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

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

Due Date	
Actual Date	10 October 2016
Document Author/s	Yuri Demchenko, Adam Belloum, Tomasz Wiktorski
Version	Release 1, v0.7
Dissemination level	PU
Status	Working document, request for comments
Document approved by	

Document Version Control			
Version	Date	Change Made (and if appropriate reason for change)	Initials of Commentator(s) or Author(s)
0.1	2/11/2015	Initial draft	YD
0.2	4/11/2015	Updated after comments by AB, TW, and with new figures	YD, AB, TW
0.3	6/11/2015	Minor revision and added more focused discussion questions	YD, WL
0.4	30/11/2015	Added information about DS skills based on analysed information, reference to ACM Information Technology Competency model	YD
0.5	30/12/2015	Update version for public comments	YD
0.6	10/03/2016	Re-structured for better readability, extended with the proposed taxonomy of occupations and knowledge areas	YD
0.7	4/07/2016	Updated after D2.2: Revised and added enumerated competences definitions. Data Science professional profiles span off into a separate document on the Data Science Professions family taxonomy	YD
Release 1	10/10/2016	Release 1 after ELG03 meeting discussion	YD

Document Editors: Yuri Demchenko		
Contributors:		
Author Initials	Name of Author	Institution
YD	Yuri Demchenko	University of Amsterdam
AB	Adam Belloum	University of Amsterdam
AM	Andrea Manieri	Engineering
TW	Tomasz Wiktorski	University of Stavanger
WL	Wouter Los	University of Amsterdam

Executive summary

The EDISON project is designed 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) and Data Science Model Curriculum (MC-DS). This will provide a formal basis for Data Science Profession definition and the professional certification, organizational and individual skills management and career transferability.

The definition of the Data Science Competence Framework (CF-DS) is a cornerstone component of the whole EDISON framework. CF-DS will provide a basis for Data Science Body of Knowledge (DS-BoK) and Model Curriculum (MC-DC) definitions, and further for the Data Science Professional certification. The CF-DS is defined in compliance with the European e-Competence Framework (e-CF3.0) and provides suggestions for e-CF3.0 extension with the Data Science related competences and skills.

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 individual or organizational Data Science competences benchmarking and building the customized Data Science education and training program.

This document presents ongoing results of the Data Science Competence Framework definition based on the analysis of existing frameworks for Data Science and ICT competences and skills, and supported by the analysis of the demand side for Data Scientist profession in industry and research.

The presented CF-DS defines five groups of competences for Data Science that include the commonly recognised groups Data Analytics, Data Science Engineering, Domain Knowledge (as defined in the NIST definition of the Data Scientist) and extend them with the two new groups *Data Management* and *Scientific Methods* (or Business Process management for business related occupations) that are recognised to be important for the successful work of Data Scientist but are not explicitly mentioned in existing frameworks.

The identified competences are provided as enumerated list of competence groups. It is also complemented with the related skills (including both hard and soft skills) and a list of required Big Data and analytics tools and programming languages.

The report suggests possible extensions to e-CF3.0 on the Data Science related competences. It also provides observation that e-CF model, which is primarily based on the organisational workflow, is not natively compatible with the Data Science functions that are rather data life cycle oriented.

The document provides an open list of questions for discussion, for which the EDISON project seeks the experts' contribution.

The documents is available for public discussion at the project website at http://edison-project.eu/data-science-competence-framework-cf-ds

TABLE OF CONTENTS

1 Introduction

The definition of the Data Science Competence Framework (CF-DS) is a cornerstone component of the whole EDISON framework. CF-DS will provide a basis for Data Science Body of Knowledge (DS-BoK) and Model Curriculum (MC-DC) definitions, and further for the Data Science Professional certification. It is intended that the CF-DS will be defined in compliance with the European e-Competence Framework (e-CF3.0) and will provide suggestions for e-CF3.0 extension with the Data Science related competences and skills.

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 candidates compliance with a created profile.

This document presents initial results of the Data Science Competence Framework definition based on overview and analysis of existing frameworks for Data Science and ICT competences and skills, and supported by the analysis of the demand side for Data Scientist profession in industry and research.

The presented CF-DS defines the five groups of competences for Data Science that include the commonly recognised groups Data Analytics, Data Science Engineering, Domain Knowledge (as defined in the NIST definition of Data Science) and extends them with the two new groups *Data Management* and *Scientific Methods* (or Business Process management for business related occupations) that are recognised to be important for a successful work of Data Scientist but are not explicitly mentioned in existing frameworks.

The document has the following structure. Section 3 provides an overview of existing 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, ACM Computing Classification System (2012). Furthermore, Section 3 surveys other important components for CF-DS and DS-BoK definition such as data lifecycle management models, scientific methods, and business process management lifecycle models. Section 4 presents the first version of CF-DS that includes identified competence groups and skills, and required technical knowledge of relevant Big Data platforms, analytics and data management tools, and programming languages. Section 5 contains the subset of the ACM CCS (2012) classification with the scientific disciplines (or topics) related to Data Science that are further extended with the new knowledge areas in the Data Science Body of Knowledge.

Appendices to this document contain important supplementary information: overview of known studies, reports and publications related to Data Science competences and skills; and information about data sets used for deriving the proposed CF-DS competences groups and vocabulary.

The document concludes with the suggested further development to finalise the CF-DS definition.

2 EDISON Data Science Framework

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) and Data Science Model Curriculum (MC-DS).

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
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- Data Science Professional profiles and occupations taxonomy (DSP)
- 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 Marketplace and Directory
- 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

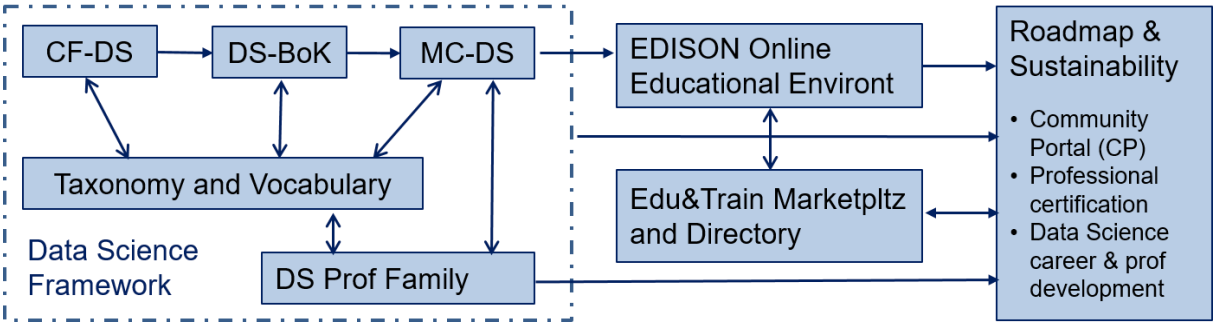


Figure 1 EDISON Data Science Framework components.

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 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. To allow easy use throughout all EDSF and in particular in DS-BoK and MC-DS, the CF-DS competences are enumerated.

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support required 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 revised and updated DS-BoK version used in this deliverable is presented in Appendix C and will be published as a next version of the DS-BoK discussion document after discussion with the EDISON Liaison Group (ELG) experts.

The MC-DS presented in this deliverable 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 DSP profiles and Data Science occupations taxonomy are defined based on and as an extension to European Skills, Competences, Qualifications and Occupations (ESCO). DSP profiles definition will create 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. It is anticipated that successful acceptance of the proposed EDSF and its core components will require standardisation and contacts with the European and international standardisation bodies and professional organisations. This work is being done by the project as a part of the Dissemination and communication activity.

The EDISON Data Science professional ecosystem illustrated in Figure 1 uses core EDSF components shape and profiles the offered services and ensure the EDISON project sustainability. In particular, CF-DS and DS-BoK are used for individual competences and knowledge benchmarking and they are instrumental for constructing personalised learning path and professional (up/re-) skilling 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 Data Science and general Computer Science and ICT competences, skills and subject domain classifications that can be, with some alignment, built upon and re-used for better acceptance from research and industrial communities... The information in this section is also complemented with the overview of individual works and publications to define required Data Science competences and skills placed in Appendix A.

3.1 NIST definition of Data Science

NIST Big Data Working Group (NBD-WG) published their first release of Big Data Interoperability Framework (NBDIF) in September 2015 [1][1] consisting of 7 volumes. Volume 1. Definitions provides a number of definitions in particular Data Science, Data Scientist and Data Life Cycle, which we will use as a starting point for our analysis:

Data science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing. Data science can be understood as the activities happening in the processing layer of the system architecture, against data stored in the data layer, in order to extract knowledge from the raw data.

Data science across the entire data life cycle incorporates principles, techniques, and methods from many disciplines and domains including data cleansing, data management, analytics, visualization, engineering, and in the context of Big Data, now also includes Big Data Engineering. Data science applications implement data transformation processes from the data life cycle in the context of Big Data Engineering.

A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data life cycle.

Data scientists and data science teams solve complex data problems by employing deep expertise in one or more of these disciplines, in the context of business strategy, and under the guidance of domain knowledge. Personal skills in communication, presentation, and inquisitiveness are also very important given the complexity of interactions within Big Data systems.

The **data life cycle** is the set of processes in an application that transform raw data into actionable knowledge.

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

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

Figure 2 from the BDIF publication provide graphical presentation of the multi-factor/multi-domain Data Science definition.

[1] Yuri Demchenko (UvA) contributed to the NBD-WG standardisation process and NBDIF authoring during work on the documents in 2013-2015

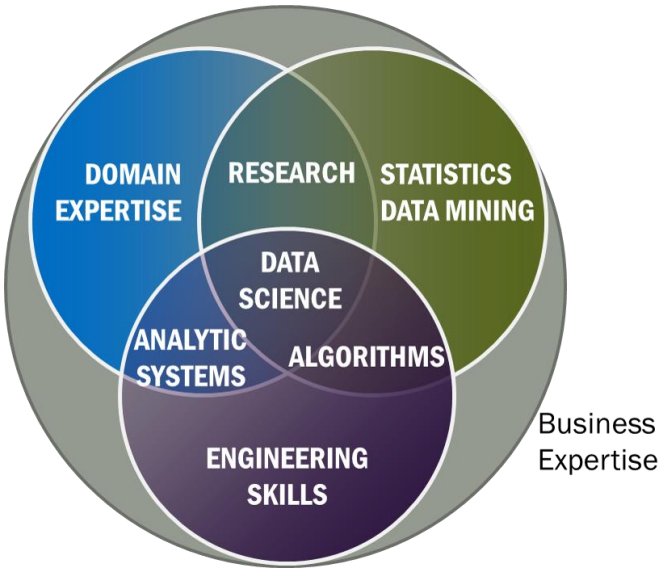


Figure 2. Data Science definition by NIST BD-WG [1].

3.2 European e-Competence Framework (e-CF)

The EDISON CF-DS development (as a part of the WP2 project activity) will follow the European e-Competences Framework (e-CF) guiding principles:

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

The European e-Competence Framework (e-CF) [2, 3, 4] was established as a tool to support mutual understanding and provide transparency of language through the articulation of competences required and deployed by ICT professionals (including both practitioners and managers).

The e-CF is structured from four dimensions:

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

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

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

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

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

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

e-competences in Dimension 1 and 2 are presented from the organisational perspective as opposed to from an individual's perspective. Dimension 3 which defines e-competence levels related to the European Qualifications Framework (EQF), is a bridge between organisational and individual competences. Table 3.1 below contains competences defined for areas A-E. For more detailed definition of e-CF3.0 dimensions 1-3 and dimension 4 refer to the original document [3].

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

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

Figure 3 illustrates the multi-purpose use of the European e-Competence Framework within ICT organisations. The e-CF has a multidimensional structure. The e-CF is a competence-based and is flexible in using for different

purposes, it can be easy adopted for organisation specific model and roles. The alternative, job-profile approach as adopted in CWA 16458 (see next section 3.3), is less flexible, making local adaptation difficult. However, combining competences from different competence areas and using them as building blocks can allow flexible job-profiles definition. This enables the derived job-profiles to be easily updated by substituting or deleting competences without the need to restructure the entire profile.

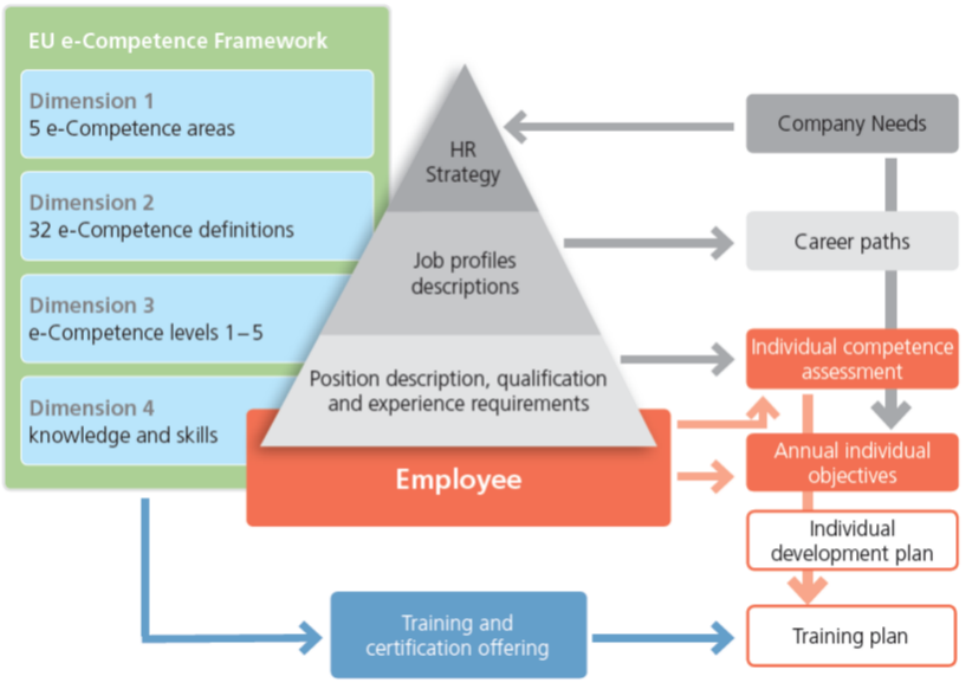


Figure 3. e-CF3.0 structure and use for definition of the job profile definition and training needs.

When using e-CF3.0 as a basis for CF-DS definition, the project will address the following additional goals:

- Link scientific research lifecycle, organizational roles, competences, skills and knowledge
- Define Data Science Body of Knowledge (DS-BoK)
- Map CF-DS competences to DS-BoK and learning outcome in related curricula
- Provide a basis for individual competence profile assessment/benchmarking and suggest necessary (self) training for intended certification

3.3 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 4 illustrates six ICT profile families and related main profiles which are non-exhaustive.

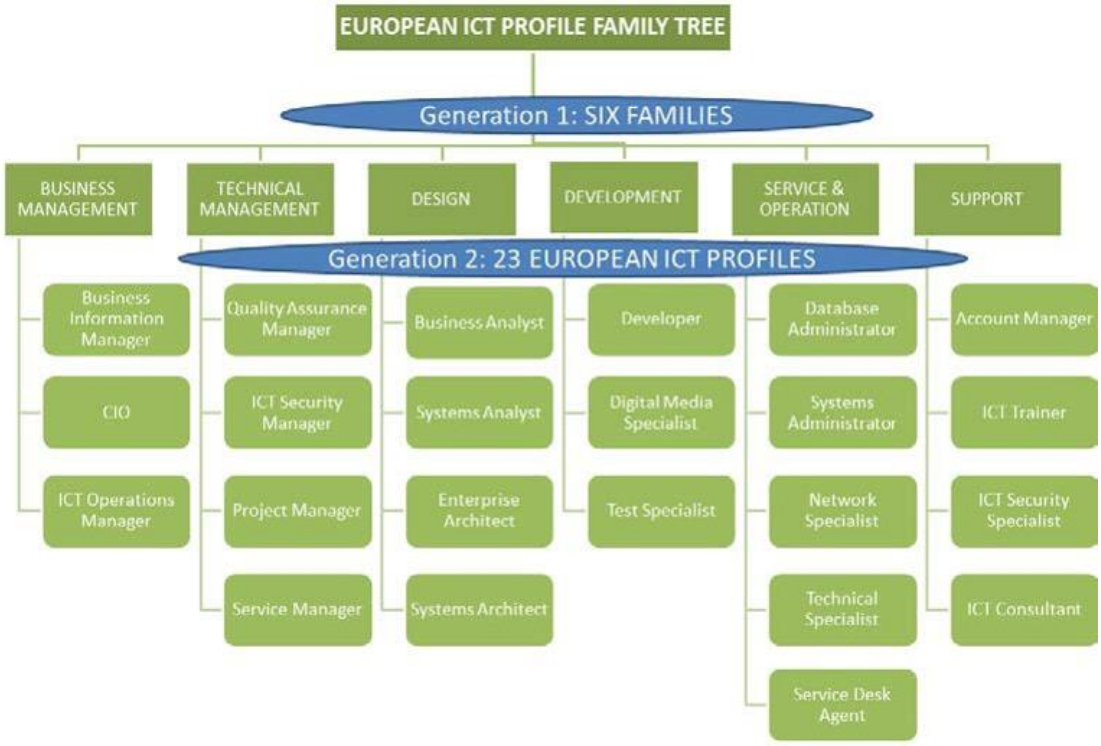


Figure 4. European ICT Profile Family Tree – Generation 1 and 2 as a shared European reference [5]

The 23 profiles constructed in CWA combined with e- competences from the e-CF3.0, 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 (as it is adopted from [5]).

Figure 5 illustrates mapping between CWA families and e-CF3.0 competence areas and also CWA ICT profiles allocation to families and competence areas.

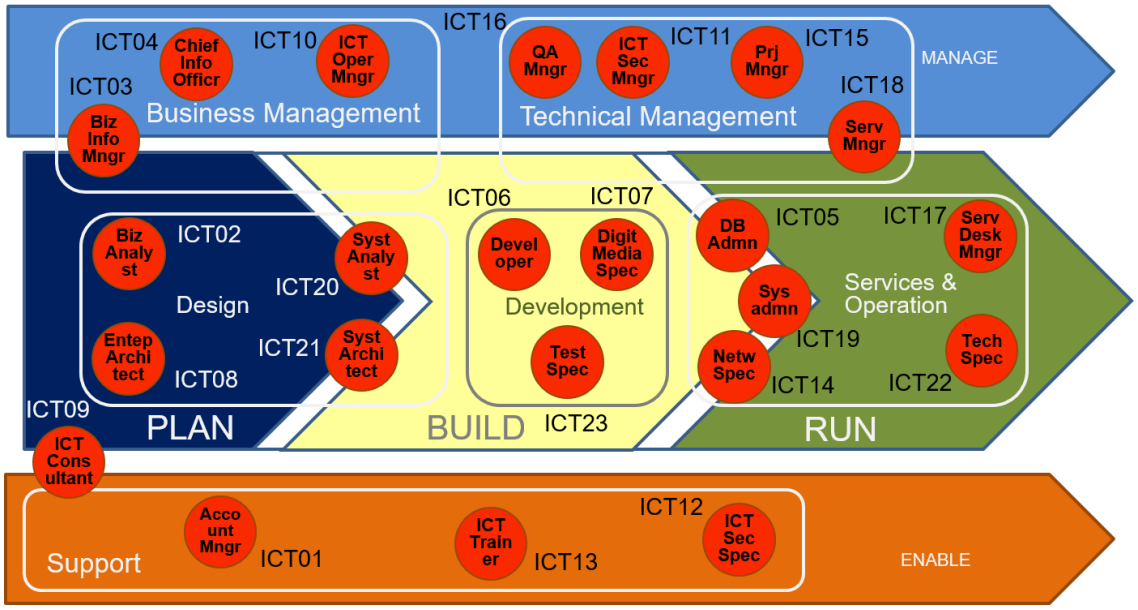


Figure 5. European ICT Professional Profiles structured by six families and positioned within the ICT Business Process (e-CF Dimension 1) (adopted from [5] and extended)

3.4 ACM Computer Science classification

The 2012 ACM Computing Classification System (CCS) [7] has been developed as a poly-hierarchical ontology that can be utilized in semantic web applications. It replaces the traditional 1998 version of the ACM Computing Classification System (CCS), which has served as the de facto standard classification system for the computing field for many years (also been more human readable). The ACM CCS (2012) is being integrated into the search capabilities and visual topic displays of the ACM Digital Library. It relies on a semantic vocabulary as the single source of categories and concepts that reflect the state of the art of the computing discipline and is receptive to structural change as it evolves in the future. ACM provides a tool within the visual display format to facilitate the application of 2012 CCS categories to forthcoming papers and a process to ensure that the CCS stays current and relevant.

However, at the moment none of Data Science, Big Data or Data Intensive Science technologies are explicitly reflected in the ACM classification. The following is an extraction of possible classification facets from ACM CCS (2012) related to Data Science what reflects multi-subject areas nature of Data Science:

Table 3.2. ACM Classification (2012) facets related to Data Science

ACM (2012) Classification facets related to Data Science
Computer systems organization Architectures Parallel architectures Distributed architectures Cloud Computing
Networks *) Network Architectures Network Protocols Network Components Network Algorithms Network Properties Network Services Cloud Computing
Software and its engineering Software organization and properties Software system structures Software architectures Software system models Ultra-large-scale systems Distributed systems organizing principles Cloud computing Abstraction, modeling and modularity Software notations and tools General programming languages Software creation and management
Mathematics of computing Probability and statistics Probabilistic representations Probabilistic inference problems Probabilistic reasoning algorithms Probabilistic algorithms Statistical paradigms Mathematical software Information theory Mathematical analysis
Information systems Data management systems Database design and models

- Data structures
- Database management system engines
- Query languages
- Information integration
- Information storage systems
- Information systems applications
  - Enterprise information systems
  - Collaborative and social computing systems and tools
  - Decision support systems
    - Data warehouses
    - Expert systems
    - Data analytics
    - Online analytical processing
  - Multimedia information systems
  - Data mining
  - Digital libraries and archives
- Information retrieval
  - Specialized information retrieval

Computing methodologies

- Artificial intelligence
  - Natural language processing
  - Knowledge representation and reasoning
  - Search methodologies
- Machine learning
  - Learning paradigms
    - Supervised learning
    - Unsupervised learning
    - Reinforcement learning
    - Multi-task learning
  - Machine learning approaches
  - Machine learning algorithms
- Modeling and simulation
  - Model development and analysis
  - Simulation theory
  - Simulation types and techniques
  - Simulation support systems

Applied computing

- Physical sciences and engineering
- Life and medical sciences
- Law, social and behavioural sciences
- Computer forensics
- Arts and humanities
- Computers in other domains
- Operations research
- Education
- Document management and text processing

Social and professional topics

- Professional topics
  - Management of computing and information systems
  - Computing education
  - Computing and business
  - Computing profession
- Computing / technology policy
- User characteristics

*) Included Networks hierarchy group is for information for general overview of the ACM CCS2012 overview. Due to complexity of this technology domain, Networks group can be considered as a domain knowledge domain in the Data Science competences and knowledge definition.

As an example, the Cloud Computing that is also a new technology and closely related to Big Data technologies, currently is classified in ACM CCS (2012) into 3 groups:

> **Networks** :: Network services :: Cloud Computing
> **Computer systems organization** :: Architectures :: Distributed architectures :: Cloud Computing
> **Software and its engineering** :: Software organization and properties :: Software Systems Structures :: Distributed systems organizing principles :: Cloud Computing

It is anticipated that based on current study the project will propose necessary extensions to correctly reflect knowledge areas and academic subjects related to Data Science (see chapter 5 for initial suggestions).

3.5 ACM Information Technology Competencies Model

The ACM Information Technology Competency Model (IT-CM) of Core Learning Outcomes and Assessment for Associate-Degree Curriculum (2014) has been developed by ACM Committee for Computing Education in Community Colleges (ACM CCECC) [8-11].

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

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

The document refers to the U.S. Department of Labour Information Technology Competency Model [19] that was one of sources that provided a foundation for the curricular guidance outlined in IT-CM report

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

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

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

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

- ACM Computing Classification System 2012
- U.S. Dept. of Labor IT 2012 Competency Model [12]
- Bloom's Revised Taxonomy [13]
- European E-Competence Framework 3.0 (Proficiency Levels 1 & 2)

3.6 Components and concepts related to CF-DS definition

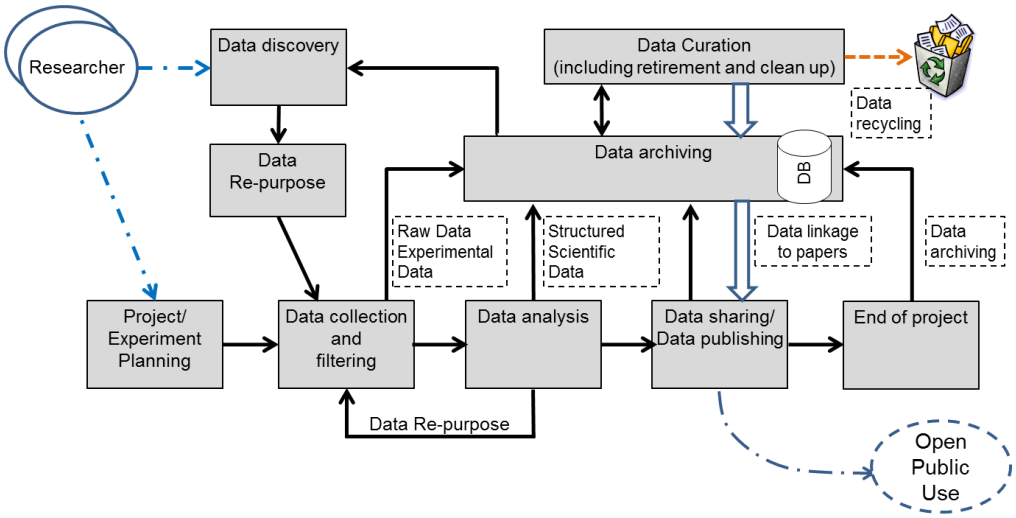
This section provides important definitions that are needed for consistent CF-DS definition in the context of organisational and business processes, e-Infrastructure and scientific research. First of all, this includes definition of typical organisational processes and scientific workflow or research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

3.6.1 Scientific Data Lifecycle Management Model

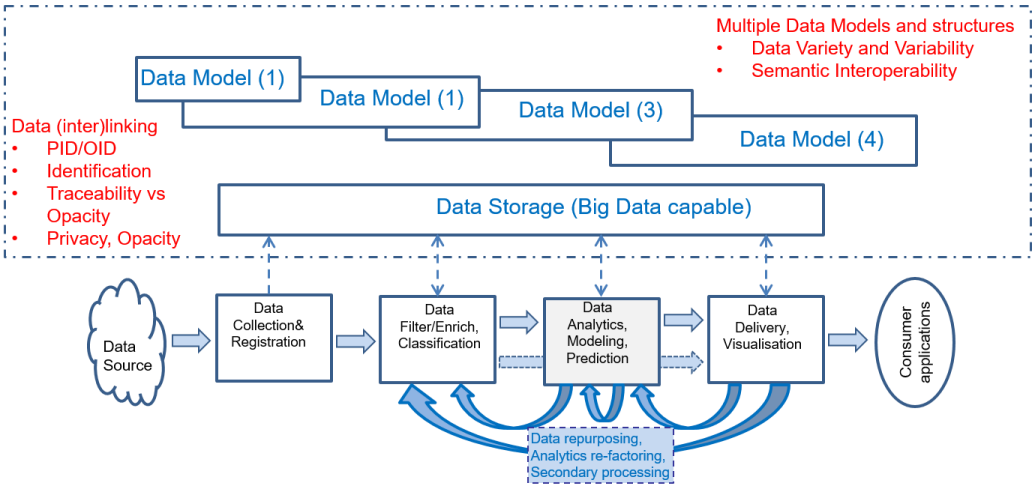
Data lifecycle is an importance component of data centric applications, which Data Science and Big Data applications belong to. Data lifecycle analysis and definition is addressed in many domain specific projects and studies. Extensive compilation of the data life cycle models and concepts is provided in the CEOS.WGISS.DSIG document [14].

For the purpose of defining the major groups of competences required for Data Scientist working with scientific applications and data analysis we will use the Scientific Data Lifecycle Management (SDLM) model [15] shown in Figure 6 (a) defined as a result of analysis of the existing practices in different scientific communities. Figure 6 (b) illustrates the more general Big Data Lifecycle Management model (BDLM) involving the main components of the Big Data Reference Architecture defined in NIST BDIF [1, 16, 17]. The proposed models are sufficiently generic and compliant with the data lifecycle study results presented in [14].

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



(a) Scientific data lifecycle management - e-Science focused



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

Figure 6. Data Lifecycle Management in (a) e-Science and (b) generic Big Data Lifecycle Management model.

Both SDLM and BDLM require data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in Scientific Data Infrastructure (SDI). Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the discussed data lifecycle models and must also be done in a secure and trustworthy way. The research data management and handling issues are extensively addressed in the work of the Research Data Alliance[2].

3.6.2 Scientific methods and data driven research cycle

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

The scientific method is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge [18, 19, 20]. Traditional steps of the scientific research were developed over time since the time of ancient Greek philosophers through modern theoretical and experimental research where experimental data or simulation results were used to validate the hypothesis formulated based on initial observation or domain knowledge study. The general research methods include: observational methods, opinion based methods, experimental and simulation methods.

The increased power of computational facilities and advent of Big Data technologies created a new paradigm of the data driven research that enforced ability of researchers to make observation of the research phenomena based on bigger data sets and applying data analytics methods to discover hidden relations and processes not available to deterministic human thinking. The principles of the data driven research were formulated in the seminal work “The Fourth Paradigm: Data-Intensive Scientific Discovery” edited by Tony Hey [21].

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

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

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

- Test Hypothesis
- Refine model and start new experiment cycle

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

The proposed CF-DS introduces research methods as an important component of the Data Science competences and knowledge and uses data lifecycle as an approach to define the data management related competences group. This is discussed in section 3.4 below.

3.6.3 Business Process Management lifecycle

New generation Agile Data Driven Enterprises (ADDE) [22] use Data Science methods to continuously monitor and improve their business processes and services. The data driven business management model allows combining different data sources to improve predictive business analytics what allows making more effective solutions, faster adaptation of services, and more specifically target different customer groups as well as do optimal resources allocations depending on market demand and customer incentives.

Similarly, to the research domain the data driven methods and technologies change how the modern business operates attempting to benefit from the new insight that big data can give into business improvement including internal processes organisation and relation with customers and market processes. Understanding Business Process Management lifecycle [23, 24] is important to identify necessary competences and knowledge for business oriented Data Science profiles.

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

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

The need for the Business Process management competences and knowledge for business oriented Data Science profiles is reflected in the proposed CF-DS definition as described in section 3.4.

4 EDISON definition of the Data Science Competence Framework (CF-DS)

This section describes the first version of the Data Science Competence Framework (CF-DS). It serves as a foundation for the definition of the Data Science Body of Knowledge and Data Science Model Curriculum. The section explains the approach used to analyse the collected data and information related to existing Data Science competence and skill frameworks. The reviewed in the previous section existing competences and skills models and definitions provided limited view and did not allow the consistent CF-DS definition that would reflect real job market demand in Data Science competences and skills. The presented here results have been achieved by combining initial desk research based on the existing publications, works and standards, and deep analysis of sample set of data from job advertisements on IEEE Jobs portal and LinkedIn jobs advertised in Netherlands, that allowed us to identify commonly recognised competence and skill groups and identify new competence groups not recognised in previous works and studies. The presented CF-DS model and taxonomy of competences will be used for extended analysis and collection of information from technology or scientific domains and different geographical regions and countries.

4.1 Relation to and use of existing framework and studies

The following describes how existing frameworks and documents were used for the analysis of initial competences and skills.

a) NIST NBDIF Data Science and Data Scientist definition [1]

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

b) European e-Competence Framework (e-CFv3.0) [3]

e-CF3.0 provided a general framework for ICT competences definition and possible mapping to Data Science competences. However, it appeared that current e-CF3.0 doesn't contain competences that reflect specific Data Scientist role in organisation. Furthermore, e-CF3.0 is built around organisational workflow while anticipated Data Scientist's role is cross-organisational bridging different organisational roles and departments in providing data centric view or organisational processes.

c) European ICT profiles CWA 16458 (2012) [4]

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

d) European Skills, Competences, Qualifications and Occupations (ESCO) [6]

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

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

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

f) O'Reilly Strata Survey (2013) [25]

It was a first extensive study on Data Scientist organisational roles, profiles and skills. Although skills are defined as very technically and technologically specific, the proposed definition of profiles is important for defining

required competence groups, in particular identification of Data Science Creative and Data Science Researcher profiles indicates an important role of scientific approach and need for research method training in Data Scientist professional education. This group of competences is included in the proposed CF-DS.

g) EC Report on the Consultation Workshop (May 2012) “Skills and Human Resources for e-Infrastructures within Horizon 2020” [26].

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

4.2 Selecting sources of information

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

- First of all, job advertisements that represent demand side for Data Scientist specialists and based on practical tasks and functions that are identified by organisations for specific positions. This source of information provided factual data to define demanded competences and skills.
- Structured presentation of Data Science related competences and skills produced by different studies as mentioned above, in particular NIST definition of Data Science that provided a basis for definition of initial 3 groups of skills, namely Data Analytics, Data Science Engineering, and Domain expertise. This information was used to correlate with information obtained from job advertisements.
- Blog articles and community forums discussions that represented valuable community opinion. This information was specifically important for defining practical skills and required tools.

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

- If referred to the category of job openings such as academic positions or industry and business related positions, the academic positions didn’t provide valuable information as they don’t specify detailed competences and skills but rather search for candidates who are capable to teach, create or support new academic courses on Data Science.
- In this initial stage we used set of Data Science job openings from IEEE Jobs portal (around 120) and LinkedIn Netherlands (around 140) collected in period of mid-September to beginning of October 2015. A number of Data Science related key words were used like Data Science, Big Data, Data Intensive technologies, data analytics, machine learning. Initial analysis of collected information allowed to make assumption that collected information from more than 250 samples was sufficiently representative for initial study, taking into account that Netherlands is one of leading countries in relation to Big Data and Data Science technologies acceptance and development. See Appendix B for more details about collected data.

4.3 EDISON approach to analysis of collected information

1) Collect data on required competences and skills
2) Extract information related to competences, skills, knowledge, qualification level, and education; translate and/or reformulate if necessary
3) Split extracted information on initial classification or taxonomy facets, first of all, on required competences, skills, knowledge; suggest mapping if necessary

4) Apply existing taxonomy or classification: for purpose of this study we used skills and knowledge groups as defined by the NIST Data Science definition (i.e. Data Analytics, Domain Knowledge, and Engineering) [1]
5) Identify competences and skills groups that don't fit into initial/existing taxonomy and create new competences and skills groups
6) Do clustering and aggregations of individual records/samples in each identified group
7) Verify the proposed competences groups definition by applying to originally collected and new data
8) Validate the proposed CF-DS via community surveys and individual interviews[3].

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

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

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

4.4 Identified Data Science Competence Groups

The results of analysis presented here provides a basis and justification for defining two (new) competence areas that have not been explicitly defined in previous studies and frameworks. In particular, the proposed CF-DS competence and skills groups include:
3 competence groups identified in the NIST document and confirmed by analysis of collected data:

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Engineering: software and infrastructure
- Subject/Scientific Domain competences and knowledge

2 new identified competence groups that are highly demanded and specific to Data Science

- *Data Management, Curation, Preservation*
- *Scientific or Research Methods for research related professions and Business Process Management for business related professions*

Data Management, curation and preservation is already being conducted within 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 policy. It is extensively addressed by the Research Data Alliance and supported numerous projects, initiatives and training programmes[4].
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 will be discussed in more detail as part of the definition of the Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS).

New identified competence areas provide a better basis for defining education and training programmes for Data Science related jobs, re-skilling and professional certification.

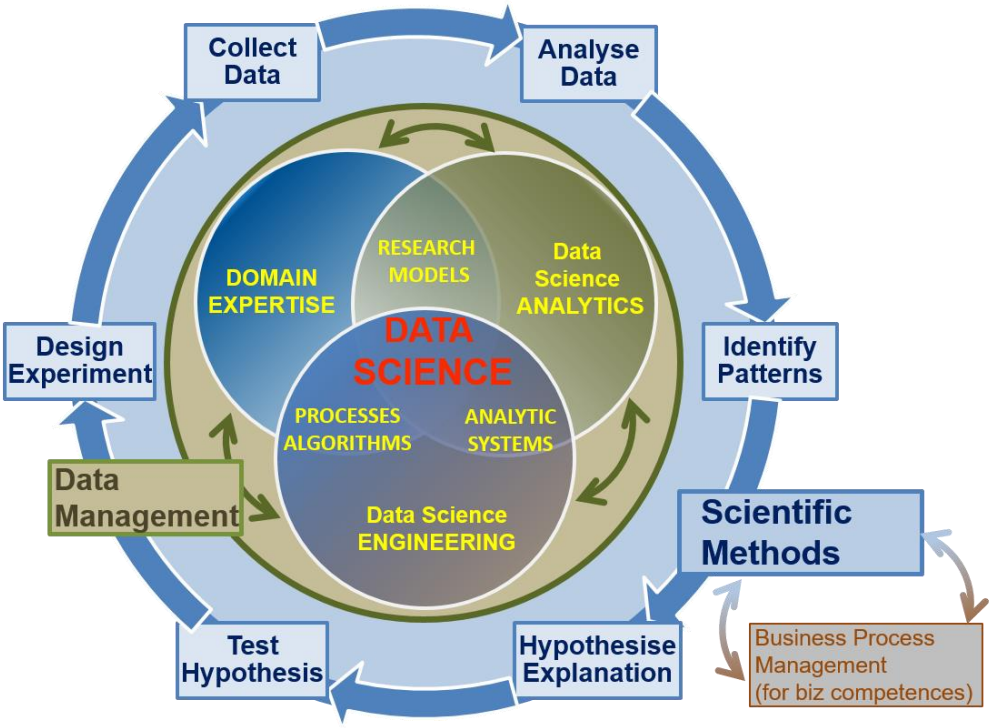
Table 4.1 provides an example of competences definition for different groups that are extracted from the collected information. It indicates that all identified groups are demanded by companies and they have expected place in the corporate structure and activities. The presented competences are enumerated to

[3] This activity will be done at the next stage and results will be reported in the final deliverable D2.3
[4] Research Data Alliance Europe https://europe.rd-alliance.org/

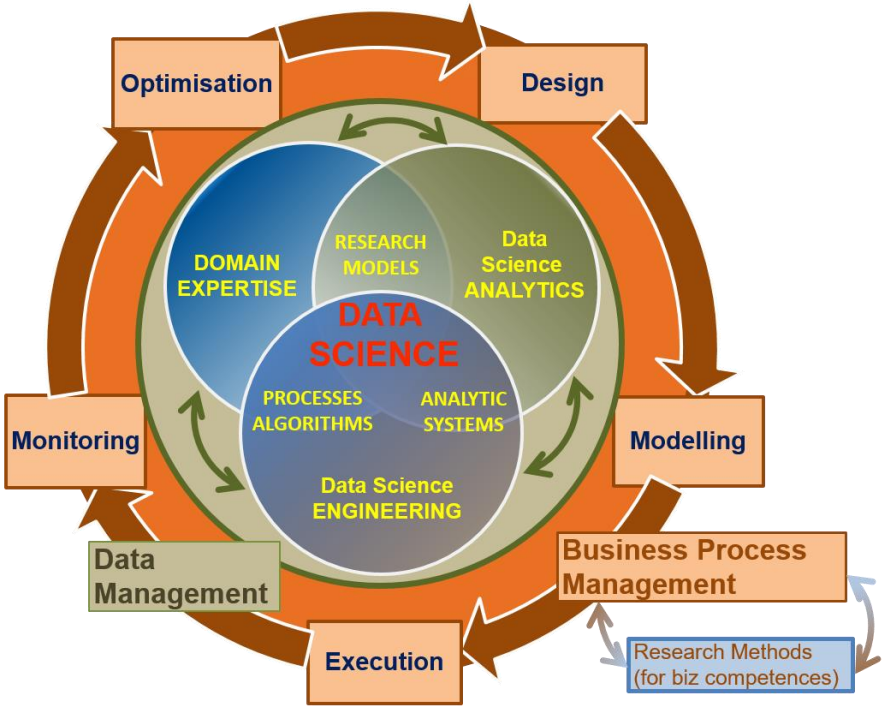
Table 4.1. Competences definition for different Data Science competence groups

Data Science Analytics (DSDA)	**Data Management (DSDM)**	**Data Science Engineering (DSENG)**	**Scientific/ Research Methods (DSRM)**	**DS Domain Knowledge, e.g., Business Apps (DSDK)**
Use appropriate statistical techniques and predictive analytics on available data to deliver insights and discover new relations	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	Use engineering principles to research, design, develop and implement new instruments and applications for data collection, analysis and management	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	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, organisational roles and relations
DSDA01 Use predictive analytics to analyse big data and discover new relations	DSDM01 Develop and implement data strategy, in particular, in a form of Data Management Plan (DMP)	DSENG01 Use engineering principles to research, design, prototype, data analytics applications, or develop structures, instruments, machines, experiments, processes, systems	DSRM01 Create new understandings and capabilities by using the scientific method (hypothesis, test, and evaluation) or similar engineering research and development methods	DSDK01 Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
DSDA02 Use appropriate statistical techniques on available data to deliver insights	DSDM02 Develop and implement relevant data models, including metadata	DSENG02 Develop and apply computational solutions to domain related problems using wide range of data analytics platforms	DSRM02 Direct systematic study toward a fuller knowledge or understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSDK02 Use data to improve existing services or develop new services
DSDA03 Develop specialized analytics to enable agile decision making	DSDM03 Collect and integrate different data source and provide them for further analysis	DSENG03 Develops specialized data analysis tools to support executive decision making	DSRM03 Undertakes creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications	DSDK03 Participate strategically and tactically in financial decisions that impact management and organizations
DSDA04 Research and analyse complex data sets, combine different sources and types of data to improve analysis.	DSDM04 Develop and maintain a historical data repository of analysis results (data provenance)	DSENG04 Design, build, operate relational non-relational databases	DSRM04 Ability to translate strategies into action plans and follow through to completion.	DSDK04 Provides scientific, technical, and analytic support services to other organisational roles
DSDA05 Use different data analytics platforms to process complex data	DSDM05 Ensure data quality, accessibility, publications (data curation)	DSENG05 Develop solutions for secure and reliable data access	DSRM05 Contribute to and influence the development of organizational objectives	DSDK05 Analyse customer data to identify/optimise customer relations actions
DSDA06 Visualise complex and variable data.	DSDM06 Manage IPR and ethical issues in data management	DSENG06 Prototype new data analytics applications	DERM06 Apply ingenuity to complex problems, develop innovative ideas	DSDK06 Analyse multiple data sources for marketing purposes

Figures 7 (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 Data Management competences and skills and Scientific/Research Methods or Business Processes 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 7. 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 Scientific Methods typically include the following stages (see section 3.6.2 for reference to existing scientific methods definitions):

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

There are a number 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 in section 3.6.3):

- 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 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.

In summary, consolidation of the presented initial version of CF-DS was challenging task due to variety of information and required expertise. It is anticipated that it will undergo further improvement involving external and subject domain experts via EDISON Liaison Group and community involvement. Other project activities will provide feedback on necessary improvements and details.

It is a challenging task to include all required subjects and knowledge into education and training programs. It will require search for new approaches in Data Science education what will be a subject for subsequent EDISON project activities and work items.

4.5 Identified Data Science Skills

Another outcome of the current analysis of Data Science job advertisements and numerous blog articles[5] is the identification of required/demanded Data Science skills (also referred to as Big Data skills) that can split on 3 groups:

- General Data Science skills or required (provable) experience
- Knowledge and experience with Big Data hardware and software platforms
- Programming language: general and those having extended statistics libraries that are generally related to Data Science Engineering skills but in many cases are treated as a separate group of skills.

It is essential to mention that for such complex professional domain as Data Science the minimum required experiences with related methods, tools or platform is 2-3 years, some companies explicitly require experience up to 5 years.

Table 4.2 lists identified Data Science skills related to the following groups

- Data Analytics and Machine Learning

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

- Data Management/Curation (including both general data management and scientific data management)
- Data Science Engineering (hardware and software) skills
- Scientific/Research Methods
- Personal, inter-personal communication, team work (also called social intelligence or soft skills)
- Application/subject domain related (research or business)

The Data Analytics and Machine Learning group is the most populated what reflect real picture of required skills primary in this area as a basis for Data Science methods.

Table 4.2. Identified Data Science skills

Skill Groups	Data Analytics and Machine Learning	Data Management/ Curation	Data Science Engineering (hardware and software)	Scientific/ Research Methods	Personal/Inter-personal communication, team work	Application/subject domain (research or business)
1	Artificial intelligence, machine learning	Manipulating and analysing complex, high-volume, high-dimensionality data from varying sources	Design efficient algorithms for accessing and analysing large amounts of data	Analytical, independent, critical, curious and focused on results	Communication skills	Recommender or Ranking system
2	Machine Learning and Statistical Modelling	Data sources and techniques for data improvement for better business insight and customer focus	Big Data solutions and advanced data mining tools	Confident with large data sets and ability to identify appropriate tools and algorithms	Inter-personal intra-team and external communication	Data Analytics for commercial purposes
3	Machine learning solutions and pattern recognition techniques	Data models and datatypes	Multi-core/distributed software, preferably in a Linux environment	Flexible analytic approach to achieve results at varying levels of precision	Network of contacts in Big Data community	Data sources and techniques for business insight and customer focus
4	Supervised and unsupervised learning	Experience of working with large data sets	Databases, database systems, SQL and NoSQL	Exceptional analytical skills and interest in data science		Mechanism Design and/or Latent Dirichlet Allocation
5	Data mining	(non)relational and (un)-structured data	Statistical analysis languages and tooling			Game Theory
6	Markov Models, Conditional Random Fields	Cloud based data storage and data management	Cloud powered applications design			Copyright and IPR
7	Logistic Regression, Support Vector Machines	Data management planning				
8	Predictive analysis and statistics (including Kaggle platform)	Metadata annotation and management				
9	(Artificial) Neural Networks	Data citation, metadata, PID (*)				
10	Statistics					

11	Natural language processing					
12	Computer Simulation					

(*) Persistent Identifier (PID), Data Types Registries, and Data Management Policies are outcome and products by Research Data Alliance (RDA) [27]

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

Table 4.3 lists identified required skills and knowledge of the Big Data platforms (hardware and software) divided into groups:

- Big Data Analytics platforms
- Math& Stats tools
- Databases
- Data Management and Curation platform
- Data and applications visualisation

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

Table 4.3. Required skills and knowledge of popular Big Data platforms and tools (hardware and software) [6]

	Big Data Analytics platforms	Math& Stats tools	Databases	Data/ applications visualization	Data Management and Curation platform
1	Big Data Analytics platforms	Advanced analytics tools (R, SPSS, Matlab, etc)	SQL and relational databases	Data visualization Libraries (D3.js, FusionCharts, Chart.js, other)	Data modelling and related technologies (ETL, OLAP, OLTP, etc)
2	Big Data tools (Hadoop, Spark, etc)	Data Mining tools: RapidMiner, others	NoSQL Databases	Visualisation software (D3, Processing, Tableau, Gephi, etc)	Data warehouses platform and related tools
3	Distributed computing tools a plus (Spark, MapReduce, Hadoop, Hive, etc.)	Matlab	NoSQL, Mongo, Redis	Online visualization tools (Datawrapper, Google Charts, Flare, etc)	Data curation platform, metadata management (ETL, Curator's Workbench, DataUp, MIXED, etc)
4	Real time and streaming analytics systems (like Flume, Kafka, Storm)	Python	NoSQL, Teradata		Backup and storage management (iRODS, XArch, Nesstar, others)
5	Hadoop Ecosystem/platform	R, Tableau R	Excel		
6	Spotfire	SAS			
7	Azure Data Analytics platforms (HDInsight, APS and PDW, etc)	Scripting language, e.g. Octave			
8	Amazon Data Analytics platform (Kinesis, EMR, etc)	Statistical tools and data mining techniques			

[6] The presented here Big Data platforms and tools are examples of the most popular platforms and tools and are not exhaustive. Please search for general and domain specific other general and domain specific reviews and inventories, for example: Data Science Knowledge Repo https://datajobs.com/data-science-repo/

9	Other cloud based Data Analytics platforms (HortonWorks, Vertica LexisNexis HPCC System, etc)	Other Statistical computing and languages (WEKA, KNIME, IBM SPSS, etc)			

*) Highlighted are cloud based and online data analytics and data management platforms that are becoming increasingly popular for enterprise and business applications.

The following programming languages with extended data analysis and statistics libraries are identified as important for Data Scientist (and typically identified in job descriptions as requiring several years of practical experience)[7]:

- Python
- Scala
- pandas (Python Data Analysis Library)
- Julia
- Java and/or C/C++ as general applications programming languages
- Git versioning system as a general platform for software development
- Scrum agile software development and management methodology and platform

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

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

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

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

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

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

4.7 Other results and recommendations

4.7.1 Data Scientist inter-personal skills

Our analyses clearly identified and confirmed importance of inter-personal skills (recently also referred to as a social intelligence) for successful Data Scientist work due to cross-organisational character of their work and responsibilities. Some job advertisements also mention that it is expected that Data Scientist is able and will intend to educate the colleagues and staff members on the basic concepts and knowledge in data analytics and statistics. We intend to treat this as a wider set of characteristics than traditionally referred to as team worker. The ideal Data Scientist is expected to bring and spread new knowledge and ensure that all benefit and contribute to the work related to data collection, analysis and exploitation as main responsibility of Data Scientist.

4.7.2 Data Scientist mission/expectation in organisation

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

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

4.7.3 Relation between Data Scientist and Subject Domain specialist

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

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

- Model (and data types)
- Methods
- Processes
- Domain specific data and presentation/visualization methods (?)
- Organisational roles and relations

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

- Translate subject domain Model, Methods, Processes into abstract data driven form
- Implement computational models in software, build required infrastructure and tools
- Do (computational) analytic work and present it in a form understandable to subject domain
- Discover new relations originated from data analysis and advice subject domain specialist

- Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data

Figure 8 illustrates relations between subject domain components and those mapped to Data Science domain which is abstract, formalised and data driven.

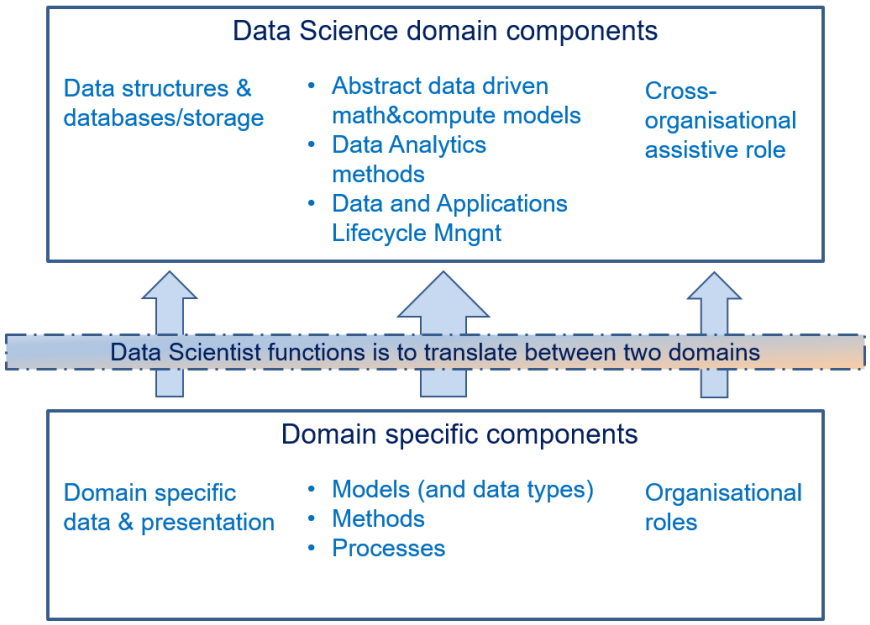


Figure 8. Relations between subject domain and Data Science domain and role of Data Scientist.

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

4.7.4 Needs for general Data Science literacy in organisations

Following observation that organisations expect that Data Scientist will bring general Big Data and Data Science knowledge to organisation, we can identify a need for the general Data Science literacy in organisation. This means that management and all workers need to obtain general knowledge on data management, curation, data presentation and structures, understand data analytics and other tools that are used by Data Science.

This should motivate general Data Science literacy training in organisations what should be a responsibility of the management. Such training should also focus on a general ways of data presentation and visualisation and effective communication and task formulation for data analytics.

4.8 Usage example: RDA IG-ETRD definition of competences and skills for e-Infrastructure management (ICT/technical)

Research Data Alliance Interest Group on Education and Training on Handling of Research Data (IG-ETRD) conducts ongoing community research to define competences and skills for the following 4 areas related to e-Infrastructure and research data handling [28]:

1. Digital libraries
2. Research management (administrative)
3. Research Infrastructure (e-Infrastructure) management and operation (technical)
4. Researcher

Research Infrastructure (e-Infrastructure) management and operation is considered as closely related to ICT profiles and related e-CF competences, however having specific focus and stakeholders from scientific and research area and community.

Initial results suggested refactoring of some e-CF3.0 competence areas and individual competences and adding new competences as shown in Table 4.5. The competences are ordered by relevance and essential competences are highlighted in bold.

Table 4.5. Proposed RI management (ICT/technical) competence profile based on e-CF3.0

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

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

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

5 Conclusion and further developments

This deliverable presents the first versions of the Data Science Competence Framework, Data Science Body of Knowledge and Data Science Taxonomy that all together provide the basis all further developments in the project, namely Data Science Model Curriculum, Data Science Certification scheme and EDISON Online Education Environment. The presented versions have been created based on extensive analysis of available information that includes Data Science job market study (primarily demand side, i.e. job advertisement), existing standards, best practices, academic publications and blog articles that are posted by experts, practitioners and enthusiasts of the new technology domain and profession of Data Scientist.

The focused work on defining all the foundational components of the whole EDISON framework for consistent Data Science profession definition have been done with wide consultation and engagement of different stakeholders, primarily from research community and Research Infrastructures, but also involving industry via standardisation bodies, professional communities and directly via the project network.

5.1 Summary of findings

The provided document contains initials results of the Data Science Competences Framework definition that already revealed two competence areas that have not been explicitly defined in previous studies and definitions. In particular the proposed CF-DS competence and skills groups include:

3 competence groups identified in the NIST document and confirmed by analysis of collected data:

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Engineering: software and infrastructure
- Subject/Scientific Domain competences and knowledge

2 new identified competence groups that are highly demanded and are specific to Data Science

- *Data Management, Curation, Preservation (new)*
- *Scientific or Research Methods (new)*

The identified competences normally can be achieved in the team of Data Scientists with different profiles or roles, however each of Data Scientist profiles must be aware and have basic knowledge of other domains, in particular data management and research methods as anticipated necessary background knowledge. This creates obvious challenge to education and training future Data Scientists to find a solution for preparing future generation of Data Scientists what is one of EDISON objectives.

The proposed study suggested a number of new competences specific to Data Science the can be added to current e-CF3.0 (see section 4.6). This proposal has been presented to the CEN e-Competence Workshop at their meeting on 9 December 2015 in Paris.

5.2 Further developments to formalize CF-DS

It is anticipated that the presented here the first versions of the Data Science Competence Framework, Data Science Body of Knowledge and Data Science Taxonomy will require further development and validation by experts and communities of practice that will include the following specific tasks and activities:

Data Science taxonomy development and formalisation:

- Finalise the taxonomy definition for the Data Science related occupations 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
- Finalise the taxonomy of Data Science related knowledge areas and scientific disciplines based on ACM CCS (2012), provide suggestion for new knowledge areas and classifications classes.

CF-DS related development:

- Define a taxonomy and classification for Data Science competences, skills and knowledge areas as a basis for more formal CF-DS definition including other components of the intended CF-DS such as proficiency levels (mapped to the expected role of Data Scientist and position seniority), skills and knowledge definition

and mapping to individual competences. The intended taxonomy and classification should be compatible with existing frameworks and taxonomies such as e-CF3.0 and ACM CCS2013.

- Create a Questionnaire using CF-DS vocabulary and run a survey among target EDISON communities, first of all, among EGI community, to validate the proposed CF-DS and collect information from multiple scientific and industry domains.
- Use the constructed questionnaire to run few key interviews, primarily experts and top executives at universities and companies to understand intended use of CF-DS, identify case studies and identify necessary extensions.
- Provide guidelines for intended CF-DS usage: job profile and vacancy description generation; individual competences and skills assessment, certification profiles, and others.

6 References

[1] 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

[2] European eCompetences Framework http://www.ecompetences.eu/

[3] 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

[4] User guide for the application of the European e-Competence Framework 3.0. CWA 16234:2014 Part 2. [online] http://ecompetences.eu/wp-content/uploads/2014/02/User-guide-for-the-application-of-the-e-CF-3.0_CEN_CWA_16234-2_2014.pdf

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

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

[8] ACM and IEEE Computer Science Curricula 2013 (CS2013) [online] http://dx.doi.org/10.1145/2534860

[9] ACM Curricula recommendations [online] http://www.acm.org/education/curricula-recommendations

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

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

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

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

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

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

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

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

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

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

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

[21] 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/

[22] Demchenko, Yuri, Emanuel Gruengard, Sander Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. 1st IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science, in Proc.The 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 December 2014, Singapore.

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

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

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

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

[27] Tobias Weigel, Timothy DiLauro, Thomas Zastrow, RDA PID Information Types WG: Final Report, Research Data Alliance, 2015/07/10 [online] https://b2share.eudat.eu/record/245/files/PID%20Information%20Types%20Final%20Report.pdf

[28] Research Data Alliance Interest Group on Education and Training on Handline of Research Data (IG-ETRD) wiki [online] https://rd-alliance.org/node/971/all-wiki-index-by-group

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

[30] 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

[31] 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/

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

Acronyms

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

Appendix A. 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

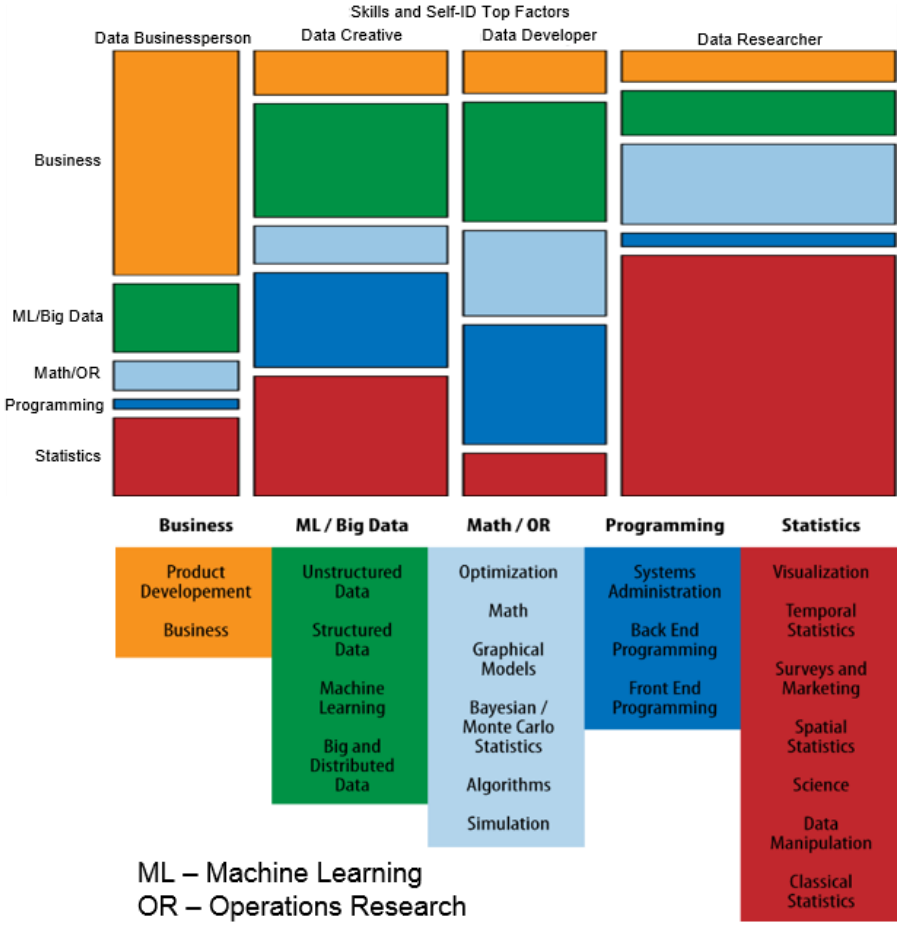


Figure A.1. Data Scientist skills and profiles according to O'Reilly Strata survey [25]

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.2. Skills and Human Resources for e-Infrastructures within Horizon 2020

The Report on the Consultation Workshop (May 2012) “Skills and Human Resources for e-Infrastructures within Horizon 2020” [26] summarises the outcomes of a consultation workshop that was organised by DG INFSO "GÉANT and e-Infrastructures" unit to consult the stakeholders on their views of approaching these challenges. The workshop discussions highlighted cross-cutting challenges of

i) new and changed skills needs which combine technical and scientific skills and require interdisciplinary thinking and communication;
ii) recognizing new job profiles and tasks rising from the emergence of computing intensive and data-driven science with integral role of e-infrastructures;
iii) need for effective European level collaboration and coordination to avoid duplication of efforts and join the forces for developing high quality human capital for e-infrastructures

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

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

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

The nine areas identified as having potentially the most significant skills gap according to RLUK report "Re-skilling for Research" (2012) [29]

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

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" [30] 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 [31]. 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.

A.5. Other studies, reports and ongoing works on defining Data Science profiles and skills

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

- LERU Roadmap for Research Data (2013) [32]
- FOSTER Project Taxonomy of training and educational resources on Open Science
- OpenAIR Scientific Information Management workflow and organisational functions
- EDSA Project Data Science skills classification and vocabulary
- THOR Project Training Program scope

Appendix B. Data used in current study of demanded Data Science competences and skills

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

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

The following are general characteristics of the collected data.

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

All working data are available in the shared project storage area on Google Drive and can be presented on demand.