Project Title	Fostering FAIR Data Practices in Europe
Project Acronym	FAIRsFAIR
Grant Agreement No	831558
Instrument	H2020-INFRAEOSC-2018-4
Topic	INFRAEOSC-05-2018-2019 Support to the EOSC Governance
Start Date of Project	1st March 2019
Duration of Project	36 months
Project Website	www.fairsfair.eu

D7.3 FAIR Competence Framework for Higher Education (Data Stewardship Professional Competence Framework)

Work Package	WP7, FAIR Competences for Higher Education
Lead Author (Org)	Yuri Demchenko (UvA)
Contributing Author(s) (Org)	Lennart Stoy (EUA), Claudia Engelhardt (UGOE), Vinciane Gaillard (EUA)
Due Date	28.02.2021
Date	24.02.2021
Version	1.0
DOI	https://doi.org/10.5281/zenodo.4562088

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Abstract

This report presents a proposed FAIR Competence Framework for Higher Education (FAIR4HE) that is defined as a part of the general Data Stewardship Professional Competence Framework (CF-DSP) presented in the deliverable. The proposed CF-DSP defines the set of competences that extend the competences initially defined in the EDISON Data Science Framework (EDSF). The proposed competence framework is defined based on a recent job market analysis for the Data Steward and related professions. The presented CF-DSP has been validated against existing Data Stewardship competence frameworks defined primarily for the research community or practitioners. CF-DSP provides the competences definition structure that allows easy mapping to a Body of Knowledge and set of Learning Outcomes that can be used for defining academic curricula. The presented CF-DSP has been discussed with, and incorporated feedback from, several community events organised by the FAIRsFAIR project.

Versioning and contribution history

Version	Date	Authors	Notes
0.0	01.09.2020	Yuri Demchenko (UvA)	ToC with draft notes
0.1	27.10.2020	Yuri Demchenko (UvA), Lennart Stoy (EUA), Vinciane Gaillard (EUA)	First draft of M7.7
0.2	09.11.2020	Yuri Demchenko (UvA)	Revised for M7.5 intermediate
0.3	12.11.2020	Yuri Demchenko (UvA)	Revised for M7.5
0.4	25.01.2021	Yuri Demchenko (UvA)	Final draft for internal WP7 review
0.5	02.02.2021	Yuri Demchenko (UvA), Lennart Stoy (EUA), Claudia Engelhardt (UGOE)	Revision based on WP7 internal review
0.6	21.02.2021	Laura Molloy (CODATA), Ari Asmi (UH)	Internal review
0.7	22.02.2021	Lennart Stoy (EUA), Vinciane Gaillard (EUA)	Reviewers comments consolidated and partly addressed
1.0	26.02.2021	Yuri Demchenko (UvA) Lennart Stoy (EuA)	Content ready

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Disclaimer

FAIRsFAIR has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no. 831558 The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Abbreviations and Acronyms

BABOK	Business Analytics Body of Knowledge
CF-DS	Data Science Competence Framework (part of EDSF)
CF-DSP	Data Stewardship Professional Competence Framework (part of FAIRsFAIR Project)
CODATA	Committee on Data of the International Science Council
CS-BoK	Computer Science Body of Knowledge (specified jointly by ACM and IEEE)
DAMA/DAMAI	Data Management Association International
DMBOK	Data Management Body of Knowledge by DAMAI
DOI	Digital Object Identifier
DSBA	Data Science Business Analytics domain knowledge area in DS-BoK
DS-BoK	Data Science Body of Knowledge
DSP4LS	Data Steward Professional Competence Framework for Life Sciences (used in this document to refer to ELIXIR Data Steward Competence Framework)
e-CF	European e-Competence Framework (currently EU standard EN 16234-1: 2019)
EDSF	EDISON Data Science Framework
ELIXIR	ELIXIR is an intergovernmental organisation that brings together life science resources from across Europe.
EOSC	European Open Science Cloud
EOSCpilot	European Open Science Cloud for Research pilot project (www.eoscpilot.eu)
FAIR4S	FAIR Competence Framework for Stewardship
FAIR	Findable, Accessible, Interoperable, Reusable data properties
FAIRsFAIR	Fostering FAIR Data Practices in Europe (https://fairsfair.eu/)
FOSTER	Facilitate Open Science Training For European Research project (https://www.fosteropenscience.eu/)
GDPR	General Data Protection Regulation

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

GO FAIR	GO FAIR is a bottom-up, stakeholder-driven and self-governed initiative that aims to implement the FAIR data principles (https://www.go-fair.org/)
HEIs	Higher Education Institutions
LO	Learning Objectives (as it is used in some competences frameworks Learning Outcomes (as it is used in EDSF and ACM/IEEE Curricula Guidelines)
MC-DS	Data Science Model Curriculum
PID	Persistent Identifier
PM-BoK	Project Management Professional Body of Knowledge
RDA	Research Data Alliance
RDM	Research Data Management
RI	Research Infrastructure

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Executive Summary

Skills for FAIR data and Open Science are cornerstones for the wide implementation of the FAIR principles in research communities and the establishment of European Open Science Cloud (EOSC). The availability of skilled researchers and other professional staff, such as data stewards, within research organisations, private companies and other organisations is an important component in this endeavour.

The professionalisation of data stewardship, as well as its adoption within domain-specific contexts, requires clear definition of knowledge, skills and competences for these professional functions. The given report proposes a framework *FAIR Competence Framework for Higher Education (FAIR4HE)* for core competences related to the four areas of Data Management and Governance, Data Engineering, Research Methods & Project Management, and Domain-related competences.

In doing so, the report extends and updates core competences initially defined in the EDISON Data Science Framework (EDSF) to address the additional requirements posed by the increasing relevance of the FAIR principles. By building on EDSF, FAIR4HE is placed within the general context of data science professions.

The proposed competence framework was defined following a job market analysis for the Data Steward and related professions. It has been validated against existing Data Stewardship competence frameworks defined primarily for the research community or practitioners.

The underlying Competence Framework for Data Stewardship (CF-DSP) provides a competences definition structure that allows easy mapping to a Body of Knowledge and Learning Outcomes that can be used for defining academic curricula. The presented CF-DSP has been discussed with and incorporated feedback from several community events organised by the project.

In addition, the FAIRsFAIR project extended the initial Body of Knowledge underlying EDSF. Together with CF-DSP, it can be used to define Data Stewardship university curricula and courses that respond to the needs of a given community or target stakeholders.

The report provides a basis for building consensus on defining Data Stewardship competences and a corresponding Body of Knowledge and Model Curriculum that can be adopted by wider groups of adopters. There are a variety of organisations and initiatives available that address different aspects of FAIR and Data Stewardship competences, training and curricula. For this purpose, the report outlines recommendations and future actions in the areas of adoption, sustainability and dissemination of the proposed FAIR4HE framework.

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Table of Contents

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

1. Introduction

Professionalising Data Stewardship and wide implementation of the FAIR data principles and culture by all communities working with data is a priority at the present time as both science and industry undergo a digital transformation where the data management, data quality and data literacy of staff are crucial for organisations to achieve effectiveness, economy and competitiveness.

Adopting and realising the potential of the European Digital Single Market will require building robust data infrastructures for research and industry as well as connecting them for effective knowledge transfer. A skilled and data-literate workforce capable of developing, operating and using data-driven services and processes both in industry and in science is an important component of ongoing science and industry digital transformation.

The IDC Study on European Data Market estimated a gap in supply and demand for data related professions by 2020 at 800,000 workers[1]. The critical situation with supplying data related workers is reported in studies related to the European Research Area. The High Level Expert Group on EOSC estimated the need for "core data experts" for EOSC would "likely exceed half a million within a decade"[2].

In response to this urgent and growing need for FAIR data competences, FAIRsFAIR aims to support Higher Education Institutions (HEIs) to increase their capacity to equip more graduates in data science and other disciplines with FAIR data competences, in line with the FAIR data Expert Group recommendations to professionalise data science and stewardship and "to coordinate, systemise and accelerate the pedagogy and availability of training for data skills, data science and data stewardship"[3].

The overall objective of the FAIRsFAIR project is to accelerate the realization of the goals of the EOSC by opening up and sharing all knowledge, expertise, guidelines, implementations, new trajectories, courses and education on FAIR matters. It seeks to establish a level playing field for all European member states (and beyond) when it comes to contributing data to scientific and scholarly communities and to re-using data from scientists and scholars elsewhere.

To recognise the importance of introducing FAIR principles and culture in conducting scientific research at an early stage of professional education and career, the project devotes special attention in the framework of FAIRsFAIR WP7, Task 7.3 to define a FAIR competence profile

[1] Final results of the European Data Market study measuring the size and trends of the EU data economy, ECIDC, March 2017 [online] Available at https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

[2] Realising the European Open Science Cloud, First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud, 2016 [online] Available at https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

[3] Turning FAIR data into reality: Final report of the European Commission Expert Group on FAIR Data [online] Available at https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

for Higher Education that can be introduced at the early stages of university education/programs.

The objective of Task 7.3. is to develop a FAIR data competence framework for Higher Education (FAIR4HE) which is complementary to or as an extension to existing and adopted data science and other competence frameworks. The FAIR4HE framework should help implement existing FAIR and Data Stewardship competence frameworks in academic curricula. This work focuses primarily on Data Science and Data Stewardship programs but is also applicable to other disciplines that are important in fostering a FAIR data culture throughout different scientific disciplines, research communities and professions.

The FAIR data competence framework includes FAIR data competences, which can be acquired via the higher education, as well as FAIR data competences for practitioners dealing with research data management (e.g. data stewards, research infrastructure managers) that can be delivered via professional and vocational training.

1.1 Scope

This report presents the progress towards the proposed *FAIR Competence Framework for Higher Education (FAIR4HE)* that can be used by universities and professional training organisations to develop their own education and training programs adopted to their own scientific domain that will bring the FAIR related competences into the Data Stewardship and Data Science curricula as well as into more focused training on Research Data Management.

The proposed FAIR4HE framework includes a Data Stewardship Professional Competence Framework and recommended content of the Data Stewardship Body of Knowledge that can be used for developing a special course and/or included as a part of other courses typically taught at universities such as Professional Issues, Research Methods or Data Management and Governance in specific academic and scientific domains.

This deliverable takes forward the outcome of the deliverable D7.2 “Briefing on FAIR Competences and Synergies”[4] and refers to this deliverable in many places, in particular to the extensive overview of existing FAIR and Data Stewardship competence and skill frameworks and practices in different Research Infrastructures and organisations.

The presented FAIR4HE framework intends to further build a consensus in defining the Data Stewardship and FAIR competences among already existing works and propose a common general view on how to design the framework in a way that enables the advancement of a FAIR data culture in university curricula.

[4] D7.2 “Briefing on FAIR Competences and Synergies”, FAIRsFAIR Project Deliverable, September 2020 [online] Available at https://doi.org/10.5281/zenodo.4009006

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

1.2 Approach

In defining FAIR4HE, Task 7.3 follows the EDISON project methodology that was used for defining the EDISON Data Science Framework (EDSF)[5] and later used in several European projects dealing with Data Science and Data Stewardship competences and education curricula definition.

1. The Data Stewardship competences and skills together with the corresponding knowledge topics are defined in the EDSF as a Data Stewards professional profile [DSPP], which is a part of the whole Data Science professional family . More details on the specific competences and skills for Data Stewards are provided in other existing frameworks such as EOSCpilot FAIR4S, ELIXIR Data Stewardship Competency Framework for Life Sciences (referred to in this document as DSP4LS), and DeIC Data Stewardship curricula recommendations.

2. To establish a common ground for defining Data Stewardship and FAIR competences, the project has analysed the job market to identify what competences and skills are required/expected from the Data Stewards and how the FAIR competences are present in the posted job vacancies.

3. Based on the job market analysis the base EDSF Data Stewardship competence profile (defined in EDSF) was revised, extended and enriched with the complementary competences, skills and knowledge topics defined in other frameworks. The proposed FAIR4HE framework is aligned with competences and skills definitions in the above mentioned frameworks.

4. The constructed competence framework is can be used for defining the corresponding Body of Knowledge and Model Curriculum learning units that can be directly mapped to existing or required educational courses or training modules, which can be used for education or training.

Further development of the FAIR and Stewardship model curriculum will be done in Task 7.4, “Development of FAIR model courses and curricula.”

1.3 Process and community contribution

Task 7.3 follows a process of internal discussions within FAIRsFAIR and outward-looking events to gather feedback from other stakeholders, experts and communities.

Following initial discussions as part of a Focus Group on the topic of “Teaching (FAIR) Data Management and Data Stewardship” on 19 November 2019 (organised through FAIRsFAIR Task 7.1), an implementation plan was developed by UvA which detailed the interim steps of work to be done within Task 7.3.

[5] Refer to EDSF summary in section 2.7.

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

A cross-WP discussion was held on 8 May 2020 to gather feedback from project partners within FAIRsFAIR. To understand what the community expects from a FAIR4HE, the WP7 and Task T7.3 held the FAIR4HE Design Workshop[6] on 8-9 October 2020. The workshop involved developers of the main existing Data Stewardship and FAIR competence frameworks together with representatives from the wider community interested in adopting FAIR principles in research and education.

The workshop aimed to exchange experience and best practices from different projects, programs and initiatives and identify complementary components that have been taken for improving the FAIR4HE definition, to identify challenges with adoption of FAIR competences in the university environment and to address them in the proposed framework. The workshop program is included in Annex A.

A further workshop presenting the main outcomes of M7.5 and gathering additional community feedback was held on 30 November 2020 as part of a session at the FAIR Convergence Symposium under the leadership of UvA and EUA entitled "Data Stewardship and FAIR Competences in Academic Curricula."[7]

1.4 Input from other FAIRsFAIR work packages and external activities

Task T7.3 activity and FAIR4HE development were done in close cooperation with other WPs in the FAIRsFAIR project. Contributions were also received from the RDA Interest Group on Data Stewardship Professionalisation[8] and the EDISON Community Initiative that currently maintains the EDISON Data Science Framework[9].

1.5 Report Structure

Chapter 2 revisits selected existing frameworks to extract and learn from best practices in defining the Data Stewardship and FAIR related competences that are implemented in the proposed FAIR4HE framework. It is noted that most of them refer to the EDISON Data Science Framework (EDSF) and methodology, which are used in the proposed FAIR4HE. The chapter thus provides an overview of EDSF and the methodology that provides a basis for extracting the essential competences, skills and topics from the job market analysis and further using this information for defining the Body of Knowledge and Model curriculum.

Chapter 3 applies the EDSF methodology to define detailed competences for Data Stewards and FAIR competences. The basis for Chapter 3 is a recent job market analysis leading to an

[6] See Appendix A for the Program and FAIR4HE Design Workshop, 8-9 Oct 2020 [online] Available at https://docs.google.com/document/d/1lVLSJL41wtZa40Y7rHoYQahjNGXQ02dsk45goniR9sA/edit#heading=h.89nh0hqdfd91

[7] Data Stewardship and FAIR Competences in Academic Curricula, Part of International FAIR Convergence Symposium [online] Available at https://conference.codata.org/FAIRconvergence2020/sessions/222/

[8] RDA IG on Professionalising Data Stewardship [online] https://www.rd-alliance.org/groups/professionalising-data-stewardship-ig

[9] EDISON Community Initiative [online] Available at https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

identification of important competences and knowledge demanded from Data Stewards. It is noted that besides general data management and FAIR data competences there are other related competences that are required from the Data Stewards to implement and maintain FAIR data in organisations, during the whole data lifecycle.

Chapter 4 provides suggestions for defining the Data Stewardship and FAIR related Body of Knowledge (DSP-BoK) that includes knowledge topics identified in the job vacancies analysis, which at the current stage is defined as a subset of the general Data Science Body of Knowledge (DS-BoK).

Chapter 5 presents recommendations and suggested steps for the effective implementation of the FAIR data principles as part of Data Science or Data Stewardship curricula, part of which will be taken up in Task 7.4 FAIR Competences Adoption Handbook for Universities and Task 7.5 workshops program of the FAIRsFAIR project.

Chapter 6 provides a short summary and conclusion of the presented work, suggestions for further development.

2. Overview of Existing Data Stewardship and FAIR competence Frameworks

This chapter provides a short summary of the existing frameworks and standards that essentially contributed to the proposed definition of FAIR4HE and are required for understanding the FAIR4HE alignment with other frameworks and developments.

The proposed summary is based on the FAIRsFAIR Deliverable 7.2 “Briefing on FAIR Competences and Synergies” that provided an overview of various FAIR-data related competence frameworks and training initiatives, forming a basis for the definition of the FAIR4HE framework. The following frameworks are analysed to extract and map identified competences, skills, knowledge topics:

- EOSCpilot FAIR4S Data Stewardship Competence Framework
- ELIXIR Data Stewardship Competence Framework (DSP4LS)
- DeIC and DM Forum: Report on National Coordination of Data Steward Education in Denmark
- FOSTER Open Science Learning outcomes
- GO FAIR Data Principles and Maturity Framework
- DAMA BoK (2007) DAMAI Data Management Body of Knowledge
- EDISON Data Science Framework (EDSF) and EDISON Community Initiative

2.1 EOSCpilot FAIR4S Framework

The EOSCpilot project defines data stewardship as a shared responsibility of professional groups involved in different data management activities: data management and curation, data science and analytics, data services engineering and domain research. The EOSCpilot

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

deliverable "D7.5: Strategy for Sustainable Development of Skills and Capabilities"[10] describes the comprehensive FAIR4S framework that defines six skill profiles grouped around the research data lifecycle stages and four professional groups (researchers, data scientists, data advisors, and data services providers) involved into different aspects of data management, data curation and related services provisioning. The defined FAIR4S is primarily focused on the EOSC services as they were defined in the EOSCpilot project 2017-2019.

The total 31 individual competences and capabilities that are defined in FAIR4S are grouped into the following groups around typical processes and stages in the research data lifecycle[11]:

- Plan and design: Plan stewardship and sharing of FAIR outputs,
- Capture and process: Reuse data from existing sources,
- Integrate and analyse: Use or develop FAIR research tools/services,
- Appraise and preserve: Prepare and document data/code to make outputs FAIR,
- Publish and release: Publish FAIR outputs on recommended repositories,
- Expose and discover: Recognise, cite and acknowledge contributions.

The FAIR4S framework defined two templates for describing the Skills profiles and Role profiles. The Skills profile template includes knowledge, skills and attitude (that can also be treated as aptitude) for three levels of proficiency Basic, Intermediate, Expert. The template also includes a list of professional groups and roles to which the competence group applies. The Role profile includes the list of suggested skills, an explanation of their importance and suggestions where these skills can be learned.

Applicability and use for FAIR4HE

The FAIR4S framework provided a valuable analysis of the FAIR competences for Data Stewardship from the point of view of the EOSC projects. It defines competences as a combination of *knowledge*, *skills* and *attitude*, an approach that has been used in other frameworks and also used in the proposed FAIR4HE/CF-DSP. The definition of the three levels of proficiency is important for defining learning outcomes when developing academic and training curricula.

2.2 ELIXIR Data Stewardship Competency Framework

The ELIXIR Data Stewardship Competency Framework for life sciences[12] (hereafter referred to as DSP4LS – Data Steward Profession for Life Sciences) is the most complete of the reviewed frameworks. It defines the competencies, skills and knowledge related to Data Stewardship as a distinct profession in the modern data driven science ecosystem and the life

[10] EOSCpilot Deliverable D7.5 Strategy for sustainable development of skills and capabilities [online] Available at https://eoscpilot.eu/content/d75-strategy-sustainable-development-skills-and-capabilities

[11] The intermediate EOSCpilot deliverable D7.3 (2018) contained 59 individual competences that included both data lifecycle groups and general activities groups. [online] Available at https://eoscpilot.eu/sites/default/files/eoscpilot-d7.3.pdf

[12] Towards FAIR Data Steward as profession for the Life Sciences, Final report ZonMw & ELIXIR-NL projects (Oct 3, 2019) [online] Available at https://doi.org/10.5281/zenodo.3471707

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

sciences in particular. The framework allows translating the Data Stewards organisational responsibilities and tasks, together with required knowledge, skills and abilities into practical learning objectives that provide a basis for developing tailored training. In this way, the framework provides a strong foundation for professionalizing Data Stewardship.

The DSP4LS starts from defining the Data Steward Roles and Competence Profiles in the following three areas:

- Policy: institute and policy focused
- Research: project and research focused
- Infrastructure: data handling and e-infrastructure focused

For all Data Steward roles the eight competence areas are defined: Policy/strategy; Compliance; Alignment with FAIR data principles; Services; Infrastructure; Knowledge management; Network; Data archiving. In the extended definition, for each competence the following attributes are defined:

- Activities and tasks (in the organisational context)
- Knowledge, Skills and Abilities
- Learning Objectives (LO) formulated as "*after successful completing training you will be able to [..]*"

Applicability and use for FAIR4HE

The DSP4LS provides the most complete definition of the Data Stewardship competences for three profiles defining main responsibilities and organisational roles of the Data Stewards with focus on Policy, Research, and Infrastructure. The defined eight competence areas reflect the whole spectrum of the activities conducted by Data Stewards in organisations and research processes. The presented detailed definition of the Learning Objectives can be directly used for Data Stewardship curriculum definition.

2.3 DeIC Data Stewardship curricula recommendations/principles

The Danish e-Infrastructure Cooperation (DeIC) and Danish National Forum for Research Data Management (DM Forum) Report on National Coordination of Data Steward Education in Denmark[13] provided valuable recommendations on defining Data Stewardship curricula, primarily aligned with the Danish research environment. The report is based on the strong evidence base derived from the LinkedIn profiles analysis (74 profiles analysed during March 2019) and Job vacancies database in Denmark analysis (119 vacancies of Data Scientists and Data Stewards analysed during March-April 2019) and an extensive overview and analysis of existing competence frameworks and educational programmes for Data Science and Data Stewardship. In addition, the community feedback was collected via a Questionnaire that received 86 complete responses (and 42 partial responses).

[13] The Danish e-Infrastructure Cooperation (DeIC) and Danish National Forum for Research Data Management (DM Forum) Report on National Coordination of Data Steward Education in Demark [online] Available at https://www.deic.dk/sites/default/files/Data%20Steward%20Education%20in%20Denmark_0.pdf

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The Data Stewardship competences are defined in six competence groups comprising 22 competences related to: Open Science, Data Collection and Data Processing, Data publishing and data preservation, and competences related to research data lifecycle phases: Planning phase, Active research phase, and Dissemination/publication phase.

The report defined the four roles for Data Stewards: Administrator; Analyst; Developer; Agent of change.

The report proposed three modes for Data Stewards education (based on the prospective student/learner background and entry level):

- Student with Bachelor degree
- Student with PhD and equivalent
- Continuing and professional education

The competences identified for the Data Stewardship curricula are grouped in three groups reflecting main Data Stewards responsibility areas in organisations:

i) Open Science
   - Open Science policies
   - Data management plans
   - Rights, licenses

ii) Data collection and data processing
   - Data and Source Search and Data Collection
   - Data storage (in connection with data collection, data storage and storage of active data in project process)
   - Data processing
   - Open Reproducible Research

iii) Data publishing and data preservation
   - Data archiving (finished data) and long-term storage
   - Data publishing
   - "Scientific publishing / scholarly communication"
   - Open Access publishing

and other competences related to research data lifecycle phases: Planning phase, Active research phase, and Dissemination/publication phase.

Applicability and use for FAIR4HE

The proposed Data Stewardship education format and curriculum design approaches provide a valuable example of the knowledgeable community approach to defining Data Stewardship competences and designing curriculum profiles for selected groups of learners such as master students, PhD students and practitioners.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2.4. GO FAIR Metadata Management Requirements and FAIR Data Maturity Model

The GO FAIR initiative[14] which is devoted to promotion and sustainable adoption of the FAIR data principles[15] provides recommendations on FAIR metadata management[16] that can be used for linking between general requirements to FAIR implementation and underlying technology and infrastructure and consequently for defining technical expertise areas. These requirements are compiled in Table 1 and later discussed when defining the required FAIR competences, skills and knowledge topics.

Table 1. FAIR metadata requirements

Findable	• F1 (meta)data are assigned a globally unique and persistent identifier; • F2 data are described with rich metadata; • F3 metadata clearly and explicitly include the identifier of the data it describes; • F4 (meta)data are registered or indexed in a searchable resource.
Interoperable	• I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation; • I2. (meta)data use vocabularies that follow FAIR principles; • I3. (meta)data include qualified references to other (meta)data.

Accessible	• A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
	  ○ A1.1 the protocol is open, free, and universally implementable;
	  ○ A1.2 the protocol allows for an authentication and authorization procedure, where necessary;
	• A2 metadata are accessible, even when the data are no longer available.

Reusable	• R1 meta(data) are richly described with a plurality of accurate and relevant attributes; • R1.1 (meta)data are released with a clear and accessible data usage license; • R1.2 (meta)data are associated with detailed provenance; • R1.3 (meta)data meet domain-relevant community standard.

[14] GO FAIR Initiative [online] Available at https://www.go-fair.org/go-fair-initiative/

[15] The FAIR Guiding Principles for scientific data management and stewardship, March 2016, Scientific Data 3(160018 (2016)) [online] Available at https://doi.org/10.1038/sdata.2016.18

[16] EOSCpilot deliverable "D7.5: Strategy for Sustainable Development of Skills and Capabilities" [online] Available at https://eoscpilot.eu/sites/default/files/eoscpilot-d7.5-v1.1.pdf

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The FAIR Data Maturity Model[17] which was developed and is maintained by the RDA community[18] provides a set of compliance indicators to assess the level of implementation of the FAIR principles and can be used for defining policy, research and infrastructure related competences in Data Stewardship and data management.

Applicability and use for FAIR4HE

The GO FAIR definition of metadata requirements and the RDA FAIR Data Maturity Model provides valuable insight into required infrastructure technologies and technical competences needed for consistent implementation of the FAIR data principles in research, industry and business. This information has been used in defining the necessary engineering competences for Data Stewards.

2.5 DAMA DMBOK: Data Governance and Stewardship

The Data Management Body of Knowledge (DMBOK) Framework by Data Management Association International (DAMAI) is an industry standard summarizing best practices in Data Management[19]. It is a valuable document that provides a basis for setting up organisational policy and structure to ensure consistent data management and governance. The DMBOK is directly used for training and certification of several data management and governance professions and roles. It goes into depth about the Knowledge Areas that make up the overall scope of data management.

The DMBOK defines 11 main Knowledge Areas and several additional areas related to technologies used. Each Knowledge Area is provided with a detailed context diagram that includes: Definition, Goals, Inputs, Activities, Deliverables, Suppliers, Participants, Consumers, Tools, Technics and Metrics – that can be used as a direct guidance for organisations setting up their data management and governance structure.

The Data Governance and Stewardship Knowledge Area is defined as central for the whole DMBOK. The DMBOK also explains the relation between Data Governance and Data Management, where Data Governance is focused on a Legal and Judicial view ("Do right things") and Data Management is dealing with Executive issues ("Do things right"). This also defines staffing of the Data Governance Office: Chief Data Steward, Executive Data Steward, Coordinating Data Steward, Business Data Steward roles. Data Management functions are performed by the Chief Information Officer office that includes Data Architects, Data Analysts, Coordinating Data Stewards and technical Data Steward roles.

[17] FAIR Data Maturity Model [online] Available at https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v0.90.pdf

[18] RDA Data maturity model Working Group [online] Available at https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

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

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Data Management principles according to DMBOK provide a good summary of best practices that can be included in data management curricula and training:

- Data is an asset with unique properties
- The value of data can and should be expressed in economic terms
- Managing data means managing the quality of data
- It takes Metadata to manage data
- It takes planning to manage data
- Data management requirements must drive Information Technology decisions
- Data management is cross-functional; it requires a range of skills and expertise
- Data management requires an enterprise perspective
- Data management must account for a range of perspectives
- Data management is lifecycle management
- Different types of data have different lifecycle characteristics
- Managing data includes managing the risks associated with data
- Effective data management requires leadership commitment

Data Steward organisational roles

The Data Steward is a core role to execute the organisational Data Governance and Data Management Policy: define, implement, embed. They typically belong to the Chief Data Officer office. The DMBOK defines the core Data Steward activity as follows:

Creating and managing core Metadata: Definition and management of business terminology, valid data values, and other critical Metadata. Documenting rules and standards: Definition/documentation of business rules, data standards, and data quality rules. High quality data are often formulated in terms of rules rooted in the business processes that create or consume data. Stewards help surface these rules and ensure their consistent use.

Managing data quality issues: Stewards are often involved with the identification and resolution of data related issues or in facilitating the process of resolution.

Executing operational data governance activities: Stewards are responsible for ensuring that, day-to-day and project-by-project, data governance policies and initiatives are adhered to. They should influence decisions to ensure that data is managed in ways that support the overall goals of the organization.

The importance of having a devoted Data Steward in the organisation is recognised by a remark in the first version of the DMBOK1 (2009): “Best Data Steward is not made but found”.

Applicability and use for FAIR4HE

The DMBOK provides a reference model and approach for defining the baseline Research Data Management and Data Stewardship Body of Knowledge and Knowledge Areas. DMBOK is used in the EDISON Data Science Framework for defining Data Management and Governance Knowledge Area Group (KAG) that is extended with the topics related to Research Data Management; further extension should include FAIR data related knowledge topics. It is

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

important to align the definition of the FAIR4HE and Data Stewardship competence framework with the DMBOK as an industry standard, understanding that the majority of university graduates will work in the industry. The industry best practices can provide also a contribution to improving the definition of Data Stewardship for the research area.

2.6 EDISON Data Science Framework and Data Steward Professional Profile Definition

2.6.1 Components of EDISON Data Science Framework (EDSF Release 4)[20]

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

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- DSPP - Data Science Professional Profiles and occupations taxonomy
- Data Science Taxonomy and Scientific Disciplines Classification

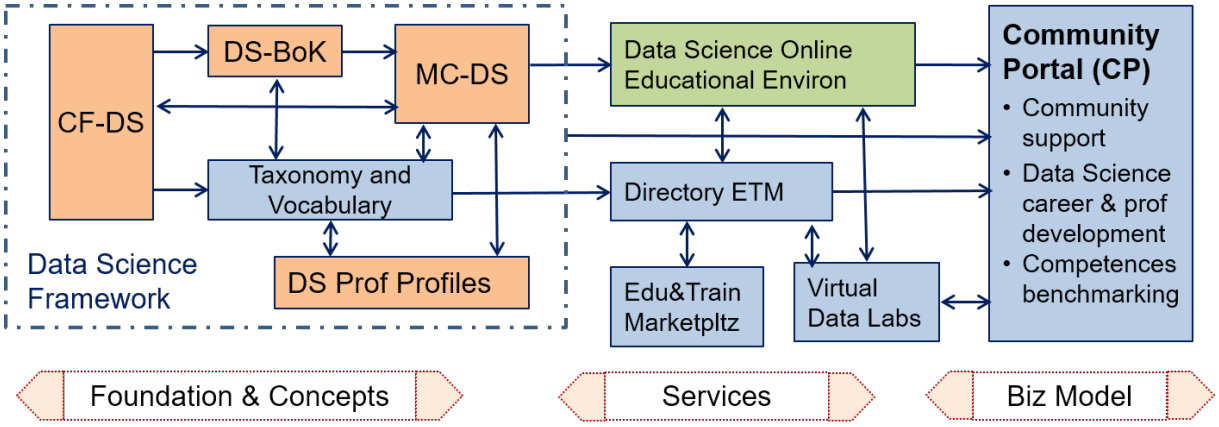


Figure 1 EDISON Data Science Framework components and related services.

The CF-DS provides the overall basis for the whole EDSF. The core CF-DS includes the common competences required for Data Scientists in different work environments in industry and in research and through the whole career path. The ongoing CF-DS development includes coverage of domain specific competences and skills based on the contribution of domain and subject matter experts.

[20] EDISON Data Science Framework (EDSF). [online] Available at https://github.com/EDISONcommunity/EDSF

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. The 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. The DS-BoK incorporates best practices in Computer Science and domain specific BoKs and includes KAs and KUs defined where possible based on the Classification Computer Science (CCS2012)[21], components taken from other BoKs and proposed new KAs/KUs to incorporate new technologies used in Data Science and its recent developments.

The Data Science Model Curriculum (MC-DS) is built based on the 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 the DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome[22] 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 Data Science Professional Profiles and occupations taxonomy (DSPP) is defined as an extension to the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy[23] using the ESCO top classification groups. The DSPP definition provides an 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.

2.6.2 Data stewards and data management related professional profiles

The definitions of the Data Science Professional Profiles including a set of Data Management profiles and that of 'Data Steward' in particular, is one of the EDSF components described in a separate document EDSF Part 4[24]. The DSPP are defined in accordance with and as a proposed extension to the ESCO Taxonomy which is a European standard for European Skills, Competences and Occupations. The DSPP definition can be instrumental in defining organisational roles in Data Science and Data Management. It can also be used for defining

[21] CCS2012, The 2012 ACM Computing Classification System. Available at http://www.acm.org/about/class/class/2012

[22] Refer to the EDSF documentation for full information about MC-DS and DS-BoK

[23] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at https://ec.europa.eu/esco/portal/#modal-one

[24] EDISON Data Science Framework, Part 4. Data Science Professional Profiles. Available at https://github.com/EDISONcommunity/EDSF/blob/master/EDISON_DSPP-release3-v07.pd f

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

education and training profiles for students and for practitioners to acquire the necessary competences and knowledge for specific professional profiles or occupations. When linked to the Competence Framework and Body of Knowledge it can be used for professional certification or career path building.

Figure 2 illustrates the existing ESCO hierarchy and the proposed new Data Science classification groups and corresponding new Data Science related profiles. The table in the figure illustrates what CF-DS competence groups are relevant to each profile by indicating competence relevance from 0 to 5 (0 – not relevant, 5 – very important).

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

- Managers (for managerial roles);
- Professionals (for Data Science and Analytics, Data Management and Stewardship, infrastructure and data centre engineering roles);
- Technicians and associate professionals (for operators, facility administrators and technicians)
- Optionally, some data management occupations can be also placed into the General and Keyboard Clerks group such as data entry clerks and user support workers.

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

- Data Science Services/Infrastructure Managers
- Data Science Professionals
- Data handling/management professionals that include Data Stewards, Digital Data Curators, Data Librarians
- Database and infrastructure professionals
- Technicians and associate professionals
- Data and information entry and access

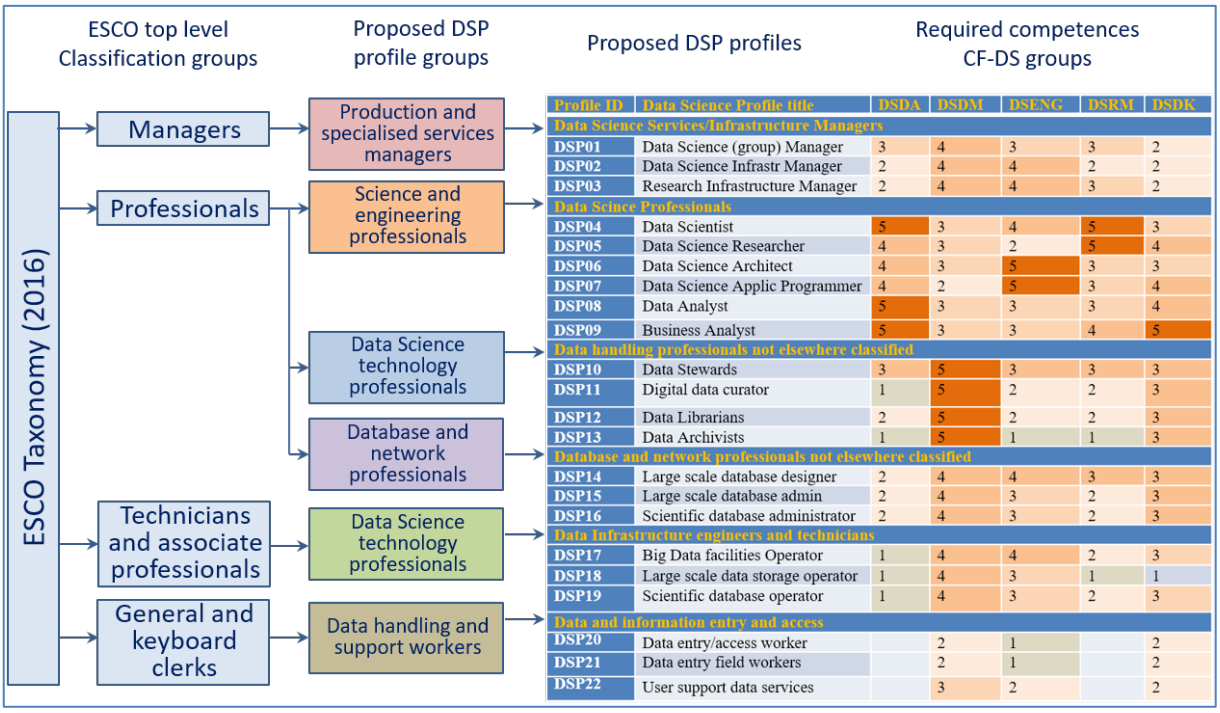
A group of occupations related to Data Stewardship, data curation, data archives and libraries are currently placed in the 3rd proposed group of professionals in the ESCO hierarchy:

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

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

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

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558



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

Figure 2. Proposed Data Science related extensions to the ESCO classification hierarchy and corresponding DSPP by classification groups (EDSF Part 4 [14]).

2.6.3. Using EDSF as a basis for defining FAIR4HE and aligning with other frameworks

The EDSF provides a complete framework and methodology for defining the whole ecosystem of Data Science competences, professional profiles, Body of Knowledge and Model Curriculum where Data Stewardship is one of the professional profiles with defined competences, skills, knowledge, and recommended Learning Objectives and Learning Units. EDSF can be used for defining an organisational structure for Data Science and Data Stewardship tasks and activities management using the definition of the Data Science Professional Profiles that support the whole data lifecycle management in a research organisation.

The EDSF and its Data Stewardship profile can be used for aligning existing Data Stewardship and FAIR competences and consolidating them into the intended FAIR4HE competence framework. It is also foreseen that future EDSF releases can accommodate the FAIR4HE competence framework as a special profile.

The following chapters describe the proposed Data Stewardship and FAIR4HE Competence framework in detail.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

3. Data Stewardship and FAIR Competences Definition

This section describes the proposed *Data Stewardship Competence Framework* (hereafter referred to as CF-DSP for the Data Stewardship Professional to distinguish from CF-DS acronym used in EDSF) with a focus on supporting FAIR data principles that will serve as a foundation for the definition of other FAIR4HE components.

The presented CF-DSP provides a full view of Data Stewardship competences, skills and knowledge. It incorporates different elements of existing Data Stewardship and FAIR competence frameworks definitions that primarily cover the research data management competences while typical industry related data steward vacancies require advanced skills for heterogeneous industrial and business data management.

3.1 Job market analysis for demanded key competence

A preliminary analysis was done of data collected from job advertisements on popular job search and employment portals indeed.com, IEEE Jobs portal and LinkedIn Jobs, where indeed.com provided the largest number of the advertised Data Steward vacancies. The collected data were used to extract information on competences, skills and knowledge demanded from prospective Data Steward candidates. The following sections explain what approach was used for the analysis of vacancies and how the extracted information was mapped to the structure of the competences definition.

3.1.1 Method and context

The EDSF was used as a basis for defining the initial set of the Data Stewardship competences, with the following revision and extension of the individual competences specific to Data Stewardship and FAIR data principles identified from the collected data. The full EDISON/EDSF methodology used for the initial definition of Data Science competences is explained in Appendix A.

When applying this methodology to the current analysis of Data Stewardship competences, the initial identification of the competence groups was not required.

The assumption was that the Data Steward competences would have the same structure as the whole Data Science Professional family, namely the competence groups Data Science Analytics (acronym DSDA as defined in EDSF or short DSA), Data Science Engineering (DSENG or DSE), Data Management (DSDM or DM), Research Methods and Project Management or Business Process Management (DSRMP or RMP), and Domain Knowledge (DSDK or DK). The benefit of this assumption is that the majority of current university curricula (in fields related to Data Science and Data Stewardship) already contain the above mentioned courses[25], and

[25] Tomasz Wiktorski, et al, Model Curricula for Data Science EDISON Data Science Framework, Proc. The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong. [online] Available at https://ieeexplore.ieee.org/document/8241134

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

that it will be easy to further map identified competences and knowledge to typical academic courses and/or learning units.

A typical job vacancy has the following structure and contains the following information that can be mapped to different components of a competence definition (such as Competence, skills, knowledge, education, proficiency level):

- Job/position name, sometimes provided with the description of organisational roles and relations;
- Functions/responsibilities and abilities which can be mapped to competences, if competences are not explicitly defined (job vacancies usually use the term 'skills' instead of 'competences');
- Skills and experiences, also including experience with tools and programming languages that all can be directly mapped to skills;
- Required knowledge or expected familiarity with named technologies or theories. This can be mapped to knowledge topics;
- Education, certification and proficiency level – can be mapped to proficiency level that indicates mastering a certain level of a specific competence; but this information is rarely specified in the typical job vacancy.

It is also important to clarify the relations between competences, skills and knowledge as illustrated in Figure 3 and used in the EDSF (which itself was adopted from the European e-Competences Framework (e-CF) and the corresponding standard EN 16234-1: 2019):

- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results.
- Competence includes/is supported by the **knowledge** that is obtained from education or (self-)training and by **skills** that are acquired as a result of practical experience.
- Professional profiles suggest necessary competences, skills and knowledge and ensure the ability to perform organisational functions

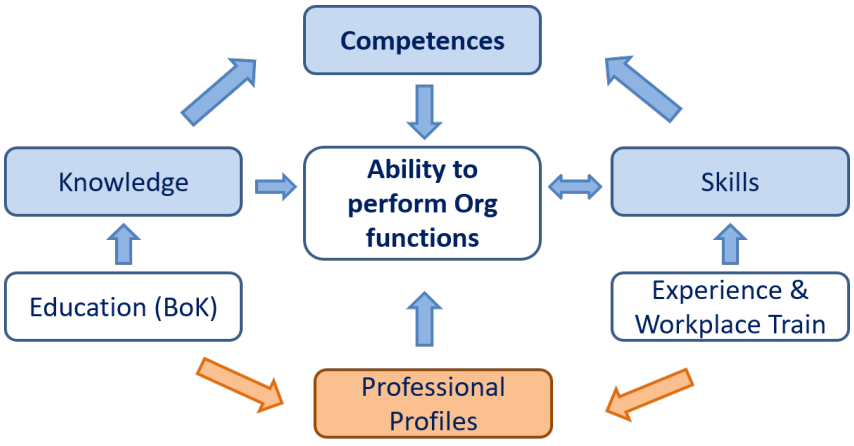


Figure 3. Relation between competences, skills and knowledge

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

3.1.2 Collected data

This section provides a summary of the information extracted from the Data Stewards vacancies analysis.

The following are general characteristics of the collected data:

- Period data collected: 30 August – 1 September 2020
- Sites: Indeed.com – NL, UK, DE, USA: monsterboard.nl - NL
- Days vacancy open: >50% more than 30 days
- Data Steward and related vacancies discovered:
  - NL – 51, UK – 30+, DE ~20, US – 300+
- Information collected/downloaded:
  - Key skills snapshot – for all or first 200 for USA
  - Full vacancy texts – approx. 40 in total
- Detailed analysis of sample vacancies
  - NL, UK – 20, US - 6
- Number of companies and organisations posted Data Steward related jobs – more than 50

Mapping of vacancies information to competences, skills and knowledge items was done manually using simple text extraction and content ordering in Excel. Appendix B provides details on the approach used for analysing data. The Excel workbooks are available in the WP7 T7.3 folder[26].

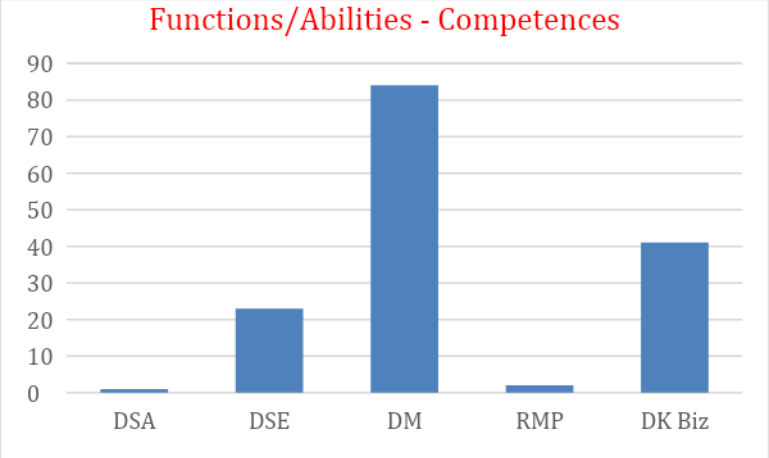
3.1.3 Identified competences, skills and knowledge and their mapping to CF-DS

One of the goals of the undertaken analysis was to identify what competence groups are demanded on the job market and if they can be mapped to competences defined in CF-DS/EDSF. If confirmed that the EDSF can be further used for the FAIR4HE and CF-DSP competences definition, this will bring the benefits of using the EDSF approach for linking competences to intended Learning Outcomes and model curricula.

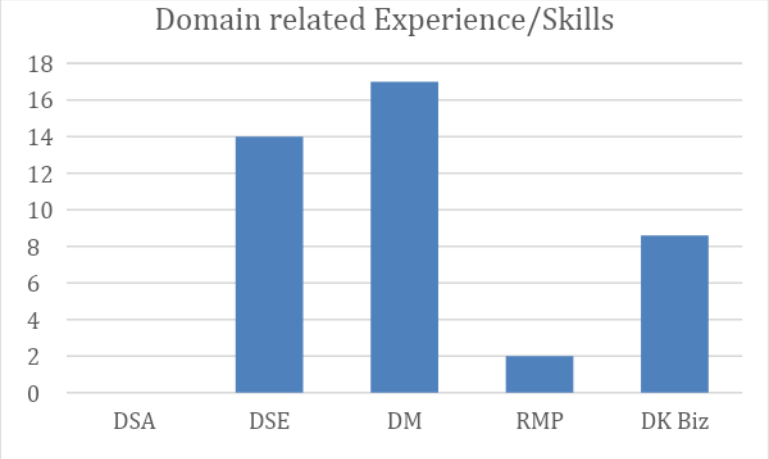
The diagrams below illustrate what types/groups of competences are required in the Data Steward vacancies. Figure 4 (a) illustrates the mapping of functions/abilities to competence groups, Figure 4 (b) maps experiences and skills to skill groups.

[26] The authors intend to publish cleaned data for reference purposes and in compliance with FAIR data principles. At the time of writing this deliverable the data are being prepared for publication.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558



(a) Competences present in Data Steward vacancies



(b) Skills present in Data Steward vacancies

Figure 4. Competence and skill groups present/required in the Data Steward vacancies. Legend: DSA - Data Science Analytics, DSE - Data Science Engineering, DM - Data Management, RMP - Research Methods and Project Management or Business Process management, DK - Domain Knowledge (such as Business Analytics).

Another valuable information obtained from the vacancies analysis is the spectrum of required knowledge topics presented in Table 2 which can also be classified into the same competence and knowledge groups as shown in Figure 5.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Table 2. Knowledge topics obtained from the Data Steward vacancies analysis

Competence/Knowledge Group	**Extracted knowledge topics**
Data Management	• Data Management techniques • FAIR data principles • Data Management and Data Governance principles • Data integrity • Metadata, PID and linked data • Ontology and Semantics • FAIR metrics and Maturity framework, FAIR certification • Data compliance regulations and standards • Data privacy law • GDPR • Ethics
Research Methods	• Research methods (general and domain related) • Project management
Data Analytics	• Data analysis and visualisation tools • Data lifecycle, lineage, provenance
Domain knowledge and Business processes	• Business process management • Marketing • Banking financial services and data management • Multilevel Bill of Materials • Data Warehouses • Version control system • Master Data Management (MDM) and Reference Data
Data Science Engineering	• Visual Basic for Applications (VBA) and interface design • WebAPI use for data access, collection and publishing • DevOps, Agile, Scrum methods and technologies • Data formats, standards • Data modeling (SQL and EDBMS, NoSQL) • Modern data infrastructure: Data registries, Data Factories, Semantic storage, SQL/NoSQL

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

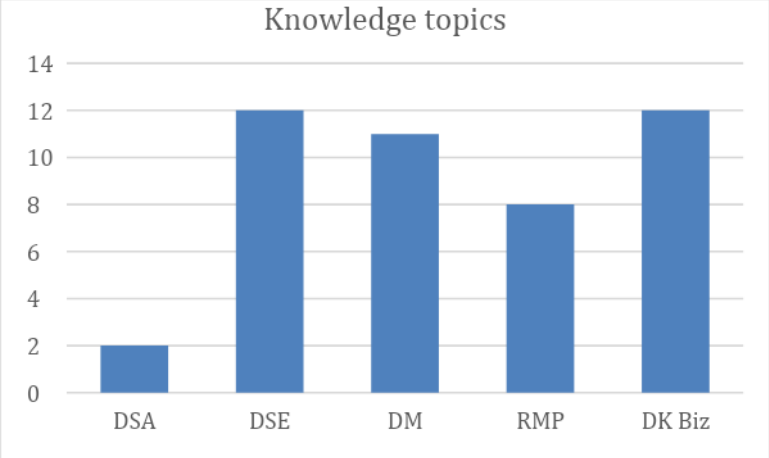


Figure 5 Knowledge topics present/required in the Data Steward vacancies. Legend: DSA - Data Science Analytics, DSE - Data Science Engineering, DM - Data Management, RMP - Research Methods and Project Management or Business Process management, DK - Domain Knowledge.

A more detailed analysis of competences is done in section 3.4 and included the extraction of individual competences and their comparison to the current definition of the competences in EDSF and existing Data Stewardship and FAIR competence frameworks (refer to section 3.4).

3.1.4 Outcome of the job vacancies analysis and further steps

The following conclusions and assumptions can be done based on the initial vacancies analysis:

- The published Data Stewards vacancies demonstrated a variety of competences, skills and knowledge required from the candidates.
- The extracted competences can be successfully mapped to the competence groups defined for the Data Science professional family that includes Data Stewards.
- The presented analysis confirms the applicability of EDSF to the analysis and further structured development of the intended FAIR4HE and Data Stewardship Competence Framework.

The most populated competence group is Data Management which reflects the nature of the Data Steward profession and responsibilities. Two other well populated groups are Domain Knowledge and Data Science Engineering what reflect another side of the Data Steward profession to act as a bridge between ICT teams operating data facilities and domain specialists. This demonstrates the need for related knowledge at the level sufficient for coordination and communication. This fact is clearly reflected in the distribution of required knowledge topics.

The collected and extracted set of competences, skills and knowledge topics was used for detailed competences analysis and mapping to current definitions and vocabulary in EDSF and necessary updates and extensions/additions will be suggested. This information is presented in the next section.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

3.2. Technological and organisational aspects of the FAIR data principles implementation

Besides collecting information from the job market, we can also look at the technological and organisational aspects of the implementation of the FAIR data principles in a typical research organisation as described in section 2.4 and 2.6. Table 3 correspondingly provides the mapping of the FAIR metadata requirements to required technological and management domains: standardisation, policy, infrastructure, and tools or platforms. Table 4 links the typical data management lifecycle stages to organisational roles related to organisational data management and governance.

Table 3. FAIR metadata requirements

FAIR metadata requirements and technology aspects	Standardisation	Policy	Infra structure	Tools
Findable				
F1. (meta)data are assigned a globally unique and persistent identifier	x		X	
F2. data are described with rich metadata		x		X
F3. metadata clearly and explicitly include the identifier of the data it describes		x		x
F4. (meta)data are registered or indexed in a searchable resource		x	x	X
Accessible				
A1. (meta)data are retrievable by their identifier using a standardized communications protocol ● A1.1 the protocol is open, free, and universally implementable ● A1.2 the protocol allows for an authentication and authorization procedure, where necessary	x		x	X
A2 metadata are accessible, even when the data are no longer available		X	X	
Interoperable				
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	x	x		X
I2. (meta)data use vocabularies that follow FAIR principles	X	X		X
I3. (meta)data include qualified references to other (meta)data		X		X

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Reusable				
R1. (meta)data are richly described with a plurality of accurate and relevant attributes ● R1.1 (meta)data are released with a clear and accessible data usage license ● R1.2 (meta)data are associated with detailed provenance ● R1.3 (meta)data meet domain-relevant community standards		x		X

Table 4. Data Lifecycle Management and organisational role

Data lifecycle stage	**Data Steward key activities**	**Other roles involved**
Data collection	Data model and metadata definition and implementation	Researchers Data engineers Data entry workers
Data preservation and curation	Data storage facility Data quality control Data integration	Data curators Data custodians/archivists
Data analysis	General contacts with data analytics team (not special tasks)	Data Scientists Data Architects Application developers
Data publication	Data publications in open repositories and archiving services, metadata Ensure data discoverability and findability	Data curators
Data governance and data management	Develop Data Governance Policy and Data Management Plan Ensure FAIR data principles implementation Coordinate and monitor data governance and management implementation Interact with ICT team and data infrastructure services Organise and conduct necessary training for data management policy	Chief Data Officer Data quality managers Data Controller (GDPR)
Data infrastructure and tools	Define and communicate requirements to data infrastructure and tools, coordinate their implementation Organise necessary training for tools and services	Infrastructure engineers Database managers/ engineers Master data managers

Effect on FAIR and Data Stewardship competences

The implementation of the FAIR data principles in the operation and practice of research infrastructures requires a strong technical base, infrastructure services and tools. This is a task for ICT and data infrastructure services/teams; but Data Stewards need to be aware of these

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

technologies and tools and maintain a link between ICT and data policy, providing also the necessary training to researchers and data workers.

3.3 Defining a Competence Framework for Data Stewardship and FAIR Data Principles (CF-DSP)

As the basis for elaborating the Data Stewardship and FAIR data Competence Framework (CF-DSP), we used the Data Science Competence Framework (CF-DS) defined in EDSF. This allows us to benefit from other EDSF components such as the Body of Knowledge and Model Curriculum. In this context, we treat the CF-DSP for Data Stewardship as a profile or subset of the more general CF-DS for the Data Science professional family.

The data collected and classified from the Data Stewards job vacancies are used for identifying the set of individual competences that match with the CF-DS competence groups. Based on this, original CF-DS competences are revised and/or extended, new competences are suggested to create a consistent Data Stewardship Competence Framework that reflects the current job market demand for Data Stewards and their essential competences. The final definition of the CF-DSP will be composed of the essential competences identified in this analysis.

It is also important to note that in the current, market-based definition of Data Steward competences and skills, the primary focus lies on data management skills (DSDM group), understanding of the required data management platforms and infrastructure (DSENG group) and domain-related or organisational competences (DSDK or DSBA group). General understanding of research methods and project management competences is required, whereas Data Science and Analytics competences (DSDA group) may only be required at the level of general literacy. The following tables (Tables 5 to 8) list the original CF-DS competence groups together with the suggested changes and extensions to individual competences for the intended/proposed CF-DSP profile.

3.3.1. Data Management and Governance competence group (DSDM)

As a consequence of the wide recognition by organisations of the importance of quality data management, almost all individual competences have been updated (see Table 5). It is also important to mention that the growing adoption of the Data Steward profession as an important organisational role and wide adoption of the FAIR data principles motivate the addition of three competences into CF-DSP. These are:

- DSDM07: Manage Data Management/Data Stewards team, coordinate related activity between organisational departments, external stakeholder to fulfill Data Governance policy requirements
- DSDM08: Develop organisational policy and coordinate activities for sustainable implementation of the FAIR data principles

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

- DSDM09: Specify requirements in terms of and supervise the organisational infrastructure for data management (and archiving), maintain the pool of data management tools

Table 5. CF-DS competence group Data Management (DSDM) and suggested extensions for CF-DSP

Data Management (DSDM)	**Relevance and proposed changes and extensions** (posted as revised text and bulleted extensions)
DSDM Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	**DSDM – extended, relevant** Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing, • ensure compliance with FAIR data principles.
DSDM01 Develop and implement data strategy, in particular in the form of data management policy and Data Management Plan (DMP)	DSDM01 – extended, essential Develop and implement data management and governance strategy, in particular in the form of Data Governance Policy and Data Management Plan (DMP) • Ensure compliance with standards and best practices in Data Governance and Data Management
DSDM02 Develop and implement relevant data models, define metadata using common standards and practices for different data sources in a variety of scientific and industry domains	DSDM02 – extended, essential Develop and implement relevant data models, define metadata using common standards and practices for different data sources in a variety of scientific and industry domains. • Ensure metadata compliance with FAIR requirements • Be familiar with the metadata management tools
DSDM03 Integrate heterogeneous data from multiple sources and provide them for further analysis and use	DSDM03 – extended, essential Integrate heterogeneous data from multiple sources and provide them for further analysis and use • Perform data preparation and cleaning • Match/transfer data models of individual datasets
DSDM04 Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)	DSDM04 – extended, highly essential Maintain historical information on data handling, including reference to published data and corresponding data sources • Publish data, metadata and related metrics • Perform and maintain data archiving • Develop necessary archiving policy, comply with Open Science and Open Access policies if applicable • Maintain data provenance and ensure continuity through the whole data lifecycle, ensure data provenance

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

DSDM05 Ensure data quality, accessibility, interoperability, compliance to standards, and publication (data curation)	DSDM05 – extended, essential Develop policy and metrics for data quality management, maintain data quality and compliance to standards, perform data curation • Interact/Collaborate with data providers and data owners to ensure data quality
DSDM06 Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management	DSDM06 – extended, essential Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management, address legal issues if necessary. • Ensure GDPR compliance in data management and access • Develop data access policies and coordinate their implementation and monitoring, including security breaches handling
None	DSDM07* - added new, essential Manage Data Management/Data Stewards team, coordinate related activity between organisational departments, external stakeholder to fulfill Data Governance policy requirements, provide advice and training to staff. Define domain/organisation specific data management requirements, communicate to all departments and supervise/coordinate their implementation. Coordinate/supervise data acquisition.
None	DSDM08* - added new, essential Develop organisational policy and coordinate activities for sustainable implementation of the FAIR data principles and Open Science, define corresponding requirements to data infrastructure and tools, ensure organisational awareness.
None	DSDM09* - added new, essential Specify requirements to and supervise the organisational infrastructure for data management and (and archiving), maintain the park for data management tools, provide support to staff (researchers or business developers), coordinate solving problems.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

3.3.2. Data Engineering competence group (DSENG)

Table 6 describes the relevance of, and proposed changes to DSENG competences to align them with the changes applied to corresponding Data Stewardship competences. Updates/extensions were added to the competences DSENG03-DSENG06 to reflect FAIR related requirements to data infrastructure, data management tools and metadata management during the whole data lifecycle.

Table 6. CF-DS competence group Data Science Engineering (DSENG) and suggested extensions for CF-DSP

Data Science Engineering (DSENG)	**Relevance and proposed changes and extensions** (posted as revised text and bulleted extensions)
DSENG Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle.	DSENG – no changes, generally relevant Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle.
DSENG01 Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis and visualisation	DSENG01 – no changes, low relevance Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis and visualisation
DSENG02 Develop and apply computational and data driven solutions to domain related problems using wide range of data analytics platforms, with a special focus on Big Data technologies for large datasets and cloud based data analytics platforms	DSENG02 – no changes, low relevance Develop and apply computational and data driven solutions to domain related problems using wide range of data analytics platforms, with a special focus on Big Data technologies for large datasets and cloud based data analytics platforms
DSENG03 Develop and prototype specialised data analysis applications, tools and supporting infrastructures for data driven scientific, business or organisational workflow; use distributed, parallel, batch and	DSENG03 – extended, relevant Develop and prototype specialised data analysis applications, tools and supporting infrastructures for data driven scientific, business or organisational workflow; use distributed, parallel, batch and streaming processing platforms, including online and cloud based

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

streaming processing platforms, including online and cloud based solutions for on-demand provisioned and scalable services	solutions for on-demand provisioned and scalable services • Develop new tools and applications, ensure support of the data FAIRness requirements by existing and new tools and applications
DSENG04 Develop, deploy and operate large scale data storage and processing solutions using different distributed and cloud based platforms for storing data (e.g. Data Lakes, Hadoop, HBase, Cassandra, MongoDB, Accumulo, DynamoDB, others)	DSENG04– extended, essential Develop, deploy and operate data infrastructure, including data storage and processing facilities, using different distributed and cloud based platforms. • Implement requirements for data storage facilities to comply with the data management policies and FAIR data principles in particular.
DSENG05 Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	DSENG05– extended, relevant Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection, ensure standards and corresponding data protection regulation compliance, in particular GDPR. • Define and implement (coordinate) data access policies for different stakeholders and organisational roles
DSENG06 Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	DSENG06– extended, essential Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform), OLTP, OLAP processes for large datasets • Define, implement and maintain data model, reference data, master data definitions, implement consistent metadata

3.3.3. Research Methods and Project Management competence group (DSRMP)

The Research Methods and Project Management competences are important for Data Stewards in supporting research projects in an organisation, to work effectively with the domain related researchers and to serve as a link between the researchers and other roles during the whole cycle of the research process and corresponding data lifecycle. Minor extensions were added to DSRMP03 and DSRMP05.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Table 7. CF-DS competence group Research Methods and Project Management (DSRMP) and suggested extensions for CF-DSP

Research Methods and Project Management (DSRMP)	**Relevance and proposed changes and extensions** (posted as revised text and bulleted extensions)
DSRMP Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals	DSRMP – revised, generally relevant Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals • Base research on collected scientific facts and collected data
DSRMP01 Create new understandings by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods	DSRMP01 – no changes, generally relevant Create new understandings, discover new relations by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods
DSRMP02 Direct systematic study toward the understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSRMP02 – no changes, generally relevant Direct systematic study toward the understanding of the observable facts, and discovers new approaches to achieve research or organisational goals
DSRMP03 Analyse domain related research process model, identify and analyse available data to identify research questions and/or organisational objectives and formulate sound hypothesis	DSRMP03- extended, essential Analyse domain related research process model, identify and analyse available data to identify research questions and/or organisational objectives and formulate sound hypothesis • Link domain-related concepts and models to general/abstract Data Science concepts and models,
DSRMP04 Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and uses this knowledge to devise new applications, contribute to the development of organizational objectives	DSRMP04 – no changes, generally relevant Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and use this knowledge to devise new (data-driven) applications, contribute to the development of organizational or project objectives

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

DSRMP05 Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	DSRMP05 – extended, essential Design experiments which include data collection (passive and active) for hypothesis testing and problem solving • Work with Data Science, Data Stewardship and data infrastructure teams to develop project/research goals.
DSRMP06 Develop and guide data driven projects, including project planning, experiment design, data collection and handling	DSRMP06 – no changes, essential Develop and guide data driven projects, including project planning, experiment design, data collection and handling

3.3.4. Domain related competence (DSDK/DSBA)

Domain-related knowledge and competences are important for Data Stewards as one of their roles is to support organisational (and project) data management during the whole data lifecycle and correspondingly through all business process or research process stages. Our job vacancies analysis indicated the importance for Data Stewards to understand and know main organisational and business processes with a focus on data management, provenance and quality.

Analysis of the Data Steward positions in the context of the organisational needs, both for the research and the business domain, identified necessary extensions that can be applied to the initial definitions in EDSF CF-DS and also a need for specific activities related to the coordinating role of Data Steward in data management and governance:

- DSBA07: Coordinate intra-organisational activities related to data analytics, data management and data provenance/lineage along all data flow stages.

We use the business related domain competence group DSDA as it is well represented in the business related Data Steward positions and has a well-defined focus on organisational needs. Table 8 summarises the proposed extensions and defines a new competence DSBA07.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Table 8. CF-DS competence group Domain Knowledge (Organisational specific and Business related, DSBA) and suggested extensions for CF-DSP

Domain related Competences (DSDK): Applied to Business Analytics (DSBA)	**Relevance and proposed changes and extensions** (posted as revised text and bulleted extensions)
DSDK Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations	DSDK – no changes, generally relevant Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
DSBA01 Analyse information needs, assess existing data and suggest/identify new data required for specific business context to achieve organizational goal, including using social network and open data sources	DSBA01 – extended, relevant for organisation processes and data Analyse information needs, assess existing data and suggest/identify new data required for specific business context to achieve organizational goal, including using social network and open data sources • Data management and Quality Assurance of organisational data assets
DSBA02 Operationalise fuzzy concepts to enable key performance indicators measurement to validate the business analysis, identify and assess potential challenges	DSBA02 – extended, relevant for organisation processes and data Operationalise fuzzy concepts to enable key performance indicators measurement to validate the business analysis, identify and assess potential challenges • Specify requirements/develop data models for organisational data
DSBA03 Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends	DSBA03 – extended, generally relevant Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends • Ensure data availability and quality for BA/BI needs

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

DSBA04 Analyse opportunity and suggest the use of historical data available at organisation for organizational processes optimization	DSBA04 – extended, relevant for organisation processes and data Analyse opportunity and suggest the use of historical data available at organisation for organizational processes optimization • Coordinate implementation of FAIR data principles for collected data, ensure proper lineage and provenance of collected data
DSBA05 Analyse customer relations data to optimise/improve interaction with the specific user groups or in the specific business sectors	DSBA05 – no changes, relevant for organisation processes and data Analyse customer relations data to optimise/improve interaction with the specific user groups or in the specific business sectors
DSBA06 Analyse multiple data sources for marketing purposes; identify effective marketing actions	DSBA06 – no changes, relevant for organisation processes and data Analyse multiple data sources for marketing purposes; identify effective marketing actions
none	DSBA07 – added, essential Coordinate intra organisational activities related to data analytics, data management and data provenance/lineage along all data flow stages, ensure data FAIRness

3.3.5 Data Steward professional and transversal skills

It is evident that the new profession of Data Stewards and the emerging FAIR data management culture will create new types of professional transversal skills (often referred to as soft skills), which can be defined using such concepts as attitude or aptitude (referring to such concepts introduced in the FAIR4S competence framework, section 2.1).

It is important to compile such skills related to Data Stewardship and FAIR principles. The workplace skills for Data Scientists defined in EDSF can provide an example and a basis for the definition of such skills for Data Stewards.

Although transversal or soft skills are not usually included directly in academic curricula, they can be a part of Professional and Academic skills training that is established at many universities.

For reference purposes, Appendix C provides an example of how such skills are defined in the EDSF for Data Science Professionals.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

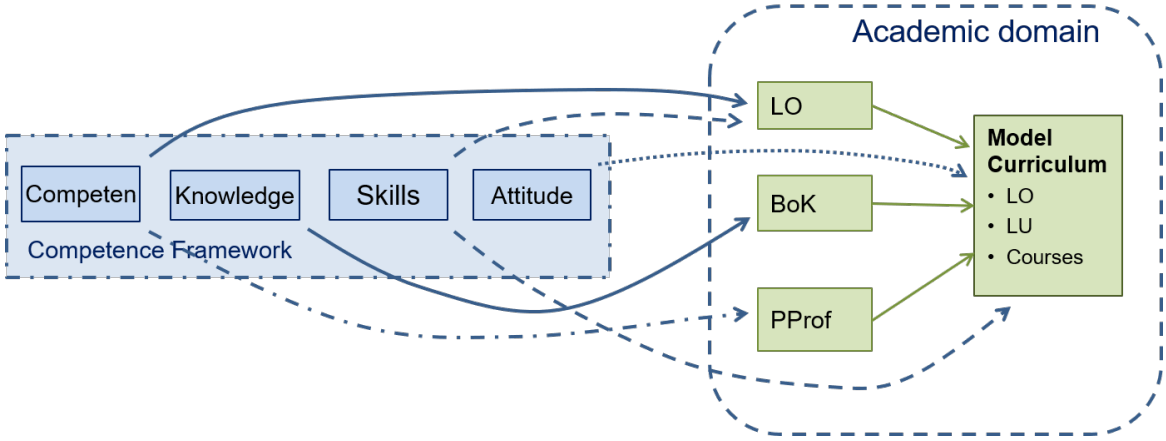
3.4 Comparing/Mapping CF-DSP to other Competence Frameworks

This section compares the proposed Data Steward Professional Competence Framework (CF-DSP) with the existing frameworks discussed in chapter 2 with the goal to provide alignment between the proposed CF-DSP and other frameworks. This will simplify educational and training courses exchange, re-use and blending. This is especially important when designing academic curricula for universities and vocational education where existing training materials and courses can be included as self-study and practical study materials.

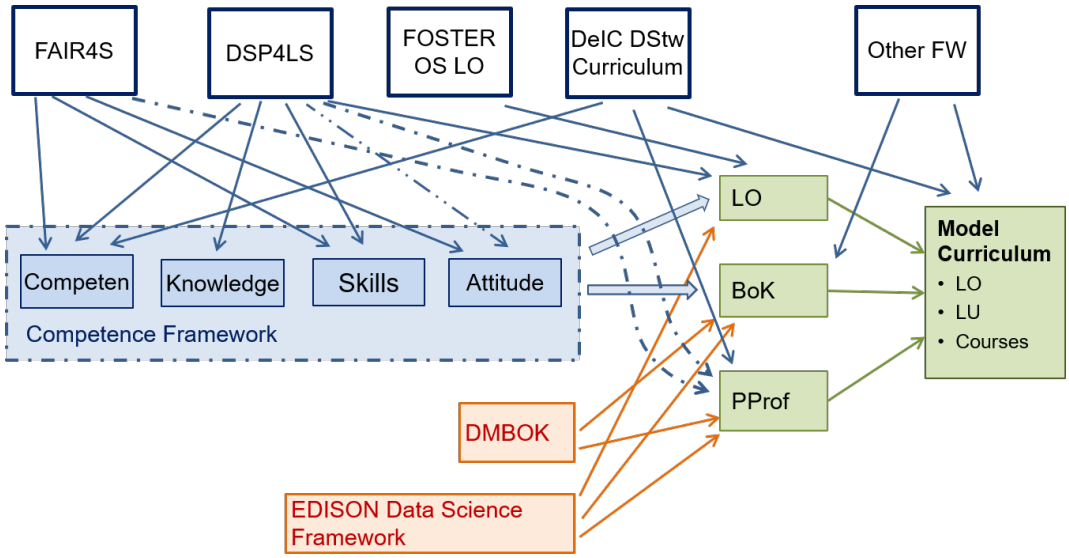
The comparison was made by mapping the CF-DSP components to similar components in other frameworks such as competence groups, individual competences, responsibilities, capabilities, skills and knowledge topics. In fact, the mapping presented is the result of an iterative process during which an initial mapping has been done for the initial set of DSP competences to clarify the initial set, discover necessary extensions and incorporate these into the current CF-DSP.

The mapping and alignment presented below has been done for the following frameworks: FAIR4S, ELIXIR Data Stewardship Competency Framework (DSP4LS), DeIC Data Stewardship Curriculum recommendations, and Foster Learning Objectives for Open Science. Figure 6 below illustrates (a) the general relation between different components of the entire professional framework for Data Stewardship and (b) the link between different components in existing frameworks and CF-DSP.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558



(a) Relations between competence framework components and academic curriculum components



(b) Contribution of different Data Stewardship frameworks to the definition of the proposed CF-DSP and further definition of the Model Curriculum

Figure 6. Relations and mapping between CF-DSP and other frameworks Legend: LO – Learning Outcomes; LU – Learning Units; BoK – Body of Knowledge; PProf – Professional Profiles. Dotted lines reflect indirect/implied relations via Data Steward organisational roles described in FAIR4S and DSP4LS.

To allow constructive comparison and mapping between frameworks, the competences definitions in individual frameworks were decomposed into individual components (statements) and enumerated for better comparison. Individual components of competences, skills, knowledge topics, learning objectives – whatever relevant to different frameworks, were enumerated for better components grouping and mapping to the corresponding CF-DSP competence groups. A number of individual components in each group was used to assess the relevance of the proposed competences. This was further used for improving the definition of the individual competences in CF-DSP.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The figures 7-9 below illustrate the mapping between CF-DSP competences and related definitions in other competence frameworks, where vertical axis presents the count of the relevant individual competences, skills or knowledge in referred frameworks corresponding to CF-DSP competences. The following general approach has been used when mapping existing frameworks to the proposed CF-DSP:

- Most of the frameworks include only definition of competences, which are also profiled for different organisational roles (such as Policy, Research and Infrastructure). However, when applying for education and training purposes, the competences should be defined in a more general form and preferably be linked to the Body of Knowledge and well established academic disciplines.
- In many cases, the defined competences and Learning Objectives (LO)[27] can be linked to education or training curricula, but others can rather be achieved through career and work experience. It is also understood that the future graduates may possess a good knowledge on the subject and have a set of necessary competences for the junior role but time is needed to gain workplace experience and corresponding knowledge.
- Most of the existing frameworks do not provide direct links between responsibilities and competences, such as are illustrated in Figure 3 and Figure 6 and implemented in EDSF. This kind of linking was applied during the mapping exercise.
- When analysing the ELIXIR Data Stewardship Competence Framework for Life Sciences (DSP4LS) the defined combined set of Skills, Knowledge and Attitude (SKA) was broken down into individual elements related to competences and knowledge. Similar mapping was done for the Learning Objectives defined in DSP4LS.

Competences defined in both FAIR4S by EOSCpilot and the DeIC Data Stewardship curriculum correspond with the main competence groups defined in CF-DSP, as shown in Figure 7 and Figure 8 respectively. Both frameworks show the majority of required competences in the groups DSDM – Data Management and DSENG – Data infrastructure, services and tools. Moreover, FAIR4S also shows the importance of the Data Science and Analytics competences. However, it is motivated by the needs of data quality assurance and tools development. This is usually done by analytics and engineering teams in coordination with the Data Steward who defines user needs and requirements. FAIR4S also identifies the importance of general research methods competences and domain related competences, as well as the importance of professional skills or attitudes. The DeIC Data Stewardship framework similarly confirms the importance of general research methods and project management, including different levels (basic, intermediate, expert) to work effectively with scientific and data publications.

The FOSTER Learning Objectives for Open Science[28] provide an important view on the competences that are required for Open Science, which are closely related to the intended

[27] This analysis uses the original term Learning Objectives as it is used in referred frameworks DSP4LS and FOSTER, while in EDSF, ACM/IEEE Curricula guidelines the term Learning Objectives is used.
[28] The FOSTER Learning Objectives for Open Science, 23 February 2015 [online] available at https://doi.org/10.5281/zenodo.608586

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

application of the FAIR data principles. Figure 9 illustrates the mapping of the FOSTER LOs to extended competences in the DSDM group: DSDM01 – DSDM09, with more stress on the organisational policy and compliance in DSDM06, DSDM08, DSDM09. It also indicates the importance of research methods in general.

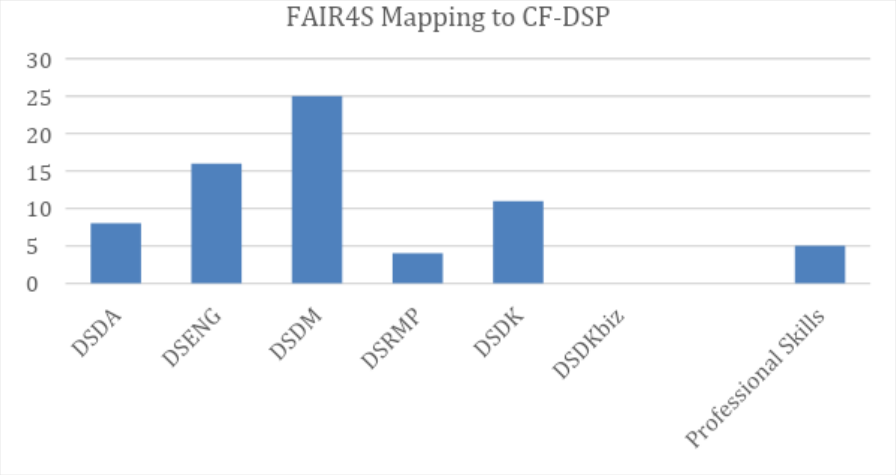


Figure 7. Mapping FAIR4S competences to CF-DSP competence groups (refer to Table 5 for extended DSDM competences definition)

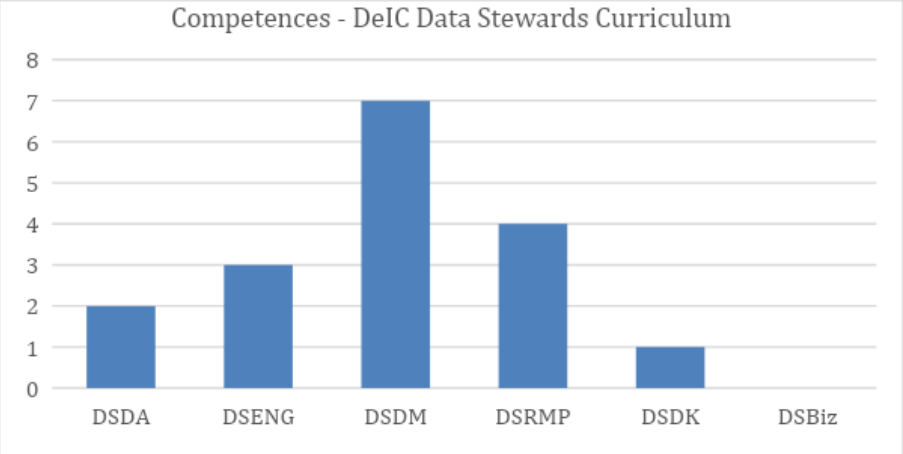


Figure 8. Mapping DeIC Data Stewardship Curriculum competences to CF-DSP competences groups

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

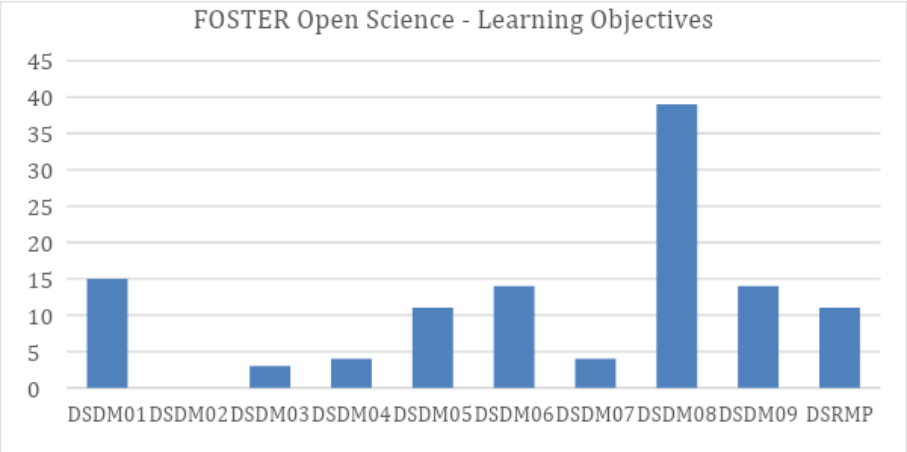


Figure 9. Mapping FOSTER Learning objectives to CF-DSP DSDM Competences (refer to Table 5 for extended DSDM competences definition for CF-DSP and to for FOSTER Learning Objectives for Open Science)

The ELIXIR Data Stewardship Competence Framework for Life Sciences (DSP4LS) provides a detailed inventory of the Activities, Tasks, and Knowledge, Skills, and Abilities for the three organisational roles Research, Infrastructure and Policy; the DSP4LS framework also defines the extended list of Learning Objectives which are aimed for graduate level Data Stewards to develop necessary competences that are grouped in the following competence areas: Policy/strategy; Compliance; Alignment with FAIR data principles; Services; Infrastructure; Knowledge management; Network; Data archiving.

Figures 10 and 11 illustrate the mapping of the DSP4LS Activities, Tasks, and Knowledge, Skills, and Abilities to the proposed CF-DSP competences that include all competences DSDM01-DSDM09 in the data management group, DSENG04-DSENG06 of the engineering group, and also to DSRMP and DSDK that indicate expected general competences and knowledge in these groups. Similarly to the previous diagrams, the vertical axis presents the count of the relevant individual activities and tasks, and knowledge, skills, abilities in DSP4LS corresponding to CF-DSP competences.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

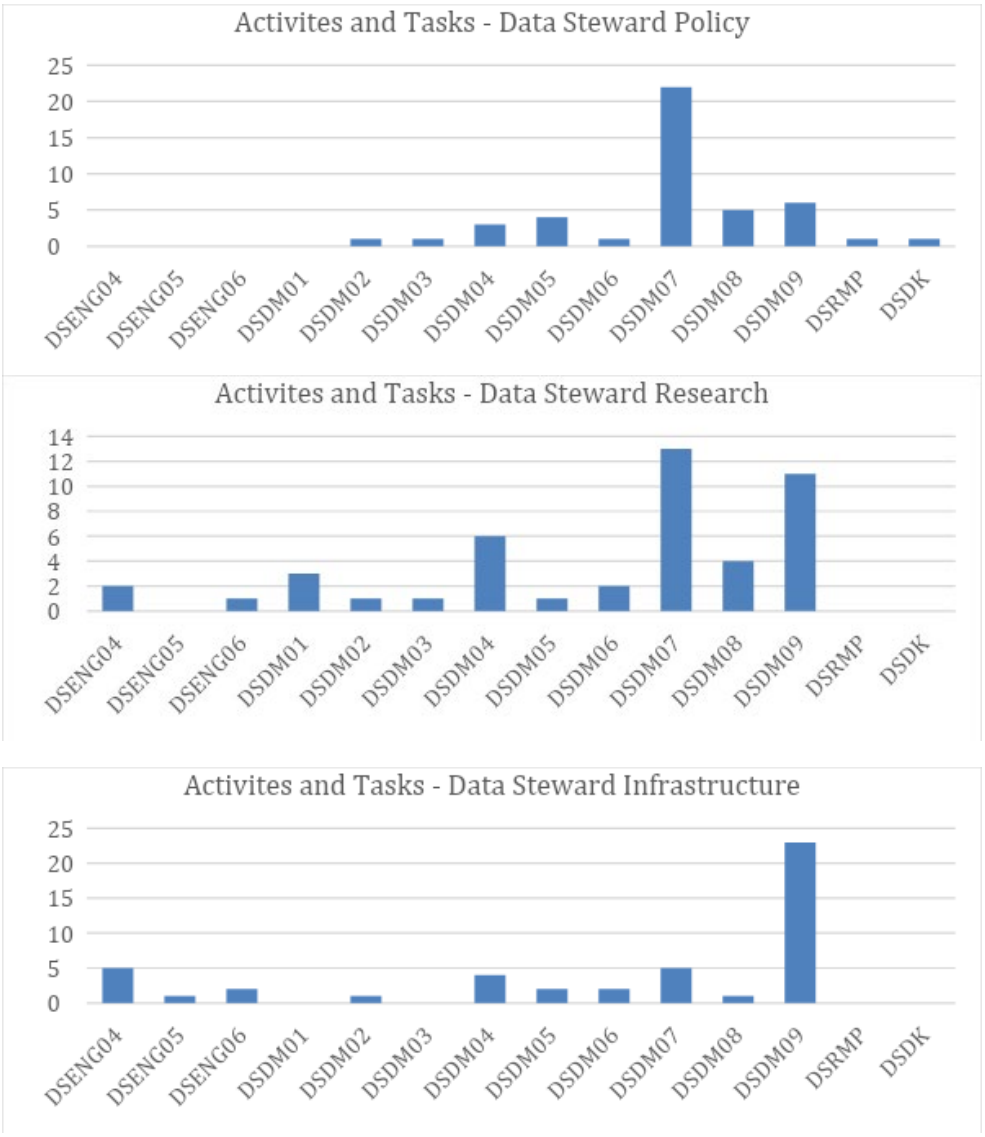


Figure 10. Mapping ELIXIR DSP4LS Activities and Tasks to selected CF-DSP competences

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

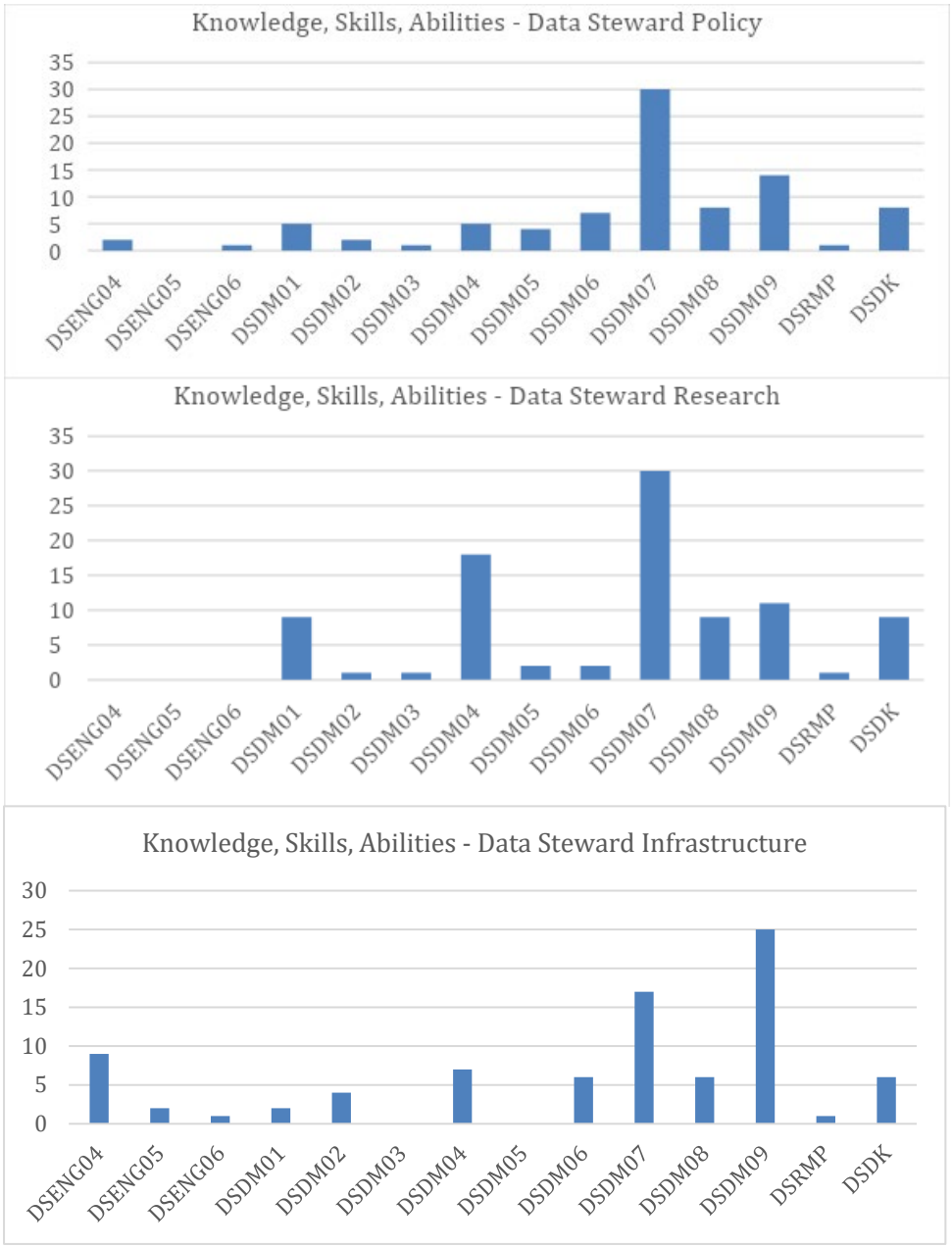


Figure 11. Mapping ELIXIR DSP4LS KSA (Knowledge, Skills, Abilities) group to selected CF-DSP competences

The analysis of the presented mapping allows us to conclude that in general DSP4LS includes competences from the DSDM group and selected competences from the DSENG group of the proposed CF-DSP. Both frameworks can be effectively used for defining Data Stewardship curricula and to address FAIR data competences. At the same time, the FAIR4HE and CF-DSP competence frameworks can benefit from the advanced development of the DSP4LS framework and its ongoing implementation in the ELIXIR Data Stewardship training programme.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

3.5. Summary on defining the Data Stewardship and FAIR competence framework for Higher Education

The detailed analysis and mapping of existing competences allowed us to identify important competences and knowledge topics that are required for the successful work of Data Stewards in different roles and organisations. Based on the analysis in section 3.3, the consolidated definition of the proposed CF-DSP competences is given in Appendix E.

The definition of the proposed CF-DSP is based on the general Data Science Competence Framework created in the EDISON project. Necessary changes and extensions have been added to the initial CF-DS definition to create the proposed Data Stewardship competence profile.

Further analysis and mapping should allow more precise linking between the proposed and existing frameworks, also creating a basis for future exchange of the practical implementation experience. To do this effectively, an ontology for Data Stewardship competences should be developed.

4. Defining Data Stewardship and FAIR Body of Knowledge

The Body of Knowledge is an important element linking competence framework with academic curricula. A Body of Knowledge defines a set of Knowledge Areas (KA) and Knowledge Units (KU) that need to be included in a curriculum to achieve the intended Learning Outcome (LO) and defines a set of academic disciplines that can be taught in a curriculum. The definition of the Body of Knowledge is typically based on the classification of the scientific disciplines, such as Classification Computer Science (CCS).

This section provides information about the Data Stewardship Body of Knowledge (hereafter referred to as DSP-BoK), which is defined as a subset or a profile of the general Data Science Body of Knowledge (DS-BoK) defined in EDSF. The DSP-BoK is extended with the FAIR data related knowledge topics as well as with knowledge topics supporting the coordination role of the Data Steward in the organisational data governance and management.

The DSP-BoK inherits the benefits of the DS-BoK definition that is based on an overview and analysis of existing bodies of knowledge that are relevant to Data Science and required to fulfill the competences and skills identified in CF-DS. DS-BoK adopts essential knowledge elements from multiple BoKs, such as DMBOK, BABOK, CS-BOK (see Table 9), and introduces a number of new Knowledge Units that reflect the practice in academic and professional training courses by universities and professional training organisations.

The 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

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

profiles (and correspondingly 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 exams and training.

4.1 Data Science Body of Knowledge Areas and Knowledge Units

The Data Science Body of Knowledge realized in EDSF is structured by knowledge area groups (KAG) corresponding with CF-DS competence groups:

- 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 Data Management and Governance knowledge area group (KAG3 DSDM) is a key and distinguishing KAG for DSP-BoK. It includes general principles and concepts in data management and stewardship, data management and governance policies and procedures, data storage systems, data modeling and data warehouses, data libraries and archives. It is extended with the FAIR data principles and other knowledge topics related to the Data Steward role in organisations.

The KAG3-DSDM group includes most of the KAs from the DAMA DMBOK however extends it with KAs related to RDA recommendations, community data management models (Open Access, Open Science, Open Data, etc.) and general Data Lifecycle Management, which is used as a central concept in many data management related education and training courses. For the DSP-BOK, FAIR data related knowledge area and knowledge units must be included as an additional knowledge area.

The following are the commonly defined Data Management and Governance Knowledge Areas:

- 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) Data libraries and archives

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Other knowledge areas are sufficiently defined in the original EDSF DS-BoK such as KAG1-DSDA, KAG2-DSENG, KAG4-DSRMP and KAG5-DSDK for domain related knowledge areas.

4.2. Defining a DSP BoK profile

Table 9 provides the general structure of KAG3-DSDM and relevant Knowledge Areas from other KAG. Extensions of the original DS-BoK are proposed based on the recent Data Stewards job market analysis in chapter 2. It contains the definition of the Knowledge Areas and Knowledge Units that need to be added to properly address Data Stewardship and FAIR data principles. Knowledge Units (KU) corresponding to suggested KAs are defined from different sources: existing BoK, CCS2012, and from practices in designing academic curricula and corresponding courses by universities and professional training organisations[29].

Further work on the DSP-BoK will include a mapping of identified knowledge topics to the corresponding Knowledge Units defined in the original DS-BoK. The DSP-BoK defined at this stage will undergo further development and will be updated based on feedback on the model curricula and courses that will be designed in the context of the FAIRsFAIR Task T7.4 activity.

Table 9. DS-BoK Knowledge Area Groups (KAG) and Knowledge Areas (KA) related to the Data Stewardship DSP-BoK

KA Groups	Suggested additional Knowledge Areas (KA)	Knowledge Areas from existing BoK, CCS2012 scientific subject groups and exiting DS&FAIR frameworks
KAG2-DSENG: Data Science Engineering	KA02.01 (DSENG.01/BDIT) Big Data Infrastructure and Technologies KA02.04 (DSENG.04/SEC) Data and Applications security KA02.07 (DSENG.07/IS) Information systems (to support data driven decision making)	ACM CS-BoK selected KAs: IM - Information Management Data and Information systems related scientific subjects from CCS2012: 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)	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,

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

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

	KA03.06 (DSDM.05/DLIB) Digital libraries and archives	(10) Metadata, and (11) Data Quality. RDA recommendations on FAIR Data Principles
KAG4-DSRMP: Research Methods and Project Management	KA04.01 (DSRMP.01/RM) Research Methods KA04.02 (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 KA are more business focused and related to KAG5-DSBA. However, its specific topics related to data management can be reflected in the KAG1-DSDA

Referred bodies of knowledge:

- ACM/IEEE CS-BoK - ACM and IEEE Computer Science Curricula 2013 (CS2013) [online] Available at http://dx.doi.org/10.1145/2534860
- DMBOK – Data Management Body of Knowledge (DM-BoK) by Data Management Association International (DAMAI) [online] Available at http://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf
- BABOK - Business Analytics Body of Knowledge (BABOK) [online] Available at http://www.iiba.org/babok-guide.aspx
- PM-BoK - Project Management Professional Body of Knowledge (PM-BoK) [online] Available at http://www.pmi.org/PMBOK-Guide-and-Standards/pmbok-guide.aspx

4.3 Using CF-DSP and DSP-BoK for Data Stewardship curriculum definition

The DSP-BoK, together with the CF-DSP, can be used to define Data Stewardship university curricula and courses that respond to the needs of a given community or target stakeholders. In this case, the required competences are expressed in the form of intended learning outcomes that, together with the knowledge topics, define the knowledge units from the BoK that need to be included in the curricula.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Appendix D presents an example of how the customised curriculum can be designed using the EDSF toolkit.

5. Recommendations on implementation of the FAIR4HE framework

5.1 Opportunities for synergies, harmonisation and collaboration

The proposed FAIR4HE and CF-DSP intend to provide a basis for building consensus on defining Data Stewardship competences and a corresponding Body of Knowledge and Model Curriculum that can be adopted by wider groups of adopters from the research community, academia and industry. There are a variety of organisations and initiatives available that address different aspects of FAIR and Data Stewardship competences, training and curricula. Training materials are available for a variety of disciplines, and various university programs already offer data management courses.

The present framework intends to harmonise the existing developments and contribute to the development and implementation of a commonly recognised competence framework for Data Stewardship and FAIR data.

5.2 Recommendations

Based on the current development and numerous community discussions, we propose the following recommendations for the promotion and initial implementation of FAIR4HE and the proposed Data Stewardship Competence Framework

Adoption

- Recommendation A1: Publish FAIR4HE on community recognised Open Access platforms
- Recommendation A2: Identify pioneers/champion universities to implement the CF-DSP and FAIR4HE competence framework. Cooperate with other project activities and EUA in identifying and working with champion universities. Facilitate experience exchange between champion universities in implementing Data Stewardship curricula/programs.
- Recommendation A3: Create a catalogue of university programmes and courses offering Data Stewardship education and training (in addition to general FAIR Training materials catalogue)
- Recommendation A4: Submit the proposed CF-DSP and FAIR4HE framework to the RDA IG on Professionalising Data Stewardship for wider community discussion and contribution.
- Recommendation A5: Involve the community and solicit contribution to further define the Body of Knowledge for Data Stewardship and FAIR data principles. Establish

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

contacts with DAMA International to include FAIR data principles and competences in the new edition of the DMBOK.

- Recommendation A6: Provide contribution and cooperate with the terms4FAIRskills initiative[30] on defining terms related to education on FAIR data and Stewardship and the related ontology.

Sustainability

- Recommendation S1: Recognise the importance of maintaining the proposed CF-DSP and FAIR4HE framework, select and appoint the maintainer organisation that will assure sustainability of the framework and its eventual revision after the project ends.
- Recommendation S2: Cooperate with EUA to facilitate the implementation of the Data Stewardship curricula and FAIR data principles in university curricula.

Dissemination

- Recommendation D1: FAIRsFAIR will develop training and guidance materials on using FAIR4HE
- Recommendation D2: Deliver training on core competences through the provision of a virtual network and, where possible, through a limited number of face-to-face training sessions.

[30] Terms4FAIRskills Initiative [online] Available at https://terms4fairskills.github.io/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

6. Conclusions

This deliverable presents the proposed Data Stewardship and FAIR data competence framework for Higher Education (FAIR4HE). The suggested Data Stewardship Body of Knowledge is an input for further work on a Data Stewardship Model Curriculum, which can be used by universities for defining customised curricula for Data Stewardship and FAIR data principles.

The presented FAIR4HE framework uses the EDISON Data Science Framework methodology to identify essential Data Stewardship competences and knowledge based on a job market analysis. The proposed definition of competences has been mapped to and verified by the existing frameworks and curriculum recommendations such as EOSCpilot FAIR4S, ELIXIR Data Steward Competency Framework, DeIC Data Stewardship Curriculum recommendations, and the Foster Learning Objectives for Open Science. This exercise showed that the proposed CF-DSP reflects variety of competences defined in those frameworks and that it can be used as common framework for further defining the Body of Knowledge and Model Curriculum for Data Stewardship and FAIR data principles benefitting from the existing EDSF ecosystem for skills management and curriculum development.

While this deliverable presents a consistent definition of the Data Stewardship and FAIR competence framework for Higher Education. Its, practical implementation will require cooperation of multiple parties - in particular the research community, universities and policy makers.

A set of recommendations is proposed in this deliverable covering adoption, sustainability and dissemination that will all work together to implement necessary competences in the professional practice and culture. It is understood that university education will play an important role, first, in adopting FAIR competences and culture, and second, in creating the emerging Data Steward profession. This should include not only doctoral (PhD) training but also include Research Data Management and FAIR data courses at the Bachelor and Master levels.

The proposed framework has been presented and discussed at several community forum organised and co-organised by the project such as FAIR4HE Design Workshop on 8-9 October 2020 and a session on Data Stewardship and FAIR Competences in Academic Curricula as part of the CODATA FAIR Convergence Symposium on 30 November 2020.

At the same time, this deliverable can contribute to existing initiatives dealing with the Research Data Management training, e.g. EOSC WG Training and the Research Data Alliance IG on Professionalising Data Stewardship. Cooperation with selected national initiatives throughout the remaining time in FAIRsFAIR will also provide an important means of adoption and outreach for establishing Data Stewardship practice and culture.

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

References

[1] Final results of the European Data Market study measuring the size and trends of the EU data economy, ECIDC, March 2017 [online] https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

[2] Realising the European Open Science Cloud: First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud, European Commission, 2016 [online] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

[3] Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data, 2018 [online] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf

[4] D7.2 "Briefing on FAIR Competences and Synergies", FAIRsFAIR Project Deliverable, September 2021 [online] https://doi.org/10.5281/zenodo.4009006

[5] FAIR4HE Design Workshop, 8-9 Oct 2020, Program [online] https://docs.google.com/document/d/1lVLSJL41wtZa40Y7rHoYQahjNGXQ02dsk45goniR9sA/edit#heading=h.89nh0hqdfd91

[6] Workshop "Data Stewardship and FAIR Competences in Academic Curricula," part of CODATA FAIR Convergence Symposium [online] https://conference.codata.org/FAIRconvergence2020/sessions/222/

[7] RDA IG on Professionalising Data Stewardship [online] https://www.rd-alliance.org/groups/professionalising-data-stewardship-ig

[8] RDA IG on Education and Training on handling of research data (IG ETHRD) [online] https://www.rd-alliance.org/groups/education-and-training-handling-research-data.html

[9] Towards FAIR Data Steward as profession for the Life Sciences, Final report ZonMw & ELIXIR-NL projects (Oct 3, 2019) [online] https://doi.org/10.5281/zenodo.3471707

[10] The Danish e-Infrastructure Cooperation (DeIC) and Danish National Forum for Research Data Management (DM Forum) Report on National Coordination of Data Steward Education in Demark [online] https://www.deic.dk/sites/default/files/Data%20Steward%20Education%20in%20Denmark_0.pdf

[11] GO FAIR Initiative [online] https://www.go-fair.org/go-fair-initiative/

[12] EOSCpilot deliverable "D7.5: Strategy for Sustainable Development of Skills and Capabilities" The FAIR Guiding Principles for scientific data management and stewardship, March 2016, Scientific Data 3(160018 (2016)) [online] https://doi.org/10.1038/sdata.2016.18

[13] FAIR Data Maturity Model [online] https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v0.90.pdf

[14] RDA Data maturity model Working Group [online] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

[15] 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

[16] EDISON Data Science Framework (EDSF). [online] Available at https://github.com/EDISONcommunity/EDSF

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

[17] The Data Science Framework, A View from the EDISON Project, Editors Juan J. Cuadrado-Gallego, Yuri Demchenko, Springer Nature Switzerland AG 2020, ISBN 978-3-030-51022-0, ISBN 978-3-030-51023-7

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

[19] European Skills, Competences, Qualifications and Occupations (ESCO) framework. [online] https://ec.europa.eu/esco/portal/#modal-one

[20] The FOSTER Learning Objectives for Open Science, 23 February 2015 [online] https://doi.org/10.5281/zenodo.608586

[21] Terms4FAIRskills Initiative [online] https://terms4fairskills.github.io/Announcement.html

[22] P21's Framework for 21st Century Learning [online] http://www.p21.org/storage/documents/P21_framework_0515.pdf

[23] Tomasz Wiktorski, Yuri Demchenko, Adam Belloum, Model Curricula for Data Science EDISON Data Science Framework, Proc. 4th IEEE STC CC Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2017), part of The 9th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2017), 11-14 Dec 2017, Hong Kong. [online] https://ieeexplore.ieee.org/document/8241134

[24] Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Data and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kingdom, ISBN 978-1-4503-6186-6/19/03.

[25] e-CF3.0, 2016 European e-Competence Framework 3.0. A common European Framework for ICT Professionals in all industry sectors. CWA 16234:2014 Part 1. [online] http://ecompetences.eu/wpcontent/uploads/2014/02/European-e-CompetenceFramework-3.0_CEN_CWA_16234-1_2014.pdf

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Appendix A. FAIR4HE Design Workshop Programme 8-9 October 2020

Agenda - Day I (8 October 2020) - 14:00-17:00 CEST

14:00–14:15	**Welcome and introduction of participants** (Lennart Stoy, Yuri Demchenko)
14:15–15:00	**Session 1: Background and objectives** ● Overview of FAIRsFAIR work on FAIR in higher education (Lennart Stoy) ● Introduction to FAIRsFAIR work on "FAIR data competence framework" for higher education (Yuri Demchenko)
15:00-15:15	*Break*
15:15-16:30	**Session 2: Inputs from related projects and activities** ● National Coordination of Data Steward Education in Denmark (Lorna Wildgaard, KB) Slides: https://tinyurl.com/y52fcvhm ● The Minimal EOSC Skills Set and FAIR data management competencies (Vinciane Gaillard, EOSC WG on Skills and Training) Slides: ● Professionalizing data stewardship: competences, training and education (Mijke Jetten, DTL, ELIXIR-NL) https://doi.org/10.5281/zenodo.4073094 ● FAIR Data Management for Industry: Overview IDSA activity, DM-BoK, MATES project (Yuri Demchenko, University of Amsterdam) Slides: ● FAIR4S by EOSCpilot project: Data Stewardship Competency Framework (Angus Whyte, DCC) Slides: *Q&A/Discussion with participants*
16:30–17:00	**Wrap up Day I** ● The relevance of inputs to the FAIR competence framework for HE? How to build consensus on FAIR4HE Framework?

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Agenda - Day II (9 October 2020) - 10:00-13:15 CEST

10:00-10:45	**FAIRsFAIR and FAIR4HE methodology** (Yuri Demchenko) ● Data Stewardship, FAIR principles and organisational/infrastructure context ● Snapshot Job market analysis for Data Stewardship and related professions ● Overview EDISON/EDSF methodology: How to apply for definition of the FAIR4HE competences (CF-FAIR4HE) and Body of Knowledge (FAIR4HE-BoK) *Q&A/Discussion with participants*
10:45-11:00	*Break*
11:00–11:15	**Introduction to breakout discussions on FAIR4HE competence Framework** (Yuri Demchenko) ● Summary from day 1 presentations and discussions ● Proposed discussion topics
11:15-12:00	**Breakout discussions: FAIR4HE Curriculum alignment and forms of delivery**
12:00-12:15	*Break*
12:15-13:00	**Plenary discussion on implementation aspects, advocacy** ● Promoting and training for universities: to be continued in Task T7.5 ● Outreach to different programmes and validation with different disciplines ● Discussion possible early adopters and feedback ● Role of FAIRsFAIR project, EOSC, EUA, RDA
13:00-13:15	**Concluding remarks and closing**

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Appendix B - Job Market Analysis: Demand for Data Stewards and required competence and skills

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

B.1. Selecting sources of information

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

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

The following are general characteristics of the collected data:

- Period data collected 30 August – 1 September 2020
- Sites Indeed.com – NL, UK, DE, USA: monsterboard.nl - NL
- Days vacancy open: >50% more than 30 days
- Data Steward and related vacancies discovered:
  - NL – 51, UK – 30+, DE ~20, US – 300+
- Information collected/downloaded
  - Key skills snapshot – for all or first 200 for USA
- Full vacancy texts – approx. 40 in total
- Detailed analysis of sample vacancies

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

- NL, UK – 20, US - 6
- Number of companies and organisations posted Data Steward related jobs – more than 50

B.2. EDISON approach to analysis of collected information

The following approach was used for analysing the job advertisement data

1) Collect data on required competences and skills
2) Extract information related to competences, skills, knowledge, qualification level, and education; translate and/or reformulate if necessary
3) Split extracted information on initial classification or taxonomy facets, first of all, on required competences, skills, knowledge; suggest mapping if necessary
4) Apply existing taxonomy or classification: for purpose of this study we used skills and knowledge groups as defined by the EDSF definition of Data Science and Data Stewardship (i.e. Data Analytics, Data Engineering, Data Management, Research Methods, Domain Knowledge)
5) Identify competences and skills groups that do not fit into initial/existing taxonomy and create new competences and skills groups
6) Do clustering and aggregations of individual records/samples in each identified group
7) Verify the proposed competences groups definition by applying to originally collected and new data
8) Validate the proposed competence framework via community surveys and individual interviews.

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

B.3. Regular Job Market analysis

It is beneficial to do regular job market analysis, at least two times a year, job market analysis to review the relevance of currently defined competences, identify new demands and necessary adjustments, also update and extend demanded skills on Data Management and Data Analytics platforms, tools and technologies.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Appendix B - Data Scientist Workplace skills (aka "soft" or transversal skills)

This Appendix provides a definition of the Data Scientist workplace skills for reference purposes as they can be directly adopted for Data Steward Professionals.

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

B.1. Data Science Professional or Attitude skills (Thinking and acting like a Data Scientist)

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

Table B.1 lists the Data Science professional (or attitude) skills which are identified by the Data Science practitioners and educators (refer to EDSF CF-DS document). Although some of the skills are common the 21st century skills, it is important to provide the whole list of skills because it can serve as can provide a guidance for future Data Scientists what skills are expected from them and need to be developed along their career.

Table B.1. Data Science Professional skills (aka Thinking and acting like Data Scientist)

Skill ID	Skill definition
DSPS	General group definition: Thinking and acting like a Data Scientist
DSPS01	Accept/be ready for iterative development, know when to stop, comfortable with failure, accept the symmetry of outcome (both positive and negative results are valuable)
DSPS02	Ask the right questions
DSPS03	Recognise what things are important and what things are not important

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

DSPS04	Respect domain/subject matter knowledge in the area of data science
DSPS05	Data driven problem solver and impact-driven mindset
DSPS06	Recognise value of data, work with raw data, exercise good data intuition
DSPS07	Good sense of metrics, understand importance of the results validation, never stop looking at individual examples
DSPS08	Be aware about power and limitations of the main machine learning and data analytics algorithms and tools
DSPS09	Understand that most of data analytics algorithms are statistics and probability based, so any answer or solution has some degree of probability and represent an optimal solution for a number of variables and factors
DSPS10	Working in agile environment and coordinate with other roles and team members
DSPS11	Work in multi-disciplinary team, ability to communicate with the domain and subject matter experts
DSPS12	Embrace online learning, continuously improve your knowledge, use professional networks and communities
DSPS13	Story Telling: Deliver actionable result of your analysis
DSPS14	Attitude: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
DSPS15	Ethics and responsible use of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

B.2. 21st Century skills (aka Soft Skills)

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

Table B.2 lists the 21st Century skills defined based on the recommendations of the OECD Report on industry digitalisation[31], and P21's Framework for 21st Century Learning[32].

[31] Going Digital in a Multilateral World, Meeting of the OECD Council at Ministerial Level, Paris, 30-31 May 2018, OECD Report, 2018 [online] https://www.oecd.org/going-digital/C-MIN-2018-6-EN.pdf

[32] P21's Framework for 21st Century Learning [online] http://www.p21.org/storage/documents/P21_framework_0515.pdf

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Table B.2. The 21st Century workplace skills

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

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Appendix D. Example Designing Customisable Data Science and Data Steward Curriculum Using Ontology for Data Science Competences and Body of Knowledge

This section discusses an approach to building a customizable Data Science curriculum for different types of learners based on using the ontology of the EDISON Data Science Framework (EDSF) developed in the EU funded Project EDISON and widely used by universities and professional training organisations. It is based on the published papers by the authors of this deliverable [17], [23] and [24].

The education and training of Data Scientists requires a multi-disciplinary approach combining wide view of the Data Science and Analytics foundation with deep practical knowledge in domain specific areas. In modern conditions with the fast technology change and strong skills demand, the Data Science education and training should be customizable and delivered in multiple form, also providing sufficient data labs facilities for practical training.

D.1. EDSF Toolkit and Practical Uses of EDSF

EDSF was developed with the view of multiple practical uses for the whole range of tasks faced by universities, professional training organisations, companies and certification bodies related to Data Science education, training and capacity management. The following are the intended practical applications of EDSF:

- Academic curriculum design for general Data Science education and individual learning path construction for customizable training and career development
- Professional competence benchmarking, including CV or organisational profiles matching
- Professional certification of Data Science Professionals
- Vacancy construction tool for job advertisement (for HR) using controlled vocabulary and Data Science Taxonomy
- Data Science team building and organisational roles specification

The EDSF toolkit was developed to support the above mentioned applications and ensure their compatibility. It contains a number of API, ontologies and datasets representing different components of the EDSF and mapping between them. EDSF Toolkit is an ongoing development and available as Open Source at the EDSF github project [16].

D.2. EDSF Data Model and Ontology

The EDSF data model represents all the complex relations between the EDSF components such as competences, knowledge, skills, professional profiles, proficiency levels, and organisational roles, that exist in real life organisations. Initial EDSF definition followed the 4 parts structure as describes in section III. Initial definition of EDSF was made in the form of Excel workbooks and table what provided a good way of documenting but was difficult to use for practical applications.

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

D.2.1. EDSF Data Model

Currently, the EDSF toolkit contains a number of datasets representing different components of the EDSF and a mapping between them. Future EDSF development will formally define the ontologies related to the EDSF components and related dictionaries.

Figure D.1 illustrates the relation between different data sets and ontologies comprising EDSF. The CF-DS is structured along four dimensions (similar to European e-Competence Framework e-CFv3.0 [25]) that include (1) competence groups, (2) individual competences definition, (3) proficiency levels, and (4) corresponding knowledge and skills. In this context, each individual competence includes a set of required knowledge topics and a set of skills type A and skills type B. Such a CF-DS structure allows for competence based curriculum design where competences can be defined based on the professional profile (see EDSF Part 4 DSPP [16] for mapping between professional profiles and competences) or target leaners group when designing a full curriculum, or based on competence benchmarking for tailored training to address identified competences and knowledge gaps.

When a set of required competences is defined together with the required ranking or proficiency level, the set of required knowledge topics can be extracted and ordered according to proficiency level and relevance (or benchmark score) for further mapping to DS-BoK Knowledge Areas and Knowledge Units. The set of KAs and KUs defined for a specific competence set define the structure of the curriculum that further can be mapped to Model Curriculum Learning Units defined as individual courses and KAG related courses groups.

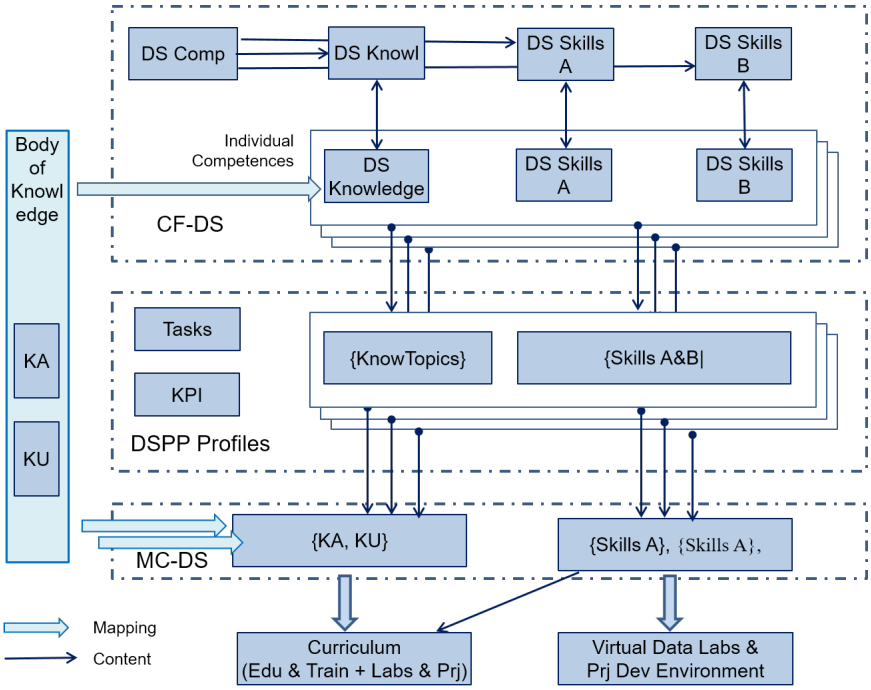


Figure D.1: EDSF Data Model and customised curriculum design for target professional group(s)

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

At the same time, required proficiency level is scored for each KA and KU, which will define a mastery levels and corresponding learning outcome for the targeted education or training. The following mastery levels are defined (using workplace terminology that can be easy mapped to mastery levels defined in MC-DS):

A - Awareness
- 1) Understand Terminology
- 2) Understand Principles
- 3) Apply principles
- 4) Understand Methods

U - Use/Application
- 5) Apply basics
- 6) Supervised use
- 7) Unsupervised Use

P - Professional/Expert
- 8) Development of applications using wide range of technologies
- 9) Supervise project development, team of professionals,

where borderline mastery levels 4 and 7 actually belong to both higher level and lower level groups.

D.2.2. Definition of the EDSF Ontology

In the next EDSF Release 4 (EDSF2020) [16], the CF-DS and DS-BoK will be expressed in a form of ontology that is linked also to DSPP profiles definition. The ontology provides an effective format for representing rich relations between EDSF components in a form of instance, classes and properties, it also allows easy design of APIs and benefitting from existing APIs (e.g. for Python and Java).

CF-DS ontology is a core ontology linking all EDSF entities, classes and properties. It includes ontologies for all individual competences defined for the main competence groups DSDA, DSENG, DSDM, DSRMP (refer to section III) defined as subclasses. Each competence is represented as an instance of the class to which it belongs (e.g. DSDA01 is an instance of DSDA subclass). Each competence instance includes the following properties:

- Knowledge that are required for competences, defined as knowledge topics and linked to Knowledge Units (KU) in the DS-BoK
- Skills related to the knowledge topics (defined in CF-DS as Skills type A)
- Skills related to practical experience including programming, tools and platforms (defined in CF-DS as Skills type B)

Figure D.2 illustrates the relation between different data sets and ontologies comprising EDSF, in particular it illustrates example of the DSDA01 competence that is defined as “Effectively use variety of data analytics techniques, such as Machine Learning, Data Mining, Prescriptive and Predictive Analytics, for complex data analysis through the whole data lifecycle”. The DSDA01 properties include knowledge topics KDSDA*, Skills Group A SDSDA* and Skills Group B SDSA*.

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

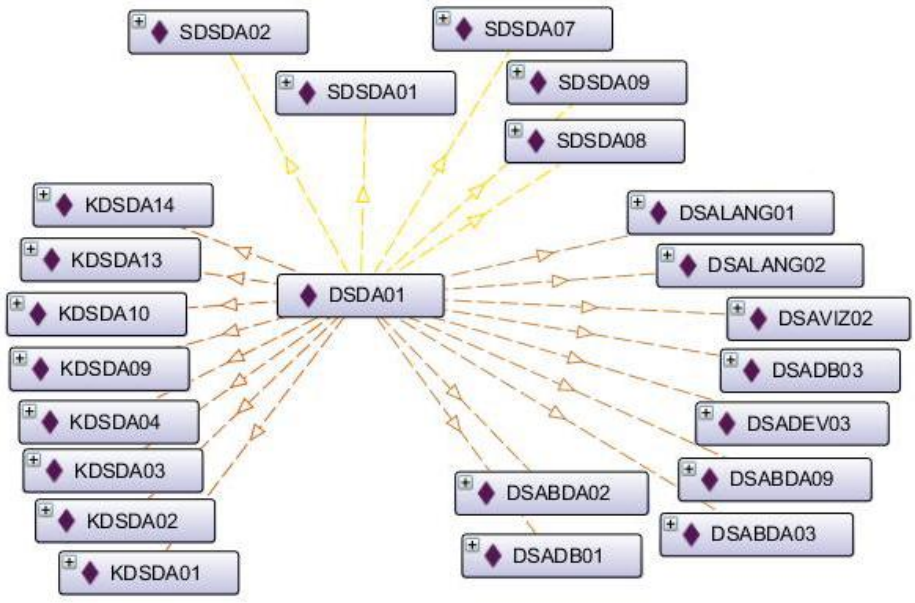


Figure D.2: Example DSDA01 Competence and its properties.

The Protégé ontology editor was used for ontology design and management. It allows creating and managing an ontology through an intuitive graphic interface and permits to export the ontology in a large number of formats. In this project RDF/OWL format is chosen in order to query the ontology using the Python module, OwlReady2.

D.3. Data Science Curriculum Design using EDSF Ontology

This section describes the workflow of using EDSF for curriculum design for selected/intended set of competences that are required for (1) a specific Data Science professional profile defined based on DSPP document, or (2) individual training program defined based on competence assessment and identified gaps. The individual competence assessment can be done based on CV matching against the intended job position or professional profile. It can be also done based on the certification exam or just self-assessment questionnaire. The outcome of this process is either level of matching or competence gap that can be used for suggesting necessary training program or tailored curriculum. As a part of the EDSF Toolkit development the authors have tested different methods for CV and job vacancy/profile matching using Doc2Vec document embedding and PV-DBOW training algorithms (available in the genism Python libraries).

When a set of required competences is defined together with the required ranking or proficiency level, the set of required knowledge topics can be extracted from individual competences (note, there exist multiple links from competence instances to single knowledge topic) and ordered according to proficiency level and relevance (or benchmark score) for further mapping to DS-BoK Knowledge Areas and Knowledge Units. The set of KAs and KUs defined for a specific competence set define the structure of the curriculum that further can

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

be mapped to the Model Curriculum Learning Units defined as individual courses and KAG related courses groups, otherwise it can be used directly as advice for constructing curriculum by the program or course manager.

At the same time, the required proficiency level is scored for each KA and KU, which will define a mastery levels and corresponding learning outcome for the targeted education or training. When using MC-DS as a template for designing customised curriculum, the proficiency levels (using scale 0 to 9) can be easily mapped to 3 mastery levels defined in MC-DS): Familiarity, Usage, Assessment (refer to EDSF Model Curriculum MC-DS [16]). Collected Skills type B linked to intended competences will provide advice on the required hands on training and practical project development tasks and development platform.

When using EDSF ontology, it is a routine task to extract all required knowledge topics, map them to KA/KU and define relevance score by querying ontology with a few lines of code using OwlReady2 Python module that allows manipulating ontology classes, instances and properties transparently.

Figure D.3 illustrates an example of relations between EDSF components when extracting required Knowledge Units for DSDA group of competences for DSP04 – Data Scientist professional profile (refer to DSPP for details). It shows that the following competences are required with the corresponding relevance/weight: DSDA01 = 9; DSD02 = 9; DSDA04 = 7. Required Knowledge Units are defined through the mapping knowledge topics KDSDA* to KU (using DS-BoK) and weighted based on average relevance by competences.

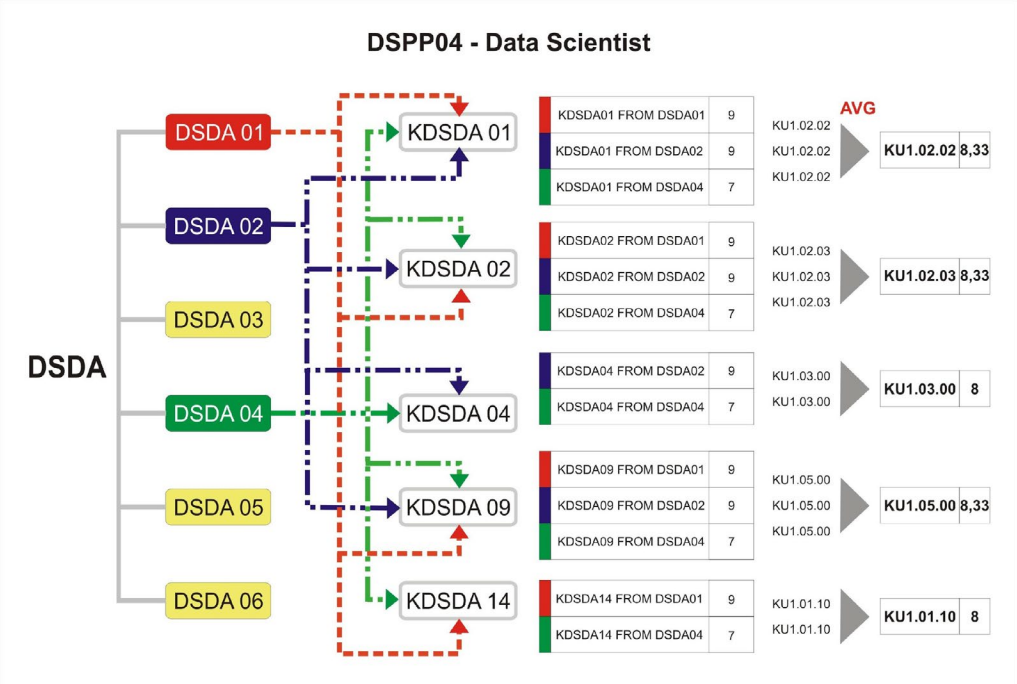
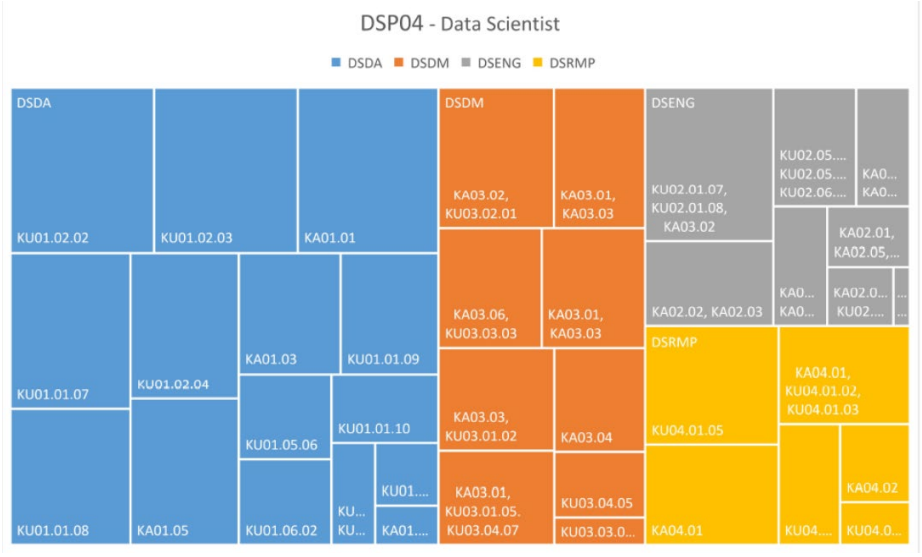


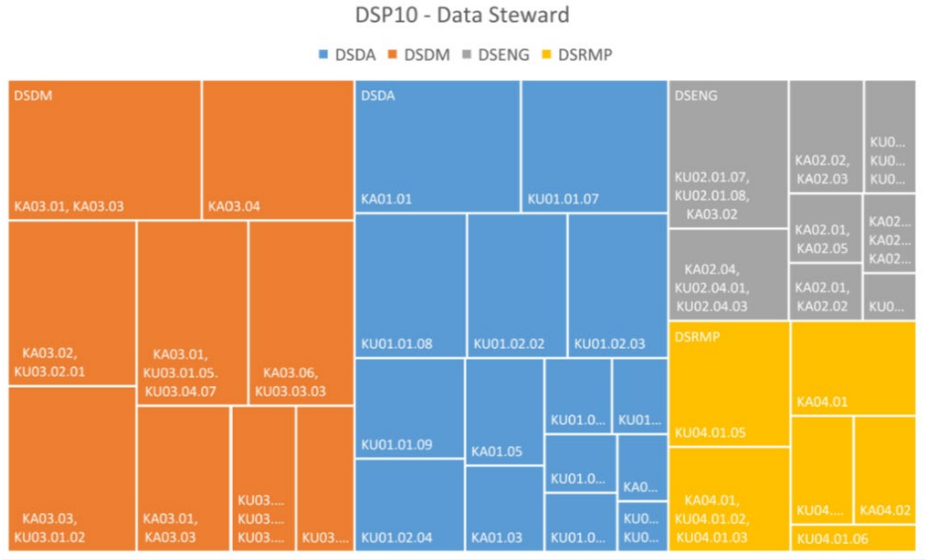
Figure D.3: Extracting required Knowledge Units from EDSF ontology.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The same process is applied to other competence groups relevant to specific professional profile or competence gap. Figure D.4 (a) and (b) shows an example of the suggested curriculum structure for two professional profiles: DSP04 – Data Scientist and DSP10 – Data Steward. The diagrams reflect the relative structure of the curriculum where Data Scientist has the major part of the Data Analytics courses (DSDA - blue) followed by necessary knowledge in Data Management (DSDM - orange), and Data Steward curriculum must focus on the Data Management courses (DSDM – orange), followed by basic knowledge in Data Analytics (DSDA – blue).



(a) Data Scientist curriculum structure



(b) Data Steward curriculum structure

Figure D.4: Example curriculum structure for DSP04 – Data Scientist and DSP10 – Data Steward.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The EDSF Toolkit and its outcome provide advice on the suggested curriculum structure that can be adjusted to the real condition of the teaching or training institution depending on the available teaching staff and lab base. It is also important that the courses are correctly ordered and necessary pre-requisite knowledge are specified. When using 3rd party educational platforms providers and cloud based data labs, the presented application can provide a specification for the required educational platform.

D.4. Defining Knowledge Units to include in the curriculum

Once the suggested Knowledge Units have been obtained, it is possible to combine them into educational courses, map them to courses defined in MC-DS or to existing courses, which are typically defined according to DS-BoK Knowledge Areas or Knowledge Units. EDSF ontology defines for these purposes the Course class, whose instances are directly connected to the KUs through the object property course. This allows for collecting relative scores for all KUs linked to the required curriculum.

When moving to a practical curriculum and courses design, it is important to define the courses relevance and their priority or sequence. The suggested courses content can be defined by KUs grouping based on their ontological similarity and difference. In a simple view, this defines the courses that need to be attended to achieve intended learning outcome and collect the necessary number of credits, in a classical education model. However, this doesn't solve the problem of the efficient programme planning or learning path design, what is especially important when designing a curriculum for workplace training, vocational education or self-education.

The Course class in the EDSF ontology can be used to calculate the course weight based on the integral score of the component KUs. This can be done by querying the ontology that will produce the list of associated courses for the required competence profile, sorted in descending order by weight. The course weight is calculated based on collecting all individual KU's scores linked to required competences, given the multiple relations and mapping between competences, knowledge topics in CF-DS, Knowledge Units in DS-BoK, and Learning Units in MC-DS. The course weights are normalized to 0-9 scale and aligned with the related competences relevance.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Appendix E. Aggregated CF-DSP competences (based on the section 3.3 analysis)

Table E.1. Data Stewardship Competence Groups (CF-DSP)

Data Management (DSDM)	Data Science Engineering (DSENG)	Data Science Research Methods and Project Management (DSRMP)	Data Science Domain Knowledge (DSDK) as Business Process Management (DSBA)
DSDM – extended, relevant Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing, ensure compliance with FAIR data principles.	DSENG – no changes, generally relevant Use engineering principles and modern computer technologies to research, design, implement new data analytics applications; develop experiments, processes, instruments, systems, infrastructures to support data handling during the whole data lifecycle.	DSRMP – revised, generally relevant Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals • Base research on collected scientific facts and collected data	DSDK – generally relevant Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations

FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

DSDM01 – extended, essential Develop and implement data management and governance strategy, in particular, in the form of Data Governance Policy and Data Management Plan (DMP) Ensure compliance with standards and best practices in Data Governance and Data Management	DSENG01 – no changes, low relevance Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis and visualisation	DSRMP01 – generally relevant Create new understandings, discover new relations by using the research methods (including hypothesis, artefact/experiment, evaluation) or similar engineering research and development methods	DSBA01 – relevant for organisation processes and data Analyse information needs, assess existing data and suggest/identify new data required for specific business context to achieve organizational goal, including using social network and open data sources • Data management and Quality Assurance of organisational data assets
DSDM02 – extended, essential Develop and implement relevant data models, define metadata using common standards and practices for different data sources in a variety of scientific and industry domains. • Ensure metadata compliance with FAIR requirements • Be familiar with the metadata management tools	DSENG02 – no changes, low relevance Develop and apply computational and data driven solutions to domain related problems using wide range of data analytics platforms, with a special focus on Big Data technologies for large datasets and cloud based data analytics platforms	DSRMP02 – generally relevant Direct systematic study toward the understanding of the observable facts, and discovers new approaches to achieve research or organisational goals	DSBA02 – relevant for organisation processes and data Operationalise fuzzy concepts to enable key performance indicators measurement to validate the business analysis, identify and assess potential challenges • Specify requirements/develop data models for organisational data

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

DSDM03 – extended, essential Integrate heterogeneous data from multiple sources and provide them for further analysis and use • Perform data preparation and cleaning • Match/transfer data models of individual datasets	DSENG03 – extended, relevant Develop and prototype specialised data analysis applications, tools and supporting infrastructures for data driven scientific, business or organisational workflow; use distributed, parallel, batch and streaming processing platforms, including online and cloud based solutions for on-demand provisioned and scalable services • Develop new tools and applications, ensure support of the data FAIRness requirements by existing and new tools and applications	DSRMP03- extended, essential Analyse domain related research process model, identify and analyse available data to identify research questions and/or organisational objectives and formulate sound hypothesis • Link domain related concepts and models to general/abstract Data Science concepts and models,	DSBA03 – generally relevant Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends • Ensure data availability and quality for BA/BI needs
DSDM04 – extended, highly essential Maintain historical information on data handling, including reference to published data and corresponding data sources • Publish data, metadata and related metrics • Perform and maintain data archiving • Develop necessary archiving policy, comply with Open	DSENG04– extended, essential Develop, deploy and operate data infrastructure, including data storage and processing facilities, using different distributed and cloud based platforms. • Implement requirements for data storage facilities to comply with the data management policies and FAIR data principles in particular.	DSRMP04 – generally relevant Undertake creative work, making systematic use of investigation or experimentation, to discover or revise knowledge of reality, and use this knowledge to devise new applications (data driven), contribute to the development of organizational or project objectives	DSBA04 – relevant for organisation processes and data Analyse opportunity and suggest the use of historical data available at organisation for organizational processes optimization • Coordinate implementation of FAIR data principles for collected data, ensure proper lineage and provenance of collected data

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Science and Open Access policies if applicable • Perform data provenance and ensure continuity through the whole data lifecycle, ensure data provenance			
DSDM05 – extended, essential Develop policy and metrics for data quality management (e.g. Altmetrix), maintain data quality and compliance to standards, perform data curation Interact/Collaborate with data providers and data owners to ensure data quality	DSENG05– extended, relevant Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection, ensure standards and corresponding data protection regulation compliance, in particular GDPR. • Define and implement (coordinate) data access policies for different stakeholders and organisational roles	DSRMP05 – extended, essential Design experiments which include data collection (passive and active) for hypothesis testing and problem solving • Work with Data Science, Data Stewardship and data infrastructure teams to develop project/research goals.	DSBA05 – relevant for organisation processes and data Analyse customer relations data to optimise/improve interaction with the specific user groups or in the specific business sectors
DSDM06 – extended, essential Develop and manage/supervise policies on data protection, privacy, IPR and ethical issues in data management, address legal issues if necessary. • Ensure GDPR compliance in data management and access	DSENG06– extended, essential Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load) and ELT	DSRMP06 – extended, essential Develop and guide data driven projects, including project planning, experiment design, data collection and handling	DSBA06 – relevant for organisation processes and data Analyse multiple data sources for marketing purposes; identify effective marketing actions

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

• Develop data access policies and coordinate their implementation and monitoring, including security breaches handling	(Extract, Load, Transform), OLTP, OLAP processes for large datasets • Define, implement and maintain data model, reference data, master data definitions, implement consistent metadata		
DSDM07* - added new, essential Manage Data Management/Data Stewards team, coordinate related activity between organisational departments, external stakeholder to fulfill Data Governance policy requirements, provide advice and training to staff. Define domain/organisation specific data management requirements, communicate to all departments and supervise/coordinate their implementation. Coordinate/supervise data acquisition.			DSBA07 – added, essential Coordinate intra organisational activities related to data analytics, data management and data provenance/lineage along all data flow stages, ensure data FAIRness
DSDM08* - added new, essential Develop organisational policy and coordinate activities for sustainable implementation of the FAIR data principles and Open Science, define corresponding requirements to			

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

data infrastructure and tools, ensure organisational awareness.			
DSDM09* - added new, essential Specify requirements to and supervise the organisational infrastructure for data management and (and archiving), maintain the park for data management tools, provide support to staff (researchers or business developers), coordinate solving problems.			

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558