



Research Data Management and FAIR Competences in University Curriculum

Yuri Demchenko

University of Amsterdam

FAIRsFAIR Doctoral Education Workshop

26 May 2021



FAIRsFAIR
Fostering Fair Data Practices in Europe



EDISON
building the data
science profession



Outline

- Demand for Data Management competences and Data Stewardship
- Snapshot Job market analysis for Data Stewardship and related professions
 - Evidence based and community driven: Job Market and Landscape
- Existing Data Stewardship framework and competences mapping
- Proposed Data Stewardship Professional Competence Framework (CF-DSP)
- Ongoing development: Body of Knowledge, Model Curriculum
- Discussion



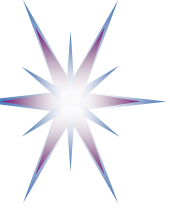
Research Data Management and FAIR Data Principles - Demand for new Competences

- Critical importance of research data sharing
 - European Open Science Cloud as Federated Data Infrastructure
 - Industry recognises importance of Data Management and Data Quality
 - International Data Management Association since 2007
- Research Data sharing and FAIR principles in European policy on European Research Area
 - FAIR data principle: Findable, Accessible, Interoperable, Reusable
- Multiple governmental and community initiatives
 - Research Data Alliance (RDA) since 2012 focused on different aspects of the Research Data sharing



FAIR is an Overloaded Concept (and term)

- Primarily FAIR is (set of) principles for sustainable Research Data Management (RDM) and Open Science
- FAIR is an initiative
- FAIR is a key policy area of EOSC
- FAIR data management is part of Data Management Plan (DMP) and required by Horizon Europe and many national funding bodies
- FAIR impose a number of requirements to research Infrastructure
- Universities should play important role in FAIR and RDM adoption
 - Still slow adoption at all levels: Bachelor, Master, Doctoral, teachers

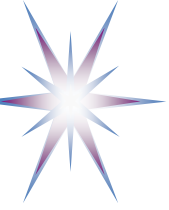


FAIR from the technical point of view

- Findable
 - Metadata and PDI – infrastructure and tools
 - Registries and handles resolution, API
 - Policies and SLA
- Accessible
 - Repositories and data storage: infrastructure and management
 - Policy and access control: infrastructure and API management
 - Data access protocols
 - Usage Policy and Sovereignty
 - Data protection, compliance, privacy and GDPR
- Interoperable
 - Standard data formats
 - Metadata and API
 - FAIR maturity level and certification
- Reusable
 - Data provenance and lineage
 - Preservation
 - Metadata, PID and API – linked or embedded into datasets

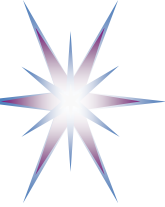
This motivates Data Stewards' interaction with both **Data Analytics and Applications developers** roles and **Data Infrastructure** roles

- Consequently related competences from Data Stewards are needed

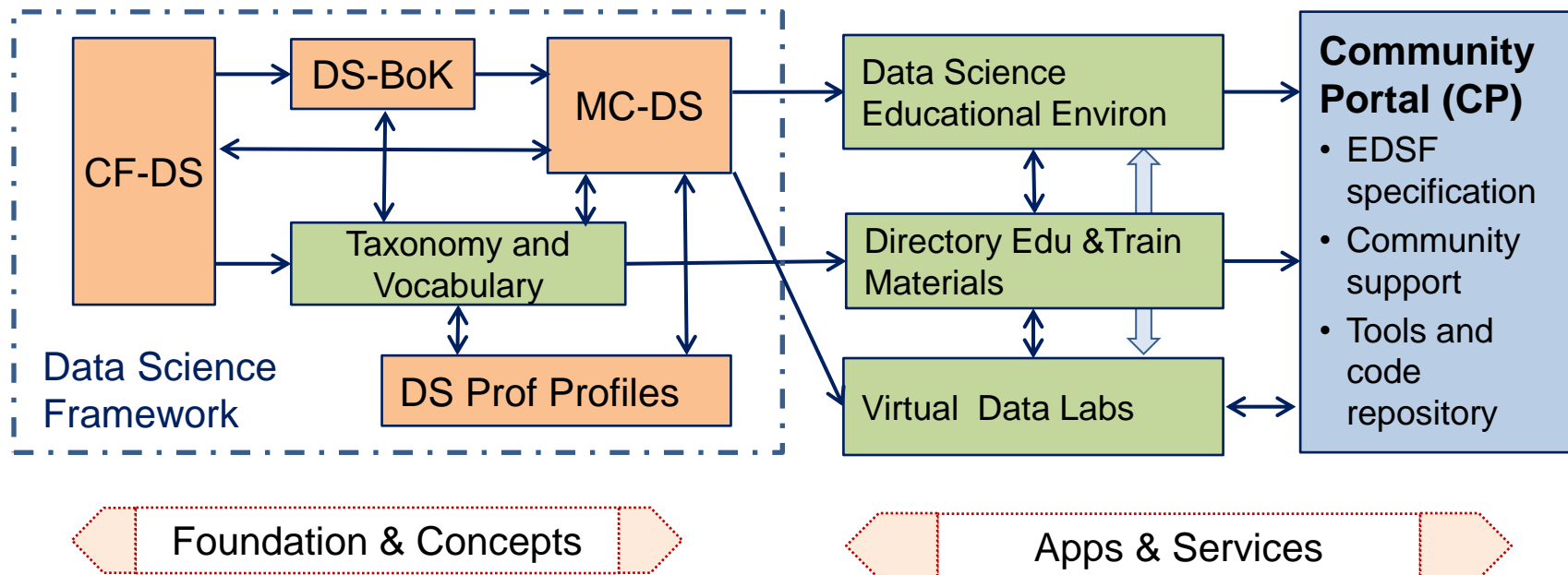


Methodology: How to put all together?

- EDISON Data Science Framework (EDSF) and community based maintenance
 - CF-DS Competence groups: DSDA, DSENG, DSDM, DSRMP, DSDK structure and mapping
 - Curriculum Design approach: From Competences and Body of Knowledge to Model Curriculum and Learning Units
- Data Management and Governance curriculum design for Data Stewardship and FAIR principles
 - EDSF Data Management Body of Knowledge (DSDM-BoK)
 - DAMA Body of Knowledge (DAMA BoK)
 - Data Stewardship – Existing Frameworks
 - Research Data Management (RDM) best practices
- Methodology: Evidence based and Community driven
 - Job market analysis
 - Mapping to and consensus with existing initiatives and frameworks



EDISON Data Science Framework (EDSF) – Core components and community maintained services

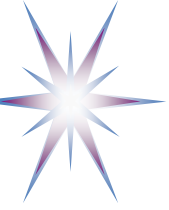


EDISON Framework core components and documents

- CF-DS – Data Science Competence Framework (Part 1)
- DS-BoK – Data Science Body of Knowledge (Part 2)
- MC-DS – Data Science Model Curriculum (Part 3)
- DSPP – Data Science Professional profiles (Part 4)
- Data Science Taxonomies and Scientific Disciplines Classification

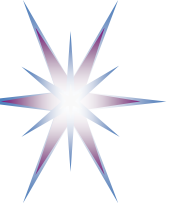
Applications and Services

- Virtual Data Science Labs
- Data Science Educational Environment
- Directory of edu & train resources
- Community Portal – currently github



Data Stewards – Job market review

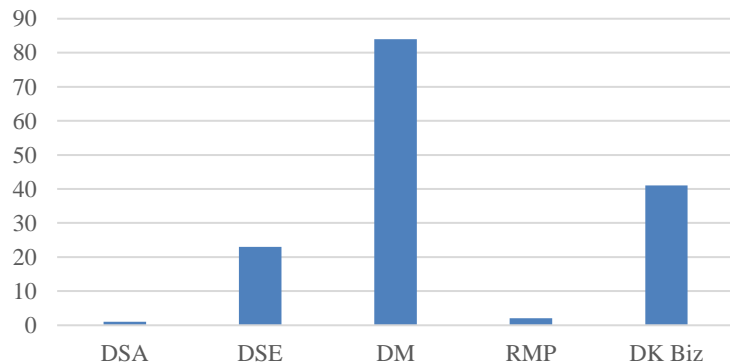
- Date 30 August – 1 September 2020
- Indeed.com – NL, UK, DE, USA
- Days open: >50% more than 30 days
- Data Steward and related vacancies
 - NL – 51, UK – 30+, DE ~20, US – 300+
 - Key skills snapshot
- Sample vacancies detailed analysis
 - NL, UK – 12, US - 6



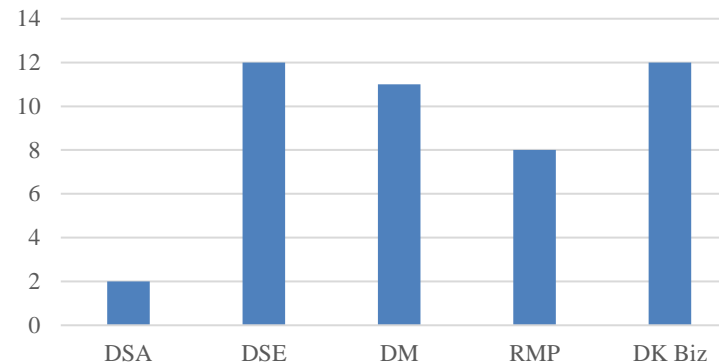
Vacancies profile – By Data Science Competence Groups

Wide range of Competences: Responsibility, Functions, activities

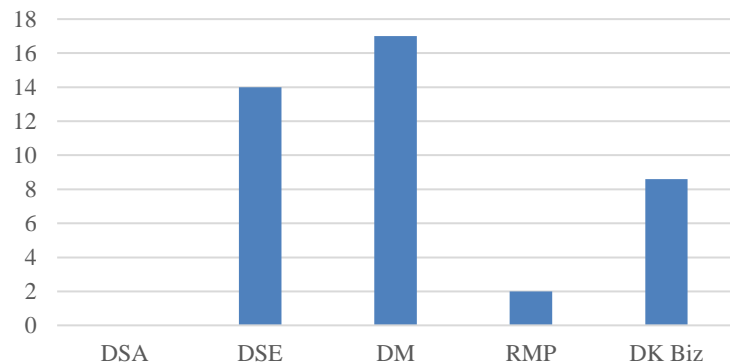
Functions/Abilities - Competences



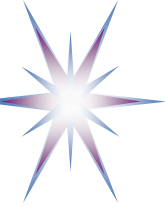
Knowledge topics



Required Experience/skills



DSA – Data Science and Analytics
DSE – Data Science Engineering
DM – Data Management and Governance
RMP – Research Methods and Project Management
DK Biz – Domain Knowledge, particular Business domain



Important Knowledge Items extracted from Job vacancies (indeed.com – NL, DE, UK, US, Sept 2020)

- 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
- Project management
- 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 analysis and visualisation tools
- Data lifecycle, lineage, provenance
- 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



Proposed CF-DSP Competences DSDM01-DSDM04 as extension to CF-DS (1)

Proposed DSP Competence

DSDM

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 management and governance strategy, in particular, in a 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 variety of scientific and industry domains.

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

DSDM03

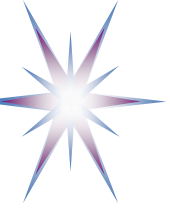
Integrate heterogeneous data from multiple sources and provide them for further analysis and use

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

DSDM04

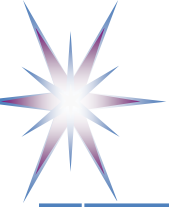
Maintain historical information on data handling, including reference to published data and corresponding data sources

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



Proposed CF-DSP Competences DSDM05-DSDM09 as extension to CF-DS (2)

Proposed DSP Competence	
DSDM05	Develop policy and metrics for data quality management (e.g. Altmetrix), maintain data quality and compliance to standards, perform data curation <ul style="list-style-type: none">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, address legal issues if necessary. <ul style="list-style-type: none">Ensure GDPR compliance in data management and accessDevelop data access policies and coordinate their implementation and monitoring, including security breaches handling
DSDM07* (new)	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* (new)	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* (new)	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.



Proposed CF-DSP Competences DSENG01-DSENG03 as extension to CF-DS (1)

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.

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

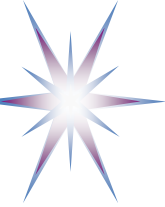
DSENG02 – no changes, low relevance

Develop and apply computational and data driven solutions to domain related problems using wide range of data analytics platforms, with the 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 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



Proposed CF-DSP Competences DSEN04-DSENG06 as extension to CF-DS (2)

Relevance and proposed changes and extensions (posted as revised text and bulleted extensions)

DSENG04

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

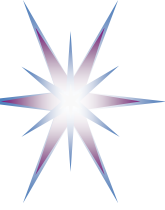
Consistently apply data security mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection, ensure standards and corresponding data protection regulation compliance, in particular GDPR.

- Define and implement (coordinate) data access policies for different stakeholders and organisational roles

DSENG06

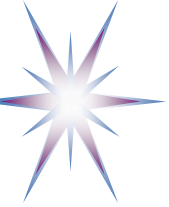
Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform), OLTP, OLAP processes for large datasets

- Define, implement and maintain data model, reference data, master data definitions, implement consistent metadata

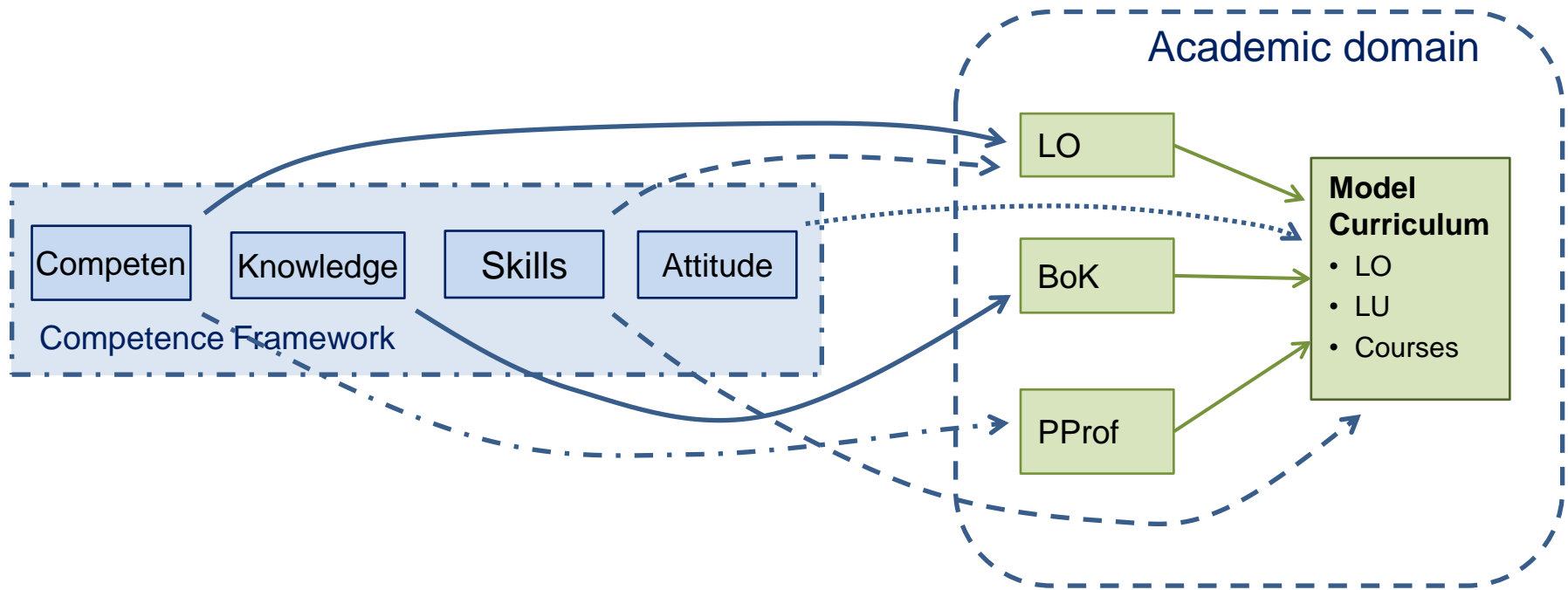


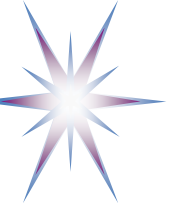
FAIR4HE & CF-DSP via Data Stewardship Professionalisation – Pillars and Cooperation (existing frameworks)

- EOSCpilot FAIR4S Data Stewardship Competence Framework
- ELIXIR Data Stewardship Competence Framework
- DeIC and DM Forum: Report on National Coordination of Data Steward Education in Denmark
- DAMA BoK (2007) – DAMAI Data Management Body of Knowledge
- EDISON Data Science Framework (EDSF) and EDISON Community Initiative

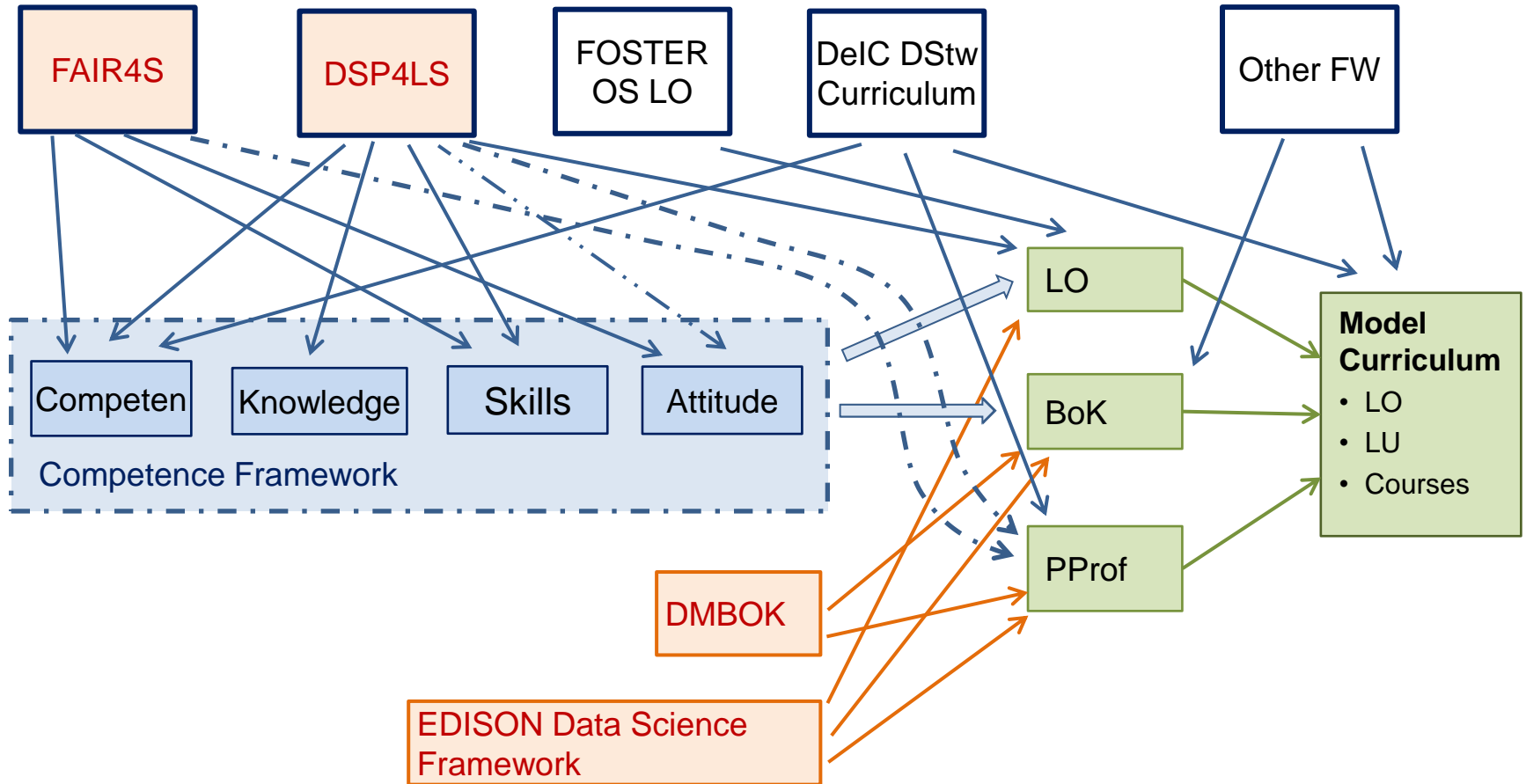


Relations between Competences, Skills, Knowledge/BoK, Professional Profiles





CF-DSP and Existing Frameworks Mapping

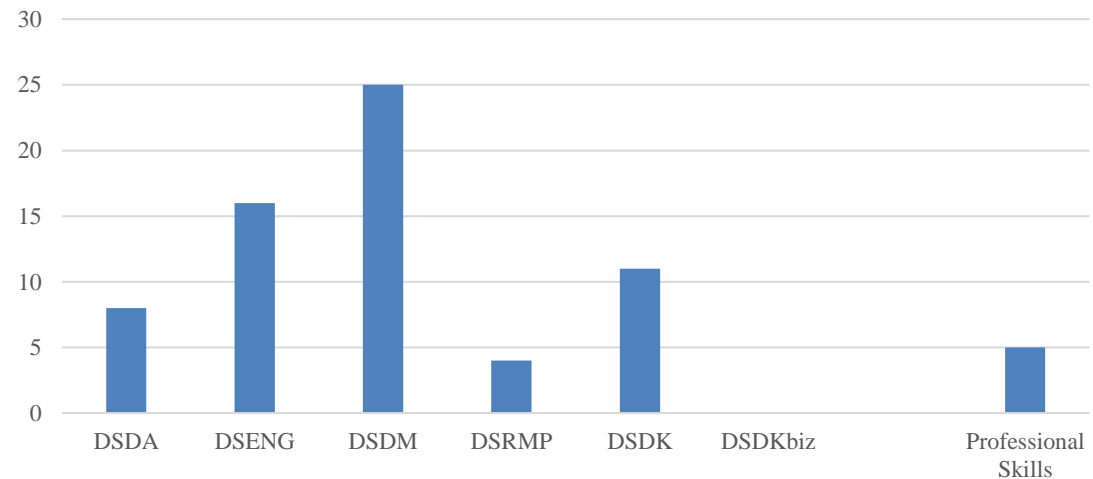




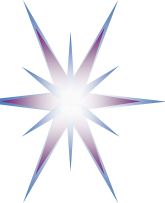
EOSCpilot FAIR4S Data Steward Competences

- Organisational capabilities for sustaining FAIR data across projects
- Stewardship skills to deliver FAIR data from projects
- Data Stewardship Roles and **Shared responsibility**:
 - Data Stewards and researchers
- 59 competences grouped in
 - 3 general groups
 - Govern and assess
 - Scope and resource
 - Advise and enable
 - 6 Data (curation) lifecycle process stages
 - Plan and design
 - Capture and process
 - Integrate and analyse
 - Apprise and preserve
 - Publish and release
 - Expose and discover

FAIR4S Mapping to CF-DSP



[ref] EOSCpilot D7.5 Strategy for sustainable development of skills and capabilities



ELIXIR - Data Stewardship Competency framework (courtesy ELIXIR Project)

Data Steward Roles and Competence Profiles

- **Policy:** institute and policy focused
- **Research:** project and research focused
- **Infrastructure:** data handling and e-infrastructure focused
- **Activities – Knowledge, Skills, Abilities – Learning Objectives**

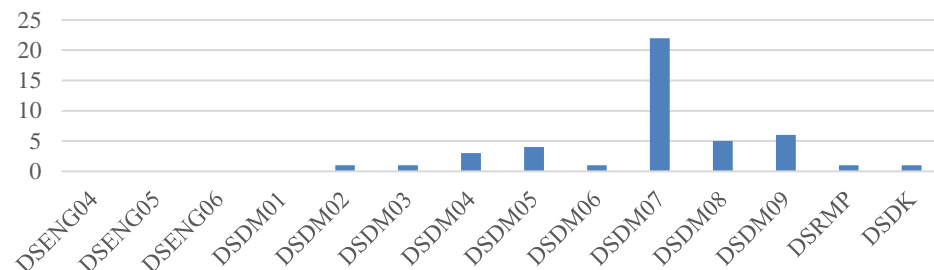
Competence groups

- 1) Policy/Strategy
- 2) Compliance
- 3) Alignment with FAIR data principles
- 4) Services
- 5) Infrastructure
- 6) Knowledge Management
- 7) Network
- 8) Data sharing

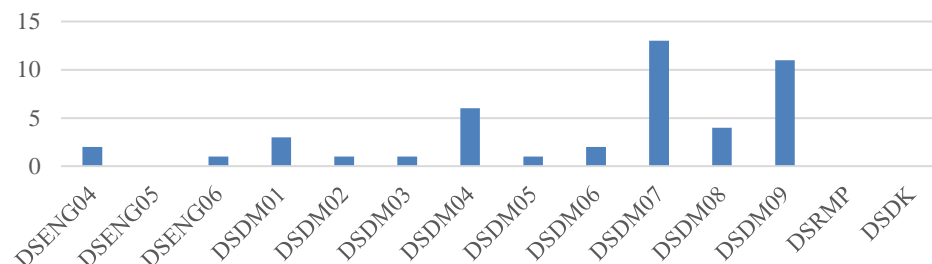
ZonMw & ELIXIR-NL funded project “Towards FAIR Data Steward as profession for the Life Sciences”

○Final report (Oct 3, 2019): <https://doi.org/10.5281/zenodo.3471707>

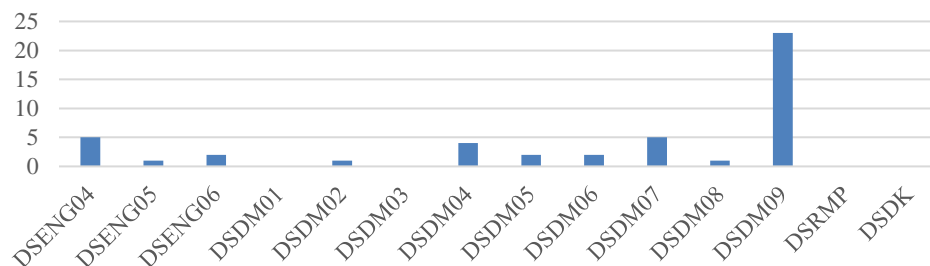
Activites and Tasks - Data Steward Policy

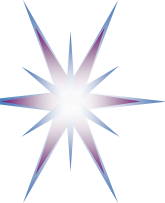


Activites and Tasks - Data Steward Research



Activites and Tasks - Data Steward Infrastructure





ELIXIR - Data Stewardship Competency framework (courtesy ELIXIR Project)

Data Steward Roles and Competence Profiles

- **Policy:** institute and policy focused
- **Research:** project and research focused
- **Infrastructure:** data handling and e-infrastructure focused
- **Activities – Knowledge, Skills, Abilities – Learning Objectives**

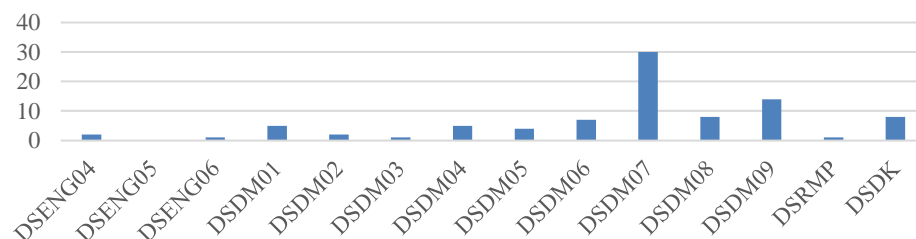
Competence groups

- 1) Policy/Strategy
- 2) Compliance
- 3) Alignment with FAIR data principles
- 4) Services
- 5) Infrastructure
- 6) Knowledge Management
- 7) Network
- 8) Data sharing

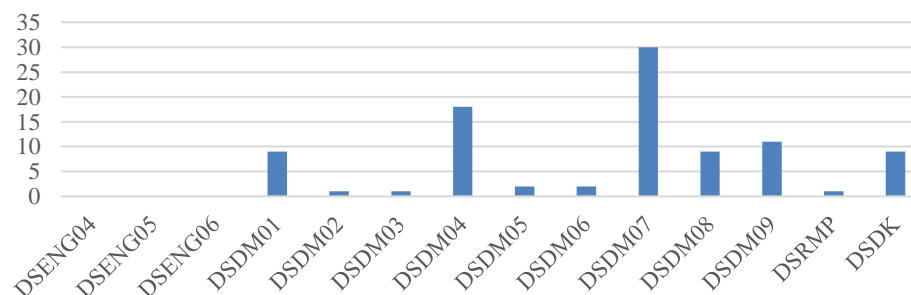
ZonMw & ELIXIR-NL funded project “Towards FAIR Data Steward as profession for the Life Sciences”

○Final report (Oct 3, 2019): <https://doi.org/10.5281/zenodo.3471707>

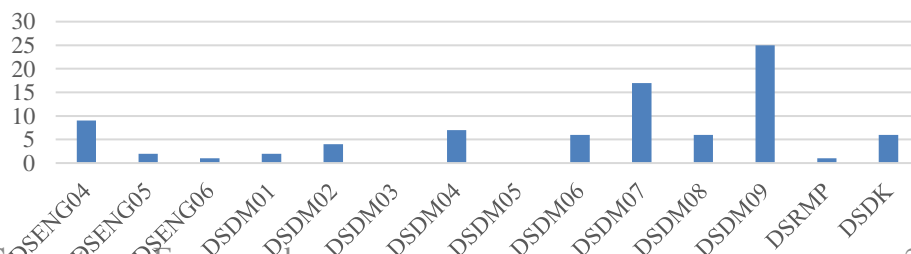
Knowledge, Skills, Abilities - Data Steward Policy

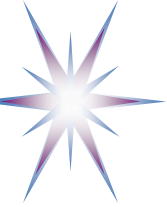


Knowledge, Skills, Abilities - Data Steward Research



Knowledge, Skills, Abilities - Data Steward Policy





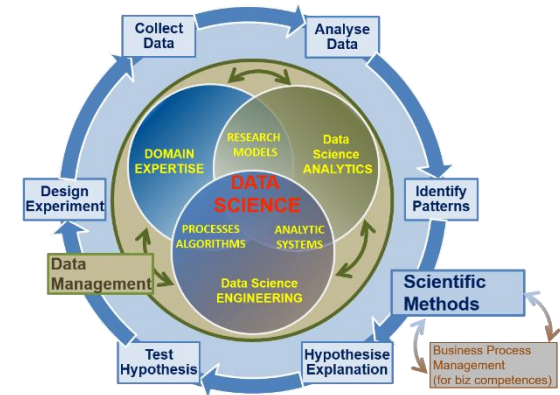
Ongoing Developments (FAIRsFAIR Project and RDA)

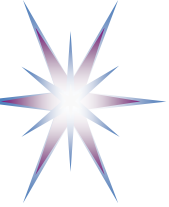
- Data Stewardship Body of Knowledge
- Data Stewardship and FAIR Model curriculum
- Related courses

EDSF: Data Science Body of Knowledge (DS-BoK)

DS-BoK Knowledge Area Groups (KAG)

- KAG1-DNA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Research Methods and Project Management group*
- KAG5-DSBA: Business Analytics and Business Intelligence
- **KAG* - DSDK**: *Data Science domain knowledge to be defined by related expert groups*





KAG3-DSDM: *Data Management group: data curation, preservation and data infrastructure*

DM-BoK version 2 “Guide for performing data management”

– 11 Knowledge Areas

(1) Data Governance

(2) Data Architecture

(3) Data Modelling and Design

(4) Data Storage and Operations

(5) *Data Security*

(5a) *Data compliance, Data Privacy, GDPR*

(6) Data Integration and Interoperability

(7) *Documents and Content*

(8) Reference and Master Data

(9) Data Warehousing and Business Intelligence

(10) *Metadata*

(11) Data Quality

Other Knowledge Areas motivated by *RDA, European Open Data initiatives, European Open Data Cloud*

(12) *PID, linked data, data registries*

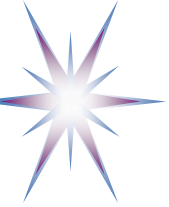
(13) *Data Management Plan*

(14) *Open Science, Open Data, Open Access, ORCID*

(15) *Responsible data use, Ethics*

(16)* *Data Sovereignty (and Indigenous data protection)*

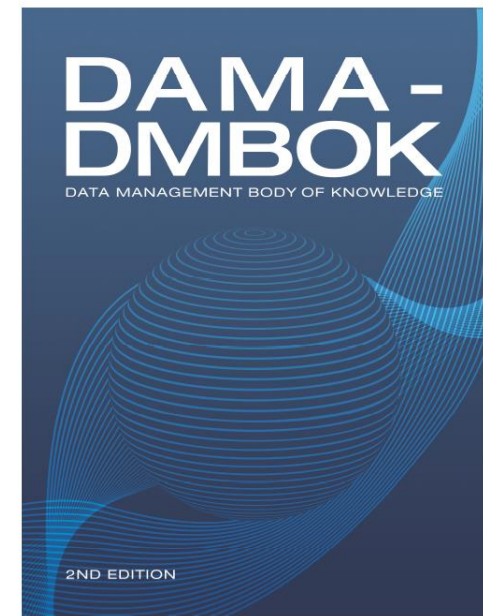
- Highlighted in red: Considered (Research) Data Management literacy (minimum required knowledge)



The DAMA-DMBOK Framework

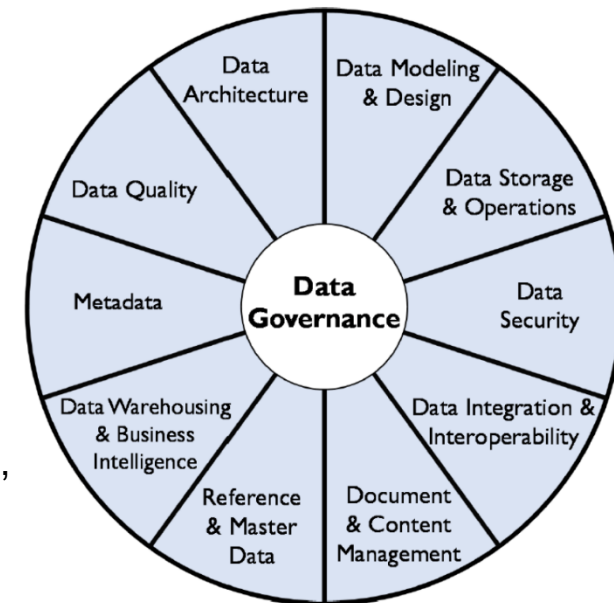
The DAMA-DMBOK Framework goes into depth about the Knowledge Areas that make up the overall scope of data management.

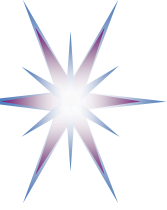
- DAMA-DMBOK Guidelines describe DMBOK and provide recommendations for implementation
- The DAMA Wheel – 11 Knowledge Areas
- The Environmental Factors hexagon
- The Knowledge Area Context Diagram



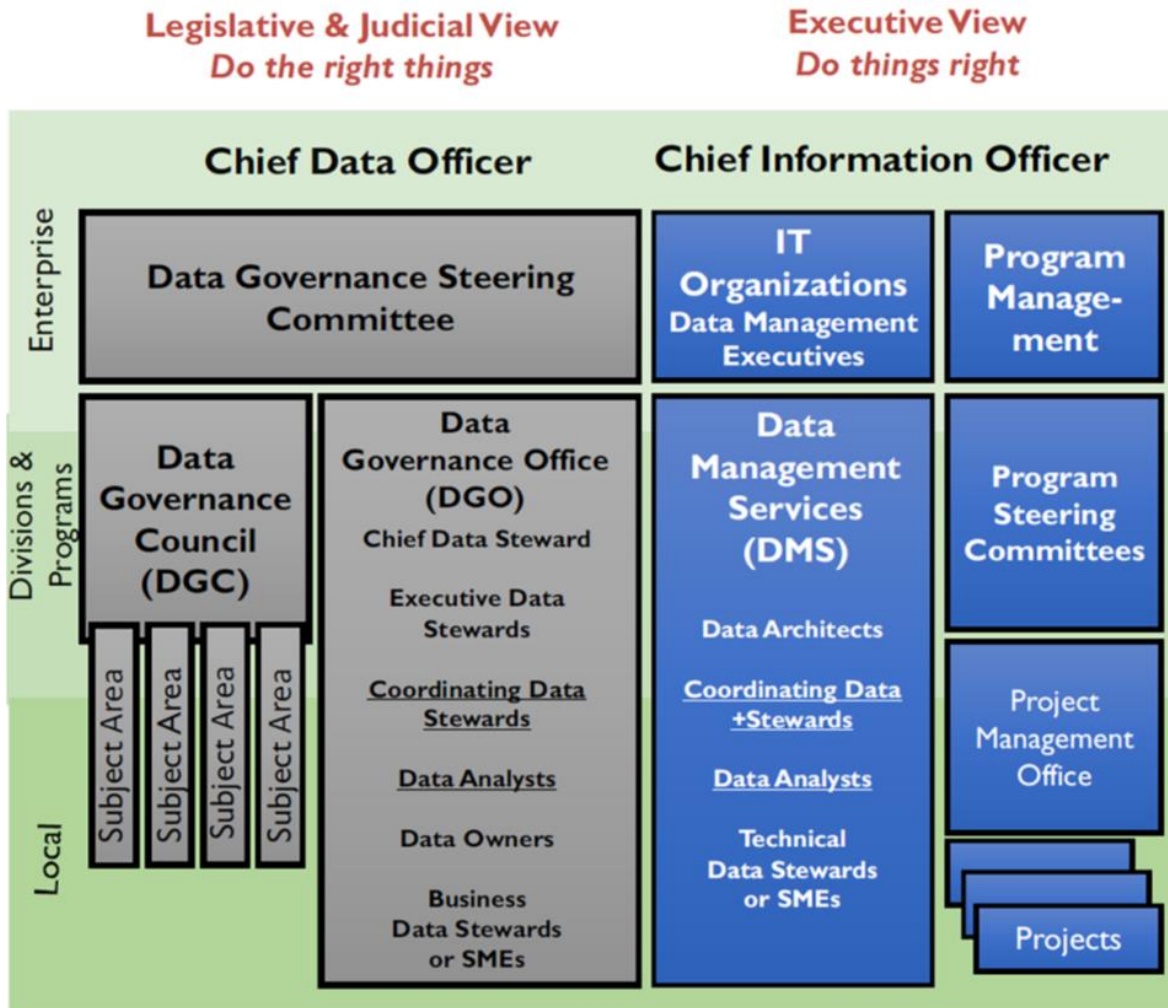
Technics Publications
BASKING RIDGE, NEW JERSEY

[ref] DAMA – DMBOK: Data Management Body of Knowledge, 2nd Edition, 2017. DAMA International, Technics Publications Llc, 626 pp





DMBOK: Data Governance Organisation Parts



- Separation of governance responsibilities
- Multi-layer
- CDO
- CIO
- Councils

Data Governance Office (DGO)

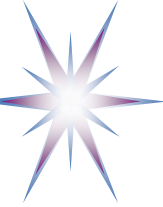
- Chief Data Steward
- Executive Data Steward
- Business Data Steward or SME



Data Stewardship (according to DM-BOK)

- **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.

“Best Data Steward is not made but found” DMBOK1 (2009)



Course: Research Data Management and Stewardship (RDMS) (1)

A. Use cases for data management and stewardship

- Preserving the Scientific Record
- Data Lifecycle and Provenance

B. Data Management elements (organisational 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
- Ethical issues



Course: Research Data Management and Stewardship (RDMS) (2)

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 and Open Data (Definition, Standards, Open Data use and reuse, open government 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 topics: DMP, Metadata, Data Formats, Data publishing, etc



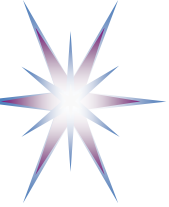
Discussion

- Proposed Data Stewardship Profession competences as extension to CF-DS
- Approach for consolidating existing frameworks



Additional information

- FAIR data principles, technical context and organisational roles



FAIR Data Principles: Metadata Management (GO FAIR recommendations)

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:

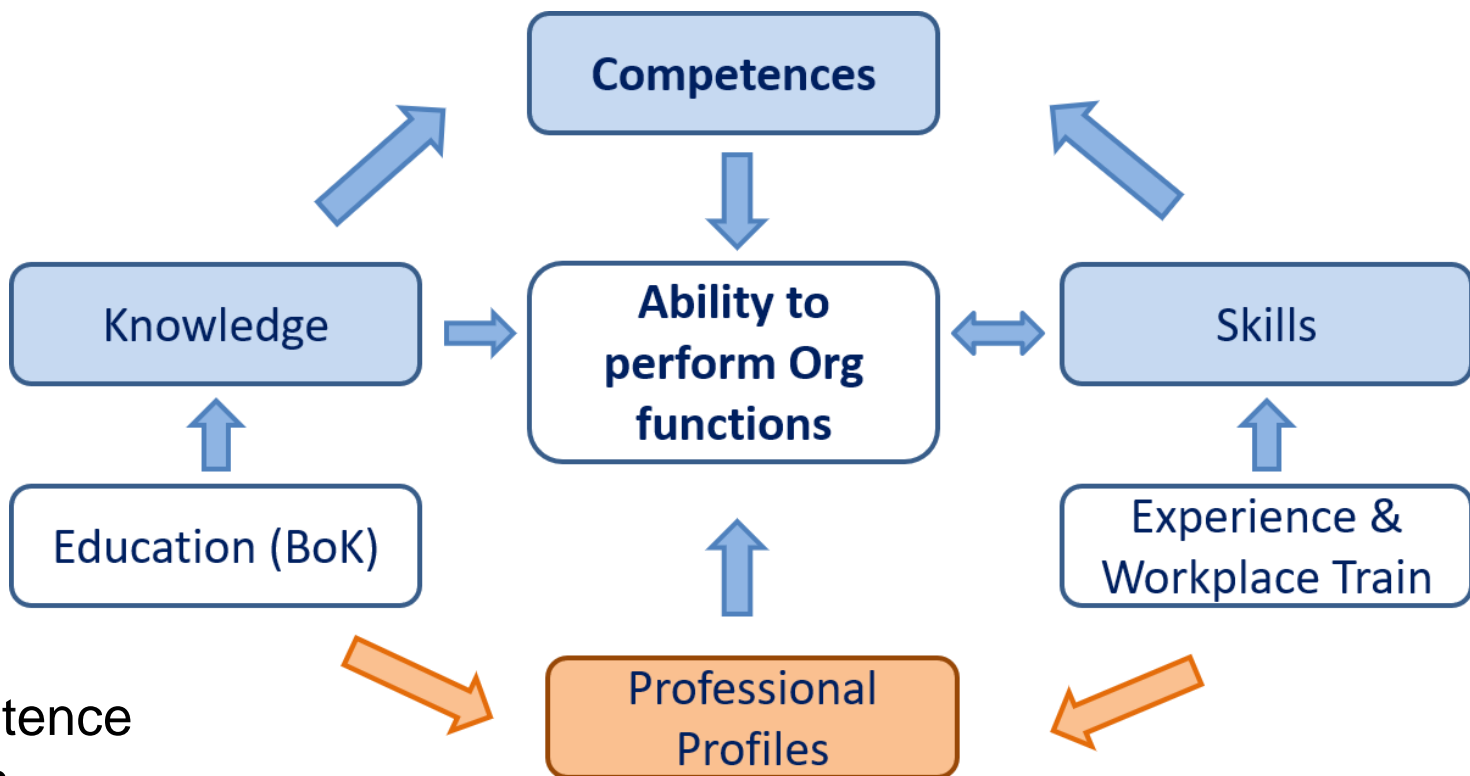
- A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
 - A1.1 the protocol is open, free, and universally implementable;
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary;
- A2 metadata are accessible, even when 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 standards;

Vacancies Analysis: Competences Map to Knowledge and Skills

- **Competence** is a demonstrated ability to apply knowledge, skills and attitudes for achieving observable results

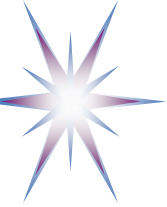


Competence
vs
Competency



FAIR adoption and Ecosystem Sustainability Elements

- FAIR must be accepted by all roles in organisational data management and governance process
 - FAIR must be endorsed by top management C-level
 - Roles and responsibilities to be defined and staffed
 - Inter-role functions as factor for modern agile organisations
- FAIR must be adopted for the whole Research/industrial Data lifecycle
- FAIR must be practiced by all participants along data lifecycle and specifically started from the data producers i.e. researchers or facility operators or sale agents
- FAIR must be supported by infrastructure and tools
- FAIR must be embedded into applications development
- Organisational capability and capacity management
- Education and training – To enable them all
 - Basic academic and professional education + continuous education



EDISON Project (2015-2017) and EDISON Data Science Framework (EDSF)

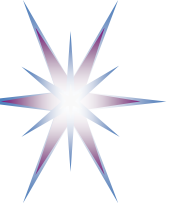
- EDISON project website - <http://edison-project.net/>
- EDISON Data Science Framework (EDSF) – main outcome of the project
- Currently maintained by EDISON Community Initiative, coordinated by UvA
- EDSF Release 3 published in 2018 – Currently active
- EDSF Release 4 Design Workshop – 20 Nov 2019, UvA
 - EDSF Release 4 (EDSF2020) to be published by the end of 2020 (initially planned January 2020)



FAIR Data Management and Organisational Roles

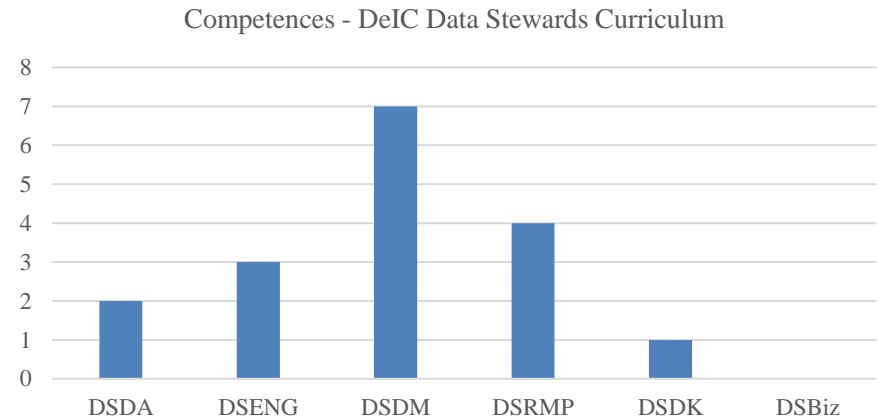
FAIR data principles to be adopted cross organisation for the whole data lifecycle

- Data collection
 - Researchers, Data Engineers, data entry workers
- Data preservation and curation
 - Data curators, Data Custodians/Archivists
- Data Analysis
 - Data Scientists, Data Architects, Application developers
- Data publication, sharing access
 - Data Stewards, Data Curators
- Data Governance and Data management
 - Data Stewards and CDO
 - Data policy and data delivery agreements
- Data Infrastructure and tools for data storage and handling
 - Storage, database engineers/managers
 - Metadata and PID services, Master data and Reference data



DeIC 2020 and DM Forum: Report “National Coordination of Data Steward Education in Denmark”

- Four roles for Data Stewards
 - Administrator
 - Analyst
 - Developer
 - Agent of change
- Competences defined: 6 competence groups, 22 competences
 - Open Science policies
 - Data management plans
 - Regulations, licenses
 - 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 (Including methodology)
 - Data archiving (finished data) and long-term storage
 - Data publishing
 - Scientific publishing / scholarly contribution
 - Open Access publishing



[ref] https://www.deic.dk/sites/default/files/Data%20Steward%20Education%20in%20Denmark_0.pdf



Course: Data Management and Governance (DMG) in Enterprise

- Introduction. Big Data Infrastructure and Data Management and Governance.
- Data Management concepts. Data management frameworks: DAMA Data Management framework, the Amsterdam Information Model. Extensions for Big Data and Data Science.
- Enterprise Data Architecture. Data Lifecycle Management and Service Delivery Model. Data management and data governance activities and roles.
- Data Science Professional profiles and organisational roles, Skills management and capacity building.
- Data Architecture, Data Modelling and Design. Data types and data models. Data modeling. Metadata. SQL and NoSQL databases overview. Distributed systems: CAP theorem, ACID and BASE properties.
- Enterprise Big Data infrastructure and integration with enterprise IT infrastructure. Data Warehouses. Distributed file systems and data storage.
- Big Data storage and platforms. Cloud based data storage services: data object storage, data blob storage, Data Lakes (services by AWS, Azure, GCP).
 - Trusted storage, blockchain enabled data provenance.
- FAIR data principles and Data Stewardship, Data Quality assessment and maturity model. Data repositories, Open Data services, public services.
- Big Data Security and Compliance. Data Privacy and GDPR. Data security and data protection. Security of outsourced data storage. Cloud security and compliance standards and cloud provider services assessment.