EDISON Data Science Framework:
Part 2. Data Science Body of Knowledge (DS-BoK)
Release 4 (EDSF04 or EDSF2022)

EDISON Community Initiative
(Maintaining the H2020 EDISON project outcome)

Release Date	31 December 2022
Document Editor/s	Yuri Demchenko
Version	Release 4, v07
Status	Working document, request for comments

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

EDSF Release 4: Part 2. Data Science Body of Knowledge (DS-BoK)

Document Version Control

	Version	Date	Change Made (and if appropriate reason for change)	Initials of Commentator(s) or Author(s)
Release 1	03	10/10/2016	Release 1 after ELG03 meeting discussion	YD
Release 2	04	03/07/2017	Release 2 document (updated after multiple discussions and comments, ELG04 comments)	YD
Pre-Release 3	05	07/09/2018	Pre-release 3. The definition and content of the DS-BoK revised and extended	YD
Release 3	06	31/12/2018	Release 3. Document updated based on received comments and feedback from practical implementation	YD, TW
Release 4	06	31/12/2022	Release 4. Document updated based on received comments and feedback from practical implementation. Knowledge Area Group related to Data Management and Governance revised and extended based on the FAIRsFAIR project contribution.	YD, JJCG

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

Contributors

Document Editors: Yuri Demchenko		
Author Initials	Name of Contributor	Institution
YD	Yuri Demchenko	University of Amsterdam
AB	Adam Belloum	University of Amsterdam
AM	Andrea Manieri	Engineering
TW	Tomasz Wiktorski	University of Stavanger
JJCG	Cuadrado Gallego Juan José	Alcala University

Acknowledgement

Gábor Kismihók, Center of Job Knowledge Research, University of Amsterdam
Oleg Chertov, Professor, Sikorski National Technical University of Ukraine “Kyiv Polytechnic Institute”
Luca Comminiello, Erasmus+ Master, University of Perugia

Executive summary

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

The EDISON Data Science Framework (EDSF) includes the four main components: Data Science Competence Framework (CF-DS), Data Science Body of Knowledge (DS-BoK), Data Science Model Curriculum (MC-DS), Data Science Professional Profiles (DSPP), which are extended with new Part 5. Use cases and guidelines. The EDSF provides a conceptual basis for the Data Science Profession definition, targeted education and training, professional certification, organizational capacity building, and organisation and individual skills management and career transferability.

The Data Science Body of Knowledge defines in a structured way the knowledge topics required for Data Science professionals to work efficiently with their work tasks and further develop their professional competences and skills. The definition of the Data Science Body of Knowledge provides a basis for defining the Data Science Model Curriculum and further can be used for the Data Science professional certification.

The presented DS-BoK defines six groups of Knowledge Areas (KAG) that are linked to the identified competence groups defined in CF-DS: KAG-DSA Data Analytics; KAG-DSDM Data Management, KAG-DSENG Data Science Engineering, KAG-DSRMPM Research Methods and Project Management; and KAG-DSBPM Business Process Management as an addressed domain knowledge area. Knowledge Areas Groups include few Knowledge Areas. Knowledge Areas (KA) are defined in most cases by existing scientific subjects according to ACM Computing Classification System (ACM CCS2012) or commonly used academic subjects. New KAs are introduced based on existing and relevant Data Science technology areas. Knowledge Areas are composed of a number of Knowledge Units (KU) which are currently the lowest component of the DS-BoK. Defining the domain knowledge groups both for science and business will be a subject for further DS-BoK development in tight cooperation with domain specialists.

The proposed EDSF and DS-BoK in particular, are intended to provide guidance and a basis for universities and education practitioners to define their Data Science curricula and select necessary courses, on the one hand, and for companies to better define a set of required competences, knowledge and skills for their specific industry domain in their search for Data Science talents, on the other hand.

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

TABLE OF CONTENTS

1 Introduction

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

This document presents the Data Science Body of Knowledge (DS-BoK) Release 4, revised and updated after the Release 3 publication in December 2018, based on feedback from multiple practical implementations by champion universities that cooperated with the EDISON project and incorporating comments and suggestions from experts and community discussions. The definition of some DS-BoK knowledge areas and knowledge units has been improved based on the feedback from the FAIRsFAIR project as a result of the Data Stewardship Professional Competence Framework definition and the development of the Data Stewardship and FAIR training curricula [6].

The main goal of the presented Data Science Body of Knowledge is to propose a consistent Data Science Body of Knowledge that would consolidate existing scattered standards, practices and resources and respond to requirements from multiple stakeholders to create sustainable Data Science competences and skills management ecosystem.

The presented DS-BoK definition is based on an overview and analysis of existing bodies of knowledge that are relevant to required competences and knowledge for Data Science and required to fulfill the identified in CF-DS competences and skills.

The presented DS-BoK defines the six Knowledge Areas Groups (KAG) that are linked to the identified competence groups defined in CF-DS: KAG-DSA Data Analytics; KAG-DSDM Data Management, KAG-DSENG Data Science Engineering, KAG-DSRMP Research Methods and Project Management; and KAG-DSBPM Business Process Management. Knowledge Areas are composed of a number of Knowledge Units (KU) which are currently the lowest component of the DS-BoK. Defining the domain knowledge groups both for science and business will be a subject for further DS-BoK development in tight cooperation with domain specialists.

DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs and KUs defined where possible based on the Classification Computer Science (CCS2012), components taken from other BoKs and proposed new KAs/KUs to incorporate new technologies used in Data Science and their recent developments.

The definition of the Data Science Body of Knowledge provides a basis for defining the Data Science Model Curriculum and further for the Data Science professional certification.

DS-BoK is maintained by the University of Amsterdam as a part of the community shared EDISON Initiative. Further work will be required to develop consistent DS-BoK that can be accepted by the academic community and professional training community.

The presented document has the following structure. Section 2 provides an overview of the EDISON Data Science Framework and related components of the Data Science professional ecosystem. Section 3 provides an overview of existing BoKs related to Data Science knowledge areas. Section 3 also includes other important components for the DS-BoK definition, such as data lifecycle management models, scientific methods, and business process management lifecycle models. Section 4 describes the proposed DS-BoK structure and provides the definition of the main components KAGS, Kas, and KUs. Section 5 provides a summary of the achieved results and suggests further development.

Appendices to this document contain important supplementary information: detailed information about reviewed bodies of knowledge related to identified Data Science knowledge areas, taxonomy of the Data Science knowledge areas and scientific disciplines built as a subset of the ACM CCS (2012) classification.

2 EDISON Data Science Framework

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

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

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

The proposed framework provides a basis for other components of the Data Science professional ecosystem[1] , such as

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

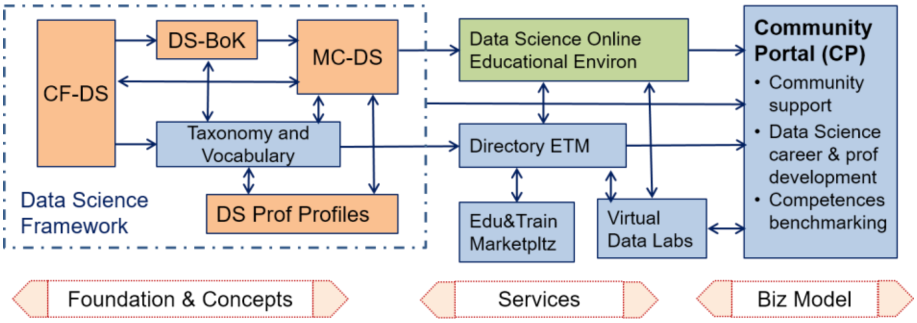

Figure 2.1. EDISON Data Science Framework components.

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

The CF-DS provides the overall basis for the whole EDSF. The core CF-DS includes common competences required for the successful work of a Data Scientist in different work environments in industry and in research and throughout the whole career path. The future CF-DS development may include coverage of the domain specific competences and skills by involving domain and subject matter experts, which may be published as separate CF-DS profiles[2].

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

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

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. Knowledge Areas are composed of a number of Knowledge Units (KU) which are currently the lowest component of the DS-BoK. DS-BoK incorporates best practices in Computer Science and domain specific BoK's and includes KAs and KUs defined where possible based on the Classification Computer Science (CCS2012) [7], components taken from other BoKs and proposed new KAs/KUs to incorporate new technologies used in Data Science and their recent developments.

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

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

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

It is anticipated that successful acceptance of the proposed EDSF and its core components will require standardisation and interaction with the European and international standardisation bodies and professional organisations. This work is being done as a part of the EDSF sustainability support by the EDISON community initiative provided by the University of Amsterdam[3].

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

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

3 Overview of BoKs relevant to DS-BoK

The following BoK's have been reviewed to provide a basis for initial definition of the DS-BoK:

- ACM Computer Science Body of Knowledge (ACM CS-BoK) [7, 9, 10]
- ICT professional Body of Knowledge (2015) [12]
- CEN EN 17748-2:2022 Foundational Body of Knowledge for the ICT Profession (ICT BoK) [13]
- Software Engineering Body of Knowledge (SWEBOK) [15]
- Business Analytics Body of Knowledge (BABOK) [16]
- Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [17]
- Project Management Professional Body of Knowledge (PM-BoK) [18]

The following sections provide a short description and analysis of each body of knowledge. These allowed to identify what components of the existing BoKs can be re-used to construct a consistent Data Science Body of Knowledge that should support competence groups defined in CF-DS. The DS-BoK should also reflect the data-lifecycle management where different organisational roles, functions, competences and knowledge are required.

The presented analysis allowed to identify what existing BoK's can be used in the DS-BoK definition or mapped to ensure knowledge transferability and education programmes compatibility. From this initial analysis the relevant best practices have been identified to structure the DS-BoK and provide a basis for defining the Data Science professional certification scheme.

3.1 ACM Computer Science Body of Knowledge (CS-BoK)

In the ACM-CS2013-final report [9, 10] the Body of Knowledge is defined as a specification of the content to be covered in a curriculum that serves as an implementation of the BoK. The ACM-BoK describes and structures the knowledge areas needed to define a curriculum in Computer Science, it includes 18 Knowledge Areas (where 6 KAs are newly introduced in ACM CS2013):

AL - Algorithms and Complexity
AR - Architecture and Organization
CN - Computational Science
DS - Discrete Structures
GV - Graphics and Visualization
HCI - Human-Computer Interaction
IAS - Information Assurance and Security (new)
IM - Information Management
IS - Intelligent Systems
NC - Networking and Communications (new)
OS - Operating Systems
PBD - Platform-based Development (new)
PD - Parallel and Distributed Computing (new)
PL - Programming Languages
SDF - Software Development Fundamentals (new)
SE - Software Engineering
SF - Systems Fundamentals (new)
SP - Social Issues and Professional Practice

Knowledge areas should not directly match a particular course in a curriculum (this practice is strongly discouraged in the ACM report), often courses address topics from multiple knowledge areas. The ACM-CS2013-final report distinguishes between two types of topics: Core topics subdivided into "Tier-1" (that are mandatory for each curriculum) and "Tier-2" (that are expected to be covered at 90-100% with minimum advised 80%), and elective topics. The ACM classification suggests that a curriculum should include all topics in Tier 1 and all or almost the topics in Tier 2. Tier 1 and Tier 2 topics are defined differently for different programmes and specialisations. To be complete, a curriculum should cover, in addition to the topics of Core Tier 1 and 2, a significant amount of elective material. The reason for such a hierarchical approach to the structure of the Body of Knowledge is a useful way to group related information, not as a structure for organizing material into courses.

The ACM for computing Education in Community Colleges [11] defines a BoK for IT outcome-based learning/education, which identifies 6 technical competency areas and 5 workplace skills. While the technical areas are specific to IT competences and specify a set of demonstrable abilities of graduates to perform some specific functions, the so-called workplace skills describe the ability of the student/trainee to:

(1) function effectively as a member of a diverse team,
(2) read and interpret technical information,
(3) engage in continuous learning,
(4) professional, legal, and ethical behaviour, and
(5) demonstrate business awareness and workplace effectiveness

The CS-BoK uses ACM Computing Classification System (CCS), which is standard and widely accepted, what makes it a good basis for using it as basis for building DS-BoK and providing necessary extensions/KAs related to identified Data Science competence groups (see section 3.4) which majority require background knowledge components from the general CS-BoK.

3.2 ICT Foundational Body of Knowledge (2015) [12]

The ICT Foundational Body of Knowledge (hereafter referred to as ICT-BoK2015) [12] is an effort promoted by the IT Professionalism Europe (ITPE) network of stakeholders committed to the advancement of IT professionalism (https://itprofessionalism.org/). The goal is to define and organise the core knowledge of the ICT discipline. In order to foster the growth of digital jobs in Europe and to improve ICT Professionalism a study has been conducted to provide the basis of the ICT BoK framework. The framework consists of four building blocks which are also found in other professions:

i) body of knowledge (BoK);
ii) competence framework;
iii) education and training resources; and
iv) code of professional ethics.

A competence framework already exists and is documented in the e-Competence Framework (now in its version 3.0 and promoted by CEN). However, an ICT Body of Knowledge that provides the basis for a common understanding of the foundational knowledge an ICT professional should possess is not yet available.

The ICT-BoK2015 is structured in 5 *Process Groups*, defining the various phases of the project development or organisational workflow: *Initiating*, *Planning*, *Executing*, *Monitoring and Controlling*, *Closing*.

The ICT-BoK2015 aims to inform about the level of knowledge required to enter the ICT profession and acts as the first point of reference for anyone interested in working in ICT. ICT-BoK2015 does not refer to Data Science competences explicitly, but the identified ICT processes can be applied to data management processes both in industry and academia in the context of well-defined and structured projects.

3.3 CEN EN 17748-2:2022 Foundational Body of Knowledge for the ICT Profession (ICT BoK) [13]

Recently published “CEN EN 17748-2:2022 Foundational Body of Knowledge for the ICT Profession (ICT BoK)” [13] actually supersedes The ICT Foundational Body of Knowledge (ICT-BoK) (2015). It was developed by “CEN/TC 428 - ICT Professionalism and Digital Competences” [14][4] , whose responsibility includes all aspects of standardization related to maturing the ICT Profession in all sectors, public and private. This includes, at a minimum, activity related to four major building blocks of ICT Professionalism:

(1) competences (standardization of a common language of digital and ICT Professional competences, skills and knowledge applied in all domains),
(2) education and certification,

[4] CEN EN 17748-2:2022 is provided on the paid basis by the CEN webshop, however you can find a cheaper version from other countries that actually sell just refolded version of the CEN standards.

(3) Code of Ethics, and
(4) Body of Knowledge (BoK).

Each knowledge Unit is labeled and can be viewed from four index conceptual attributes:

- Knowledge domains (7 in total)
- EN 16234-1 (e-CF) competences (41 in total)
- EN 16234-1 (e-CF) competence areas (5 in total)
- CWA 16458 (European ICT Professional Profiles) (30 in total)

Mapping is provided between all four conceptual/index views.

The ICT BoK knowledge domains include:
1 - Transversal Knowledge
2 - Behavioural Knowledge
3 - Architecture
4 - Network
5 – Software
6 – Data
7 – Business

The Knowledge Unit structure includes:

- Common knowledge (mostly referred to transversal knowledge view),
- Base knowledge (related to ICT professional profiles),
- Specialised Knowledge aspect is provided as a reference to external knowledge or BoKs.

The Transversal Aspects are relevant to the common knowledge content of the Foundational Body of Knowledge. Transversal knowledge represents the knowledge components of transversal aspects that are articulated in the EN 16234-1 (e-CF): T1 Accessibility, T2 Ethics, T3 ICT legal issues, T4 Privacy, T5 Security, T6 Sustainability, T7 Usability.

During development, the CEN/TC 428 team reviewed the EDSF and DS-BoK, in particular, and consulted with the EDSF team on methodology and possible mapping. It resulted in the adoption of some aspects of the BoK definition and EDSF if referenced in connection tot the Data domain knowledge 6.

3.4 Software Engineering Body of Knowledge (SWEBOK) [15]

The Software Engineering Body of Knowledge (SWEBOK) is an international standard ISO/IEC TR 19759:2015[5] specifying a guide to the generally accepted Software Engineering Body of Knowledge. The Guide to the Software Engineering Body of Knowledge (SWEBOK Guide) has been created through cooperation among several professional bodies and members of industry and is published by the IEEE Computer Society. The standard can be accessed freely from the IEEE Computer Society (http://www.computer.org/web/swebok/v3).[6]

The published version of SWEBOK V3 has the following 15 knowledge areas (KAs) within the field of software engineering: and 7 additional disciplines are recognized as linked and providing important background knowledge that is beneficial for Software Engineering:

SWEBOK Knowledge Areas

- Software requirements
- Software design
- Software construction
- Software testing
- Software maintenance
- Software configuration management

Additional linked disciplines

- Computer engineering
- Systems engineering
- Project management
- Quality management
- General management
- Computer science

[5] ISO/IEC TR 19759:2015 Software Engineering - Guide to the software engineering body of knowledge (SWEBOK)

[6] SWEBOK can be also accessed from http://www4.ncsu.edu/~tjmenzie/cs510/pdf/SWEBOKv3.pdf

- Software engineering management
- Software engineering process
- Software engineering models and methods
- Software quality
- Software engineering professional practice
- Software engineering economics
- Computing foundations
- Mathematical foundations
- Engineering foundations
- Mathematics

3.5 Business Analysis Body of Knowledge (BABOK) [16]

BABOK Guide was first published by the International Institute of Business Analysis (IIBA) as a draft document version 1.4, in October 2005, for consultation with the wider business analysis and project management community to document and standardize generally accepted business analysis practices. Current version 3 was released in April 2015.

The Business Analysis Body of Knowledge provides an interesting example of the business oriented body of knowledge that covers important for Data Science knowledge domain. BABOK is published in a Guide to the Business Analysis Body of Knowledge (BABOK Guide). It is the globally recognized standard for the practice of business analysis. BABOK Guide reflects the collective knowledge of the business analysis community and presents the most widely accepted business analysis practices.

BABOK Guide recognizes and reflects the fact that business analysis is continually evolving and is practiced in a wide variety of forms and contexts. It defines the skills, knowledge, and competencies required to perform business analysis effectively. It does not describe the processes that people will follow to do business analysis.

BABOK Guide includes chapters on:

- Business Analysis Key Concepts: define important terms that are the foundation of the practice of business analysis.
- Knowledge Areas: represents the core content of *BABOK Guide* and contains the business analysis tasks that are used to perform business analysis.
- Underlying Competencies: describes the behaviours, characteristics, knowledge, and personal qualities that help business analysts be effective in their job.
- Techniques: describes 50 of the most common techniques used by business analysts.
- Perspectives (new to version 3): describes 5 different views of business analysis (Agile, Business Intelligence, Information Technology, Business Architecture, and Business Process Management).

BABOK Guide organises business analysis tasks within 6 knowledge areas. The knowledge areas logically organize tasks but do not specify a sequence, process, or methodology. Each task describes the typical knowledge, skills, deliverables, and techniques that the business analyst requires to be able to perform those tasks competently.

The following knowledge areas of BABOK Guide are defined:

- Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts.
- Elicitation and Collaboration: describes the tasks used to prepare for and conduct elicitation activities and confirm the results.
- Requirements Life Cycle Management: describes the tasks used to manage and maintain requirements and design information from inception to retirement.
- Strategy Analysis: describes the tasks used to identify the business need, address that need, and align the change strategy within the enterprise.

- Requirements Analysis and Design Definition: describes the tasks used to organize requirements, specify and model requirements and designs, validate and verify information, identify solution options, and estimate the potential value that could be realized.
- Solution Evaluation: describes the tasks used to assess the performance of and value delivered by a solution and to recommend improvements on increasing value.

BABOK knowledge areas organisation by tasks allows easy linking to Business Analysis competences what approach can be used in the intended DS-BoK.

3.6 Data Management Body of Knowledge (DM-BoK) by DAMAI [17]

The Data Management Association International (DAMAI) was founded in 1988 in the US with the aim: (i) to provide a non-profit, vendor-independent association where data professionals can go for help and assistance; (ii) to provide the best practice resources such as the DM-BoK and DM Dictionary of Terms; (iii) to create a trusted environment for DM professionals to collaborate and communicate.

The DM-BoK version2 “Guide for performing data management” is structured in 11 knowledge areas covering core areas in data management:

(1) Data Governance,
(2) Data Architecture,
(3) Data Modelling and Design,
(4) Data Storage and Operations,
(5) Data Security,
(6) Data Integration and Interoperability,
(7) Documents and Content,
(8) Reference and Master Data,
(9) Data Warehousing and Business Intelligence,
(10) Metadata, and
(11) Data Quality.

Each KA has section topics that logically group activities and is described by a context diagram. There is also an additional Data Management section containing topics that describe the knowledge requirements for data management professionals. Each context diagram includes: Definition, Goals, Processes, Inputs, Supplier roles, Responsible, Stakeholder, Tools, Deliverables, and Metrics (See Appendix A).

When using DM-BoK for defining the Data Management knowledge area for DS-BoK (DSDM) it needs to be extended with the recent data modelling technologies and Big Data management platforms that address generic Big Data properties such as Volume, Veracity, Velocity. The DS-BoK should also include the widely accepted by the research community FAIR data principles and Data Stewardship best practices in research data management [6].

3.7 Project Management Professional Body of Knowledge (PM-BoK) [18]

The PM-BoK is maintained by the Project Management Institute (PMI) the provides research and education services to Project Managers through publications, networking opportunities in local chapters, hosting conferences and training seminars, and providing accreditation in project management. PMI, exploit volunteers and sponsorships to expand project management's body of knowledge through research projects, symposiums and surveys, and shares it through publications, research conferences, and working sessions. The “A Guide to the Project Management Body of Knowledge" (PM-BoK) has been recognized by the American National Standards Institute (ANSI), and in 2012 ISO adapted the project management processes from the PMBOK Guide 4th edition (see Appendix A).

The PMI-BoK defines five Process Groups related to project management:

- Initiating - Processes to define and authorize a project or project phase
- Planning - Processes to define the project scope, objectives, and steps to achieve the required results.
- Executing - Processes to complete the work documented within the Project Management Plan.

- Monitoring and Controlling - Processes to track and review the project progress and performance. This group contains the Change Management.
- Closing - Processes to formalize the project or phase closure.

The nine Knowledge Areas are linked to the Process Groups:

- Project Integration Management - Processes to integrate various parts of the Project Management.
- Project Scope Management - Processes to ensure that all of the work required is completed for a successful Project and manages additional "scope creep".
- Project Time Management - Processes to ensure the project is completed in a timely manner.
- Project Cost Management - Processes to manage the planning, estimation, budgeting and management of costs for the duration of the project.
- Project Quality Management - Processes to plan, manage and control the quality and to provide assurance the quality standards are met.
- Project Human Resource Management - Processes to plan, acquire, develop and manage the project team.
- Project Communications Management - Processes to plan, manage, control, distribute and final disposal of project documentation and communication.
- Project Risk Management - Processes to identify, analyse and management of project risks.
- Project Procurement Management - Processes to manage the purchase or acquisition of products and services, or result to complete the project.
- Project Stakeholder Management – Process to identify stakeholders, determine their requirements, expectations and influence

Each Process Group contains processes within some or all of the Knowledge Areas. Each of the 42 processes has Inputs, Tools and Techniques, and Outputs. (It is not the scope of this analysis to enter into the details of each process).

4 Data Science Body of Knowledge (DS-BoK) definition

The presented DS-BoK definition is based on an overview and analysis of existing bodies of knowledge that are relevant to Data Science and required to fulfill the identified in CF-DS competences and skills. This is also enriched by analysis of the practice in academic and professional training courses development by universities and professional training organisations.

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

4.1 General Approach and Structure of DS-BoK

The DS-BoK contains the following Knowledge Area groups (KAG) that follow the competence groups defined in CF-DS [1]:

- KAG1-DSDA: Data Analytics group including Data Analytics methods, Machine Learning, statistical methods, and data visualisation
- KAG2-DSENG: Data Science Engineering group including software engineering, database and Big Data technologies
- KAG3-DSDM: *Data Management group, including data curation, preservation and data modeling*
- KAG4-DSRMP: *Research Methods and Project Management*
- KAG5-DSBA: Business Analytics (strongly based on KAG1-DSDA)
- KAG*-DSDK: Placeholder for the Data Science Domain Knowledge groups to include domain specific knowledge

The subject domain related knowledge group (scientific or business) KAG*-DSDK is recognized as essential for the practical work of Data Scientist what in fact, means not professional work in a specific subject domain but understanding the domain related concepts, models and organisation (refer to CF-DS section 4.8 [1]) and corresponding data analysis methods and models. These knowledge areas will be a subject for future development in tight cooperation with subject domain specialists.

It is also anticipated that due to the complexity of the Data Science domain, the DS-BoK will require a wide spectrum of background knowledge, first of all in mathematics, statistics, logic and reasoning as well as general computing, and cloud computing in particular. Similar to the ACM CS2013 curricula approach, background knowledge can be required as an entry condition or must be studied as elective courses.

The proposed DS-BoK re-uses where possible existing BoK's, taking necessary KA and KU definitions and combining them into defined above DS-BoK knowledge area groups. The following BoKs were used and/or mapped to the selected DS-BoK knowledge groups:

- ACM Computer Science Body of Knowledge (ACM CS-BoK) [7, 9, 10, 11]
- Software Engineering Body of Knowledge (SWEBOK) [15]
- Business Analytics Body of Knowledge (BABOK) [16]
- Data Management Body of Knowledge (DM-BoK) by DAMAI [17]
- Project Management Professional Body of Knowledge (PM-BoK) [18]

4.2 Data Science Body of Knowledge Areas and Knowledge Units

Table 4.1 provides a consolidated view of the identified Knowledge Areas in the Data Science Body of Knowledge. The table contains a detailed definition of the KAG1-DSDA, KAG2-DSENG, KAG3-DSDM groups that are well supported by existing BoK's and academic materials. General suggestions are provided for KAG4-DSRMP, KAG5-DSBA groups that correspond to newly identified competences and knowledge areas and require additional study of existing practices and contribution from experts in corresponding scientific or business domains.

The KAG1-DSDA Data Analytics knowledge area group is key and distinguishing KAG for DS-BoK. It includes different methods and algorithms, primarily statistical, machine learning and data mining, to enable data processing, modelling, analysis and inspection with the goal of discovering useful information, providing insight and recommendations, and supporting decision-making. The following are commonly defined Data Science Analytics Knowledge Areas:

- KA01.01 (DSDA.01/SMA) Statistical methods, including Descriptive statistics, exploratory data analysis (EDA) focused on discovering new features in the data, and confirmatory data analysis (CDA) dealing with validating formulated hypotheses;
- KA01.02 (DSDA.02/ML) Machine learning and related methods for information search, image recognition, decision support, classification;
- KA01.03 (DSDA.03/DM) *Data mining* is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes;
- KA01.04 (DSDA.04/TDM) Text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data;
- KA01.05 (DSDA.05/PA) Predictive analytics focuses on the application of statistical models for predictive forecasting or classification.
- KA01.06 (DSDA.06/BA) Business Analytics and Business Intelligence covers data analysis that relies heavily on aggregation and different data sources and focuses on business information;
- KA01.07 (DSDA.07/MSO) Computational modelling, simulation and optimisation

The KAG2-DSENG group includes selected KAs from ACM CS-BoK and SWEBOK and extends them with new technologies and engineering technologies and paradigms such as cloud based, agile technologies and DevOps that are promoted as continuous deployment and improvement paradigms and allow organisations to implement agile business and operational models.

The KAG3-DSDM group includes most of KAs from DM-BoK, however, extended it with KAs related to RDA recommendations, community data management models (Open Science, Open Access, Open Data, FAIR data principles, etc.) and general Data Lifecycle Management that is used as a central concept in many data management related education and training courses.

Table 4.2 provides a detailed definition of DS-BoK Knowledge Areas and Knowledge Units. Knowledge Units (KU) corresponding to suggested KAs are defined from different sources: existing BoK, CCS2012, and practices in designing academic curricula and corresponding courses by universities and professional training organisations[7].

The presented DS-BoK high-level content is not exhaustive at this stage and will undergo further development based on feedback from MC-DS implementation.

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

Table 4.1. DS-BoK Knowledge Area Groups and corresponding Knowledge Areas

KA Groups	**Suggested DS Knowledge Areas (KA)**	**Knowledge Areas from existing BoK and CCS2012 scientific subject groups**
KAG1-DSDA: Data Science Analytics	KA01.01 (DSDA.01/SMDA) Statistical methods for data analysis KA01.02 (DSDA.02/ML) Machine Learning KA01.03 (DSDA.03/DM) Data Mining KA01.04 (DSDA.04/TDM) Text Data Mining KA01.05 (DSDA.05/PA) Predictive Analytics KA01.06 (DSDA.06/MODSIM) Computational modelling, simulation and optimisation	There is no formal BoK defined for Data Analytics. Data Science Analytics related scientific subjects from CCS2012: CCS2012: Computing methodologies CCS2012: Mathematics of computing CCS2012: Computing methodologies
KAG2-DSENG: Data Science Engineering	KA02.01 (DSENG.01/BDIT) Big Data Infrastructure and Technologies KA02.02 (DSENG.02/DSIAPP) Infrastructure and platforms for Data Science applications KA02.03 (DSENG.03/CCT) Cloud Computing technologies for Big Data and Data Analytics KA02.04 (DSENG.04/SEC) Data and Applications security KA02.05 (DSENG.05/BDSE) Big Data systems organisation and engineering KA02.06 (DSENG.06/DSAPPD) Data Science (Big Data) applications design KA02.07 (DSENG.07/IS) Information systems (to support data driven decision making)	ACM CS-BoK selected KAs: AL - Algorithms and Complexity AR - Architecture and Organization (including computer architectures and network architectures) CN - Computational Science GV - Graphics and Visualization IM - Information Management PBD - Platform-based Development (new) SE - Software Engineering (can be extended with specific SWEBOK KAs) SWEBOK selected KAs • Software requirements • Software design • Software engineering process • Software engineering models and methods • Software quality Data Science Analytics related scientific subjects from CCS2012: CCS2012: Computer systems organization CCS2012: Information systems CCS2012: Software and its engineering
KAG3-DSDM: Data Management	KA03.01 (DSDM.01/DMORG) General principles and concepts in Data Management and organisation KA03.02 (DSDM.02/DMS) Data management systems KA03.03 (DSDM.03/EDMI) Data Management and Enterprise data infrastructure KA03.04 (DSDM.04/DGOV) Data Governance KA03.05 (DSDM.05/BDST0R) Big Data storage (large scale) KA03.06 (DSDM.05/DLIB) Digital libraries and archives	DM-BoK selected KAs (1) Data Governance, (2) Data Architecture, (3) Data Modelling and Design, (4) Data Storage and Operations, (5) Data Security, (6) Data Integration and Interoperability, (7) Documents and Content, (8) Reference and Master Data, (9) Data Warehousing and Business Intelligence, (10) Metadata, and (11) Data Quality. Data Science Analytics related scientific subjects from CCS2012: CCS2012: Information systems

KA Groups	**Suggested DS Knowledge Areas (KA)**	**Knowledge Areas from existing BoK and CCS2012 scientific subject groups**
KAG4-DSRMP: Research Methods and Project Management	KA04.01 (DSRMP.01/RM) Research Methods KA04.01 (DSRMP.02/PM) Project Management	There are no formally defined BoK for research methods PMI-BoK selected KAs • Project Integration Management • Project Scope Management • Project Quality • Project Risk Management
KAG5-DSBPM: Business Analytics	KA05.01 (DSBA.01/BAF) Business Analytics Foundation KA05.02 (DSBA.02/BAEM) Business Analytics organisation and enterprise management	BABOK selected KAs *) • Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts. • Requirements Analysis and Design Definition. • Requirements Life Cycle Management (from inception to retirement). • Solution Evaluation and improvements recommendation.

*) BABOK KAs are more business focused and related to KAG5-DSBA; however, its specific topics related to data analysis can be reflected in the KAG1-DSDA

Table 4.2. Detailed definition of the DS-BoK and suggested Knowledge Units (KU)

Knowledge Area Groups (KAG)	**Knowledge Areas (KA)**	**Knowledge Unit (KU)**	**Suggested Knowledge Units (KU)**	**Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)**
KAG1-DSDA: Data Science Analytics	KA01.01 DSDA.01/SMDA Statistical methods for data analysis	KU1.01.00	General overview and main concepts in Statistical methods for data analysis	**CCS2012: Mathematics of computing** • Discrete mathematics ○ Graph theory ○ Probability and statistics ○ Probabilistic representations ○ Probabilistic inference problems ○ Probabilistic reasoning algorithms ○ Probabilistic algorithms • Statistical paradigms • Mathematical software • Information theory • Mathematical analysis
		KU1.01.01	Probability & Statistics	
		KU1.01.02	Statistical paradigms (regression, time series, dimensionality, clusters)	
		KU1.01.03	Probabilistic representations (causal networks, Bayesian analysis, Markov nets)	
		KU1.01.04	Frequentist and Bayesian statistics	
		KU1.01.05	Probabilistic reasoning	
		KU1.01.06	Exploratory and confirmatory data analysis	
		KU1.01.07	Quantitative analytics	
		KU1.01.08	Qualitative Analytics	
		KU1.01.09	Data preparation and preprocessing	
		KU1.01.10	Performance analysis	
		KU1.01.11	Markov models, Markov networks	
		KU1.01.12	Operations research	
		KU1.01.13	Information theory	
		KU1.01.14	Discrete Mathematics and Graph Theory	
		KU1.01.15	Mathematical analysis	
		KU1.01.16	Mathematical software and tools	
KAG1-DSDA: Data Science Analytics	KA01.02 DSDA.02/ML Machine Learning	KU1.02.00	General overview and main concepts in Machine Learning	**CCS2012: Computing methodologies** • Artificial intelligence ○ Machine learning ○ Learning paradigms ▪ Supervised learning ▪ Unsupervised learning ▪ Reinforcement learning ▪ Multi-task learning • Machine learning approaches ○ Machine learning algorithms
		KU1.02.01	Machine Learning theory and algorithms	
		KU1.02.02	Supervised Machine Learning	
		KU1.02.03	Unsupervised Machine Learning	
		KU1.02.04	Reinforced learning	
		KU1.02.05	Classification methods	
		KU1.02.06	Design and Analysis of Algorithms	
		KU1.02.07	Game Theory & Mechanism design	
		KU1.02.08	Artificial Intelligence	
		KU1.01.02	Statistical paradigms (regression, time series, dimensionality, clusters)	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
		KU1.01.03	Probabilistic representations (causal networks, Bayesian analysis, Markov nets)	
		KU1.01.04	Frequentist and Bayesian statistics	**CCS2012: Theory of computation** • Design and analysis of algorithms ○ Data structures design and analysis • Theory and algorithms for application domains ○ Machine learning theory ○ Algorithmic game theory and mechanism design • Semantics and reasoning
		KU1.01.05	Probabilistic reasoning	
		KU1.01.08	Performance analysis	
KAG1-DSDA: Data Science Analytics	KA01.03 DSDA.03/DM Data Mining	KU1.03.00	General overview and main concepts in Data Mining	**CCS2012: Theory of computation** • Design and analysis of algorithms ○ Data structures design and analysis • Theory and algorithms for application domains ○ Machine learning theory ○ Algorithmic game theory and mechanism design • Semantics and reasoning
		KU1.03.01	Data mining and knowledge discovery	
		KU1.03.02	Knowledge Representation and Reasoning	
		KU1.03.03	CRISP-DM and data mining stages	
		KU1.03.04	Anomaly Detection	
		KU1.03.05	Time series analysis	
		KU1.03.06	Feature selection, Apriori algorithm	
		KU1.03.07	Graph data analytics	
		KU1.01.08	Performance analysis	
		KU1.02.01	Machine Learning theory and algorithms	
		KU1.02.02	Supervised Machine Learning	
		KU1.02.03	Unsupervised Machine Learning	
		KU1.02.04	Reinforced learning	
		KU1.02.05	Classification methods	
KAG1-DSDA: Data Science Analytics	KA01.04 DSDA.04/TDM Text Data Mining	KU1.04.00	General overview and main concepts in Text Data Mining	**CCS2012: Computing methodologies** • Artificial intelligence ○ Natural language processing ○ Knowledge representation and reasoning ○ Search methodologies
		KU1.04.01	Text analytics including statistical, linguistic, and structural techniques to analyse structured and unstructured data	
		KU1.04.02	Data mining and text analytics	
		KU1.04.03	Natural Language Processing	
		KU1.04.04	Predictive Models for Text	
		KU1.04.05	Retrieval and Clustering of Documents	

Knowledge Area Groups (KAG)	**Knowledge Areas (KA)**	**Knowledge Unit (KU)**	**Suggested Knowledge Units (KU)**	**Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)**
		KU1.04.06	Information Extraction	
		KU1.04.07	Sentiments analysis	
KAG1-DSDA: Data Science Analytics	KA01.05 DSDA.05/PA Predictive Analytics	KU1.05.00	General overview and main concepts in Predictive Analytics	
		KU1.05.01	Predictive modeling and analytics	
		KU1.05.02	Inferential and predictive statistics	
		KU1.05.03	Machine Learning for predictive analytics	
		KU1.05.04	Regression and Multi Analysis	
		KU1.05.05	Generalised linear models	
		KU1.05.06	Time series analysis and forecasting	
		KU1.05.07	Deploying and refining predictive models	
KAG1-DSDA: Data Science Analytics	KA01.06 DSDA.06/MODSIM Computational modelling, simulation and optimisation	KU1.06.00	General overview and main concepts in Computational modeling, simulation and optimisation	**CCS2012: Computing methodologies** • Modeling and simulation ○ Model development and analysis ○ Simulation theory ○ Simulation types and techniques ○ Simulation support systems
		KU1.06.01	Modelling and simulation theory and techniques (general and domain oriented)	
		KU1.06.02	Operations research and optimisation	
		KU1.06.03	Large scale modelling and simulation systems	
		KU1.06.04	Network oprtimisation	
		KU1.06.05	Risk simulation and queueing	
KAG1-DSDA: Data Science Analytics	KA01.07 *) DSDA.07/DAVIZ Data Analytics Visualisation and Story Telling	KU1.07.01	Data Analytics Visualisation Methods	
		KU1.07.02	Data Analytics Visualisation Tools and Software (desktop and cloud based)	
		KU1.07.03	Story telling best practices, dashboards and reports design	
KAG2-DSENG: Data Science Engineering	KA02.01 DSENG.01/BDI Big Data Infrastructure and Technologies	KU2.01.00	General overview and main concepts in Big Data Infrastructure and Technologies	**CCS2012: Computer systems organization** • Architectures ○ Parallel architectures ○ Distributed architectures • Networks *) ○ Network Architectures ○ Network Services ○ Cloud Computing
		KU2.01.01	Computer systems organisation for Big Data applications, CAP, BASE and ACID theorems	
		KU2.01.02	Parallel and Distributed Computer Architecture	
		KU2.01.03	High Performance and Cloud Computing	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
		KU2.01.04	Clouds and scalable computing	
		KU2.01.05	Cloud based Big Data platforms and services	
		KU2.01.06	Big Data (large scale) storage and filesystems (HDFS, Ceph, etc)	
		KU2.01.07	NoSQL databases	
		KU2.01.08	Computer networks for high-performance computing and Big Data infrastructure	
		KU2.01.09	Computer networks: architectures and protocols	
		KU2.01.10	Big Data Infrastructure management and operation	
KAG2-DSENG: Data Science Engineering	KA02.02 DSENG.02/DSIAPP Infrastructure and platforms for Data Science applications	KU2.02.00	General overview of infrastructure and platforms for Data Science applications	Proposed new KA for DS-BoK • Infrastructure and platforms for Data Science applications group: • CCENG - Cloud Computing Engineering (infrastructure and services design, management and operation) • CCAS - Cloud based applications and services development and deployment • BDA – Big Data Analytics platforms (including cloud based) • BDI - Big Data Infrastructure services and platforms, including data storage infrastructure
		KU2.02.01	Big Data Infrastructure: services and components, including data storage infrastructure	
		KU2.02.02	Big Data analytics platforms and tools (including Hadoop, Spark, and cloud based Big Data services)	
		KU2.02.03	Large scale cloud based storage and data management	
		KU2.02.04	Cloud based applications and services operation and management	
		KU2.02.05	Big Data and cloud based systems design and development	
		KU2.02.06	Data processing models (batch, steaming, parallel)	
		KU2.02.07	Enterprise information systems	**CCS2012: Information systems** • Information storage systems • Information systems applications
		KU2.02.08	Data security and protection	
KAG2-DSENG: Data Science Engineering	KA02.03 DSENG.03/CCT Cloud Computing technologies for	KU2.03.00	General overview of Cloud Computing technologies and their use for Big Data and Data Analytics	**DSDA Extension group for CCS201** **Theory of computation**
		KU2.03.01	Cloud Computing architecture and services	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
	Big Data and Data Analytics	KU2.03.02	Cloud Computing Engineering (infrastructure and services design, management and operation)	• DSA Extension point: Algorithms for Big Data computation **Mathematics of computing** • DSA Extension point: Mathematical software for Big Data computation **Computing methodologies** • DSA Extension point: New DSA computing **Information systems** • DSA Extension point: Big Data systems (e.g. cloud based) **Information systems applications** • DSA Extension point: Big Data applications DSA Extension point: Doman specific Data applications
		KU2.03.03	Cloud based applications and services operation and management	
KAG2-DSENG: Data Science Engineering	KA02.04 DSENG.04/SEC Data and Applications security	KU2.04.00	General overview and main concepts in Data and applications security	
		KU2.04.01	Infrastructure, applications and data security	
		KU2.04.02	Data encryption and key management, blockchain based technologies	
		KU2.04.03	Access Control and Identity Management	
		KU2.04.04	Security services management, including compliance and certification	
		KU2.04.05	Data anonymisation	
		KU2.04.06	Data privacy	
KAG2-DSENG: Data Science Engineering	KA02.05 DSENG.05/BDSE Big Data systems organisation and engineering	KU2.05.00	General overview and main principles of Big Data systems organisations and Engineering	**CCS2012: Software and its engineering** • Software organization and properties ○ Software system structures • Software architectures ○ Software system models ○ Distributed systems organizing principles ▪ Cloud computing ▪ Grid computing • Software notations and tools ○ General programming languages ○ Software creation and management
		KU2.05.01	Big Data systems organisation and design	
		KU2.05.02	Big Data algorithms for large scale data processing	
		KU2.05.03	Big Data Analytics	
		KU2.05.04	Big Data analytics platforms and tools (including Hadoop, Spark, and cloud based Big Data services)	
		KU2.05.05	Big Data algorithms for data ingest, pre-processing, and visualisation	
		KU2.05.06	Big Data systems for application domains	
		KU2.05.07	Big Data software (systems) architectures	
		KU2.05.08	Requirements engineering and software systems development	
		KU2.05.09	Large and ultra-large scale software systems organisation	
		KU2.05.10	DevOps and cloud enabled applications development	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
		KU2.05.11	Big Data Infrastructure management and operation	
KAG2-DSENG: Data Science Engineering	KA02.06 DSENG.06/DSAPPD Data Science (Big Data) applications design	KU2.06.00	General overview and main approaches to Data Science (Big Data) applications design	**SWEBOK selected KAs** • Software requirements • Software design • Software construction • Software testing • Software maintenance • Software configuration management • Software engineering management • Software engineering process • Software engineering models and methods • Software quality • Agile development technologies • Methods, platforms and tools • DevOps and continuous deployment and improvement paradigm
		KU2.06.01	Data analytics, data handling software requirements and design	
		KU2.06.02	Applications engineering management	
		KU2.06.03	Software engineering models and methods	
		KU2.06.04	Software quality assurance	
		KU2.06.05	Programming languages for Big Data analytics: R, python, Pig, Hive, others	
		KU2.06.06	Models and languages for complex interlinked data presentation and visualisation	
		KU2.06.07	Agile development methods, platforms and tools	
		KU2.06.08	DevOps and continuous deployment and improvement paradigm	
KAG2-DSENG: Data Science Engineering	KA02.07 DSENG.07/IS Information systems (to support data driven decision making)	KU2.07.00	General overview and basic architectures of Information systems to support data driven decisions)	**CCS2012: Information systems** • Information systems applications ○ Decision support systems ▪ Data warehouses ▪ Expert systems ▪ Data analytics ▪ Online analytical processing ○ Multimedia information systems ○ Data mining
		KU2.07.01	Decision Analysis and Decision Support Systems	
		KU2.07.02	Predictive analytics and predictive forecasting	
		KU2.07.03	Data Analysis and statistics	
		KU2.07.04	Data warehousing and Data Mining	
		KU2.07.05	Data Mining	
		KU2.07.06	Multimedia information systems	
		KU2.07.07	Enterprise information systems	
		KU2.07.08	Collaborative and social computing systems and tools	
				CCS2012: Information systems • Information systems applications

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
				○ Enterprise information systems ○ Collaborative and social computing systems and tools
KAG3-DSDM: Data Management	KA03.01 DSDM.01/DMORG General principles and concepts in Data Management and organisation	KU3.01.00	Overview of general principles, concepts and practices in Data Management and organisation	**Proposed new KA for DS-BoK** General Data Management KA's • Data Lifecycle Management • Data archives/storage compliance and certification
		KU3.01.01	Overview Data type, data type registries, data formats	
		KU3.01.02	Metadata, metadata formats, metadata standards, metadata registries	
		KU3.01.03	Data Lifecycle Management	
		KU3.01.04	Data infrastructure and Data Factories	
		KU3.01.05	Open Science, Open Data, Open Access, ORCID	New KAs to support RDA recommendations and community data management models (Open Access, Open Data, etc) • Data type registries, PIDs • Data infrastructure and Data Factories • New KAs to follow RDA and ERA community developments
		KU3.01.06*)	Metadata registries, publishing metadata	
		KU3.01.07*)	Persistent Identifiers (PID), Open Researcher and Contributor ID (ORCID), Research Organization Registry (ROR)	
		KU3.01.08*)	Ethical principles and data privacy	
		KU3.01.09*)	FAIR metadata management, tools for FAIR metadata management	
		KU3.01.10*)	FAIR metadata management, tools for FAIR metadata management	
		KU3.01.11	Data infrastructure compliance and certification	
KAG3-DSDM: Data Management	KA03.02 DSDM.02/DMS Data management systems	KU3.02.00	General overview and main architectural components in Data management systems	**CCS2012: Information systems** • Data management systems ○ Database design and models ○ Data structures ○ Database management system engines ○ Query languages ○ Database Administration
		KU3.02.01	Data architectures (OLAP, OLTP, ETL)	
		KU3.02.02	Data Modelling, Databases and Database Management Systems	
		KU3.02.03	Data structures	
		KU3.02.04	Data Models and Query Languages	
		KU3.02.05	Database design and models	
		KU3.02.06	Database Administration	

Knowledge Area Groups (KAG)	**Knowledge Areas (KA)**	**Knowledge Unit (KU)**	**Suggested Knowledge Units (KU)**	**Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)**
		KU3.02.07	Enterprise Data Warehouses, architectural components and popular platforms	○ Middleware for databases ○ Information integration • CCS2012: Theory of computation ○ Database theory
		KU3.02.08	Middleware for databases	
		KU3.02.09*)	Master Data Management, Data Dictionaries	
		KU3.02.10*)	FAIR data management requirements and compliance	
		KU3.02.11*)	User data management tools and user support	
KAG3-DSDM: Data Management	KA03.03 DSDM.03/EDMI Data Management and Enterprise data infrastructure	KU3.03.00	General overview and main components in enterprise infrastructure for data management	**DM-BoK selected KAs** (1) Data Governance, (2) Data Architecture, (3) Data Modelling and Design, (4) Data Storage and Operations, (5) Data Security, (6) Data Integration and Interoperability, (7) Documents and Content, (8) Reference and Master Data, (9) Data Warehousing and Business Intelligence, (10) Metadata, and (11) Data Quality.
		KU3.03.01	Data management, including Reference and Master Data	
		KU3.03.02	Data Warehousing and Business Intelligence	
		KU3.03.03	Data storage and operations	
		KU3.03.04	Data archives/storage compliance and certification	
		KU3.03.05	Metadata, linked data, provenance	
		KU3.03.06	Data infrastructure, data registries and data factories	
		KU3.03.07	Data security and protection	
		KU3.03.08	Data backup	
		KU3.03.09	Data anonymisation	
		KU3.03.10*)	Personal data protection, GDPR compliance	
KAG3-DSDM: Data Management	KA03.04 DSDM.04/DGOV Data Governance	KU3.04.00	General overview and main concepts in Data Governance	DM-BoK (as above)
		KU3.04.01	Data governance, data quality, data Integration and Interoperability	
		KU3.04.02	Data Management Planning	
		KU3.04.03	Data Management Policy	
		KU3.04.04	Data interoperability	
		KU3.04.05	Data curation	
		KU3.04.06	Data provenance	
		KU3.04.07	Responsible data use, data privacy, ethical principles, IPR, legal issues	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
		KU3.04.08*)	Data quality management, best practices and frameworks, data quality metrics	
		KU3.04.09*)	Data infrastructure compliance and certification, compliance standards	
		KU3.04.10*)	Data protection policies (including personal data), data access policies, GDPR compliance	
		KU3.04.11*)	User needs analysis and definition of requirements to supporting infrastructure and tools	
		KU3.04.12*)	Data management costs, funding models, budgeting	
KAG3-DSDM: Data Management	KA03.05 DSDM.05/BDSTOR Big Data storage (large scale)	KU3.05.00	General overview and architecture components in Big Data storage	New DSENG Knowledge area: Big Data Storage • Distributed file systems • Data Lakes • Data Factories
		KU3.05.01	Big Data storage infrastructure and operations	
		KU3.05.02	Storage architectures, distributed files systems (HDFS, Ceph, Lustre, Gluster, etc)	
		KU3.05.03	Data storage redundancy and backup	
		KU3.05.04	Data factories, data pipelines	
		KU3.05.05	Cloud based storage, Data Lakes	
KAG3-DSDM: Data Management	KA03.06 DSDM.05/DLIB Data libraries, data archives	KU3.06.00	General overview of data libraries, data archives, digital libraries	**CCS2012: Information systems** • Information systems applications ○ Digital libraries and archives
		KU3.06.01	Data libraries and data archives organisation and services	
		KU3.06.02	Digital libraries organisation and services	
		KU3.06.03	Information Retrieval	
		KU3.06.04	Data curation and provenance	
		KU3.06.05	Search Engines and technologies	
		KU3.06.07*)	Trusted data repositories and certification	
KAG4-DSRMP: Research Methods and Project Management	KA04.01 DSRMP.01/RM Research Methods	KU4.01.00	Overview research methods and data driven research	**Proposed new KA for DS-BoK for DSRM related competences:** • Research methodology, research cycle (e.g. 4 steps model Hypothesis – Research Methods – Artefact – Validation) • Modelling and experiment planning • Data selection and quality evaluation
		KU4.01.01 *)	Research methods and research cycle, research questions and hypothesis evaluation	
		KU4.01.02 *)	Research types and research process models	
		KU4.01.03	Modelling and experiment planning	
		KU4.01.04 *)	Research data collection and quality assessment	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
		KU4.01.05	Data discovery (published data), data selection and use in research	• Use cases analysis: research infrastructures and projects
		KU4.01.06	Data lifecycle management and data provenance	
		KU4.01.07	Research data management plan and ethical issues	
		KU4.01.08	Use cases analysis: research infrastructures and projects	
KAG4-DSRMP: Research Methods and Project Management	KA04.01 DSRMP.02/PM Project Management	KU4.02.00	Overview research process and project management	**PMI-BoK selected KAs** • Project Integration Management • Project Scope Management • Project Quality • Project Risk Management
		KU4.02.01	Project Integration Management	
		KU4.02.02	Project Scope Management	
		KU4.02.03	Project Quality	
		KU4.02.04	Project Risk Management	
		KU4.02.05 *)	Grant application and management	
		KU4.02.06 *)	European Research Area. Open Science, Open Data, and FAIR data sharing	
KAG5-DSBPM: Business Analytics	KA05.01 DSBA.01/BAF Business Analytics Foundation	KU5.01.00	Overview Business Analytics methods and practices	**BABOK selected KAs** • Business Analysis Planning and Monitoring: describes the tasks used to organize and coordinate business analysis efforts. • Requirements Analysis and Design Definition. • Requirements Life Cycle Management (from inception to retirement). • Solution Evaluation and improvements recommendation.
		KU5.01.01	Business Analytics and Business Intelligence: Data, Models (statistical) and Decisions	
		KU5.01.02	Data driven Customer Relations Management (CRP), User Experience (UX) requirements and design	
		KU5.01.03	Operations Analytics	
		KU5.01.04	Business Process Optimization	
		KU5.01.05	Data Warehouses technologies, data integration and analytics	
		KU5.01.06	Data driven marketing technologies	
		KU5.01.07	Business Analytics Capstone	
		KU5.01.08	Econometrics methods and application for Business Analytics	
		KU5.01.09	Cognitive technologies for Business Analytics	
KAG6-DSBA: Business Analytics	KA05.02 DSBA.02/BAEM Business Analytics organisation and	KU5.02.00	Overview Business Analytics process organisation and enterprise management	
		KU5.02.01	Business processes and operations	

Knowledge Area Groups (KAG)	Knowledge Areas (KA)	Knowledge Unit (KU)	Suggested Knowledge Units (KU)	Mapping to CCS2012 and existing BoKs (DMBOK, BABOK, PMI-BoK, SWEBOK, ACM BoK)
	enterprise management	KU5.02.02	Project scope and risk management	**Proposed new KA/KU for DS-BoK** • General Business processes and operations KAs • Business processes and operations • Agile Data Driven methodologies, processes and enterprises • Use cases analysis: business and industry
		KU5.02.03	Business Analysis Planning and Monitoring	
		KU5.02.04	Requirements Analysis and Design Definition	
		KU5.02.05	Requirements Life Cycle Management (from inception to retirement)	
		KU5.02.06	Solution Evaluation and improvements recommendation	
		KU5.02.07	Agile Data Driven methodologies, processes and enterprises	
		KU5.02.08	Use cases analysis: business and industry	
		KU5.02.09 *)	Data management for BA/BI (Business Analytics, Business Intelligence), organisational models and requirements	
		KU5.02.010 *)	Data quality managemennt, FAIR data principles for organisational data	

*) KA and KU added to DS-BoK in EDSF Release 4

5 Conclusion and further developments

The presented work on defining the DS-BoK and other foundational components of the whole EDISON Data Science Framework has been done with the wide consultation and engagement of different stakeholders, primarily from the research community and European Research Infrastructures, but also involving industry experts via standardisation bodies, professional communities and directly via the project network.

5.1 Summary of the recent developments

The presented Data Science Body of Knowledge defines necessary knowledge areas and knowledge units required by the Data Science competences defined in the CF-DS document [1].

DS-BoK includes the following Knowledge Area groups (KAG):

- KAG1-DSDA: Data Analytics group including Data Analytics methods, Machine Learning, statistical methods, and data visualisation
- KAG2-DSENG: Data Science Engineering group including software engineering, database and Big Data technologies
- KAG3-DSDM: *Data Management group including data curation, preservation and data modeling*
- KAG4-DSRMP: *Research Methods and Project Management*
- KAG5-DSBA: Business Analytics (also strongly based on KAG1-DSDA)
- KAG*-DSDK: Placeholder for the Data Science Domain Knowledge groups to include domain specific knowledge

Valuable contribution was provided by the FAIRsFAIR project to the definition of the three main domains of the Data Stewardship professional domain:

- KAG3-DSDM: *Data Management group including data curation, preservation and data modelling*
    - Extended with the Open Science and FAIR related knowledge units
    - Added Data Stewardship specific knowledge aspects
- KAG4-DSRMP: *Research Methods and Project Management*
    - Extended with the knowledge required for Data Stewards effectively work in a research team and research projects
- KAG5-DSBA: Business Analytics (also strongly based on KAG1-DSDA)
    - Added knowledge aspects related to data management and data quality assurance expected from the Data Stewards

5.2 Further developments to formalize CF-DS and DS-BoK

It is anticipated that the presented ongoing development will require practical validation by experts and communities of practice that will include the following specific tasks and activities:

- Continue validating and improving the currently proposed knowledge areas and knowledge units by involving experts in the related knowledge areas, beneficially also engaging with the specific professional communities such as IEEE, ACM, DAMA, IIBA, etc.
- Formalise the taxonomy definition of the Data Science related knowledge areas and scientific disciplines based on ACM CCS (2012), provide suggestions for new knowledge areas and classifications classes.
- Collect feedback from the known pilot implementation of the EDSF and DS-BoK by the champion universities and wider community of practitioners to provide a further update to the DS-BoK.

Initial validation of the proposed DS-BoK has been done during the EDISON project lifetime by actively involving project partners and champion universities, engaging with the community of practice via workshops, seminars and active outreach activity, and soliciting feedback and contribution from the academic and professional community, including experts' interviews.

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

Future support for EDSF and DS-BoK in particular will be provided in the framework of the EDISON Community via github project space https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome.

6 References

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

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

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

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

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

[6] FAIR Competence Framework for Higher Education (Data Stewardship Professional Competence Framework), FAIRsFAIR Project Deliverable D7.3, February 2021 [online] https://zenodo.org/record/4562089#.Y6uctnbMK38

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

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

[9] ACM and IEEE Computer Science Curricula 2013 (CCS2013) [online] http://dx.doi.org/10.1145/2534860

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

[11] Information Technology Competency Model of Core Learning Outcomes and Assessment for Associate-Degree Curriculum, ACM Committee for Computing Education in Community Colleges (CCECC), 2014 [online] http://ccecc.acm.org/files/publications/ACMITCompetencyModel14October201420150114T180322.pdf

[12] The European Foundational ICT Body of Knowledge, Version 1.0, European Commission, 22 February 2015 [online] https://itprofessionalism.org/app/uploads/2021/02/The-European-Foundational_ICT-Body-of-Knowledge-2015-11.pdf

[13] CEN EN 17748-2:2022 Foundational Body of Knowledge for the ICT Profession (ICT BoK) User Guide and Methodology [online] https://www.en-standard.eu/pd-cen-tr-17748-2-2022-foundational-body-of-knowledge-for-the-ict-profession-ict-bok-user-guide-and-methodology/

[14] CEN/TC 428 - ICT Professionalism and Digital Competences [online] https://standards.cencenelec.eu/dyn/www/f?p=205:7:0::::FSP_ORG_ID:1218399&cs=16D21D7497970A5A38FB4CCE737358BFE

[15] Software Engineering Body of Knowledge (SWEBOK) [online] https://www.computer.org/web/swebok/v3

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

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

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

Acronyms

Acronym	**Explanation**
ACM	Association for Computer Machinery
BABOK	Business Analysis Body of Knowledge
CCS	Classification Computer Science by ACM
CF-DS	Data Science Competence Framework
CODATA	International Council for Science: Committee on Data for Science and Technology
CRISP-DM	Cross Industry Standard Process for Data Mining
CS	Computer Science
DigComp	Digital Competences for citizens (EU report 2017)
DM-BoK	Data Management Body of Knowledge by DAMAI
DS-BoK	Data Science Body of Knowledge
EDSA	European Data Science Academy
EOEE	EDISON Online E-Learning Environment
EOSC	European Open Science Cloud
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
FAIR	Findable, Accessible, Interoperable, Reusable data management principles to support Open Science
FAIRsFAIR	EU funded project to promote FAIR competences
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
P21C	21st Century Skills Framework
PID	Persistent Identifier
PM-BoK	Project Management Body of Knowledge
PRACE	Partnership for Advanced Computing in Europe
RDA	Research Data Alliance
SWEBOK	Software Engineering Body of Knowledge

Appendix A. Overview of Bodies of Knowledge relevant to Data Science

This section provides detailed information about existing Bodies of Knowledge relevant to the Data Science Body of Knowledge definition which are linked to or mapped to the current DS-BoK.

A.1. ICT Professional Body of Knowledge

Character	Explanation
Name of the Profession	ICT professional
Reference Community	(potentially) all ICT Professional
Leadership	Capgemini Consulting and Ernst & Young for the European Commission, Directorate General Internal Market, Industry, Entrepreneurship and SMEs
Organisation structure	N/A
Partners	N/A
Ethical Code	N/A
Estimated #members	N/A
Link to BoK	http://www.ictbok.eu/images/EU_Foundationa_ICTBOK_final.pdf
Year/Edition	2015/1st
Structure of BoK	There are 12 Knowledge Areas: ***1.*** ICT Strategy & Governance ***2.*** Business and Market of ICT ***3.*** Project Management ***4.*** Security Management ***5.*** Quality Management ***6.*** Architecture ***7.*** Data and Information Management ***8.*** Network and Systems Integration ***9.*** Software Design and Development ***10.*** Human Computer Interaction ***11.*** Testing ***12.*** Operations and Service Management Each Knowledge Area is defined by; • List of items required as foundational knowledge necessary under this Knowledge Area; • List of references to the e-Competence Framework (dimension 4: knowledge); • List of possible job profiles that require having an understanding of the Knowledge Area; • List of examples of specific Bodies of Knowledge, certification and training possibilities
Proposed use of BoK	• Education providers: as a source of inspiration for curricula design and development; • Professional Associations: to promote the Body of Knowledge to their members, ICT professionals; • HR Department and Managers within industry with a need to understand the range of knowledge and the entry level required by ICT professionals in order to improve recruiting and people development processes (together with skills and competencies).
Certification promoted	N/A

Character	Explanation
Name of the Profession	Data Management Professional Data Science Body of Knowledge (DS-BoK)
Reference Community	Mainly US Data managers, professionals and scholars. Relevant chapters in UK and Australia.
Leadership	DAMAI a Volunteer US-based organization governed by an Executive Board of Directors. Directors are voted in for a 2 year term of office and may stand for re-election
Organisation structure	The members adhere through the nearest local chapter and through that (autonomous organisations affiliated with the central associations) participate to the life of the community
Partners	US-based organisation of medium relevance that provides educational resources (Dataversity, DEBtech, IRM UK, Technics Publications) or instruments and tools (VoltDB)
Ethical Code	Yes (available for members https://www.dama.org/content/chapter-kit-behind-login)
Estimated #members	Conferences are attended by a thousand people, 16 Chapters worldwide. No references about a number of subscriptions
Link to BoK	BoK Framework • http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf DAMA International Guide to Data Management Body of Knowledge (on purchase) • https://technicspub.com/dmbok/ Other resources DAMA International Dictionary of Data Management Terms (on purchase) • https://technicspub.com/dmbok/
Edition/version	2012/v.2
Structure of BoK	The document is structured in 11 knowledge areas covering core areas in the DAMA - DMBOK2 Guide for performing data management. The 11 Data Management Knowledge Areas are: 1. **Data Governance** – planning, oversight, and control over management of data and the use of data and data-related resources. Governance covers 'processes', not 'things', hence the common term for Data Management Governance is Data Governance. 2. **Data Architecture** – the overall structure of data and data-related resources as an integral part of the enterprise architecture 3. **Data Modelling &Design** – analysis, design, building, testing, and maintenance (was Data Development in the DAMA - DMBOK 1st edition) 4. **Data Storage & Operations** – structured physical data assets storage deployment and management (was Data Operations in the DAMA-DMBOK 1st edition) 5. **Data Security** – ensuring privacy, confidentiality and appropriate access 6. **Data Integration & Interoperability** – acquisition, extraction, transformation, movement, delivery, replication, federation, virtualization and operational support (a Knowledge Area new in DMBOK2) 7. **Documents & Content** – storing, protecting, indexing, and enabling access to data found in unstructured sources (electronic files and physical records), and making this data available for integration and interoperability with structured (database) data. 8. **Reference & Master Data** – Managing shared data to reduce redundancy and ensure better data quality through standardized definition and use of data values. 9. **Data Warehousing & Business Intelligence** – managing analytical data processing and enabling access to decision support data for reporting and analysis 10. **Metadata** – collecting, categorizing, maintaining, integrating, controlling, managing, and delivering metadata

Field Code Changed

	11. **Data Quality** – defining, monitoring, maintaining data integrity, and improving data quality Each KA has section topics that logically group activities and it is described by a context diagram. There Is also an additional Data Management section containing topics that describe the knowledge requirements for data management professionals. Each context diagram includes: • *Definition*: a concise description of the Knowledge Area. • *Goals*: he desired outcomes of the Knowledge Area within this Topic. • *Process*: the list of discrete activities and sub-activities to be performed, with activity group indicators. • *Inputs*: what documents or raw materials are directly necessary for a Process to initiate or continue • *Supplier roles*: roles and/or teams that supply the inputs to the process. • *Responsible roles:* roles and/or teams that perform the process. • *Stakeholder roles*: roles and/or teams Informed or consulted on the process execution. • *Tools*: technology types used by the process to perform the function. • *Deliverables*: what is directly produced by the processes • *Consumer roles*: roles and/or teams that expect and receive the Deliverables. • *Metrics*: Measurements That quantify the success of Processes based on the Goals
Proposed use of BoK	• Informing a diverse audience about the nature and importance of data management. • Helping build consensus within the data management community. • Helping data stewards, data owners, and data professionals understand their responsibilities. • Providing the basis for assessments of data management effectiveness and maturity. • Guiding efforts to implement and improve data management knowledge areas. • Educating students, new hires, practitioners and executives on data management knowledge areas • Guiding the development and delivery of data management curriculum content for higher education. • Suggesting areas of further research in the field of data management. • Helping data management professionals prepare for Certified Data Management Professional (CDMP) data exams. • Assisting organizations in defining their enterprise data strategy.
Certification promoted	Certified Data Management Professional (CDMP) in four levels: • Associate (https://www.dama.org/content/cdmp-associate), • Practitioner (https://www.dama.org/content/cdmp-practitioner), • Master (https://www.dama.org/content/cdmp-master), • Fellow (https://www.dama.org/content/cdmp-fellow) Cost per exam: vary depending on the examination (from $220 of Associate till the 1560 for Master). Fellow is an assigned through nomination by peers and Chapter. Requirements: member of local chapter, sign/adhere to Ethical code/ proven experiences verifiable on the CV and contributions to the Association at various level

Field Code Changed

A.2. Data Management Professional Body of Knowledge

Character	Explanation
Name of the Profession	Project Management Professional
Reference Community	Industry-centered worldwide Project Managers
Leadership	ence Body of Knowledge (DS-BoK) Project Management Institute – www.pmi.org PMI is a worldwide not-for-profit professional membership association for the project, program and portfolio management profession. Founded in 1969, PMI delivers advocacy, collaboration, education and research to its members.
Organisation structure	PMI is governed by a 15-member volunteer Board of Directors. Each year PMI members elect five directors to three-year terms. Three directors elected by others on the Board serve one-year terms as officers. Day-to-day PMI operations are guided by the Executive Management Group and professional staff at the Global Operations Centre located near Philadelphia. Each member adheres through the nearest local chapter and through that (autonomous organisations affiliated with the central associations) participate to the life of the community
Partners	No specific partnership but some 1600 Registered Education Providers (R.E.P.s) and about 100 certified courses worldwide (http://www.pmi.org/learning/professional-development/global-accreditation-center.aspx)
Ethical Code	Yes (http://www.pmi.org/About-Us/Ethics/Code-of-Ethics.aspx#)
Estimated #members	700.000 in 195 countries (source www.pmi.org) [Estimated some 2,9 acting PM worldwide and some 1,5 million PM posts till 2020]
Link to BoK	http://www.pmi.org/PMBOK-Guide-and-Standards/pmbok-guide.aspx (on purchase - $46,17) other resources • Lexicon of PM terms (http://www.pmi.org/PMBOK-Guide-and-Standards/PMI-lexicon.aspx - free for members) • PMBoK in other 11 languages (Arabian, Italian, Korean, Russian, Hindi, Japanese, Portuguese, Spanish, German, French, Chinese); • **Software Extension to the PMBOK** Guide Fifth Edition (This standard, developed by PMI jointly with IEEE Computer Society, provides guidance on the management of software development projects, and bridges the gap between the traditional, predictive approach described in the PMBOK® Guide and iterative approaches such as agile more commonly used in software development) (on purchase – $37,07) External sites: http://www.projectmanagement.com/Practices/PMI-Standards/
Year/Edition	2014/5th edition
Structure of BoK	The Five Process Groups *Initiating* - Processes to define and authorize a project or project phase *Planning* - Processes to define the project scope, objectives and steps to achieve the required results. *Executing* - Processes to complete the work documented within the Project Management Plan. *Monitoring and Controlling* - Processes to track and review the project progress and performance. This group contains the Change Management. *Closing* - Processes to formalize the project or phase closure. The Nine Knowledge Areas *Project Integration Management* - Processes to integrate various parts of the Project Management. *Project Scope Management* - Processes to ensure that all of the work required is completed for a successful Project and manages additional "scope creep". *Project Time Management* - Processes to ensure the project is completed in a timely manner.

Field Code Changed

	Project Cost Management - Processes to manage the planning, estimation, budgeting and management of costs for the duration of the project. *Project Quality Management* - Processes to plan, manage and control the quality and to provide assurance the quality standards are met. *Project Human Resource Management* - Processes to plan, acquire, develop and manage the project team. Project Communications Management - Processes to plan, manage, control, distribute and final disposal of project documentation and communication. *Project Risk Management* - Processes to identify, analyse and management of project risks. *Project Procurement Management* - Processes to manage the purchase or acquisition of products and service, or result to complete the project. Each Process Group contains processes within some or all of the Knowledge Areas. Each of the 42 processes has Inputs, Tools & Techniques and Outputs. (It is not the scope of this analysis to enter into the details of each process).
Proposed use of BoK	It provides project managers with the fundamental practices needed to achieve organizational results and excellence in the practice of project management. It's a competence framework to support PM practices. It's used also as "one of the books" to pass the examination.
Certification promoted	Several certification other than the basic about Project Management Professional in correspondence of specific roles that the PM may adopt in the carrier or depending on the type of project (http://www.pmi.org/certification.aspx): CAPM – Certified Associate Project Management PMP – Project Management Professional PgMP – Program Management Professional PfMP – Portfolio Management Professional PMI–PBA – PMI-Professional Business Analyst PMI-ACP – PMI Agile Certified Professional PMI-RMP – PMI Risk Management Professional PMI-SP – Scheduling Professional *Cost*: it may vary from the $225 of CAPM till the $900 for PgMP and PfMP of non-Members; *Requirements*: general Education (Secondary school or Degree) + Experience on the field of certification + specific Education on the field of certification.

A.3. Project Management Professional Body of Knowledge

Appendix B. Subset of ACM/IEEE CCS2012 for Data Science (as defined in DS-BoK Release 1)

This Appendix provides historical information about the subset of the ACM/IEEE CCS2012 taxonomy used in the DS-BoK Release 1. This information is provided for those who build their Data Science curriculum definition on the previous DS-BoK version. The new DS-BoK Release 3 version has a whole set of generically defined knowledge areas and knowledge units that can be partly mapped to CCS2012 but primarily based on the knowledge topics defined in CF-DS document.

The defined below subset of ACM CCS (2012) classification can provide a basis for its future extension with a new classification group related to Data Science and individual disciplines that are missing in the current ACM/IEEE classification.

B.1. ACM Classification Computer Science (2012) structure and Data Science related Knowledge Areas

The 2012 ACM Computing Classification System (CCS) [6] 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 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:

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

Taxonomy is required to consistently present information about scientific disciplines and knowledge areas related to Data Science. Taxonomy is an important component to link such components as Data Science competences and knowledge areas, Body of Knowledge, and corresponding academic disciplines. From the practical point of view, the taxonomy includes the vocabulary of names (or keywords) and the hierarchy of their relations.

The presented here initial taxonomy of Data Science disciplines and knowledge areas is based on the 2012 ACM Computing Classification System (ACM CCS (2012)). Refer to the initial analysis of ACM CCS (2012) classification and a subset of data related disciplines in the DS-BoK Release 1. Table B.1 below includes ACM CCS (2012) subsets/subtrees that contain scientific disciplines that are related to Data Science Knowledge Area groups as defined in DS-BoK Release 1, which are compatible with the DS-BoK Release 2 and later:

- KAG1-DSDA: Data Analytics group, including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSENG: Data Science Engineering group, including Software Engineering and infrastructure engineering
- KAG3-DSDM: Data Management group, including data curation, preservation and data infrastructure

Two other groups, KAG4-DSRMP: Research Methods and Project Management and KAG5-DSBPM: Business Process Management cannot be mapped to ACM CCS (2012) and their taxonomy is defined based on other bodies of knowledge. It is important to notice that ACM CCS (2012) provides a top level classification entry "Applied computing" that can be used as an extension point for domain related knowledge area group KAG6-DSDK.

The following approach was used when constructing the proposed taxonomy:

- ACM CCS (2012) provides almost full coverage of Data Science related knowledge areas or disciplines related to KAG1, KAG2, and KAG3. The following top level classification groups are used:
- Theory of computation
- Mathematics of computing
- Computing methodologies
- Information systems
- Computer systems organization
- Software and its engineering
- Each of KAGs includes subsets from a few ACM CCS (2012) classification groups to cover theoretical, technology, engineering and technical management aspects.
- Extension points are suggested for possible future extensions of related KAGs together with their hierarchies.
- KAG3-DSDM: Data Management group is extended with new concepts and technologies developed by the Research Data Alliance community and documented in community best practices.

Table B.1 Data Science classification based on ACM Classification (2012)

DS-BoK Knowledge Groups *)	ACM (2012) Classification facets related to Data Science			
Data Science Analytics (DSDA)	Theory of computation			
		Design and analysis of algorithms		
			Data structures design and analysis	
		Theory and algorithms for application domains		
			Machine learning theory	
			Algorithmic game theory and mechanism design	
			Database theory	
		Semantics and reasoning		
Data Science Analytics (DSDA)	Mathematics of computing			
		Discrete mathematics		
			Graph theory	
		Probability and statistics		
			Probabilistic representations	
			Probabilistic inference problems	
			Probabilistic reasoning algorithms	
			Probabilistic algorithms	
			Statistical paradigms	
		Mathematical software		
		Information theory		
		Mathematical analysis		
Data Science Analytics (DSDA)	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	
Data Science Analytics (DSDA)	Information systems			
		Information systems applications		
			Decision support systems	
				Data warehouses
				Expert systems
				Data analytics
				Online analytical processing

DS-BoK Knowledge Groups *)	ACM (2012) Classification facets related to Data Science
	Multimedia information systems Data mining
Data Science Analytics (DSDA) EXTENSION POINT	Theory of computation DSA Extension point: Algorithms for Big Data computation Mathematics of computing DSA Extension point: Mathematical software for Big Data computation Computing methodologies DSA Extension point: New DSA computing Information systems DSA Extension point: Big Data systems (e.g. cloud based) Information systems applications DSA Extension point: Big Data applications DSA Extension point: Doman specific Data applications
Data Science Data Management (DSDM)	Information systems Data management systems Database design and models Data structures Database management system engines Query languages Database administration Middleware for databases Information integration
Data Science Data Management (DSDM)	Information systems Information systems applications Digital libraries and archives Information retrieval Document representation Retrieval models and ranking Search engine architectures and scalability Specialized information retrieval
Data Science Data Management (DSDM) EXTENSION POINT	Information systems Data management systems Data types and structures description Metadata standards Persistent identifiers (PID) Data types registries
Data Science Engineering (DSE)	Computer systems organization Architectures Parallel architectures Distributed architectures
Data Science Engineering (DSENG)	Networks **) Network Architectures Network Services Cloud Computing
Data Science Engineering (DSENG)	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 Grid computing Abstraction, modeling and modularity Real-time systems software Software notations and tools General programming languages Software creation and management
Data Science Engineering (DSENG)	Computing methodologies Modeling and simulation Model development and analysis Simulation theory Simulation types and techniques Simulation support systems
Data Science Engineering (DSENG)	Information systems Information storage systems Information systems applications

DS-BoK Knowledge Groups *)	ACM (2012) Classification facets related to Data Science
	Enterprise information systems
	Collaborative and social computing systems and tools
Data Science Engineering (DSENG)	Software and its engineering
	Software organization and properties
	DSE Extension point: Big Data applications design
EXTENSION POINT	Data Analytics programming languages
	Information systems
	DSE Extension point: Big Data and cloud based systems design
	Information systems applications
	DSA Extension point: Big Data applications
	DSA Extension point: Doman specific Data applications
DS Domain Knowledge (DSDK)	Applied computing
	Physical sciences and engineering
	Life and medical sciences
EXTENSION POINT	Law, social and behavioral sciences
	Computer forensics
	Arts and humanities
	Computers in other domains
	Operations research
	Education
	Document management and text processing

*) All Acronyms for classification groups and DS-BoK Knowledge Area Groups are brought in accordance to CF-DS-competence groups

**) Due to the important role of the Internet and networking technologies, basic knowledge about networks are required. However, as a technology domain, Networks knowledge area group should be considered a domain specific knowledge area in the general Data Science competences and knowledge definition.

Appendix H. Subset of ACM/IEEE CCS2012 for Data Science (as defined in the DS-BoK)

This Appendix provides historical information about subset of the ACM/IEEE CCS2012 taxonomy that provided the initial structure for the DS-BoK that was further extended with the full set of knowledge areas and knowledge units related to Data Science that can partly be mapped to CCS2012.

The defined below subset of ACM CCS (2012) classification can provide a basis for future CCS2012 extension with a new classification group related to Data Science and individual disciplines that are missing in the current ACM/IEEE classification.

The 2012 ACM Computing Classification System (CCS) [6] 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 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.

As an example, Cloud Computing is also a new technology and closely related to Big Data technologies currently 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

Taxonomy is required to consistently present information about scientific disciplines and knowledge areas related to Data Science. Taxonomy is an important component to link such components as Data Science competences and knowledge areas, Body of Knowledge, and corresponding academic disciplines. From the practical point of view, the taxonomy includes a vocabulary of names (or keywords) and the hierarchy of their relations.

The presented ACM CCS (2012) subsets/subtrees contain scientific disciplines related to three Data Science Knowledge Area groups as they are defined in DS-BoK:

- KAG1-DSDA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSENG: Data Science Engineering group including Software Engineering and infrastructure engineering
- KAG3-DSDM: Data Management group including data curation, preservation and data infrastructure

Two other groups KAG4-DSRMP: Research Methods and Project Management and KAG5-DSDK don’t have a direct mapping to ACM CCS (2012) and their taxonomies are defined based on other domain specific bodies of knowledge. It is important to notice that ACM CCS (2012) provides a top level classification entry “Applied computing” that can be used as an extension point for domain related knowledge area group KAG6-DSDK.

The following approach was used when constructing the proposed taxonomy:

- ACM CCS (2012) provides almost full coverage of Data Science related knowledge areas or disciplines related to KAG1, KAG2, and KAG3. The following top level classification groups are used:
- Theory of computation
- Mathematics of computing
- Computing methodologies

- Information systems
- Computer systems organization
- Software and its engineering
- Each of KAGs includes subsets from a few ACM CCS (2012) classification groups to cover theoretical, technology, engineering and technical management aspects.
- Extension points are suggested for possible future extensions of related KAGs together with their hierarchies.
- KAG3-DSDM: Data Management group is extended with new concepts and technologies developed by the Research Data Alliance community and documented in community best practices.

The following lists the ACM CCS2012 classification facets related to the Data Science grouped by DS-BoK Knowledge Area Groups *):

Data Science Analytics (DSDA) related CCS2012 facets

- Theory of computation
  - Theory of computation
    - Design and analysis of algorithms
      - Data structures design and analysis
    - Theory and algorithms for application domains
      - Machine learning theory
      - Algorithmic game theory and mechanism design
      - Database theory
    - Semantics and reasoning
  - Mathematics of computing
    - Discrete mathematics
      - Graph theory
    - Probability and statistics
      - Probabilistic representations
      - Probabilistic inference problems
      - Probabilistic reasoning algorithms
      - Probabilistic algorithms
      - Statistical paradigms
    - Mathematical software
    - Information theory
    - Mathematical analysis
  - 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

- Information systems
  - Information systems applications
    - Decision support systems
    - Data warehouses
    - Expert systems
    - Data analytics
    - Online analytical processing
  - Multimedia information systems
  - Data mining

CCS2012 extension points for DSDA

- Theory of computation
  - DSA Extension point: Algorithms for Big Data computation
- Mathematics of computing
  - DSA Extension point: Mathematical software for Big Data computation
- Computing methodologies
  - DSA Extension point: New DSA computing
- Information systems
  - DSA Extension point: Big Data systems (e.g. cloud based)
  - Information systems applications
    - DSA Extension point: Big Data applications
    - DSA Extension point: Doman specific Data applications

Data Science Data Management (DSDM) related CCS2012 facets

- Information systems
  - Data management systems
    - Database design and models
    - Data structures
    - Database management system engines
    - Query languages
    - Database administration
    - Middleware for databases
    - Information integration
  - Information systems applications
    - Digital libraries and archives
  - Information retrieval
    - Document representation
    - Retrieval models and ranking
    - Search engine architectures and scalability
    - Specialized information retrieval

Data Science Data Management (DSDM) extension facets

- Information systems
  - Data management systems
    - Data types and structures description
    - Metadata standards
    - Persistent identifiers (PID)
    - Data types registries

Data Science Engineering (DSENG) related CCS2012 facets

- Computer systems organization
  - Architectures
    - Parallel architectures
    - Distributed architectures
- Networks **)
  - Network Architectures
  - 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
        - Grid computing
      - Abstraction, modeling and modularity
      - Real-time systems software
  - Software notations and tools
    - General programming languages
  - Software creation and management
- Computing methodologies
  - Modeling and simulation
    - Model development and analysis
    - Simulation theory
    - Simulation types and techniques
    - Simulation support systems
- Information systems
  - Information storage systems
  - Information systems applications
    - Enterprise information systems
    - Collaborative and social computing systems and tools

Data Science Engineering (DSENG) extension facets

- Software and its engineering
  - Software organization and properties
    - DSE Extension point: Big Data applications design
    - Data Analytics programming languages
- Information systems
  - DSE Extension point: Big Data and cloud based systems design
  - Information systems applications
    - DSA Extension point: Big Data applications
    - DSA Extension point: Doman specific Data applications

DS Domain Knowledge (DSDK) Extension Points

- Applied computing

- Physical sciences and engineering
- Life and medical sciences
- Law, social and behavioral sciences
- Computer forensics
- Arts and humanities
- Computers in other domains
- Operations research

*) All Acronyms for classification groups and DS-BoK Knowledge Area Groups are brought in accordance to CF-DS-competence groups

**) Due to important role of the Internet and networking technologies, basic knowledge about networks are required. However, as a technology domain, Networks knowledge area group should be considered as a domain specific knowledge area in the general Data Science competences and knowledge definition