Research Data Management and Data Stewardship Competences in University Curriculum

Yuri Demchenko
University of Amsterdam, The Netherlands
y.demchenko@uva.nl

Lennart Stoy
EUA, Belgium
lennart.stoy@eua.eu

***Abstract*— Skills for data governance and management are critical for wide adoption of the Open Science practices and effective use of the data in research, industry, business and other economy sectors. The FAIR (Findable – Accessible – Interoperable - Reusable) data management principles and data stewardship provide a foundation of effective research data management. The 2018 "Turning FAIR into Reality" report and other documents recommend that data skills should be more widely included in university curricula and that a concerted effort should be made to coordinate and accelerate the pedagogy for professional data roles. Throughout Europe, and beyond, many organisations, projects and initiatives work on providing training on FAIR data competences. However wider adoption of the FAIR data culture can be achieved by including FAIR competence into university curricula. This paper presents the ongoing work of the FAIRsFAIR project to develop Data Stewardship competence framework and provide recommendations for implementing in the university curricula by defining the Data Stewardship Body of Knowledge Model Curricula. The proposed approach and identified competences and knowledge items are supported by the job market analysis. The presented work is actively using the EDISON Data Science Framework as a basis for Data Stewardship competences definition and methodology for linking competences, skills, knowledge, and intended learning outcomes when designing curricula.**

Keywords—Data Steward Professional, Data Management and Governance, Research Data Management, FAIR data principles, Data Stewardship Competence Framework (CF-DSP), EDISON Data Science Framework (EDSF), Big Data.

I. INTRODUCTION

The growing importance of data in the modern data driven economy, research and industry, requires special attention to including data management and governance related topics in university education. The future specialists should understand the role and value of data in research and industry and be able to derive actionable value from data collected from research, technological process or business/social activity, be able to use open data and public data. Modern data driven research and industry require new types of specialists capable of supporting all stages of the data lifecycle from data production to data processing and actionable results delivery, visualisation and reporting, which can be jointly defined as the Data Science professions family [1, 2].

Data Management and Data Analytics are the critical aspects in digital transformation, however it requires a change of the whole organisational culture, which is often referred to as data literacy. The research community has responded to this with the formulation of the FAIR data principles that suggest data must be Findable, Accessible, Interoperable, Reusable [3].

The education and training of Data Stewards should not be limited to the general data management of FAIR principles. The presented research identified a number of competences, skills, knowledge from multiple technology and data management areas that are required from the Data Stewards for their successful work in their future organisation. Besides data related competences and knowledge, the Data Stewards are required to have an understanding of organisational processes project management (research or business, depending on organisation).

At the present time most of the existing university curricula and training programs cover a limited set of competences and knowledge areas of what is required for multiple Data Science and Data Stewardship professional profiles and organisational roles enacted by research and industry. In conditions of continuous technology development and shortened technology change cycle, Data Science and Data Stewardship education requires effective combination of theoretical, practical and workplace skills.

Recent European initiatives and projects such as the European Open Science Cloud (EOSC) [4], Research Data Alliance (RDA) [5] facilitated implementation of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles [6, 7] that allow for effective data exchange and integration across scientific domains, making scientific data a valuable resource and a growth factor for the whole digital economy and society.

The proposed work has been done in the framework of the EU funded FAIRsFAIR project [8] that recognises

the importance of establishing new profession of the Data Steward and introducing FAIR principles and culture at the early stage of professional education and career by including FAIR principles into university curricula. The FAIR competences and corresponding Knowledge Areas can be introduced as a special course and/or a part of other courses typically taught at universities such as Research Methods, Research Data Management, or Professional Issues. Research Data Management and FAIR principles are currently attributed to the emerging profession of the Data Steward.

The proposed Data Stewardship Professional Competence Framework (CF-DSP) is based on the EDISON Data Science Framework (EDSF) [2] and defines the main competences required from the Data Steward in their work in different organisations. CF-DSP also complemented with the DSP Body of Knowledge (DSP-BoK) that is defined as a subset of the Data Science Body of Knowledge. This allows reusing the whole EDSF toolkit developed for customised curriculum design [9]

The paper refers to the previous authors' works on defining the EDISON Data Science Framework (EDSF) [10] and its application of individual competences management and customised curricula design based on required competences and intended learning outcome that can be targeted for specific professional profiles including Data Stewards [11].

The paper is organized as follows. Section II provides a reference to European and international initiatives related to research data management and growing demand for the Data Stewardship profession. Section III summarises the job market analysis for Data Steward and Data Management vacancies to identify demanded competences, skills and knowledge. Section IV provides an overview of existing frameworks defining Data Stewardship and related competences, including EDSF. Section V discusses the proposed definition of the Data Stewardship Professional Competence Framework (CF-DSP) as extension to EDSF. Section VI provides suggestions about new knowledge topics to be included into the DSP Body of Knowledge. A conclusion in section VII provides a summary and refers to ongoing and future developments.

II. Research Data Management and Data Stewardship

The importance of data and research information sharing has been central in a number of European wide initiatives and projects, such as Open Access Open Data, Open Science, Open Commons. The Research Data Alliance (RDA) that was created in 2012 jointly by the National Science Foundation of USA (NSF) and European Commission, became a key community coordination body to exchange and develop best practices in research data management.

To facilitate research data sharing and implementation of the FAIR principles, European Commission started Open Research Data (ORD) Pilot [12] and currently all EU funded projects are required to develop and implement the Data Management Plan (DMP) at the initial stage of the project. Data produced in the project must be stored in the open available but secure repositories (operated by the project or using national or European data archive services. Metadata must be published, and quality of data ensured, in particular, compliance with the FAIR principles. The DMP template provided by the Commission is structured to ensure that the data produced by funded research are open and FAIR [13]

FAIR data principles and Data Stewardship are among key objectives of the European Open Science Cloud (EOSC) initiative started in 2016 as the part of the "European Cloud Initiative - Building a competitive data and knowledge economy in Europe" [14] that is targeted to capitalise on the data revolution. Under this initiative, EOSC federates existing and emerging e-Infrastructures to provide European science, industry, and public authorities with world-class data infrastructure connected to high performance computers (HPC).

The EOSC goals are to enable the Open Science Commons [15] and achieve FAIRness in research data management and in the services provided. At the present time, the EOSC projects created the foundation for research data interoperability and integration for European IRs. The Minimum Viable EOSC (MVE) achieved by the end of 2020, will create a starting point for future EOSC development [16].

III. Demand for Data Stewardship and Data Management Competences and Skills

A. Job Market Analysis

The presented study and the proposed Data Stewardship 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 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 the 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. Refer for details to the FAIRsFAIR deliverable D7.3

Data Stewardship Professional Competence Framework [17].

The following are general characteristics of the data collected from the job market:

- Period data collected 30 August – 1 September 2020
- Sites Indeed.com – NL, UK, DE, USA (large number of vacancies); monsterboard.nl, IEEE Jobs – NL (single vacancies)
- 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 analysed – approx. 40 in total
- Detailed analysis of sample vacancies
- Number of companies and organisations posted Data Steward related jobs – more than 50

B. *EDISON methodology to collect and analyse job market and competence related data [1, 2]*

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 approach has been used when 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 the 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 don't fit into the 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 outlined above process has been applied to the collected information and all steps are tracked in the two Excel workbooks provided as supplementary material which is available at the FAIRsFAIR project shared storage.

C. *Demanded Data Stewardship competences*

A preliminary analysis has been done based on data collected from the job advertisements on such popular job search and employment portals as indeed.com, monsterboard.com, IEEE Jobs portals and LinkedIn Jobs what provided sufficient amount vacancies for decisive analysis. The collected data were used to extract information on competences, skills and knowledge demanded from prospective Data Steward candidates (following EDSF methodology as explained above).

Figure 1 illustrates what competences, skills and knowledge topics have been extracted form collected vacancies data and their mapping to CF-DS competence groups.

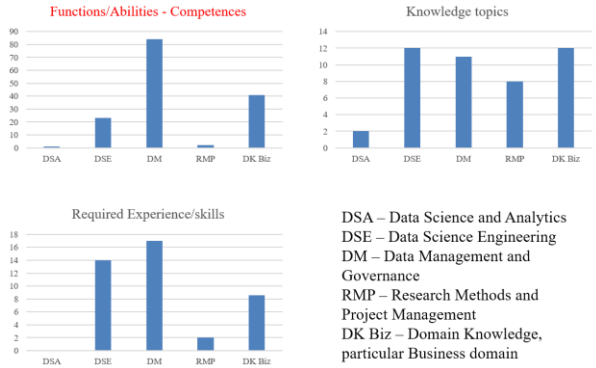

Figure 1. Data Stewardship job market analysis: identified (a) competences; (b) experience/skills; (c) knowledge topics

The analysis allows to assume that (1) EDSF and CF-DS suitable for defining Data Stewardship Professional Competence Framework as a CF-DS profile; (2) certain

extension may be needed for the general CF-DS as it is described below.

The diagram illustrates a wide variety of knowledge that is expected from the Data Stewards covering all main competence and knowledge groups defined for Data Scientists: DSENG, DSDM, DSRMP, DSDK/DS Biz, with less demand for Data Analytics, which will remain an area for the Data Scientists. This knowledge profile can be explained by the expected role of the Data Steward as coordinating multiple cross-organisational (horizontal) activities to ensure data quality, data management infrastructure operation, and promoting best practices in data management, in particular FAIR data principles.

D. The outcome of the job vacancies analysis

The following conclusions and assumptions can be made from the initial vacancies analysis:

- The published Data Stewards vacancies demonstrated the 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 framework

The most populated competence group is Data Management, what reflects the nature of the Data Steward profession and responsibilities. Two other populated groups are Domain Knowledge and Data Science Engineering what reflects another side of the Data Steward profession to act as bridge between ICT teams operating data facilities and domain specialists. This imposes 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/extracted set of competences, skills and knowledge topics will be 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.

IV. Existing Frameworks for Data Management and Data Stewardship Competences

A. EOSCpilot FAIR4S Framework

EOSCpilot project defines data stewardship as a shared responsibility of professional groups involved in data management: data management and curation, data science and analytics, data services engineering, and domain research [18]. The EOSCpilot deliverable "D7.5: Strategy for Sustainable Development of Skills and Capabilities" [18] describes the comprehensive FAIR4S framework that defines 6 skill profiles grouped around the research data lifecycle stages and 4 professional groups that include 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 at the stage of the EOSCpilot project 2017-2019.

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

- 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
- Apprise and present: 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 2 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 3 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.

B. ELIXIR Data Stewardship Competency Framework for Life Sciences (DSP4LS)

The ELIXIR Data Stewardship Competency Framework for Life Sciences [19] (hereafter referred as DSP4LS – Data Steward Profession for Life Sciences) is the complete framework that defines the competencies, skills and knowledge related to Data Stewardship as a distinct profession in the modern data driven science ecosystem and life sciences in particular. The defined 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 3 areas:

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

For all Data Steward roles, the 8 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 successfully completing training you will be able to*"

C. DeIC Data Stewardship curricula recommendations

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 [20] 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 extensive overview and analysis of existing competence frameworks and educational programmes for Data Science and Data Stewardship. The community feedback was collected via a Questionnaire that collected 86 complete responses (and 42 partial responses).

The Data Stewardship competences are defined in 6 competence groups comprising 22 competences: 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

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

The GO FAIR initiative [21] which is devoted to promoting and sustainable adoption of the FAIR data principles, provided recommendations on FAIR metadata management that can be used for linking between general requirements to FAIR implementation and underlying technology and infrastructure and consequently for defining technical expertise areas. Important to note that these requirements require both advanced data management infrastructure tools and corresponding competences from Data Engineers and Data Stewards.

E. 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 [22]. 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 the central for the whole DMBOK. The DMBOK also explains the relation between Data Governance and Data Management where Data Governance is focused on Legal and Judicial views (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 or SME roles. Data Management functions are performed by the Chief Information Officer office that include Data Architects, Data Analysts, Coordinating Data Stewards and technical Data Steward roles.

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 definition and organisational roles include the following responsibilities and activities:

- 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-today 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.

To stress the uniqueness if the Data Stewardship competences, the DMBOK reads: “Best Data Steward is not made but found” [22]

V. EDISON Data Science Framework (EDSF)

The EDISON Data Science Framework [2] provides a basis for the definition of the Data Science profession and enables the definition of other components related to Data Science education, training, organisational roles definition and skills management. EDSF provides a common semantic basis for interoperability of the different forms of the Data Science curriculum definition and education or training delivery, as well as knowledge assessment and professional certification based on the fully enumerated definition of EDSF components and individual units.

A. EDSF Components

Figure 2 below illustrates the main EDSF components and their inter-relations:

- 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

The proposed framework provides the basis for other components and services of the Data Science professional environment such as

- Data Science Education Environment (DSEE) and Virtual Data Labs (that can be cloud based and provisioned on demand)
- Directory of Education and Training Materials
- Data Science Community Portal (CP) that can provide information and community support services, such as individual competences benchmarking and personalized educational path building.

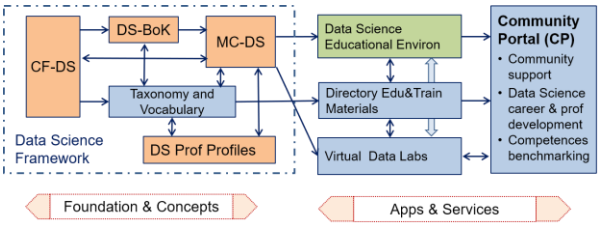

Figure 2. EDISON Data Science Framework components and Data Science Educational environment.

The CF-DS provides the overall basis for the whole framework. The CF-DS includes the core competences required for the successful work of a Data Scientist in different work environments in industry and in research and through the whole career path.

The following core CF-DS competence and skills groups have been identified (refer to CF-DS specification [2] for details):

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, others) (DSDA)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools) (DSENG)
- Data Management and Governance (including data stewardship, curation, and preservation) (DSDM)
- Research Methods and Project Methods (DSRMP)
- Domain Knowledge and Expertise (Subject/Scientific domain related)

Data Science competences must be supported by knowledge that are defined primarily by education and training, and skills that are defined by work experience correspondingly. The CF-DS defines two types of skills (refer to CF-DS [2] for full definition of the identified knowledge and skills groups):

- Skills Type A which are related to the professional experience and major competences, and
- Skills Type B that are related to wide range of practical computational skills including using programming languages, development environment and cloud based platforms.

The DS-BoK defines the Knowledge Areas (KA) for building Data Science curricula that are required to support identified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK is based on ACM/IEEE Classification Computer Science (CCS2012) [23], incorporates best practices in defining domain specific BoK's and provides reference to existing related BoK's. It also includes proposed new KA to incorporate new technologies and scientific subjects required for consistent Data Science education and training.

The MC-DS is built based on DS-BoK and linked to CF-DS where Learning Outcomes are defined based on CF-DS competences (specifically skills type A), and Learning Units are mapped to Knowledge Units in DS-BoK. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles. Three mastery (or proficiency) levels are defined for each Learning Outcome to allow for flexible curricula development and profiling for different Data Science professional profiles. The proposed Learning Outcomes are enumerated to have a direct mapping to the enumerated competences in CF-DS. The practical curriculum should be supported by a corresponding educational environment for hands-on labs and educational projects development.

Figure 3 below illustrates relations between Competence framework components: competence, skills, knowledge, attitude, and academic domain, including Body of Knowledge, Model Curriculum, Learning Outcomes, and educational profiles.

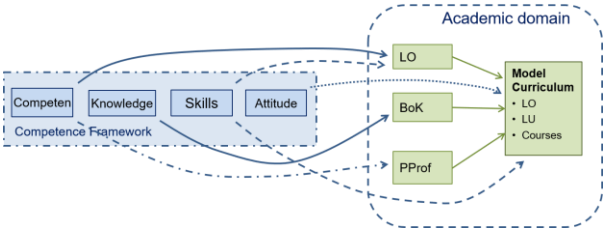

Figure 3. Relations between Competence Framework components and academic domain

The formal DS-BoK and MC-DS definition creates a basis for Data Science education and training programmes compatibility and consequently, Data Science related competences and skills transferability.

B. DSPP and Data Steward Professional Definition

The DSPP is defined as an extension to European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy [24] using the ESCO top classification groups. DSPP definition provides an important instrument to define effective organisational structures and roles related to Data Science positions and can also be used for building an individual career path and corresponding competences and skills transferability between organisations and sectors.

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

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

VI. CF-DSP: Data Stewardship and FAIR Competences as Extension for EDSF CF-DS

Based on the above analysis and community discussion around FAIRsFAIR project activity, the following provides short summary about the additional competences that are proposed to be added to the current Data Science Competence Framework (CF-DS/EDSF) competence groups (only additional full or partial definition is provided).

Data Management competence group is the most important activity for the Data Steward and FAIR principles implementation. So, it got most of the extensions. Specific extensions are suggested for the Data Science Engineering competence group and to Domain related competences to ensure link and interaction with different organisational units. In this case Data Steward would play the role of liaison, coordinator and communicator to ensure the FAIR principles are implemented and sustained.

The following are suggested extensions for the Data Management competence group (refer to the EDSF [2] and Deliverable D7.3 [17] for full list and details).

DSDM02 (extended)

- Ensure metadata compliance with FAIR requirements. Be familiar with the metadata management tools

DSDM03 (extended)

- Perform data preparation and cleaning. Match/transfer data model

DSDM04 (extended)

- 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
- Perform data provenance and ensure continuity through the whole data lifecycle, ensure data provenance

DSDM06(extended)

- Ensure GDPR compliance in data management and access

- Develop data access policies and coordinate their implementation and monitoring, including security breaches handling

DSDM07* (added new): Manage Data Science team and coordinate organisational activity

- Manage Data Management/Data Stewards team, coordinate related activity between organisational departments, external stakeholder to fulfil 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

DSDM08* (added new): Develop policy and implement FAIR principles

- 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.

DSDM09* (added new): Define requirements to data management infrastructure and liaise with IT department

- 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.

VII. Data Stewardship Body of Knowledge

The Data Stewardship Professional Body of Knowledge (DSP-BoK) can be defined as a profile or subset of the general Data Science Body of Knowledge as defined in EDSF [2]

Similarly, to extending DSDM competence group in the CF-DSP, the KAG3-DSDM group can be extended with the specific Knowledge Units (KU) related to newly defined and developed knowledge on FAIR and RDM related to FAIR metadata, Persistent Identifiers (PID) [25], Open Science, others. However, core KAG3-DSDM Knowledge Areas (KA) and Knowledge Units will remain relevant to DSP-BoK, including most of KAs from DM-BoK and KAs related to RDA recommendations.

A more detailed definition of the DSP-BoK will be developed at the next stage of the FAIRsFAIR project.

VIII. Example Research Data Management and Stewardship Course Syllabus

The Research Data Management is well defined and well supported with training materials, however in most cases, they are focused on the specific scientific domain. The outlined below course include FAIR principles and Data Stewardship related topics.

The following RDMS course example is structured along practical aspects of the research data management:

A. Use cases for data management and stewardship
   - Preserving the Scientific Record

B. Data Management elements (organizational and individual)
   - Goals and motivation for managing your data
   - Data formats, metadata, related standards
   - Creating documentation and metadata, metadata for discovery
   - Using data portals and metadata registries
   - Tracking Data Usage, data provenance, linked data
   - Handling sensitive data
   - Backing up data, backup tools and services
   - Data Management Plan (DMP)

C. Responsible Data Use (Citation, Copyright, Data Restrictions)
   - Data privacy and GDPR compliance

D. FAIR principles in Research Data Management, supporting tools, maturity model and compliance

E. Data Stewardship and organisational data management
   - Responsibilities and competences
   - DMP management and data quality assurance

F. Open Science, Open Access and Open Data (Definition, Standards, Open Data use and reuse, open governmental data)
   - Research data and open access
   - Repository and self- archiving services
   - RDA products and recommendations: PID, data types, data type registries, others
   - ORCID identifier for data and authors
   - Stakeholders and roles: engineer, librarian, researcher
   - Open Data services: ORCID.org, Altmetric Doughnut, Zenodo

G. Hands-on practice includes the following topics:
   a) Data Management Plan design
   b) Metadata and tools
   c) Selection of licenses for open data and contents (e.g. Creative Common, and Open Database)

The presented course outline has been implemented in multiple formats for Computer Science and non-IT masters as well as in the professional training programs.

IX. Conclusion and Further Developments

The presented work on the CF-DSP definition has benefitted from the wide community contribution from a number of communities such as EOSC Skills and Training Working Group, Research Data Alliance (RDA) Interest Groups on Data Stewardship Professionalisation, GO FAIR and Dutch e-Science, and

the FAIRsFAIR Project. The progress of this works has been discussed at a number of workshops and events organised by the FAIRsFAIR in cooperation with RDA and CODATA. See FAIRsFAIR deliverable D7.3 for details and overview [17].

Further CF-DSP definition will include general workplace skills, also referred to as "soft" skills or professional attitude/aptitude skills, which are becoming increasingly important in modern data driven and future data driven economy. This would include two groups of skills that are increasingly demanded by employers: Data Stewardship professional skills, and general 21st Century skills that comprise a set of skills that include critical thinking, design thinking, communication, collaboration, organizational awareness, ethics, and others. This will be done by assessing similar skill groups defined in EDSF if they can be directly used or modified.

The next stages in the FAIRsFAIR project will target the definition of the required extensions to the Body of Knowledge and development of the Data Stewardship Model Curriculum, which expectedly will reuse EDSF definition and methodology.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Horizon2020 projects FAIRsFAIR (grant number 831558), MATES (grant number 591889) and EDISON (grant n. 675419).

The authors also value wide discussions in the RDA and EOSC forums on the proposed FAIR and Stewardship competences and knowledge topics.

REFERENCES

[1] 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

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

[3] Barend Mons, et al, The FAIR Guiding Principles for scientific data management and stewardship [online] https://www.nature.com/articles/sdata201618

[4] European Open Science Cloud (EOSC) [online] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

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

[6] 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

[7] FAIRdata Forum [online] https://fairdataforum.org/

[8] FAIRsFAIR Project: Fostering FAIR data practices in Europe [online] https://www.fairsfair.eu/

[9] Yuri Demchenko, Adam Belloum, Cees de Laat, Charles Loomis, Tomasz Wiktorski, Erwin Spekschoor, Customisable Data Science Educational Environment: From Competences Management and Curriculum Design to Virtual Labs On-Demand, 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.

[10] Yuri Demchenko, Adam Belloum, Wouter Los, Tomasz Wiktorski, Andrea Manieri, Steve Brewer, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry, 3rd IEEE STC CC and RDA Workshop on Curricula and Teaching Methods in Cloud Computing, Big Data, and Data Science (DTW2016), in Proc.The 8th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2016), 12-15 December 2016, Luxembourg.

[11] Yuri Demchenko, Luca Comminiello, Gianluca Reali, Designing Customisable Data Science Curriculum using Ontology for Science and Body of Knowledge, 2019 International Conference on Big Daya and Education (ICBDE2019), March 30 - April 1, 2019, London, United Kigndom, ISBN978-1-4503-6186-6/19/03.

[12] Data management, Extension of the Open Research Data Pilot in Horizon 2020, Horizon2020 Manual [online] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[13] Data Management and Data Management Plan template [online] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[14] "European Cloud Initiative - Building a competitive data and knowledge economy in Europe" [online] https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe

[15] Tiziana Ferrari, Diego Scardaci, Sergio Andreozzi, The Open Science Commons for the European Research Area, Part of the ISSI Scientific Report Series book series (ISSI, volume 15) [online] https://link.springer.com/chapter/10.1007/978-3-319-65633-5_3

[16] SRIA Solutions for a Sustainable EOSC. A tinman report from the EOSC Sustainability Working Group, Draft 2 December 2019, https://www.eosc-nordic.eu/content/uploads/2020/03/Tinman_draft_19_compressed.pdf

[17] FAIRsFAIR Project Deliverable D7.3 Data Stewardship Professional Competence Framework, Work in Progress. To be published Feb 2021.

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

[19] 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
[20] 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
[21] GO FAIR Initiative [online] https://www.go-fair.org/go-fair-initiative/
[22] DAMA Data Management Body of Knowledge (DMBOK2), DAMA International, 2017
[23] CCS, 2012 The 2012 ACM Computing Classification System. Available at http://www.acm.org/about/class/class/2012
[24] European Skills, Competences, Qualifications and Occupations (ESCO) framework. Available at https://ec.europa.eu/esco/portal/#modal-one
[25] Persistent Identifiers, Groups of European Data Experts, 27 Nov 2017, RDA [online] https://www.rd-alliance.org/system/files/PID-report_v6.1_2017-12-13_final.pdf