DOI: https://doi.org/10.14324/000.learn.21

Case Study 20

The Emerging Role of the Data Scientist and the experience of Data Science education at the University of Amsterdam

Author: Yuri Demchenko (Senior Researcher, System and Network Engineering Research Group. University of Amsterdam)
Email: y.demchenko@uva.nl

20.1. MOTIVATION AND BACKGROUND

Modern research requires new types of specialists that are capable of supporting all stages of the research data lifecycle – from data production and input to data processing, storage, and the publishing and dissemination of scientific results, which can be jointly defined as key components of the emerging profession of Data Science (DSP).

To address this demand from research and industry, the Horizon 2020 Programme is funding the EDISON Project (Grant 675419, INFRASUPP-4-2015: CSA),[1] the goal of which is to build the Data Science profession for European research and industry. This includes the definition of Data Science and data handling-related professional profiles (or occupations), corresponding core competences and skills, the Data Science Body of Knowledge and a Model Curriculum that together comprise the EDISON Data Science Framework. This work is done with the involvement of the main stakeholders from the research community, industry, data preservation and handling community, universities and professional training organisations.

The University of Amsterdam is coordinator and a base organisation for the EDISON Project; other partners include the University of Stavanger (Norway), the University of Southampton (UK), Engineering Italy, EGI.eu, FTK (Germany), and Inmark Europe (Spain). The project benefits from multiple Data Science-related initiatives and academic activity and effective cooperation between Computer Science and multi-disciplinary departments, University Library and IT departments. It is also supported by such external initiatives as the Amsterdam Data Science Centre and Amsterdam School of Data Science (ASDS). On the other hand, all project recommendations find their practical pilot implementation at the University of Amsterdam and in cooperating organisations. This includes four Data Science and Big Data programmes, Research Data Management (RDM) training (together with the University Library), training for researchers, programmes and course catalogue services for universities and students, and advice for companies.

20.2. STAKEHOLDERS AND THEIR ROLE IN DATA SCIENCE EDUCATION

To create a foundation for the sustainable education and training of future Data Science professionals and Core Data Experts to support present and future data-driven research, the EDISON Project involves and is cooperating with multiple stakeholders, relevant bodies and communities. This includes but is not limited to the following:

- **Academic and research departments** are key for developing and teaching educational courses on Data Science and Research Data Management: four different Data Science programmes have started in the 2016-17 academic year, targeting different demand sectors in research and industry

[1] EDISON: http://edison-project.eu/; accessed 5 February 2017.

(see below for a description of the programmes). Course development, teaching and support is provided primarily by departmental staff with some facility services maintained by ICT departments.

- **The University Library** is involved in two main activities: (i) it provides basic training for researchers and contributes to the more general academic education for students in RDM; (ii) it cooperates with the ICT department in developing and implementing university-wide RDM services, infrastructure and policy.
- **The ICT department** supports Data Science education by providing and maintaining HPC facilities and services. The ICT department cooperates with the University Library in implementing RDM infrastructure and policy university-wide.

20.3. EDISON DATA SCIENCE FRAMEWORK

The EDISON Data Science Framework (EDSF),[2] is a core product of the EDISON Project that provides a basis for the definition of the whole ecosystem for education, training and professional development in core Data Science and Data Management-related competences and skills. An important component of EDSF is the Data Science professional family that provides a basis for defining customisable educational and training programmes for different target professional groups. Figure 20.1 below illustrates the main EDSF 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 and occupations taxonomy
- Data Science Taxonomy and Scientific Disciplines Classification (including Vocabulary)

The proposed framework provides a basis for other components of the Data Science professional ecosystem:

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

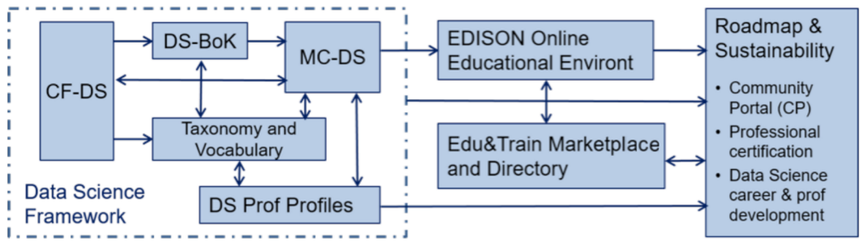

Figure 20.1 EDISON Data Science Framework components.

[2] EDISON: http://edison-project.eu/edison/edison-data-science-framework-edsf; accessed 5 February 2017.

The CF-DS includes common competences required for the successful work of Data Scientists in different work environments in industry and in research and throughout the whole career path. Future CF-DS development will include coverage of domain-specific competences and skills and will involve domain and subject matter experts.

The DS-BoK defines the Knowledge Areas and Knowledge Units for building Data Science curricula that are required to support specified Data Science competences. DS-BoK is organised by Knowledge Area Groups (KAG) that correspond to the CF-DS competence groups. DS-BoK incorporates best practices in Computer Science and domain-specific BoK's and includes KAs based on the Computing Classification System (CCS2012), components taken from other BoKs and proposed new KAs to incorporate new technologies used in Data Science and their recent developments.

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

The DSP profiles and Data Science occupations taxonomy are defined based on, and as an extension to, the European Skills, Competences, Qualifications and Occupations (ESCO) framework. The DSP profiles definition provides an instrument to create effective organisational structures and corresponding roles to support the whole data management lifecycle. For example, in the area of professional data handling/management, the following taxonomy is proposed: Professional (data handling/management): Data Stewards, Digital Data Curator, Digital Librarians, Data Archivists. DSP can also be used for building individual career paths and corresponding competences and skills transferability between organisations and sectors of the economy.

20.3.1. Data Science Competence Framework (CF-DS)

The Data Science Competence Framework (CF-DS)[3] has been built based on an extensive study of the demand and supply side of the Data Science job market, organisational structures and roles as well as existing practices and standards in the area of competences and skills management. The figure below [20.2] presents the following competences:

Three competence groups identified in the NIST document and confirmed by the analysis of collected data:

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Engineering: software and infrastructure
- Subject/Scientific Domain competences and knowledge

Two newly identified competence groups that are in high demanded and are specific to Data Science

- Data Management, Curation, Preservation (new)
- Scientific or Research Methods (new)

[3] EDISON: http://edison-project.eu/data-science-competence-framework-cf-ds; last accessed 9 February 2017.

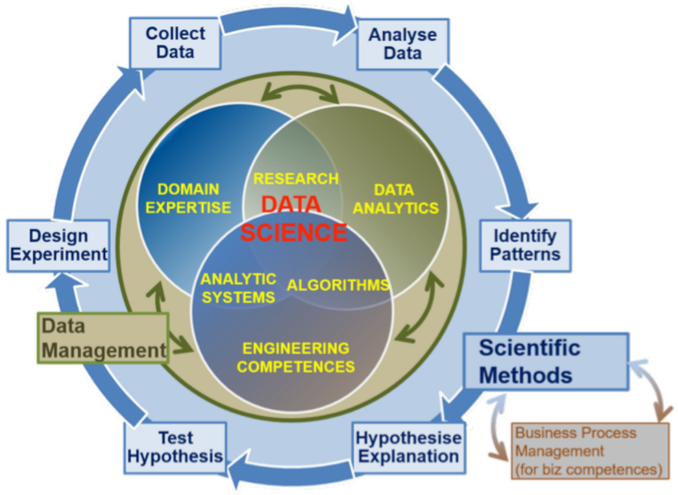

Figure 20.2. Data Science competence groups

Knowledge of scientific research methods and techniques makes the Data Scientist profession different from all previous professions. For business-related professions, a similar role belongs to business process management in areas that need to be adapted to a new data-driven agile business model, in particular, to adopt continuous data-driven business processes improvement.

Data management, curation and preservation are already included in existing (research) data-related professions such as data steward, data archivist, data manager, digital librarian, data curator, and others. Research data management is an important component of European Research Area policy. Companies also recognise the need for data management skills when they start using data-driven technologies.

The identified demand for general competences and knowledge of Data Management and Research Methods needs to be addressed in future Data Science education and training programmes, as well as being included in re-skilling training programmes. It is important to mention that knowledge of Research Methods does not mean that all Data Scientists must be talented scientists; however, they need to understand general research methods such as formulating an hypothesis, applying research methods, producing artefacts, and evaluating an hypothesis (so called 4 steps model). Research Methods training is already included into Masters programmes and for graduate students.

The identified competence areas provide a basis for defining education and training programmes for Data Science-related jobs, re-skilling and professional certification.

Other skills commonly recognised are referred to as “soft skills” or “social/professional intelligence”: inter-personal skills or team work, the ability to cooperate. In many cases, an organisation expects the Data Scientist to provide a kind of literacy advice and guidance on related data analysis and management technologies.

20.3.2. Data Science Body of Knowledge (DS-BoK)

The DS-BoK should contain the following Knowledge Area groups (KAG) that are defined after CF-DS competence groups:

- KAG1-DSDA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSENG: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: Data Management group including data curation, preservation and data infrastructure
- KAG4-DSRM: Scientific or Research Methods group
- KAG5-DSBPM: Business process management group
- KAG6-DSDKX: Data Science Domain Knowledge group, which includes domain-specific knowledge

Universities can use DS-BoK as a reference to define knowledge areas that they need to cover in their programmes depending on the primary demand groups in research or industry. Domain-specific knowledge can be acquired as a part of academic education or as postgraduate professional training at the graduate’s work place. It is also commonly recognized that KAG6-DSDKX is essential for the practical work of a Data Scientist, which means that Data Scientists need to have sufficient understanding of specific subject domain-related concepts, models, organisation and corresponding data analysis methods to effectively communicate with domain-related specialists for data collection, insight and the presentation of results.

20.3.3. Data Science Model Curriculum (MC-DS)

The initial Data Science Model curriculum provides two basic components for building customisable Data Science curricula: (1) the definition of a learning outcomes (LO) based on the CF-DS competences, including their differentiation for different proficiency levels, e.g. using Bloom’s Taxonomy, (2) definition of the Learning Units (LU) that map to the LOs for target professional groups, which need to be defined in accordance with existing academic discipline classifications such as the 2012 ACM Computing Classification System (CCS2012).[4]

20.3.4. Data Science Professional Profiles Definition (DSPP)

The proposed Data Science Professional profiles (DSPP)[5] definition is based on the analysis of the demand in research and industry in data-related professions as well as in current company practices in defining new data-related organisational roles. The identified professional profiles are classified using ESCO taxonomy[6], and necessary extensions are proposed to support the following hierarchy of data handling-related occupations (see Figures 20.3 & 20.4):

- Managers: Chief Data Officer (CDO), Data Science (group/department) 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.
- Professionals (database): Large scale (cloud) database designers and

[4] ACM: http://www.acm.org/about/class/class/2012; last accessed 9 February 2017.
[5] EDISON: http://edison-project.eu/data-science-professional-profiles-definition-dsp; accessed 5 February 2017.
[6] European Commission: https://ec.europa.eu/esco/portal/home; last accessed 9 February 2017.

administrators, scientific database designers and administrators
- Professional (data handling/management): Data Stewards, Digital Data Curator, Digital Librarians, Data Archivists
- Technicians and associate professionals: Big Data facilities operators, scientific database/infrastructure operators
- Support and clerical workers: Support and data entry workers.

The competences and skills required for different professions are defined in the DSP Profiles document in accordance with the Data Science Competence Framework (CF-DS). An example of mapping CF-DS competences to identified data handling-related occupations is provided.

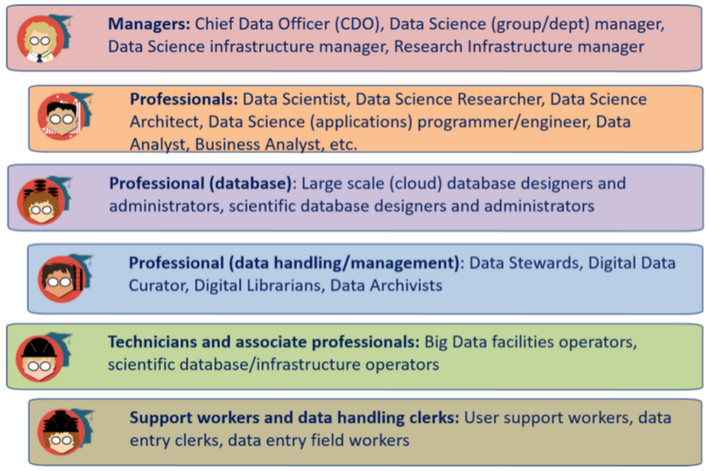

Figure 20.3 Data Science Professions family groups

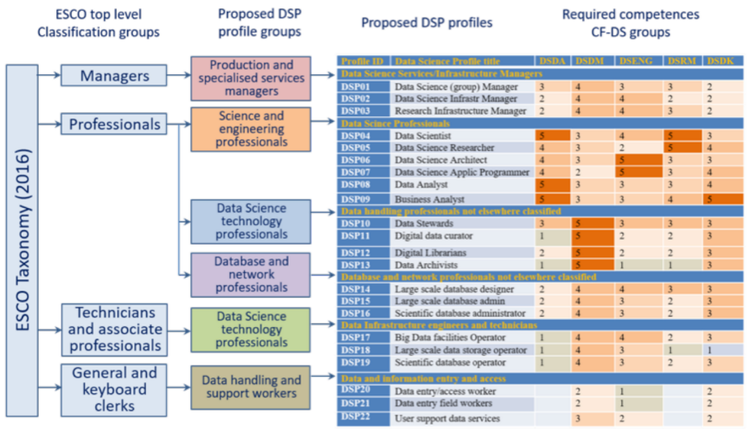

Figure 20.4. Data Science Professions family groups with hierarchy

20.4. DATA SCIENCE PROGRAMMES IMPLEMENTATION AT THE UNIVERSITY OF AMSTERDAM (UVA)

The University of Amsterdam is starting 4 new Data Science programmes and tracks that are based on/ originate from different departments, and which are aimed at different industries and target groups from Computer Science, Business Administration, and multidisciplinary studies. They are primarily intended to answer the needs of the Dutch economy (i.e. industry, research and public services) which is to a large extent international. The programmes and tracks are developed by the departments independently, but all of them use general EDISON recommendations.

i. Artificial Intelligence and Data Science (specialisation)

(http://gss.uva.nl/future-msc-students/information-sciencescontent26study-programme/ profile-data-science.html)
Track - Master

At the core of Data Science are methods for the analysis of large volumes of data. Recently much more data has become available in electronic form, and methods for the analysis and modelling of these data for prediction, classification and optimisation have become much more effective. Recent technical innovations, such as Deep Learning, provide increasingly powerful tools that make it possible to find complex patterns in very large datasets.

Much of the Master's Artificial Intelligence (AI) degree is about Data Science. The obligatory courses on Machine Learning address key technology and theory for modelling large amounts of data. The courses on Machine Learning, Natural Language Processing, Information Retrieval and Computational Intelligence all have a strong focus on data-driven methods. For the "AI courses" in the curriculum, students can choose advanced courses on these topics: Machine Learning 2, Computer Vision 2, Natural Language Processing 2, Information Retrieval 2, Deep Learning, Data Mining Techniques, Information Visualisation and Probabilistic Robotics. All these courses are about modelling data. These can be complemented by courses outside AI, for example on distributed computer systems, privacy and ethical questions, or on statistics.

Within programme: Artificial Intelligence
Organisation: UvA
Language: English
Duration: 5 months

ii. Big Data Engineering

(http://gss.uva.nl/future-msc-students/information-sciences/content28/computer-science.html)
Track - Master

In the Internet era, data is at the centre of the stage. We all continuously communicate via social networks, we expect all information to be accessible online continuously, and the world's economies thrive on data processing services where revenue is created by generating insights from raw data. These developments are enabled by a global data processing infrastructure, connecting everyone from small company computer clusters to data centres run by world-leading IT giants. In the Big Data Engineering track, you study the technology from which these infrastructures are built, allowing you to design and operate solutions for processing, analysing and managing large quantities of data. This track is part of the joint Masters in

Computer Science, in which renowned researchers from both the Vrije Universiteit Amsterdam (VU) and the University of Amsterdam (UvA) contribute their varied expertise in one of the strongest Computer Science programmes available in Europe.

Within programme: Computer Science
Organisation: UvA + VU
Language: English
Duration: 2 years

iii. MBA Business Analytics & Data Science

(http://abs.uva.nl/programmes/mba/content2/mba-big-data.html)
Track – Master MBA

This MBA in Big Data and Business Analytics is intended for hands-on Big Data specialists, for people in leadership roles working with Big Data and for Entrepreneurs. The curriculum of this MBA is highly multidisciplinary, with courses from A (analytics), B (business) and C (computer science), and with projects to practise and implement the integration of these three aspects.

Furthermore, the curriculum is a mix of state-of-the art theory taught by renowned academic professors, and it includes practical applications of this knowledge taught by people with extensive industry experience. In the curriculum, much time will be devoted to the '21st century skills' - the skills required to become successful in this age: entrepreneurship / entrepreneurial attitude, flexibility, teamwork, communication skills and ethics.

Key features:

- Two-year part-time programme (2 evenings per week);
- Balanced curriculum consisting of Business courses (e.g. strategy, finance, marketing, HRM), Analytics courses (e.g. statistics, econometrics, system optimization) and Computer Science courses (e.g. machine learning, data visualisation);
- All lecturers combine theory with practical applications;
- Silicon Valley study trip and Big Data Thesis Project will be part of the programme;
- Degree: Master of Business Administration (MBA) granted by the University of Amsterdam;

Organisation: Amsterdam Business School, UvA
Language: English
Duration: 2 years

iv. Data Science

(http://gss.uva.nl/future-msc-students/information-sciences/content/data-science.html)
Track - Master

In the one-year Data Science Master's track, you will acquire knowledge of the theories and tools used in data science. We will teach you how to use these tools for working with data in different domains, such as Healthcare, Media and Communication, Smart City, Life Sciences and Digital Humanities. Graduates have an integrated view on the possibilities and development of data science in society. Students will benefit from the strong collaboration with Amsterdam Data Science (ADS), bringing together leading researchers across the entire life cycle of data science, from expertise in machine learning and information retrieval to human computer interaction and large-scale data management.

Within programme: Information Studies
Organisation: UvA + VU
Language: English
Duration: 1 year

20.5. RESEARCH DATA MANAGEMENT EDUCATION AND TRAINING

Research Data Management training is recognised as essential for practising researchers of all scientific domains and important for academic Data Science education. It is typically covered by training programmes for postgraduates, PhD students and researchers; however it is rarely covered by existing or planned academic programmes and courses. It has been identified that to cover the wide needs of the research and academic community, the RDM curriculum and training materials must allow easy customisation and localisation to adjust to the trainees' background and local infrastructure resources, as well as to cater for the needs of specific scientific domains.

The EDISON Project has addressed RDM training and education as a priority issue in order to contribute to raising standards in general competences and skills related to working with research data and with the variety of modern data including social (network) data, environmental data and business data. The EDSF provides a basis for defining a general RDM training program that covers the major practical aspects of RDM; this can be also considered as an important component of more general data literacy training.

The proposed customisable RDM training program

The following RDM training program has been constructed based on an extensive study of existing RDM training programmes and resources, in particular collected at the Data Management Clearinghouse[7] and by the RDA US directory of RDM resources[8]. It covers most topics available in currently-available RDM training programmes and curricula, has a modular structure and provides the possibility of expanding into more specific data management topics that may be required by specific groups of practitioners.

A Research Data Management training or education programme should contain the following essential modules (allowing extension and adoption to particular target communities):

A. Use cases for data management and stewardship

- Preserving the Scientific Record

[7] DM-Clearinghouse, 2016. Data Management Training Clearinghouse. Available at https://www.sciencebase.gov/catalog/item/56d88012e4b015c306f6cffc; accessed 5 February 2017.
[8] RDA-US-RDM, 2016, *RDA US directory of RDM resources*, 2016. Available at https://docs.google.com/spreadsheets/d/10RTW-nZk0x_mpQw2VAlttcc656MV9EeCaDe2lM4umb4/edit#gid=0; accessed 5 February 2017.

B. Data Management elements (organisational and individual)

- Goals and motivation for managing your data
- Data formats
- Creating documentation and metadata, metadata for discovery
- Using data portals and metadata registries
- Tracking data usage
- Backing up your data
- Data security and integrity
- Data Management Plan (DMP) (also a part of hands on session(s))

C. Responsible Data Use Section (Citation, Copyright, Data Restrictions)

- Handling sensitive data
- Ethical issues, obtaining consents

D. Open Science, Open Access and Open Data (Definition, Standards, Open Data use and re-use, open government data)

- Research data and open access
- Repository and self-archiving services
- PID identifier for data and ORCID identifier for researchers
- Stakeholders and roles: engineer, librarian, researcher
- Open Data services: ORCID.org, Altmetric Donut, Zenodo

E. Hands on and labs:

a) DMP design
b) Metadata and tools
c) Selection of licenses for open data and contents (e.g. Creative Common and Open Database)

The proposed RDM training program has been taught at the Data Science workshop since May 2016 at Amsterdam Business School, University of Amsterdam[9] organized by the EU Erasmus+ Eduworks[10] Project. The program contained two major parts: general RDM topics, and Data Management Plan (DMP) design that was presented as a hands-on exercise. The training materials were developed jointly by the EDISON Project UvA in cooperation with the University Library and are available under a CC BY licence. Further development is expected in the framework of the proposed RDA Working Group on RDM literacy.

20.6. REQUIRED RESOURCES

A successful Data Science education programme depends on the availability of 3 key components: (1) teaching staff, (2) computing and lab facilities, and (3) a pool of experts/advisers and related topics for course and thesis projects. All three components create challenges and require advanced planning. The following offerings are made available by the relevant departments:

1. Teaching staff: Core teaching staff are provided by departments hosting the programme or track; associate teaching staff from industry provide specialised courses; local industry experts are invited to give selected lectures; leading domain researchers and experts are invited to give lectures, seminars and colloquia.
2. Computing and lab facilities: Computer classes are operated by departments and supported by ICT departments; high performance computing facilities are provided by the SURFsara Dutch research HPC facility; departments are actively using research and educational grants from the major cloud and Big Data providers such as Amazon Web Services, Microsoft Azure, IBM Watson and BlueMix to give students the opportunity to learn about leading industry platforms and applications.
3. A pool of experts and project development topics: departments maintain a network of external experts and collaborating research and technology organisations that provide advice on students' projects and host students' thesis projects.

A common problem and gap in developing consistent Data Science programmes is setting up a professional Data Management course that would cover both Research Data Management and industry data management and governance topics. The EDISON Project is cooperating with departments to develop core Data Management courses including Research Data Management courses and training for students and researchers.

20.7. COORDINATION OF RELEVANT ACTIVITIES INSIDE UVA

For coordination purposes and for the exchange of experience, UvA has created the Data Science Interest Group and a corresponding mailing list that has become an important forum for coordinating activities between departments, projects and collaborating organisations. This important role belongs to the Amsterdam Data Science Centre (ADS)[11] which is a joint initiative of 10+ companies and institutions in the Amsterdam area; the recently-established Amsterdam School of Data Science (ASDS) also has an important role to play.[12]

20.8. CONCLUSIONS

- The EDISON Data Science Framework (EDSF), the product of the EDISON Project, provides a strong background for building customisable Data Science programmes, including Research Data Management education and training programs;
- The Data Science Professional profiles (DSPP) play an important role since they define the whole spectrum of the data-related organisational roles currently present and required by research organisations and industry;
- The University of Amsterdam has created an effective cooperative and creative environment for coordinating efforts from multiple departments in establishing Data Science-related programmes. The EDISON Project provides necessary recommendations and materials for building consistent Data Science programmes;
- The University Library cooperates with the ICT department and research and teaching departments on RDM training and in setting up academic courses on Data Management;
- The EDISON Project maintains an extensive network of Data Science experts and Champion Universities that run or implement Data Science programmes in Europe;
- The current success of the EDSF makes it critically important to ensure the

9 Amsterdam Business School: http://abs.uva.nl/; accessed 5 February 2017.
10 Eduworks: http://www.eduworks-network.eu/; accessed 5 February 2017.

11 Amsterdam Data Science: http://amsterdamdatascience.nl/; accessed 5 February 2017.
12 Amsterdam School of Data Science: https://www.schoolofdatascience.amsterdam/; accessed 5 February 2017.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654139.