



Conference paper

Defining Customisable Model Curriculum for Research Data Management Training

Abstract

This paper discusses an approach to define customisable training curriculum for Research Data Management (RDM) training. It refers to the work done in the EDISON project to define the Data Science Model Curriculum (MC-DS) allows customization based on the required training profiles specifying intended learning outcomes and required competences to be acquired by students or trainees. The paper presents an example of the general RDM training program that is a part of professional Data Management curriculum and can be considered as an important component of the more general data literacy training.

Introduction

Effective research data management (RDM) is important part of the modern research and a requirement from the funding agencies in Europe, US, and elsewhere in the world. Previously, RDM services were sufficient to be provided by dedicated department in research institutions. However now with growing research digitalization and among of data produced by individual researchers, the RDM knowledge and skills are required from all researchers. Necessary RDM training needs to be provided to large amount of researchers with different backgrounds and different forms.

Although RDM training is provided by many institutions, in many cases by libraries, and many training materials are available, in many cases sufficient customization is needed to adjust materials to the trainees background and local infrastructure resources, as well as for specific scientific domain.

This paper discusses an approach to define customisable training curriculum for Research Data Management (RDM) training. It refers to the work done in the EU funded EDISON project to define the Data Science Model Curriculum (MC-DS) that would allow customization based on the required training profiles specifying intended learning outcomes and required competences to be acquired by students or trainees. This approach is based on using other components of the EDISON Data Science Framework (EDSF) such as Data Science Competence Framework (CF-DS) and Data Science Body of Knowledge (DS-BoK) briefly introduced below.

The paper presents an example of the general RDM training program that covers the major practical aspects in RDM and can be considered as an important component of the more general data literacy training.

Data Science Competence Framework and Body of Knowledge

The Data Science Competences Framework (CF-DS) is a cornerstone of the EDISON Data Science Framework and used for defining such components as Data Science Body of Knowledge (DS-BoK) and Data Science Model Curriculum (MC-DS).

The CF-DS includes the following competences groups

- Data Analytics including statistical methods, Machine Learning and Business Analytics
- Engineering: software and infrastructure
- Data Management, Curation, Preservation
- Subject/Scientific Domain competences and knowledge
- Scientific or Research Methods (for research professions) and Business Process Management (for business related professions)

The identified competence areas provide a better basis for defining education and training program for Data Science related jobs, re-skilling and professional certification. Knowledge of the scientific research methods and techniques makes the Data Scientist profession different from all previous professions. It is recommended that both RDM and research methods are included into all Data Science curricula.

The CF-DS provides a basis for the definition of the Data Science Body of Knowledge (DS-BoK), the knowledge needed by the practitioners to perform all the data related processes of their profession. The BoK typically defines the content of a curriculum and is linked to CF-DS via learning outcomes that can be defined for the specific groups of trainees.

Following the CF-DS competence group definition, the DS-BoK should contain the following 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: *Scientific or Research Methods group*
- KAG5-DSBP: Business process management group

Universities can use DS-BoK as reference to define knowledge areas that they need to cover in their programs depending on their primary demand groups in research or industry. The domain specific knowledge can be acquired as a part of the academic education or as a post-graduate professional training at the graduate's work place. It is also commonly recognized that a "fresh" Data Scientist would require 2-3 year to become proficient in his/her profession.

Example: RDM training program

The following RDM training program has been constructed based on extensive study of existing RDM training programs and resources, in particular collected at the Data Management Clearinghouse (DM-Clearinghouse, 2016) and RDA US directory of RDM resources (RDA-US-RDM, 2016). It covers most topics available in currently available RDM training programmes and curricula, has modular structure and provides possibility to extend with more specific data management topics that are required by specific groups of practitioners.

Research Data Management Program (modular organisation)

- A. Use cases for data management and stewardship
 - Preserving the Scientific Record
- 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)
- C. Responsible Data Use Section (Citation, Copyright, Data Restrictions)
 - Handling sensitive data
 - Ethical issues, consent obtaining
- D. Open Science, Open Access and Open Data (Definition, Standards, Open Data use and reuse, open government data)
 - Research data and open access
 - Repository and self- archiving services
 - ORCID identifier for data
 - Stakeholders and roles: engineer, librarian, researcher
 - Open Data services: ORCID.org, Altmetric Doughnut, Zenodo
- E. Hands on:
 - a) Data Management Plan design
 - b) Metadata and tools
 - c) Selection of licenses for open data and contents (e.g. Creative Common and Open Database)

Conclusion and Further Developments

The presented RDM training program has been taught at the Data Science workshop in May 2016 at Amsterdam Business School, University of Amsterdam (<http://abs.uva.nl/>) organized by the EU Erasmus+ Eduworks project (<http://www.eduworks-network.eu/>). The program contains two major parts: general RDM topics and Data Management Plan (DMP) design that is done as hands on exercise.

The modular training materials will be offered as Open Source under Creative Common Attribution license. The possibility of starting the community project at RDA and Data Carpentry will be investigated.

Acknowledgements

The EDISON project is supported under H20202 Grant Agreement n. 675419 by the European Commission.

Competing Interests

The authors declare that they have no competing interests.

References

EDISON Project: Building Data Science Profession [online] <http://www.edison-project.eu/>

Andrea Manieri, et al, 2015, Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists, Proc. The 7th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2015), 30 November - 3 December 2015, Vancouver, Canada

CF-DS, 2016, Data Science Competence Framework (CF-DS). EDISON draft v0.7, 4 July 2016 [online] <http://www.edison-project.eu/data-science-competence-framework-cf-ds>

Data Science Workshop, 2016, Eduworks Project, Amsterdam Business School, 23-27 May 2016. [online] <http://abs.uva.nl/news-events/events/events/content/folder/workshops/2016/05/data-science-workshop.html>

DS-BoK, 2016, Data Science Body of Knowledge (DS-BoK). EDISON draft v0.2, 4 July 2016 [online] <http://www.edison-project.eu/data-science-body-knowledge-ds-bok>

DSP, 2016, Data Science Professional profiles definition (CF-DS). EDISON draft v0.1, 11 July 2016 [online] <http://www.edison-project.eu/data-science-professional-profiles-dsp>

DM Clearinghouse, 2016, Data Management Training Clearinghouse [online] <https://www.sciencebase.gov/catalog/item/56d88012e4b015c306f6cffc>

MC-DS, 2016, Data Science Model Curriculum (MC-DS), EDISON Draft v0.1, 11 June 2016 [online] <http://www.edison-project.eu/data-science-model-curriculum-mc-ds>

RDA-US-RDM, 2016, RDA US directory of RDM resources, 2016 [online] https://docs.google.com/spreadsheets/d/10RTW-nZk0x_mpQw2VAltcc656MV9EeCaDe2IM4umb4/edit#gid=0