The Role of Dublin Core Metadata in the Expanding Digital and Analytical Skill Set Required by Data-Driven Organizations

Steve Brewer
Dublin Core Metadata Initiative – 12 July 2018
Infoculture Ltd
Outline of talk

• Introduction
• Context: digital transformation
• Data and metadata
• Data Science skills and competences: EDISON overview
• Conclusions and actions
Introduction

• World is changing
• Increasing dependency on data
• Data-driven transformations
• Skills and competences
• Compatibility and interoperability
EDISON Project value contribution and legacy:

Education and training for Data Science and data related competences

EDISON Data Science Framework (EDSF)

Yuri Demchenko, EDISON Project
University of Amsterdam

April 2018, Amsterdam
Outline of EDISON overview

- Background: Data driven research and demand for new skills
  - Foundation, recent reports, studies and facts
- EDISON Data Science Framework (EDSF)
  - Data Science competences and skills
  - Essential Data Scientist professional skills: Thinking and doing like Data Scientist
- Data Science Professional Profiles
- Data Science Body of Knowledge and Model Curriculum
- Use of EDSF and Example curricula
  - Competences assessment
  - Building Data Science team
- Roadmap recommendations
- References and additional materials

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
The Fourth Paradigm: Data-Intensive Scientific Discovery.

Riding the wave: How Europe can gain from the rising tide of scientific data.

The Data Harvest: How sharing research data can yield knowledge, jobs and growth. 
An RDA Europe Report. December 2014 

HLEG report on European Open Science Cloud 
(October 2016) 

Emergence of Cognitive Technologies 
(IBM Watson, Cortana and others)
Initiatives: GO FAIR and IFDS

• Global Open FAIR
  • Findable – Accessible – Interoperable - Reusable
• IFDS – Internet of FAIR Data and Services = EOSC
• GO FAIR implementation approach
  • GO-TRAIN: Training of data stewards capable of providing FAIR data services
  • FAIRdICT: Top Sector Health collaboration with top team ICT
• A critical success factor is availability of expertise in data stewardship
  • Training of a new generation of FAIR data experts is urgently needed to provide the necessary capacity

https://www.dtls.nl/fair-data/
https://www.dtls.nl/fair-data/go-fair/
Industry reports on Data Science Analytics and Data-enabled skills demand

• Final Report on European Data Market Study by IDC (Feb 2017)
  • The EU data market in 2016 estimated EUR 60 Bln (growth 9.5% from EUR 54.3 Bln in 2015)
    • Estimated EUR 106 Bln in 2020
    • Number of data workers 6.1 mln (2016) - increase 2.6% from 2015
    • Estimated EUR 10.4 million in 2020
  • Average number of data workers per company 9.5 - increase 4.4%
  • Gap between demand and supply estimated 769,000 (2020) or 9.8%

• PwC and BHEF report “Investing in America’s data science and analytics talent: The case for action” (April 2017)
  • 2.35 mln postings, 23% Data Scientist, 67% DSA enabled jobs
  • DSA enabled jobs growing at higher rate than main Data Science jobs

• Burning Glass Technology, IBM, and BHEF report “The Quant Crunch: How the demand for Data Science Skills is disrupting the job Market” (April 2017)
  • DSA enabled jobs takes 45-58 days to fill: 5 days longer than average
  • Commonly required work experience 3-5 yrs
PwC&BHEF: Skills that are tough to find

Faster growing jobs require both analytical and social skills


To be mapped to Competences, Knowledge, Skills and Personal (soft) Skills

EDISON 2017 Slide Deck
Challenge for Education: Sustainable ICT and Data Skills Development

- **Educate vs Train**
  - Training is a short term solution
  - Education is a basis for sustainable skills development

- **Technology focus changes every 3-4 years**
  - Study: 50% of academic curricula are outdated at the time of graduation

- **Lack of necessary skills leads to underperforming projects and organisations and lose of competitiveness**
  - Challenge: Policy and decision makers still don’t include planning human factor (competences and skills) as a part of the technology strategy

- **Need to change the whole skills management paradigm**
  - **Dynamic (self-) re-skilling**: Continuous professional development and **shared responsibility between employer and employee**
  - Professional and workplace skills and career management as a part of professional orientation

- **Millennials factor and changing nature of workforce**
• EDISON Data Science Framework (EDSF)
  • Compliant with EU standards on competences and professional occupations e-CFv3.0, ESCO
  • Customisable courses design for targeted education and training

• Skills development and career management for Core Data Experts and related data handling professions

• Capacity building and Data Science team design

• Academic programmes and professional training courses (self) assessment and design

• EU network of Champion universities pioneering Data Science academic programmes

• Engagement in relevant RDA activities and groups

• Cooperation with International professional organisations IEEE, ACM, BHEF, APEC (AP Economic Cooperation)
EDISON Data Science Framework (EDSF)

EDISON Framework components
- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- DSP – Data Science Professional profiles
- Data Science Taxonomies and Scientific Disciplines Classification
- EOEE - EDISON Online Education Environment

Methodology
- ESDF development based on job market study, existing practices in academic, research and industry.
- Review and feedback from the ELG, expert community, domain experts.
- Input from the champion universities and community of practice.
What challenges related to skills management the EDSF can help to address?

1. Guide researchers in using right methods and tools, latest Data Analytics technologies to extracting value from scientific data
2. Educate and train RI engineers dev to build modern data intensive research infrastructure and understand trends and project for future
3. Develop new data analytics tools and ensure continuous improvement (agile model, DevOps)
4. Correctly organise and manage data, make them accessible (adhering FAIR principles), education new profession of Data Stewards
5. Help managers to facilitate career dev for researchers and organise effective teams
6. Ensure skills and expertise sustain in organisation
7. Help research institutions to sustain in competition with industry and business in data science talent hunting
Competences Map to Knowledge and Skills

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

Based on the definitions by NIST SP1500 – 2015, extended by EDISON

- **A Data Scientist is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle till the delivery of an expected scientific and business value to organisation or project.**

- Core Data Science competences and skills groups
  - **Data Science Analytics** (including Statistical Analysis, Machine Learning, Business Analytics)
  - **Data Science Engineering** (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
  - **Domain Knowledge and Expertise** (Subject/Scientific domain related)

- EDISON identified 2 additional competence groups demanded by organisations
  - Data Management, Data Governance, Stewardship, Curation, Preservation
  - Research Methods and vs Business Processes/Operations

- **Data Science professional skills:** Thinking and acting like Data Scientist – required to successfully develop as a Data Scientist and work in Data Science teams
Data Science Competence Groups - Research

Data Science Competences include 5 groups
- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Research Methods and Project Management
  - Business Process Management (biz)

Scientific Methods
- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesis Explanation
- Test Hypothesis

Business Operations
- Operations Strategy
- Plan
- Design & Deploy
- Monitor & Control
- Improve & Re-design
Data Science Competences Groups – Business

Data Science Competences include 5 groups:
- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Research Methods and Project Management
  - Business Process Management (biz)

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

Business Process Operations/Stages
- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design
**Identified Data Science Competence Groups**

<table>
<thead>
<tr>
<th>Data Science Analytics (DSDA)</th>
<th>Data Science Engineering (DSENG)</th>
<th>Data Management and Governance (DSDM)</th>
<th>Research/Scientific Methods and Project Management (DSRMP)</th>
<th>Data Science Domain Knowledge, e.g. Business Analytics (DSDK/DSBPM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td><strong>Use appropriate data analytics and statistical techniques on available data to deliver insights into research problem or org. processes and support decision making</strong></td>
<td>Use engineering principles and modern computer technology to research, design, implement new data analytics applications, develop experiments, processes, instruments, systems and infrastructures to support data handling during the whole data lifecycle</td>
<td>Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.</td>
<td>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</td>
</tr>
<tr>
<td>1</td>
<td><strong>DSDA01</strong> Effectively use variety of data analytics techniques</td>
<td><strong>DSENG01</strong> Use engineering principles (general and software) to research, design, develop and implement new instruments and applications</td>
<td><strong>DSDM01</strong> Develop and implement data strategy, in particular, Data Management Plan (DMP)</td>
<td><strong>DSRMP01</strong> Create new understandings and capabilities by using scientific/research methods</td>
</tr>
<tr>
<td>2</td>
<td><strong>DSDA02</strong> Apply designated quantitative techniques</td>
<td><strong>DSENG02</strong> Develop and apply computer methods to domain related problems</td>
<td><strong>DSDM02</strong> Develop data models including metadata</td>
<td><strong>DSRMP02</strong> Direct systematic study toward a fuller knowledge or understanding of the observable facts</td>
</tr>
<tr>
<td>3</td>
<td><strong>DSDA03</strong> Pull together data from diff sources ...</td>
<td><strong>DSENG03</strong> Develop and prototype data analytics applications</td>
<td><strong>DSDM03</strong> Collect integrate data</td>
<td><strong>DSRMP03</strong> Undertakes creative work</td>
</tr>
<tr>
<td>4</td>
<td><strong>DSDA04</strong> Use diff perform techniques</td>
<td><strong>DSENG04</strong> Develop, deploy operate Big Data storage</td>
<td><strong>DSDM04</strong> Maintain repository</td>
<td><strong>DSRMP04</strong> Translate strategies into actions</td>
</tr>
<tr>
<td>5</td>
<td><strong>DSDA05</strong> Develop analytics applic</td>
<td><strong>DSENG05</strong> Apply security mechanisms</td>
<td><strong>DSDM05</strong> Visualise cmplx data</td>
<td><strong>DSRMP05</strong> Contribute to organis goals</td>
</tr>
<tr>
<td>6</td>
<td><strong>DSDA06</strong> Visualise results of analysis, dashboards</td>
<td><strong>DSENG06</strong> Design, build, operate SQL and NoSQL</td>
<td><strong>DSRMP06</strong> Develop and manage policies</td>
<td><strong>DSRMP06</strong> Develop and guide data driven projects</td>
</tr>
</tbody>
</table>
Identified Data Science Skills/Experience Groups

Skills Type A – Based on knowledge acquired

• **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods or Business Process Management
  - Application/subject domain related (research or business)

• **Group 2: Mathematics and statistics**
  - Mathematics and Statistics and others

Skills Type B – Base on practical or workplace experience

• **Group 3: Big Data (Data Science) tools and platforms**
  - Big Data Analytics platforms
  - Mathematics & Statistics applications & tools
  - Databases (SQL and NoSQL)
  - Data Management and Curation platform
  - Data and applications visualisation
  - *Cloud based platforms and tools*

• **Group 4: Data analytics programming languages and IDE**
  - General and specialized development platforms for data analysis and statistics

• **Group 5: Soft skills and Workplace skills**
  - Data Science professional skills: Thinking and Acting like Data Scientist
  - 21st Century Skills: Personal, inter-personal communication, team work, professional network
Data Science Professional Skills: Thinking and Acting like Data Scientist (1)

1. Recognise value of data, work with raw data, exercise good data intuition, use SN and Open Data
2. 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)
3. Good sense of metrics, understand importance of the results validation, never stop looking at individual examples
4. Ask the right questions
5. Respect domain/subject matter knowledge in the area of data science
6. Data driven problem solver and impact-driven mindset
7. Be aware about power and limitations of the main machine learning and data analytics algorithms and tools
8. 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 variables and factors
Data Science Professional Skills: Thinking and Acting like Data Scientist (2)

9. Recognise what things are **important** and what things are **not important** (in data modeling)
10. Working in **agile environment** and coordinate with other roles and team members
11. Work in **multi-disciplinary team**, ability to communicate with the domain and subject matter experts
12. Embrace **online learning**, continuously improve your knowledge, use **professional networks** and communities
13. **Story Telling**: Deliver actionable result of your analysis
14. **Attitude**: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
15. **Ethics and responsible use** of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)
21st Century Skills (DARE & BHEF & EDISON)

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

• **Subject domain components**
  - Model (and data types)
  - Methods
  - Processes
  - Domain specific data and presentation/visualization methods
  - Organisational roles and relations

• **Data Scientist is an assistant to Subject Domain Specialists**
  - Translate subject domain Model, Methods, Processes into abstract data driven form
  - Implement computational models in software, build required infrastructure and tools
  - Do (computational) analytic work and present it in a form understandable to subject domain
  - Discover new relations originated from data analysis and advice subject domain specialist
  - Present/visualise information in domain related actionable way
  - Interact and cooperate with different organizational roles to obtain data and deliver results and/or actionable data
Data Science and Subject Domains

Data Science domain components
- Data structures & databases/storage
- Visualisation
- Abstract data driven math & compute models
- Data Analytics methods
- Data and Applications Lifecycle Management
- Cross-organisational assistive role

Domain specific components
- Domain specific data & presentation (visualization)
- Models (and data types)
- Methods
- Processes
- Organisational roles

Data Scientist/Data Steward functions is to translate between two domains

Data Scientist role is to maintain the Data Value Chain (domain specific):
- Data Integration => Organisation/Process/Business Optimisation => Innovation
Practical Application of the CF-DS

• Basis for the definition of the Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS)
  • CF-DS => Learning Outcomes (MC-DS) => Knowledge Areas (DS-BoK)
  • CF-DS => Data Science taxonomy of scientific subjects and vocabulary

• Data Science professional profiles definition
  • Extend existing EU standards and occupations taxonomies: e-CFv3.0, ESCO, others

• Professional competence benchmarking
  • For customizable training and career development
  • Including CV or organisational profiles matching

• Professional certification
  • In combination with DS-BoK professional competences benchmarking

• Vacancy construction tool for job advertisement (for HR)
  • Using controlled vocabulary and Data Science Taxonomy
Data Science Professions Family

**Managers:** Chief Data Officer (CDO), Data Science (group/dept) manager, Data Science Infrastructure manager, Research Infrastructure manager

**Professionals:** Data Scientist, Data Science Researcher, Data Science Architect, Data Science (applications) programmer/engineer, Data Analyst, Business Analyst, etc.

**Professional (database):** Large scale (cloud) database designers and administrators, scientific database designers and administrators

**Professional (data handling/management):** Data Stewards, Digital Data Curator, Data Librarians, Data Archivists

**Technicians and associate professionals:** Big Data facilities operators, scientific database/infrastructure operators

**Support workers and data handling clerks:** User support workers, data entry clerks, data entry field workers

Icons used: Credit to [ref] https://www.datacamp.com/community/tutorials/data-science-industry-infographic
DSP Profiles mapping to ESCO Taxonomy High Level Groups

- DSP Profiles mapping to corresponding CF-DS Competence Groups
- Relevance level from 5 – maximum to 1 – minimum

<table>
<thead>
<tr>
<th>DSP Profiles</th>
<th>CF-DS Competence Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSP01 Data Science (group) Manager</td>
<td>3 4 3 3 2</td>
</tr>
<tr>
<td>DSP02 Data Science Infrastructure Manager</td>
<td>2 4 4 2 2</td>
</tr>
<tr>
<td>DSP03 Research Infrastructure Manager</td>
<td>2 4 4 3 2</td>
</tr>
</tbody>
</table>

**Data Science Professionals**

<table>
<thead>
<tr>
<th>Profile ID</th>
<th>Data Science Profile title</th>
<th>DSPA</th>
<th>DSPB</th>
<th>DSPC</th>
<th>DSPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSP04 Data Scientist</td>
<td>6 3 3 1 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP05 Data Science Researcher</td>
<td>4 3 2 5 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP06 Data Science Architect</td>
<td>4 3 5 3 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP07 Data Science Applic Programmer</td>
<td>4 2 5 3 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP08 Data Analyst</td>
<td>5 3 3 3 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP09 Business Analyst</td>
<td>5 3 3 4 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Science technology professionals**

<table>
<thead>
<tr>
<th>Profile ID</th>
<th>Data Science Profile title</th>
<th>DSPA</th>
<th>DSPB</th>
<th>DSPC</th>
<th>DSPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSP10 Data Stewards</td>
<td>3 3 3 3 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP11 Digital data curator</td>
<td>1 5 2 2 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP12 Digital Librarians</td>
<td>2 5 2 2 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP13 Data Archivists</td>
<td>1 5 1 1 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Database and network professionals**

<table>
<thead>
<tr>
<th>Profile ID</th>
<th>Data Science Profile title</th>
<th>DSPA</th>
<th>DSPB</th>
<th>DSPC</th>
<th>DSPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSP14 Large scale database designer</td>
<td>2 4 4 3 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP15 Large scale database admin</td>
<td>2 4 3 2 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP16 Scientific database administrator</td>
<td>2 4 3 2 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data handling and support workers**

<table>
<thead>
<tr>
<th>Profile ID</th>
<th>Data Science Profile title</th>
<th>DSPA</th>
<th>DSPB</th>
<th>DSPC</th>
<th>DSPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSP17 Big Data facilities Operator</td>
<td>1 4 4 2 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP18 Large scale data storage operator</td>
<td>1 4 3 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP19 Scientific database operator</td>
<td>1 4 3 2 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data and information entry and access**

<table>
<thead>
<tr>
<th>Profile ID</th>
<th>Data Science Profile title</th>
<th>DSPA</th>
<th>DSPB</th>
<th>DSPC</th>
<th>DSPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSP20 Data entry/access worker</td>
<td>2 1 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP21 Data entry field workers</td>
<td>2 1 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP22 User support data services</td>
<td>3 2 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CF-DS and Data Science Professional Profiles
### Example DS Professional Profile Definition (compliant with CWA)

<table>
<thead>
<tr>
<th>Profile title</th>
<th>Gives a commonly used name to a profile.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary statement</strong></td>
<td>Indicates the main purpose of the profile.</td>
</tr>
<tr>
<td></td>
<td>The purpose is to present to stakeholders and users a brief, concise understanding of the specified ICT Profile. It should be understandable by ICT professionals, ICT managers and Human Resource personnel. It should provide a statement of the job’s main activity.</td>
</tr>
<tr>
<td><strong>Mission</strong></td>
<td>Describes the rationale of the profile.</td>
</tr>
<tr>
<td></td>
<td>The purpose is to specify the designated job role defined in the ICT Profile.</td>
</tr>
<tr>
<td><strong>Deliverables</strong></td>
<td>Accountable (A)</td>
</tr>
<tr>
<td><strong>Main task/s</strong></td>
<td>Specifies the Profile by key deliverables.</td>
</tr>
<tr>
<td></td>
<td>The purpose is to illuminate the ICT Profiles and to explain relevance including the perspective from a non-ICT point of view.</td>
</tr>
<tr>
<td><strong>e-CF competences assigned</strong></td>
<td>Provides a list of necessary competences (from the e-CF) to carry out the mission.</td>
</tr>
<tr>
<td></td>
<td>Must include 1 up to 5 competences.</td>
</tr>
<tr>
<td></td>
<td>Level assignment is important. Can be (usually) 1 or (maximum) 2 levels.</td>
</tr>
<tr>
<td><strong>KPI Area</strong></td>
<td>Based upon KPIs (Key Performance Indicators) KPI area is a more generic indicator, congruent with the overall profile granularity level. It is deployed to add depth to the mission.</td>
</tr>
<tr>
<td></td>
<td>Not prescriptive. Non-specific measurements. Use general examples.</td>
</tr>
<tr>
<td></td>
<td>The principle is to provide KPI areas (which are stable, general and long lasting) providing users with an inspiration to enable development of specific KPI’s for specific roles</td>
</tr>
<tr>
<td></td>
<td>Must be related to the key deliverables in order to measure them.</td>
</tr>
</tbody>
</table>
EDSF for Education and Training

• Foundation and methodological base
  • Data Science Body of Knowledge (DS-BoK)
    • Taxonomy and classification of Data Science related scientific subjects
  • Data Science Model Curriculum (MC-DS)
    • Set Learning Units mapped to CF-DS Learning and DS-BoK Knowledge Areas/Units
    • Instructional methodologies and teaching models

• Platforms and environment
  • Virtual labs, datasets, developments platforms
  • Online education environment and courses management

• Services
  • Individual benchmarking and profiling tools (competence assessment)
  • Knowledge evaluation tools
  • Certifications and training for self-made Data Scientists practitioners
  • Education and training marketplace: Courses catalog and repository
Data Science Body of Knowledge (DS-BoK)

**DS-BoK Knowledge Area Groups (KAG)**

- **KAG1-DSA**: 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
Data Science Model Curriculum (MC-DS)

Data Science Model Curriculum includes

• Learning Outcomes (LO) definition based on CF-DS
  • LOs are defined for CF-DS competence groups and for all enumerated competences
  • Knowledge levels: Familiarity, Usage, Assessment (based in Bloom’s Taxonomy)

• LOs mapping to Learning Units (LU)
  • LUs are based on CCS(2012) and universities best practices
  • Data Science university programmes and courses inventory (interactive)
    http://edison-project.eu/university-programs-list

• LU/course relevance: Mandatory Tier 1, Tier 2, Elective, Prerequisite

• Learning methods and learning models (in progress)
Learning methods and learning models

- Bloom’s Taxonomy and Cognitive learning activities
  - BT application areas and limitations
- Constructive Alignment and Intended Learning Outcome (ILO)
  - ILO is formulated from the student perspective
  - Outcome Based Learning (OBL)
- Other education technologies for teaching in fast technology changing world
  - Project Based Learning (PBL)
  - Flipped classroom
  - Activating teaching and activating strategies
Bloom’s Taxonomy and Knowledge Levels for MC-DS

<table>
<thead>
<tr>
<th>Level</th>
<th>Action Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiarity</td>
<td>Choose, Classify, Collect, Compare, Configure, Contrast, Define, Demonstrate, Describe, Execute, Explain, Find, Identify, Illustrate, Label, List, Match, Name, Omit, Operate, Outline, Recall, Rephrase, Show, Summarize, Tell, Translate</td>
</tr>
<tr>
<td>Usage</td>
<td>Apply, Analyze, Build, Construct, Develop, Examine, Experiment with, Identify, Infer, Inspect, Model, Motivate, Organize, Select, Simplify, Solve, Survey, Test for, Visualize</td>
</tr>
<tr>
<td>Assessment</td>
<td>Adapt, Assess, Change, Combine, Compile, Compose, Conclude, Criticize, Create, Decide, Deduct, Defend, Design, Discuss, Determine, Disprove, Evaluate, Imagine, Improve, Influence, Invent, Judge, Justify, Optimize, Plan, Predict, Prioritize, Prove, Rate, Recommend, Solve</td>
</tr>
</tbody>
</table>
Data Science Model Curriculum (MC-DS)

Data Science Model Curriculum includes

- Learning Outcomes (LO) definition based on CF-DS
  - LOs are defined for CF-DS competence groups and for all enumerated competences
  - Knowledge levels: Familiarity, Usage, Assessment (based in Bloom’s Taxonomy)

- LOs mapping to Learning Units (LU)
  - LUs are based on CCS(2012) and universities best practices
  - Data Science university programmes and courses inventory (interactive)
    http://edison-project.eu/university-programs-list

- LU/course relevance: Mandatory Tier 1, Tier 2, Elective, Prerequisite

- Learning methods and learning models (in progress)
Data Science Engineering (KAG2-DSENG)

- KA02.01 (DSENG/BDI) Big Data infrastructure and technologies, including NOSQL databases, platforms for Big Data deployment and technologies for large-scale storage;
- KA02.02 (DSENG/DSIAPP) Infrastructure and platforms for Data Science applications, including typical frameworks such as Spark and Hadoop, data processing models and consideration of common data inputs at scale;
- KA02.03 (DSENG/CCT) Cloud Computing technologies for Big Data and Data Analytics;
- KA02.04 (DSENG/SEC) Data and Applications security, accountability, certification, and compliance;
- KA02.05 (DSENG/BDSE) Big Data systems organization and engineering, including approaches to big data analysis and common MapReduce algorithms;
- KA02.06 (DSENG/DSAPPD) Data Science (Big Data) application design, including languages for big data (Python, R), tools and models for data presentation and visualization;
- KA02.07 (DSENG/IS) Information Systems, to support data-driven decision making, with focus on data warehouse and data centers.
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*
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, metadata, data registries
13. Data Management Plan
14. Open Science, Open Data, Open Access, ORCID
15. Responsible data use

• Highlighted in red: Considered (Research) Data Management literacy (minimum required knowledge)
Outcome Based Educations and Training Model

Data Science Competence Framework (CF-DS) → DS Professional Profiles (DSP-P) → Learning Outcomes (LO) (OBE Learning Model) → Knowledge Units (KU) (DS-BoK) → Learning Units (LU) (MC-DS) → Tracks/Specialisations (based on DSP-P) (LO, LU, KU, courses/modules)

- EDSF allow for customized educational courses and training modules design

From Competences and DSP Profiles to Learning Outcomes (LO) and to Knowledge Unites (KU) and Learning Units (LU)
Individual Education/Training Path based on Competence benchmarking

- Red polygon indicates the chosen professional profile: Data Scientist (general)
- Green polygon indicates the candidate or practitioner competences/skills profile
- Insufficient competences (gaps) are highlighted in red
  - DSDA01 – DSDA06 Data Science Analytics
  - DSRM01 – DSRM05 Data Science Research Methods
- Can be use for team skills match marking and organisational skills management

[ref] For DSP Profiles definition and for enumerated competences refer to EDSF documents CF-DS and DSP Profiles.
Building a Data Science Team

- Data Science Group Manager, Data Science Architect
- Data Steward
- Data Engineer, Database Developer
- Data Analyst/Business Analyst
- Data Science Applications Developer
- Data Scientist
- Data Science Researcher
- Business Analyst
- Data Steward
- Data Entry/Support
- Researcher (Scientific domain)
- Data Facilities Operator
- Data Collection
- Data Ingest
- Data Analysis
- Results, Actionable Data
- Data Visualisation, Reporting, Storage
- Data Source (Experiment, Data Driven Application)
EDSF Data Model and API

- EDSF API provides access to all EDSF functionality
DSP04 – Data Scientist MC structure
DSP10 – Data Steward MC structure

DSP10 - Data Steward

[Diagram showing DSP10 data stewardship structure with color-coded sections for DSDA, DSDM, DSENG, DSRMP]
Roadmap recommendations: Data Science Education and Training for Europe

A. Policy recommendations for the EC and Member States
   • R1. Critical skills management and training
   • R2. Gender balance and multi-cultural environment
   • R3. Common data literacy
   • R4. Data driven technology and education divide.
   • R5. European studies on demand and role of Data Science and Analytics skills
   • R6. Include the above actions in the European Digital Single Market Scoreboard

B. Recommendations to universities and professional training organisations
   • R7. EDSF adoption by universities
   • R8. Addressing Data Science professional and workplace skills in university curricula
   • R9. Multiple delivery form
   • R10. Sharing experience, courses and instructors
   • R11. Supporting technical infrastructure for Data Science and data related education and professional training

C. Recommendations for Research (e) Infrastructures
   • R12. Critical skills management in Research (e) Infrastructures

D. Required support and contribution to the standardisation bodies
   • R13. The following are essential measures to achieve EDSF support existing and new standards.
Next steps (1) – Further development and exploitation

• EDSF Release 3 (End 2017)
  • Fully enumerated CF, BoK, MC, DSPP in machine readable format

• Development of EDSF based tools:
  • Self assessment and Market monitoring

• Certification Framework for at least two levels of Data Science competences proficiency: Associates and Professionals
  • Data Science knowledge and competences for decision makers

• Toward EDSF and Data Science profession standardisation
  • ESCO (European Skills, Competences and Occupations) taxonomy
  • CEN TC428 (European std body) – Extending current eCFv3.0 and ICT profiles towards e-CF4 with Data Science related competences
  • Work with the IEEE and ACM curriculum workshop to define Data Science Curriculum and extend current CCS2012 (Classification Computer Science 2012)
Next steps (2) – Community Activities and Initiatives

- **Data Science Manifesto** – Primarily focused on professional and ethical issues in Data Science, new type of professional
  - Professional and ethical issues as a primary focus

- Inter-universities initiative “Data Science for UN’s Sustainable Development Goals” to focus in-curricula research (projects) on UN priority goals

- Build a wider network of early adopters, champions and ambassadors in Europe and worldwide
  - EDISON team information and advice support will continue through out 2017-2018
Ongoing activities and developments

https://github.com/EDISONcommunity/EDSF

• EDISON Community Initiative and Call for contribution EDSF Release 3
  • Call for comments and contribution - Deadline 30 June 2018
  • Editorial team work June – July 2018
  • Target publication EDSF Release 3 – end July 2018
  • Call for sponsorship

• Industry digitalisation projects and data literacy skills training
  • MATES project funded by EU ERASMUS Programme - EU Maritime industry digital transformation, data skills development and Data + Ocean literacy
  • Port Rotterdam Data Management training for data driven digital transformation
  • Skills for SME: Data Science, IoT, Cybersecurity: Prepare to Data Economy and Industry 4.0

• DARE project by APEC (Asia Pacific Economic Cooperation)
  • Continuing cooperation for Asia Pacific region
  • Recommended Data Science and Analytics Competences published August 2017 - https://www.apec.org/Press/Features/2017/0620_DSA
Further development and exploitation

- Development of EDSF based tools
  - Competences benchmarking and Job market monitoring

- Inter-universities initiative “Data Science for UN’s Sustainable Development Goals” to focus in-curricula research (projects) on UN priority goals

- Toward EDSF and Data Science profession standardisation
  - ESCO (European Skills, Competences and Occupations) taxonomy
  - CEN TC428 (European std body) – Extending current eCFv3.0 and ICT profiles towards e-CF4 with Data Science related competences
  - Work with the IEEE and ACM curriculum workshop to define Data Science Curriculum and extend current CCS2012 (Classification Computer Science 2012)

- Data Science Manifesto – Primarily focused on professional and ethical issues in Data Science, new type of professional
  - Professional and ethical issues as a primary focus
Conclusion and actions

• More work needed to develop the third release of the EDISON Data Science Framework (EDSF release 3)
• More work needed to operationalise the Framework for professionals, employers and universities and other learning and training organisations
• Interesting to explore in practice how Dublin Core could be of benefit in new business sectors to expand the data-driven innovation in the digital economy
Thank you

• Information points:
  • EDSF github project - https://github.com/EDISONcommunity/EDSF
    • Component documents CF-DS, DS-BoK, MC-DS, DSPP
  • EDISON Community work area and discussions - https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome
  • Mailing list - edison-net@list.uva.nl
  • EDISON project website (still active) http://edison-project.eu/
    • EDISON Data Science Framework Release 2 (EDSF), 3 July 2017 http://edison-project.eu/edison-data-science-framework-edsf
  • Infoculture: http://infoculture-lab.com/
  • Steve Brewer: hello@infoculture-lab.com