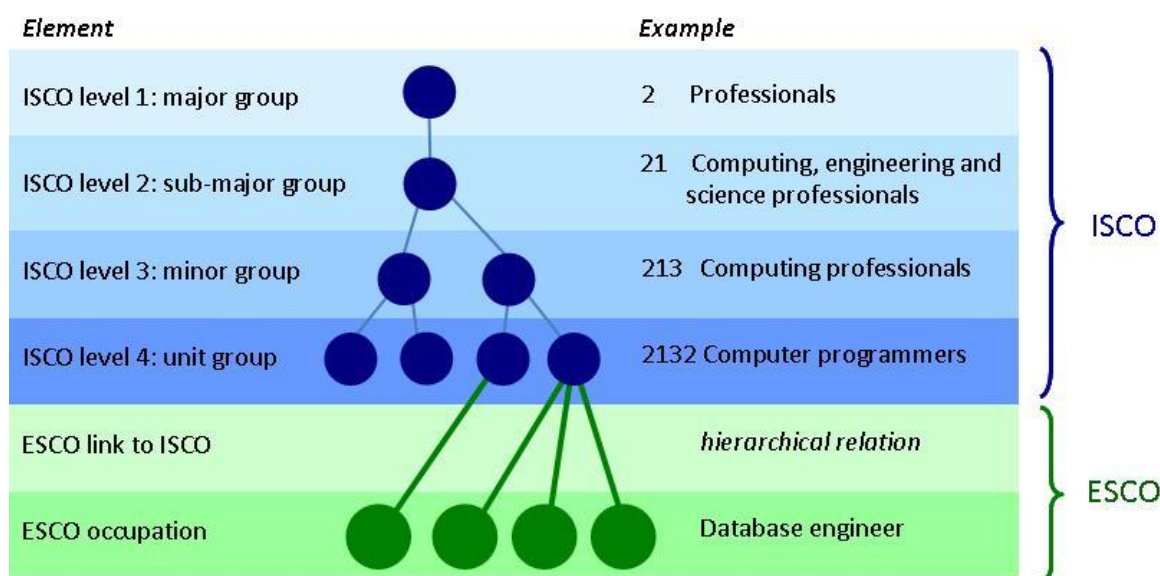


Figure 4. Relation between ESCO and ISCO: ISCO provides 4 top hierarchy levels in ESCO occupations pillar [8].

Since ISCO is a statistical classification, its occupation groups do not overlap. Each ESCO occupation is therefore mapped to only one ISCO unit group. It follows from this structure, that ESCO occupation concepts can be equal to or narrower than ISCO unit groups, but not broader. The result is a strictly mono-hierarchical structure where each element at level 2 or lower has exactly one parent. A few groups of ISCO-08 do not contain ESCO occupations. These are usually occupation groups with no economic activity in the EU, such as "water and firewood collectors".

Appendix B provides example of ISCO Top Occupation groups for Professionals > *Software and Applications Developers and Analysts >* Information and communications technology professionals > Database and network professionals. No Data Science or data related occupations are defined in ISCO.

## 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 [12] 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 [12]. 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 [142, 14] 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" [15].

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 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 profiles/roles is based on the ESCO occupations classification where proposed profiles' competences and organisational roles are defined similar to CWA 16458 ICT profiles.

Table 3 presents the proposed Data Science and Data Management related professional profiles 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 proposed new occupations don't include variety of the Data Science and Analytics enabled professions in different industry and research domains and sectors, which are becoming popular and highly demanded by modern organisations implementing data driven business model.

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:
  - o    Managers (for managerial roles);
  - o    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 related 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 [5], eCFv3.0 (future CEN standard EN 16324) [6].

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 5 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 illustrates what competence groups are relevant to each profile by indicating competence relevance from 0 to 5 (0 – not relevant, 5 – very important) where information is taken from Table 5. Figure 6 provides visual presentation of the identified DSPP and their grouping by the proposed high level classification groups.

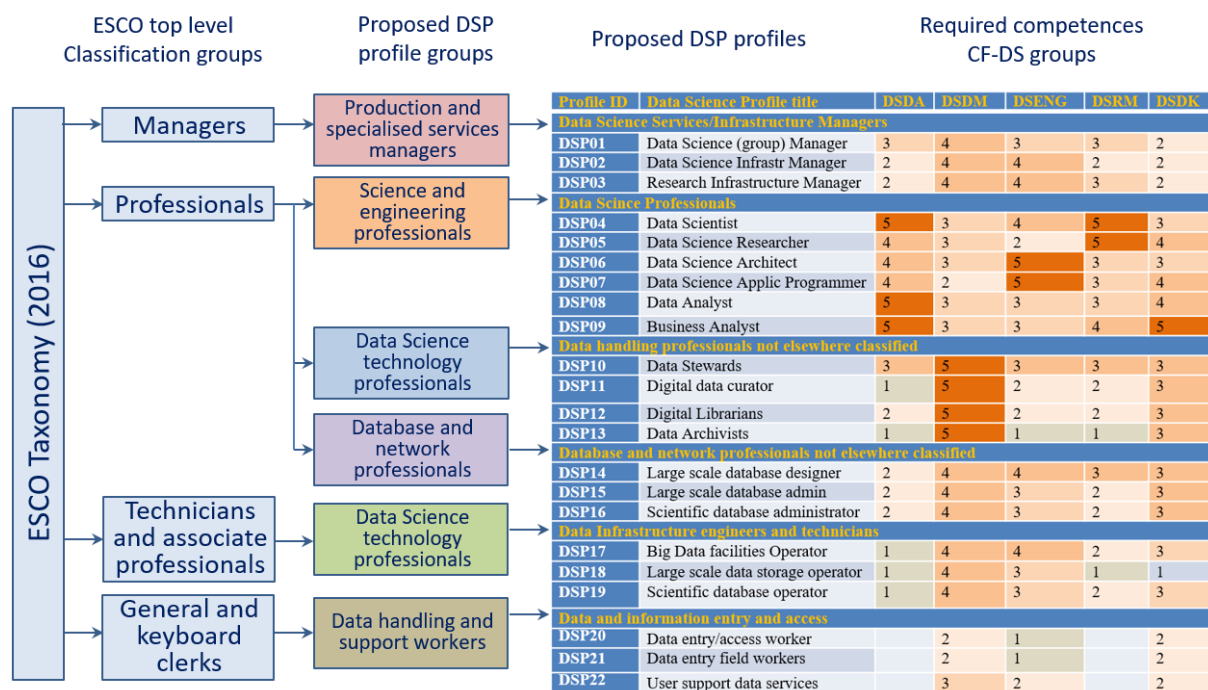| Profile ID | Data Science Profile title | DSDA | DSDM | DSENG | DSRM | DSDK |
|---|---|---|---|---|---|---|
| **Data Science Services/Infrastructure Managers** | | | | | | |
| DSP01 | Data Science (group) Manager | 3 | 4 | 3 | 3 | 2 |
| DSP02 | Data Science Infrastr Manager | 2 | 4 | 4 | 2 | 2 |
| 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 5. Proposed Data Science related extensions to the ESCO classification hierarchy and corresponding DSPP by classification groups.
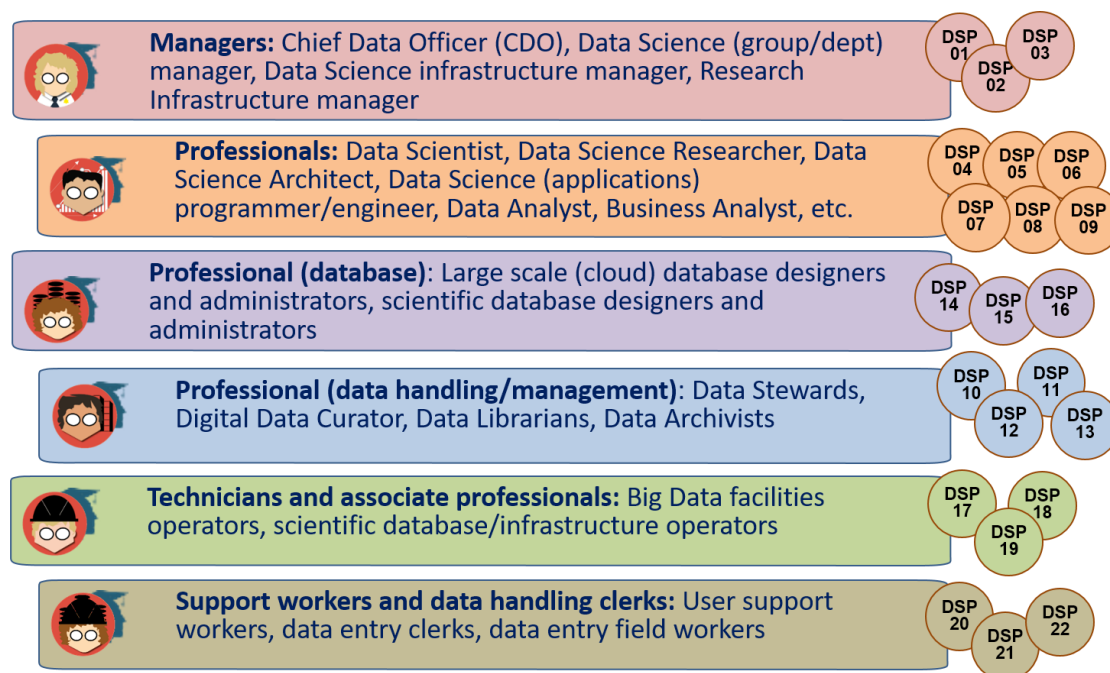


Figure 6. Data Science Professional profiles and their grouping by the proposed new professional groups compliant with the ESCO taxonomy.

**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, Data Engineer |
| **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. | Data Engineer, Data Architect |
| 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 | |

*) The proposed new occupations don't include variety of the Data Science and Analytics enabled professions in different industry and research domains and sectors.

**) 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 0 is not relevant and 5 is highly relevant).

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

**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

**Research Methods and Project Management (DSRMP) for research domain and Business Process Management (DSBPM)**
> 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

Note, Table 5 is using scale 0 to 5 for the competence groups relevance, more detailed mapping all 30 individual competences as defined in CF-DS document is provided in the EDSF/CF-DS Excel documents available from the EDSF github project repository[2]. It uses more granular mapping using 0 to 9 scale.

---

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

**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, Project Management | 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** | 0 | 2 | 1 | 0 | 2 |
| DSP21 | **Data entry field workers** | 0 | 2 | 1 | 0 | 2 |
| DSP22 | **User support data services** | 0 | 3 | 2 | 0 | 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 6: 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.**<br><br>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.**<br><br>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.**<br><br>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.**<br><br>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.**<br><br>Must include 1 up to 5 competences.<br><br>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.**<br><br>Not prescriptive. Non-specific measurements. Use general examples.<br><br>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).<br><br>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[3]

| 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<br>• Data selection | • Data analytics applications<br>• Data Analysis to support decision making | • Data Management<br>• 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)<br>• Apply Prescriptive Analytics methods to initial data insight and organisational workflow optimisation<br>• Develop effective pipeline for data preparation and preprocessing<br>• Define the whole data analysis workflow to support decision making<br>• Identify, investigate and correct problems or inconsistencies related to data analysis<br>• Develop effective visualiation and storytelling tools, create dashboards and data analytics reporting applications | | |
| Competences<br>*(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)<br>Contribution to the organisational goals fulfilment, or scientific discovery by providing actionable data insight | | |

[3] Competences are defined based on the CF-DS document. Only the most important competences are listed. Full mapping of individual profiles to competences, skills and knowledge can be derived from Table 5 in this document and EDSF taxonomy currently expressed in the form of Excel workbooks (see the EDSF github project).

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

## 6.3    Data Science Analytics enabled jobs and profiles

Recent studies by BHEF, PwC [16] and IBM, BGT and BHEF [17] 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 7 from PwC and BHEF study [16] 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
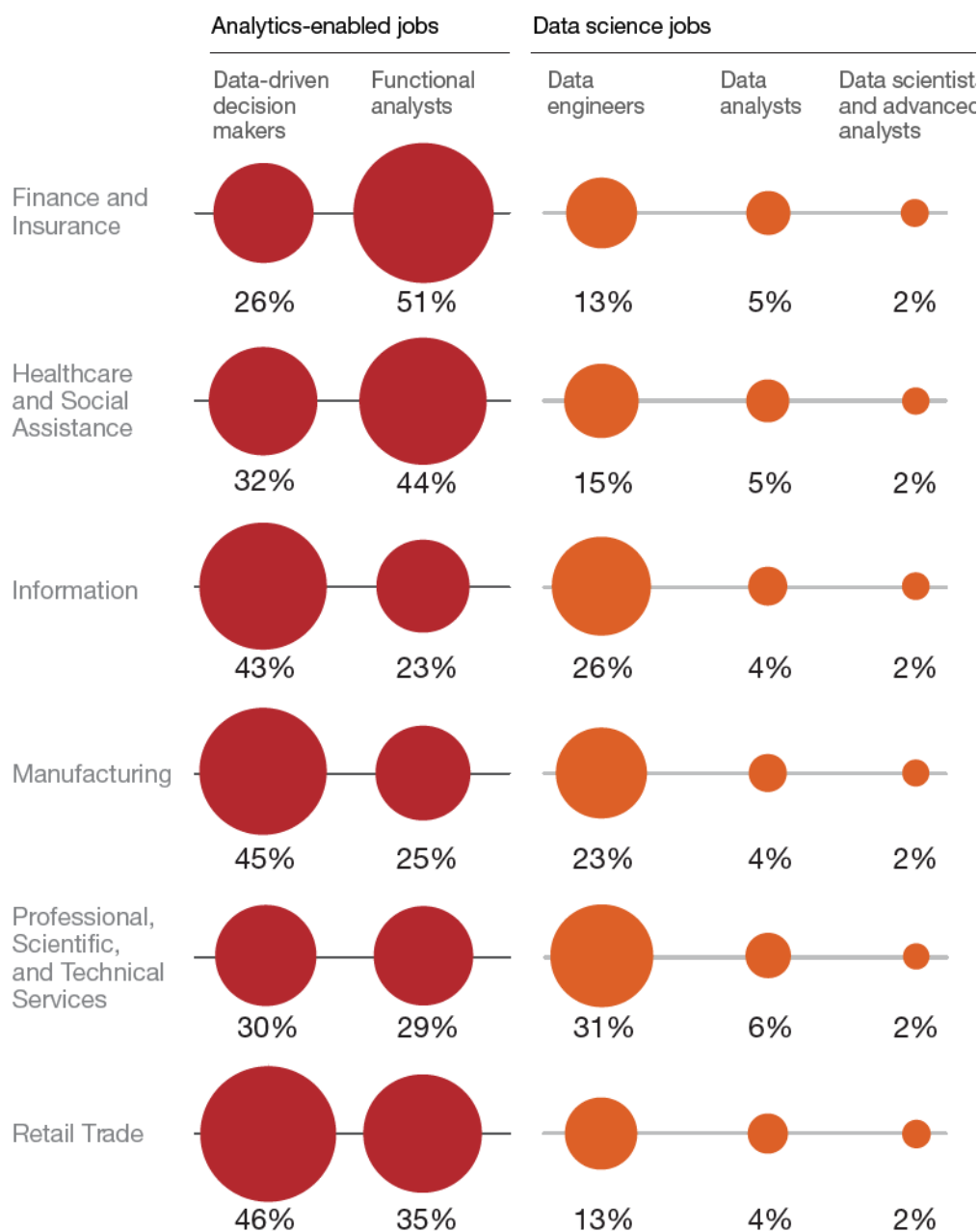


Figure 7. Strong demand for business people with analytics skills, not just data scientists in multiple industry sectors [16].

# 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 8 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.
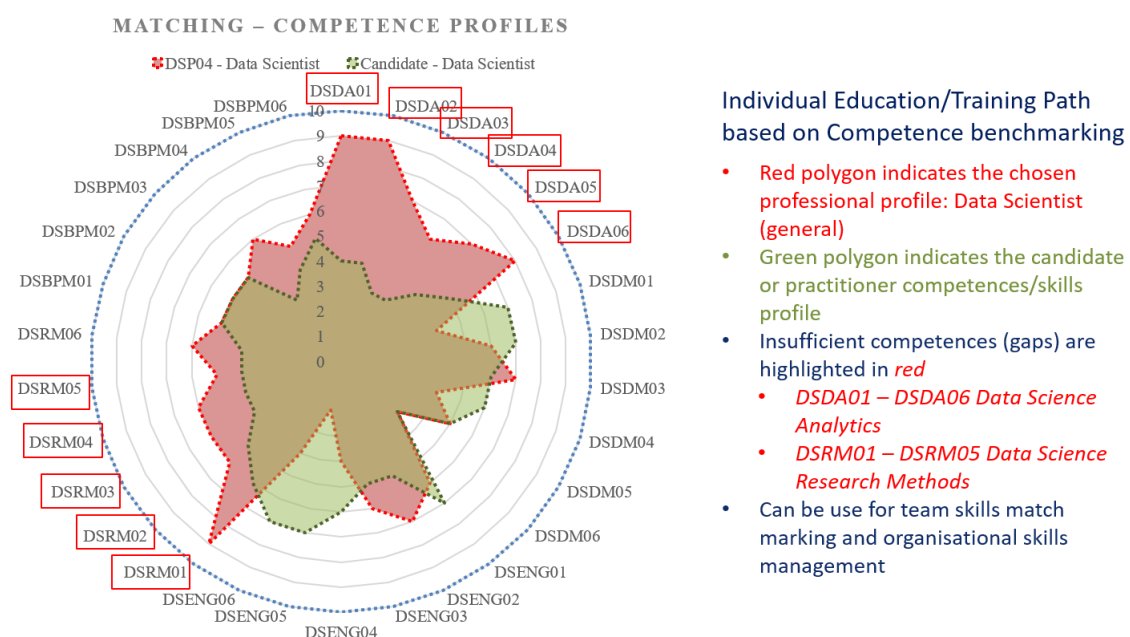


Figure 8. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

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 9 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
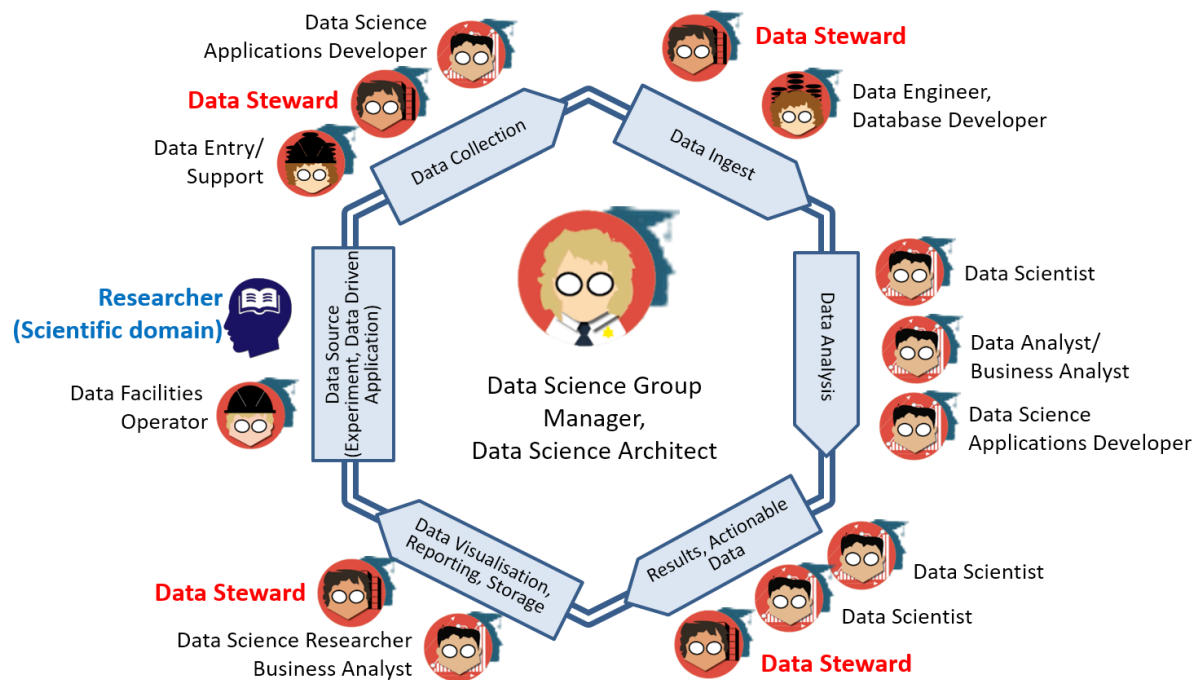
Figure 9. Matching the candidate's competences for the Data Scientist competence profile (as defined in the DSPP document [4])

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.

Similar approach to Data Science and Data Governance roles definition and team building was used in IBM enterprise consulting practice [18].

## 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 the 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 Release 3 update is based on practical use of the proposed profiles for developing educational curricula and professional training programmes for different target groups, and wide community discussion after the past EDSF Release 2 in July 2017.

The proposed document is 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 and consulting 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

## 9   References

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

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

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

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

[5]   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

[6]   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

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

[8]   ESCO handbook: European Skills, Competences, Qualifications and Occupation, European Commission, September 2017. doi:10.2767/934956 [online] https://ec.europa.eu/esco/portal/document/en/0a89839c-098d-4e34-846c-54cbd5684d24

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

[10]  ESCO Strategic Framework, May 2017 [online] https://ec.europa.eu/esco/portal/document/en/89a2ca9a-bc79-4b95-a33b-cf36ae1ac6dbInternational Standard Classification of Occupations (ISCO) Home, http://www.ilo.org/public/english/bureau/stat/isco/

[11]  ISCO-88, the International Standard Classification of Occupations, http://www.ilo.org/public/english/bureau/stat/isco/isco88/index.htm

[12]  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

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

[14]  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/

[15]  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/

[16]  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

[17]  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

[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]  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

[21]  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/

## 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 [13] 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
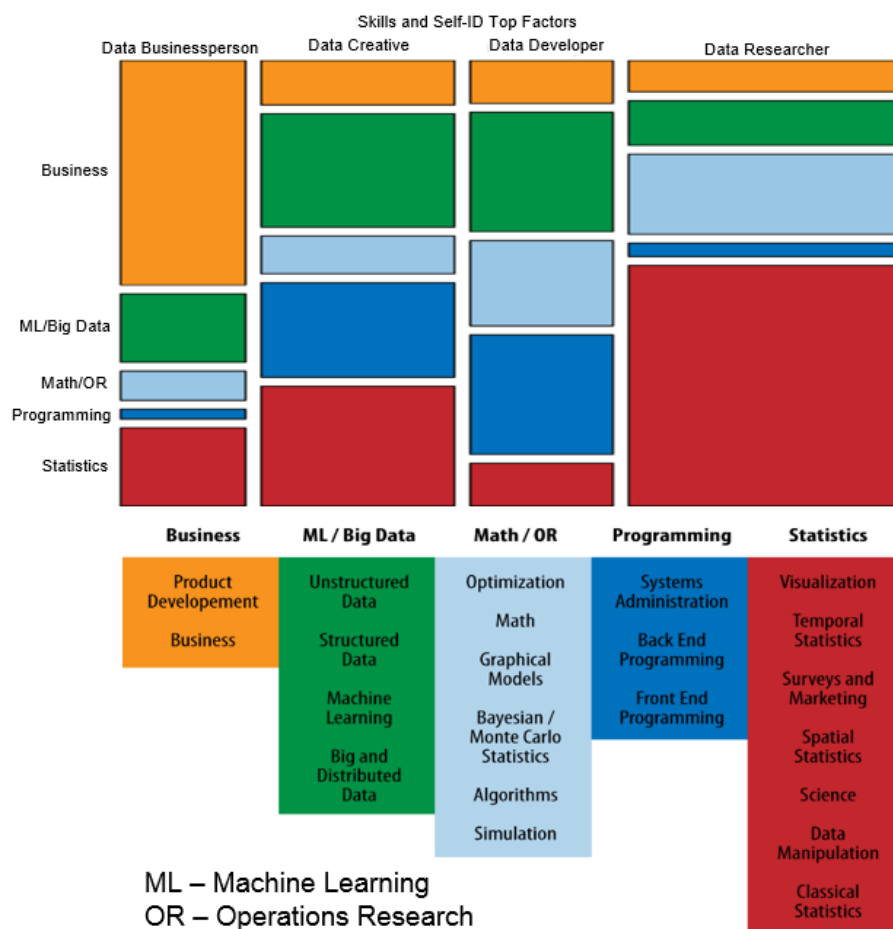- Data Creative
- Data Developer
- Data Researcher



Figure A.1. Data Scientist skills and profiles according to O'Reilly Strata survey [13]

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" [20] 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 [21]. It provides 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. ISCO: Example of ISCO Top Occupation groups

This appendix provides an excerpt from ISCO-08 document to illustrate the structure of the occupations classification.

The ISCO-08 divides jobs into 10 major groups [11]:
1. Managers
2. Professional
3. Technicians and associate professionals
4. Clerical support workers
5. Service and sales workers
6. Skilled agricultural, forestry and fishery workers
7. Craft and related trades workers
8. Plant and machine operators, and assemblers
9. Elementary occupations
10. Armed forces occupations

Major group 2 classification: Professionals

21 Science and engineering professionals
>    211 Physical and earth science professionals
>    212 Mathematicians, actuaries and statisticians
>    213 Life science professionals
>    214 Engineering professionals (excluding electrotechnology)
>    215 Electrotechnology engineers
>    216 Architects, planners, surveyors and designers

22 Health professionals
>    221 Medical doctors
>    222 Nursing and midwifery professionals
>    223 Traditional and complementary medicine professionals
>    224 Paramedical practitioners
>    225 Veterinarians
>    226 Other health professionals
>    227 Medical Assistant professionals

23 Teaching professionals
>    231 University and higher education teachers
>    232 Vocational education teachers
>    233 Secondary education teachers
>    234 Primary school and early childhood teachers
>    235 Other teaching professionals

24 Business and administration professionals
>    241 Finance professionals
>    242 Administration professionals
>    243 Sales, marketing and public relations professionals

25 Information and communications technology professionals
>    251 Software and applications developers and analysts
>    252 Database and network professionals

26 Legal, social and cultural professionals
>    261 Legal professionals
>    262 Librarians, archivists and curators
>    263 Social and religious professionals
>    264 Authors, journalists and linguists
>    265 Creative and performing artists

Classification for unit groups related to Information and communication technologies.

251 Software and Applications Developers and Analysts

2511 Systems Analysts
2512 Software Developers
2513 Web and Multimedia Developers
2514 Applications Programmers

252 Database and Network Professionals
2521 Database Designers and Administrators
2522 Systems Administrators
2523 Computer Network Professionals

351 Information and Communications Technology Operations and User Support Technicians
3511 Information and Communications Technology Operations Technicians
3512 Information and Communications Technology User Support Technicians
3513 Computer Network and Systems Technicians
3514 Web Technicians
352 Telecommunications and Broadcasting Technicians
3521 Broadcasting and Audiovisual Technicians
3522 Telecommunications Engineering Technicians

Other related occupations:

1330 Information and Communications Technology Services Managers
2152 Electronics Engineers
2153 Telecommunications Engineers

2166 Graphic and Multimedia Designers
2356 Information Technology Trainers
2434 Information and Communications Technology Sales Professionals
7422 Information and Communications Technology Installers and Servicers