

The CODATA-RDA Data Steward school

Daniel Bangert, Göttingen State and University Library, University of Göttingen

Joy Davidson, Digital Curation Centre, University of Glasgow

Steve Diggs, Scripps Institute, UCSD

Marjan Grootveld, Frans Huigens, DANS

Sanjin Muftić, University of Cape Town

Cristiana Pisoni, University of Bergamo

Hugh Shanahan, Royal Holloway, University of London

Lina Sitz, ICTP

Ana Slavec, InnoRenew CoE

Sothearith Seang, Université Côte d'Azur

Shanmugasundaram Venkataraman, Digital Curation Centre, University of Edinburgh

Abstract

Given the expected increase in demand for Data Stewards and Data Stewardship skills it is clear that there is a need to develop training, education and CPD (continuous professional development) in this area.

In this paper a brief introduction will be provided to the origins of the definitions of Data Stewardship. Also it notes the present tendency towards equivalence between Data Stewardship skills and FAIR principles. It then focuses on one specific training event – the pilot Data Stewardship strand of the CODATA-RDA Research Data Science schools that was held in Trieste in August 2019. The paper will discuss the overall curriculum for the pilot school, how it matches with FAIR4S framework, and findings from the students and instructors on how to improve the school.

Introduction

Data stewardship as a role has come into prominence over the last decade. Early references to Data stewards occur in the literature in the first decade of the century with respect to Health Data (Diamond, Mostashari, & Shirky, 2009; Rosenbaum, 2010) and had a strong emphasis on maintaining the privacy of the data sets. Since then, the role has developed into one that carries out a variety of roles for data (Peng, Privette, Kearns, Ritchey, & Ansari, 2015; Peng et al., 2018; Peng, Ge et al., 2016; Salome Scholtens et al., 2019; Sapp Nelson, 2016; Sapp Nelson, Megan R., 2017; Verheul et al., 2019).

The first formal publication of the FAIR data guiding principles (Wilkinson et al., 2016) makes an explicit connection between those principles and Data Stewardship. This paper will make explicit use of this connection in terms of defining an initial curriculum framework for Data Stewardship. Regardless of how deep that connection is, the adoption of FAIR principles as a policy goal (European Commission, 2016) and the identification that these practices at least attempt to address critical issues such as a reproducibility (Hartter, Ryan, MacKenzie, Parker, & Strasser, 2013) indicates that there will be a substantial increase in the required number of Data Stewards.

In this respect there will need to be an extensive increase in the amount of activity in this area from Educational, Training and continuous professional development (CPD) perspectives. This paper will discuss one particular initiative, the Data Steward strand of the CODATA-RDA schools in Research Data Science which is being piloted for the first time in August 2019 in cooperation with the FAIRsFAIR project. The medium-term goal of the school will not give students an introduction to Data Stewarding but instead embed them with Early Career Researchers (ECRs) with the goal of demonstrating the importance of partnership between these roles over the research lifecycle. The immediate goal for the pilot school is to deliver a draft curriculum that will be refined based on feedback from pilot participants and offered through subsequent schools delivered by CODATA/RDA and the FAIRsFAIR project (<https://www.fairsfair.eu/fair-competence-centre>).

Curriculum Framework – FAIR4S

A more detailed description of the landscape of training resources and curriculum frameworks is discussed elsewhere (Shanahan, H. et al., 2019). The most recent and relevant curriculum framework, which incorporates previous work, is FAIR4S (https://www.eoscpilot.eu/sites/default/files/fair4s_eoscpilot_skills_framework.pdf) (Whyte, A, 2019; Whyte, Angus et al., 2019). This provides a high level description of the skills necessary to make data FAIR and keep data FAIR. It also provides a description of the level of understanding of the topics required for a variety of roles, including Data Steward. The highest set of topics are: *Plan and design* (planning and design of data, research software and other outputs, including documentation); *Capture and process* (capturing and processing of data or related materials to enable research evidence to be prepared for analysis); *Integrate and analyse* (developing and applying appropriate methods to enable lines of enquiry for research); *Appraise and preserve* (developing and applying appropriate methods to appraise research outputs); *Publish and release* (describing research products and their inter-relationships and providing access to them); *Expose and discover* (ensuring processes and mechanisms for providing access to research products); *Govern and assess* (developing and maintaining legally compliant strategies, policies, and processes on outputs); *Scope and resource* (identifying the scope of research data services and stewardship activities and securing the resources to sustain these) and *Advise and enable* (managing services that enable data stewardship and open research).

The school

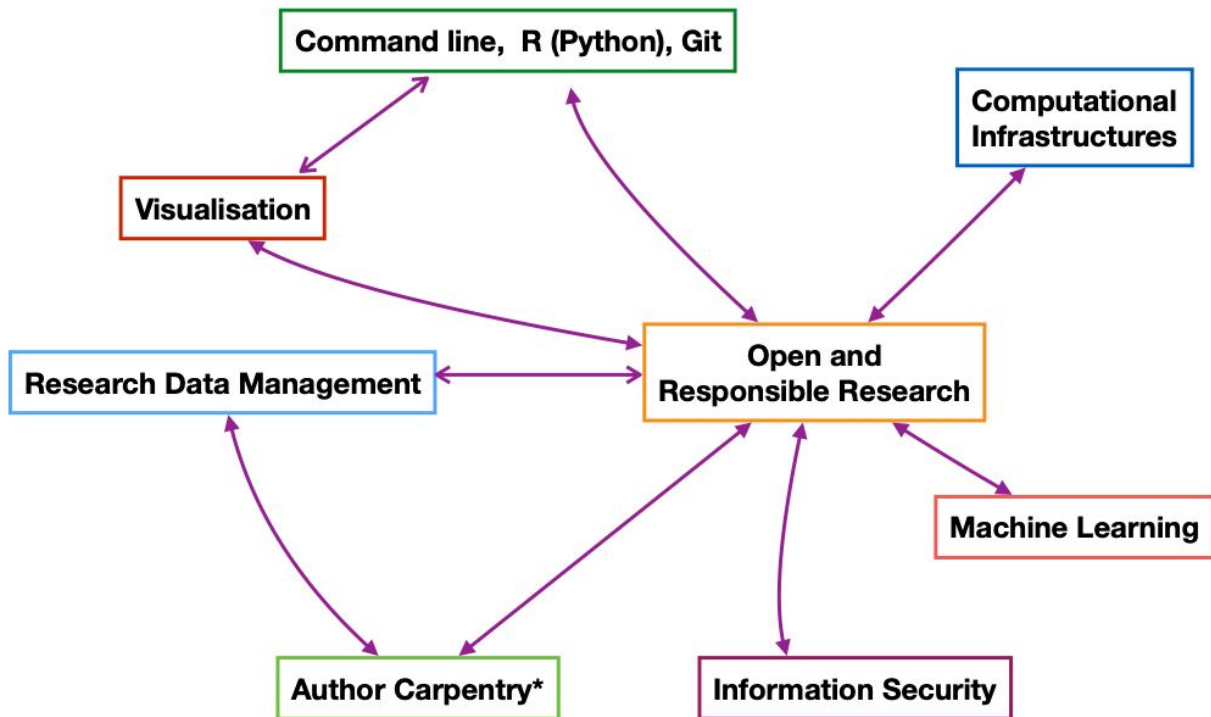
The CODATA-RDA Research Data Science schools are a series of schools that have run since 2016. The long-term goal of the schools is to create communities of ECRs that are enabled to make the most of the Data Revolution in research. This is enabled by delivering an expanding set of schools delivered regionally which provide a foundation in Data Science – skills that are independent of the domain that the ECR is based in. There is a strong emphasis on teaching practical skills with team learning and ample opportunities for reflection and discussion. Students come from a wide variety of domains including Bioinformatics and other Life Sciences, Earth and Atmospheric Sciences, High Energy Physics and others. The priority is to deliver these schools to individuals from Low and Middle Income Countries though the curriculum is applicable for students from High Income countries as well. Through the expansion of these schools the concept of providing such a curriculum (or something similar) will become embedded in Higher Education and hence such Data Science skills for ECRs will become accepted in much the same way that an understanding of basic Biostatistics is essential in the Life Sciences (Metz, 2008) or Linear Algebra in Engineering (Barry & Steele, 1993).

The school emphasises responsible research and hence distinguishes itself from the standard, Machine Learning focussed, Data Science bootcamps that tend to focus more on purely technical content. Over two weeks it delivers modules on

- Open and Responsible Research (Bezuidenhout, Louise, Quick, Rob, & Shanahan, Hugh, n.d.);
- the Carpentries introductory material (Teal et al., 2015; Wilson, 2006) on the Unix command line, Git and R;
- Research Data Management, which provides a broad introduction to the topic for ECRs;
- Author Carpentry (Caltech Library, 2017) describing the skills necessary for authorship in the 21st century;

- Visualisation, specifically the visualisation of data;
- Machine Learning;
- Information Security;
- Computational Infrastructures, providing an introduction computing beyond a laptop or desktop computer.

A diagrammatic representation of how these modules are related to each other is provided in figure 1. In December 2019 the school in San José, Costa Rica used Python as the main language rather than R. By February 2020, nine schools ran on four continents to students from over 40 countries.



*** Not usually run in regional schools**

Figure 1 Diagrammatic representation of modules run in the ECR strand of the CODATA-RDA schools.

Extension to Data Stewards

There is already a substantial overlap between this curriculum and the FAIR4S framework. Hence the CODATA-RDA school's curriculum, with some adjustment, would provide an excellent introduction to the area. The strong emphasis on building communities of researchers with a grounding in responsible research practices also presents the opportunity to embed Early Career Data Stewards and Researchers with each other. This would encourage both roles to work more closely with each other. FAIRsFAIR (<https://fairsfair.eu>) is a project that addresses the development and concrete realisation of an overall knowledge infrastructure on academic quality

data management, procedures, standards, metrics and related matters, based on the FAIR principles. One of its goals is to develop and provide a series of schools in this area and hence FAIRsFAIR is partnering with the CODATA-RDA schools to deliver training along the lines described above.

Data Steward Pilot School

In August 2019 a pilot version of the Data Steward school ran in parallel with the CODATA-RDA school at the International Centre for Theoretical Physics. A cohort of six students with a Data Stewarding background was taught in the pilot school. In table 1 the modules that were run in common with the ECRs are listed. Week 1 of the school was common for both ECRs and Data Stewards. The Data Steward specific modules were run in week 2 and are listed in table 2. A detailed description of the Data Steward Specific teaching is provided in Appendix A. In terms of omissions from the ECR school in week 2, Visualisation was dropped entirely and Machine learning scaled back to one seminar on the penultimate day of the school. The Computational infrastructures module was scaled back from a day and a half to one day. These and the modules delivered in week one were taught with the ECRs.

Table 1. Modules covered in Pilot school that are common to ECRs and Data Stewards. The matching terms from FAIR4S are abbreviated as follows. Plan and Design = PD; Capture and Process = CP; Integrate and Analyse = IA; Appraise and Preserve = AP; Publish and Release = PR; Expose and Discover = ED; Govern and Assess = GA; Scope and Resource = SR; Advise and Enable = AE.

Topic	Matching FAIR4S	Week run
Open and responsible research	AE, GA	1
Introduction to the command line	IA, CP, PD	1
R	IA, CP, PD, ED	1
Git	PR, ED	1
Author carpentry	CP, AP, PR	1
Research data management	PD, AP, ED, GA	1
Introduction to information security	GA	2
Computational infrastructures	IA, SR	2
Machine learning	IA	2

Table 2. Modules covered in Pilot school that are unique to the Data Stewards. All of these modules are run in the second week of the school. The matching terms from FAIR4S are abbreviated as follows. Plan and Design = PD; Capture and Process = CP; Integrate and Analyse = IA; Appraise and Preserve = AP; Publish and Release = PR; Expose and Discover = ED; Govern and Assess = GA; Scope and Resource = SR; Advise and Enable = AE.

Topic	Matching FAIR4S
Data steward practice	SR, AE
FAIR data	PD, AP, PR, ED, GA
Data management planning	PD, AP, PR, ED, GA, SR
Metadata	CP, ED
Accessibility	PR, ED, GA
Preservation and Publishing	PR, AP
Data policies	GA, AE
Storing data	AP, SR
Linked data	PD, PR, ED, IA

A diagrammatic representation of how the modules for the Data Steward strand of the school are related to each other is provided in figure 2, which relates the common modules for ECR and Data Steward strands with the Data Steward specific modules, and figure 3 which focuses on the relationships between the Data Steward specific modules.

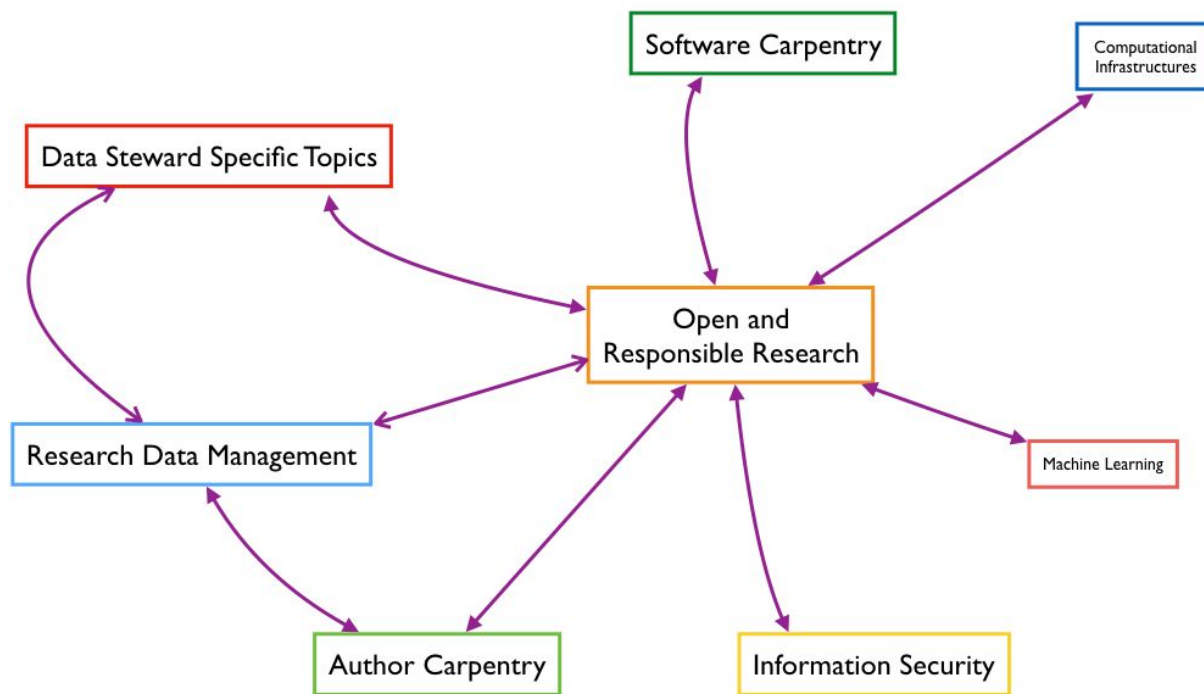


Figure 2 Diagrammatic representation of modules run in the Data Steward strand of the CODATA-RDA schools.

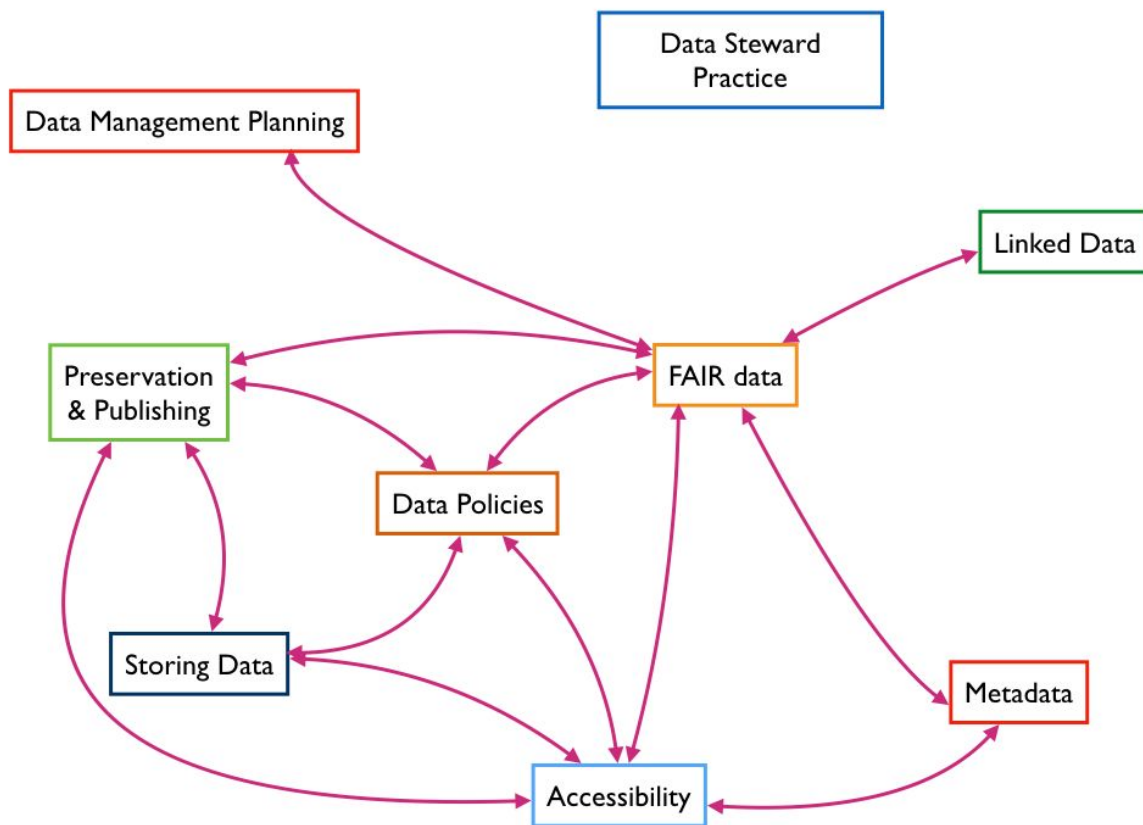


Figure 3 Diagrammatic representation of Data Steward specific modules run of the CODATA-RDA schools.

Data Steward Specific Modules

The materials taught were delivered in nine modules. The modules were largely taught in an overlapping fashion during the second week. It is important to note that the module structure described here is a *post hoc* description of the set of materials delivered. In detail the modules are:

- Data Steward Practice - providing an introduction into what Data Stewardship is and how to train individuals in the topic.
- FAIR data - specifically an explanation of the FAIR principles and the assessment of the FAIRness of data sets.
- Data Management Planning - providing the necessary skills to support researchers to develop DMPs.
- Metadata - in particular determining minimum metadata requirements.
- Accessibility - licensing of data and conditions associated with access to data.
- Preservation and publishing - providing an introduction to archiving and issues such as long term sustainability, trustworthy repositories and publishing data.

- Data policies - determining key elements involved in the development of an institutional data policy.
- Storing data - practical issues associated with uploading data and the skills to get researchers to deposit or archive their data.
- Linked Data - providing an introduction to the principles and implementation of linked data.

Assessment

The CODATA-RDA schools works on a principle of iterative improvement of its curriculum and has been shown to be a powerful technique (Wolf, 2007). This pilot in particular represents a first step in this area and will require further revision. Given the small size of the cohort more qualitative approaches were taken to understand the impact of the pilot school. Specifically, students were requested to write an action plan on their future plans and a workshop on the pilot was held.

Action plans

The students were asked to complete an action plan for what they would do after the school. Specifically they were asked to address two questions.

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?
2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

The responses of the students are in Appendix B.

A number of recurring themes emerged from these responses.

Training: Specifically to train researchers or potential data stewards at their home institution. The topics proposed to be taught ranged from practical issues such as the Carpentry material taught at the school to more general Open Science issues. Four students addressed and planned this.

Dissemination (Blogs/Talks): This is an opportunity for the students to report back to their institution of their findings from the school through seminars or reports or to discuss the school to a wider audience through web sites or blogs. Four students addressed and planned this.

Research usage: In particular this is to research how data is managed within their own institution, e.g. do researchers use repositories and hence understand what the landscape looks like. Four students addressed and planned this.

Specific references to FAIR: In 4 out of the 6 action plans there is a specific reference to FAIR principles. However, in 3 of those cases this is very much in the larger context of RDM practices in

general. The fourth is very focussed on FAIR but the relevant student has a specific FAIR-related project.

Develop or use tools/procedures: This was a commitment to either develop a procedure for their institution for use by their researchers in RDM that came out of the school's materials or to use a tool such as RISE or to be part of a team to develop a relevant tool, such as terms4FAIRskills (<https://github.com/terms4fairskills/>). Three students addressed and planned this.

Engage with external Data Steward community: This was a commitment to continue engaging with the Data Steward community beyond their own institution. Three students addressed and planned this.

Distribution of actions: The activities described above were not evenly distributed between the students. A network diagram (drawn using the geomnet R package using the mds layout algorithm) shows the connections between students and themes. As can be seen in figure 4, the themes divide into those related to dissemination and teaching and those related to research and implementation. The interest in the former set of themes is perhaps something of a surprise, namely that students are as interested in teaching and communicating what they've learnt to their colleagues and researchers as implementing Data Stewardship.

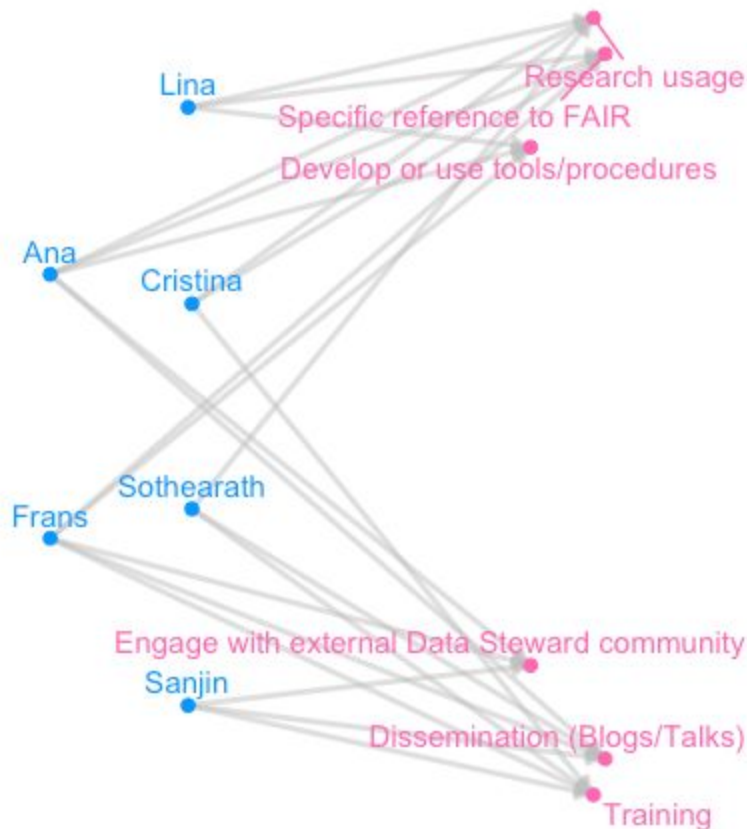


Figure 4: Network of themes and students. An edge indicates that a student referred to the theme.

Workshop

In October 2019 a workshop on the school was held at the Research Data Alliance 14th plenary in Helsinki. During the workshop the experiences of the students and instructors were discussed with a group of data professionals to get their feedback on the schools and for the students and instructors to further reflect on the school and how it could be improved. At this meeting, two themes emerged. In the first instance the data professionals were for the most part happy with the overall design of the curriculum. Secondly, it was noted that while the ECRs and Data Stewards were aware of each other during the school there wasn't an opportunity for them to a) get a better understanding of what the other group was doing in their strand, and b) that there wasn't an opportunity for the two groups to work together in their respective roles. A proposal to deal with this would be, for future schools, to create teams of ECRs at the outset of the school and to assign a Data Steward student to each team. During the school, joint exercises could then be run so the team of ECRs grasp the purpose of the Data Steward and Data Stewards see their work in the context of improving research.

Conclusions

This paper describes a pilot school to provide training for Data Stewards over a two week period. This pilot builds on a school specifically for ECRs that combines technical and responsible research. A small cohort of Data Steward students were selected to test the materials for this pilot. The students take part in a curriculum that is very similar to the ECR school but with crucial differences in the second week of the school, with tailored material delivered to only them.

Given the paucity of training in this area, particularly with respect to FAIR-related activities, there are a number of issues that have not been addressed here but would merit further research. In the first instance the term "Data Steward" has been used here rather than other related terms such as "Data Manager". Rather than attempting to tease out such roles which are likely to be evolving, a starting point was to use the reference document FAIR4S and ensure that there is a mapping between the broad categories described there and the topics covered in the school. In other words the focus is very much on the teaching of a set of skills rather than training for a specific role.

In reviewing the school there were a number of findings. In the first instance, students were as interested in disseminating and providing their own training at their institute as carrying out the roles of being a Data Steward. There was also interest in disseminating their findings through blog posts etc. It is clear that future schools should take this into consideration and consider carefully how one can provide examples of good training practice. Secondly, the Data Steward students were interested in understanding how data is being managed at their institution and hence make suggestions on how to improve it. Students understand the context of the FAIR principles. There is clear interest in using tools and engaging with the larger data stewardship community.

Feedback from data professionals from the RDA plenary workshop indicates that the curriculum is for the most part apposite. What can be improved is ensuring that there is greater cross-talk between the ECR's and Data Stewards. This would represent a unique offering in terms of training - namely an opportunity to simulate what the partnership between Researchers and Data Stewards should look like and give both groups the chance to see the purpose, challenges and opportunities that come from working with the other group. This could be achieved by creating a series of complementary exercises where ECRs and Data Stewards work together during the second week of the school and represents the next challenge. Finally the strong emphasis of the CODATA-RDA

schools on building a sense of community amongst its students will be important in ensuring that Data Stewards from this school will also feel that they are part of a community that they can reach out to.

The pilot school that ran in August 2019 represents a first step; focusing on ensuring the draft curriculum that is apposite for the Data Stewards. Adjustments will be made to the schools content and will be run again in 2020 in Trieste. Furthermore, negotiations are in place to run other instances of the school in specific research domains. The CODATA-RDA schools themselves have expanded considerably over the last five years and the authors see no reason why that could not occur for the Data Steward strand as well.

Acknowledgements

This work has been made possible by support from the FAIRsFAIR project. FAIRsFAIR “Fostering FAIR Data Practices in Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558.

References

- Barry, M. D. J., & Steele, N. C. (1993). A core curriculum in mathematics for the European engineer: An overview. *International Journal of Mathematical Education in Science and Technology*, 24(2), 223–229. <https://doi.org/10.1080/0020739930240207>
- Bezuidenhout, Louise, Quick, Rob, & Shanahan, Hugh. (n.d.). “Ethics When You Least Expect It”: A Modular Approach to Data Ethics Instruction. *Submitted for Publication*.
- Caltech Library. (2017). *AuthorCarpentry Homepage*. <https://doi.org/10.7907/z96h4ffz>
- Diamond, C. C., Mostashari, F., & Shirky, C. (2009). Collecting And Sharing Data For Population Health: A New Paradigm. *Health Affairs*, 28(2), 454–466. <https://doi.org/10.1377/hlthaff.28.2.454>
- European Commission. (2016, September 5). European Commission—PRESS RELEASES - Press release—G20 Leaders’ Communique Hangzhou Summit. Retrieved 4 July 2019, from http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm
- Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., & Strasser, C. A. (2013). Spatially

- Explicit Data: Stewardship and Ethical Challenges in Science. *PLOS Biology*, 11(9), e1001634. <https://doi.org/10.1371/journal.pbio.1001634>
- Metz, A. M. (2008). Teaching Statistics in Biology: Using Inquiry-based Learning to Strengthen Understanding of Statistical Analysis in Biology Laboratory Courses. *CBE—Life Sciences Education*, 7(3), 317–326. <https://doi.org/10.1187/cbe.07-07-0046>
- Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., & Ansari, S. (2015). A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets. *Data Science Journal*, 13, 231–253. <https://doi.org/10.2481/dsj.14-049>
- Peng, G., Privette, J. L., Tilmes, C., Bristol, S., Maycock, T., Bates, J. J., ... Kearns, E. J. (2018). A Conceptual Enterprise Framework for Managing Scientific Data Stewardship. *Data Science Journal*, 17(0), 15. <https://doi.org/10.5334/dsj-2018-015>
- Peng, Ge, Ritchey, Nancy A., Casey, Kenneth S., Kearns, Edward J., Privette, Jeffrey L., Saunders, Drew, ... Ansari, Steve. (2016). Scientific Stewardship in the Open Data and Big Data Era—Roles and Responsibilities of Stewards and Other Major Product Stakeholders. *D-Lib Magazine*, 22(5/6). <https://doi.org/10.1045/may2016-peng>
- Rosenbaum, S. (2010). Data Governance and Stewardship: Designing Data Stewardship Entities and Advancing Data Access. *Health Services Research*, 45(5p2), 1442–1455. <https://doi.org/10.1111/j.1475-6773.2010.01140.x>
- Salome Scholtens, Petronella Anbeek, Jasmin Böhmer, Mirjam Brullemans-Spansier, Marije van der Geest, Mijke Jetten, ... Celia W G van Gelder. (2019). *Life sciences data steward function matrix*. <https://doi.org/10.5281/zenodo.2561723>
- Sapp Nelson, M. (2016). Pilot Data Information Literacy Competencies Matrix Scaffolded Across Undergraduate, Graduate and Data Steward Levels. *Libraries Faculty and Staff Scholarship and Research*. Retrieved from https://docs.lib.purdue.edu/lib_fsdocs/136

- Sapp Nelson, Megan R., M. (2017). A Pilot Competency Matrix for Data Management Skills: A Step toward the Development of Systematic Data Information Literacy Programs. *Journal of EScience Librarianship*, 6(1). <https://doi.org/10.7191/jeslib.2017.1096>
- Shanahan, H., Hoebelheinrich, N., Whyte, A., Davis, R., Jones, S., & Hodson, S. (2019). Teaching FAIR. *Submitted for Publication*.
- Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., & Pawlik, A. (2015). Data Carpentry: Workshops to Increase Data Literacy for Researchers. *International Journal of Digital Curation*, 10(1), 135–143. <https://doi.org/10.2218/ijdc.v10i1.351>
- Verheul, I., Imming, M., Ringerma, J., Mordant, A., Ploeg, J.-L. van der, & Pronk, M. (2019). *Data Stewardship on the map: A study of tasks and roles in Dutch research institutes*. <https://doi.org/10.5281/zenodo.2669150>
- Whyte, A. (2019, March 31). *FAIR4S, a skills and capability framework for the European Open Science Cloud*. Presented at the Drexel-CODATA FAIR and Responsible Research Data Management (FAIR-RRDM) Workshop, Philadelphia.
- Whyte, Angus, Leenarts, Ellen, de Vries, J., Huigen, Frans, Sipos, Gergely, Dijk, E., ... Ashley, Kevin. (2019). *EOSCpilot D7.5 Strategy for Sustainable Development of Skills and Capabilities*. Retrieved from <https://eoscipilot.eu/themes/skills>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilson, G. (2006). Software Carpentry: Getting Scientists to Write Better Code by Making Them More Productive. *Computing in Science & Engineering*, 8(6), 66–69. <https://doi.org/10.1109/MCSE.2006.122>

Appendix A

Detailed topics covered that are specific for Data Stewards.

Day 1 Week 2	Module	Description
Welcome and introductions		An ice breaking opportunity for the students to get to know each other and the lecturers. The session also helps the students to understand each others' motivations for participation.
Background		An introduction to the course themes and some background on drivers for research data management, Open Science and the emergence of the FAIR principles.
What does a data steward do?	Data Steward Practice	Overview of - potential - tasks and roles, plus relevant literature
FAIR data	FAIR data	A skeptic's view: everyone loves FAIR, but what is it really? Including group discussion
FAIR assessment tools - FOSTER course and exercise	FAIR data	Students take the online tutorial of FAIR assessment tools then complete a practical exercise to assess the FAIRness of data using one of the tools presented.
Data Management Planning	Data Management Planning	Data management plans (DMPs) are increasingly part of good RDM practice, in many cases being mandated by institutions and funders and therefore it was deemed necessary to train data stewards in how to create a good DMP. This session involved a descriptive outline of DMPs and a practical exercise.

Day 2 Week 2	Module	Description
--------------	--------	-------------

Data Management Plan Activity	Data Management Planning	A creative exercise to help students to consider the motivations for developing data management plans from different stakeholder perspectives. The activity is based on the Train the Trainer Card Game 2.0 created by Gwen Franck.
Metadata and description activity	Metadata	A practical exercise to get students to consider minimum metadata requirements from the perspective of both creators and data reusers. This exercise is based on an activity developed by Sally Rumsey, University of Oxford.
Licences	Accessibility	Message: licences tell potential reusers what they can do. OpenAIRE recommendations for research data licences. Creative Commons. Software licences. Plus discussion
Repositories	Preservation and publishing	What does an archivist do? What is the OAIS model? Long-term sustainability. Preferred file formats.
Certification: FAIR data in Trustworthy repositories	Preservation and publishing	Message: trustworthy repositories make and keep data FAIR. CoreTrustSeal repository certification. Plus exercise to understand how CoreTrustSeal requirements also concern the data producer, i.e. the researcher.
RDM service development and optimisation and RISE activity	Data Management Planning	Data Stewards are well placed to contribute and aid in service building. The RISE tool was therefore introduced along with a practical session. Different institutions will have different priorities relating to what they will want to focus on but by taking stock of their current status they can then see where effort should be directed most.

Day 3 Week 2	Module	Description
Developing RDM policies and activity	Data policies	A session to introduce students to key elements that may be included in an institutional data policy and other environmental factors to consider when developing policies.
Promote and archive data	Preservation and publishing	Message: publish data in a data journal. Plus exercise to choose one data journal and give a one-minute summary of the manual/guidelines
Upload your data and practice data access	Storing data	Introduction in the B2SHARE repository. Hands-on exercise in depositing data in the training version of B2SHARE with its annotation service B2NOTE.
Exercise on data access	Accessibility	Exercise: explore these three data repositories and their conditions on data access. Are the conditions clear and feasible?
Discussion of data access challenges	Accessibility	An open Q&A session where students can ask questions about providing data access.
Designing RDM training activity	Data Steward Practice	An exercise for the students to design their RDM training activity.

(Day 4 week 2 is given over to Computational Infrastructures).

Day 5 Week 2	Module	Description
Linked Data theory and basics	Linked Data	An introduction to the aims and principles of Linked Data and the Semantic Web. Exercise: explore the Wikidata knowledge graph.
Ontologies	Linked Data	An introduction to ontologies, including how ontologies and their structures are expressed. Exercise: explore and compare selected ontologies.
Producing RDF	Linked Data	Exercise: create RDF triples and upload to Blazegraph triplestore.
SPARQL	Linked Data	Exercise: query uploaded data and explore other query services.

Summary and feedback discussion	Data Steward Practice	Summary of Linked Data in theory and practice. Revisiting of the FAIR principles. Final reflections and feedback on the Data Steward school.
---------------------------------	-----------------------	--

Appendix B

Action plans of students.

Data Stewardship Action Plans

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?

2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

Ana Slavec's data stewardship action plan

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?

At my institution (InnoRenew CoE) I plan to carry out the following activities:

- 1) Write a blog post about the data stewardship school
- 2) Have a presentation about research data management on a weekly meeting (we have a staff meeting the same time every week and every time someone else presents)
- 3) Update our data management plan and prepare a template for researchers to use in their individual projects
- 4) Collect information on what data has been collected in the life time of the institute (2.5 years) and help researchers to select the most appropriate repository for their data (in case there is no disciplinary repository Zenodo is used)
- 5) Establish a procedure to assist researchers in depositing their data and metadata to the repository

In addition to my work at the institution, what I learned in the data stewardship course is going to help me in my activities in other organisations:

- I. As part of my RDA ambassador programme I will visit several research institutes in Europe that collect data on renewable materials and products. In the first part of the presentation I will make an introduction to data management in general and in the second part I will present them RDA and what are the possibilities of being involved.
- II. Development of the programme for Open Science 2019, an event that I am organizing in November with Young Academy Slovenia (with the support of the FIT4RRI project) at the University of Maribor
- III. Skype meeting with members of the Eurodoc Open Science WG and Eurodoc's open science ambassadors (to share the experience with the school)

2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

At my institution:

- 1) Blog post will be published on the institute's website and shared on social media. I hope it will motivate other institutions to establish the position of a data steward.
- 2) The presentation will make our researchers more aware of the importance of managing their data and will motivate them to seek help in preparing their data for deposition
- 3) The template will make it easier for researchers to prepare a data management plan
- 4) Based on the overview we will be able to prepare a plan for data deposition (point 5)
- 5) I hope the outcome will be that all our data will be deposited in either a disciplinary repository or on Zenodo

In other organisations:

- I. Getting more researchers from the renewable materials domain to become RDA members and establishment of a RDA interest group (and a WG in the long term)
- II. Raise awareness of FAIR data and other aspects of open science among Slovenian researchers
- III. Motivate early career researchers involved with Eurodoc to learn more about the FAIR data aspect of open science and to share this in their communities

Cristiana Pisoni's Data Stewardship Action Plan

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?

1. A meeting with the Chief of the Library Service and the Research provost of our University to underline the benefits that our Institution could have from making the datasets produced by its researchers at least traceable and attributable to the institution, drawing their attention to some actions that could be taken at no cost, as described below

2. Within a very little group of researchers (4-5 max, preferably from the engineering campus, where I work):
 - ❖ I would like to interview them separately about the interest they would have on a mirror repository on our publication repository (<https://aisberg.unibg.it/>) with a specific section (i.e. Section 3: Research Data) to collect bibliographic data about their research datasets with a specific metadata for the link to dataset uploaded in other repositories/archives
 - ❖ Analyzing their answer in order to understand:
 - If they understand the importance of the data management
 - If there is any real interest in the pilot project: if yes, build on a Section 3: Research Data in our repository (zero cost)
 - If, in future, there will be the possibility of inserting in the global policy a section about research data so that the linking to datasets uploaded elsewhere would be mandatory

3. In collaboration with the Research Office:
 - ❖ analyze which researchers are involved in H2020/FP7 (open access pilot?) in the last three years
 - ❖ contacting them with an e-mail to understand if they are collecting/have collected the data for their research projects, asking them whether they can cooperate in a survey

- ❖ send a short survey of 4-5 questions to those who have expressed interest:
 - Where do you usually store research data?
 - Would you like to have at your disposal a help service for cleaning, data processing and data storage?
 - Do you know what data FAIRness is?
 - Are you interested in making your data findable, accessible, interoperable and reusable (FAIR)?
 - How could we help with the Research Data Management?
- 4. In collaboration with the Phd Students School, propose some courses with very practical cut, using the Author Carpentry and Library Carpentry materials
 - ❖ Persistent identifier: importance and use. How to manage ORCID and DOI
 - ❖ Publishing: licenses and their importance. Examples from our repository
 - ❖ Publishing: how to choose the best journal vs. predatory publishing
 - ❖ Research Data Management: RCR, Open Data and FAIR Data
 - ❖ Research Data Management: tools and strategies (see Venkat's and Louise's presentations)

2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

First, I hope to be able to use some of the tools and materials used during the Summer School right away:

- Rmarkdown to review the Repository Policy and to make a web page on the example of the University of Bath Research Data Policy
- Carpentries' Materials to prepare courses for Phd Students

I think that also the dialogue with other internal structures (Research Office and Phd school) could lead to a greater awareness of the importance of managing all aspects of research in a coordinated way, starting from attribution and credits.

Speaking with research would let me know if there is the awareness of the growing importance of Data Research Management. Whether positive or negative the awareness would be, I would look in every way I can to spread the knowledge of the concepts deepened during the Summer School, asking for the possibility to insert some ad hoc news on Library Service Homepage as well as on Repository Homepage.

I hope that also the Phd courses will spread a more conscious awareness of data importance and of Open Data and FAIR Data concepts.

I am also considering the possibility of participating in online courses to certify my path towards the acquisition of the skills of data steward, but I would like to ask you some guidance to choose the best way to do this.

Actually I am not sure even if the costs are worth the benefits, since it is not clear to me if a certification like the one proposed by <https://ecm.elearningcurve.com> or <https://dama.org/content/cdmp-status> can really be useful.

Sanjin Muftić Data Stewardship Action Plan

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?

1. Re-design the Website

a. We have begun work on redoing our department website - Digital Library Services - DLS and we have had lots of discussions around the content. This debate has been centered around how much info is necessary, how open should we be, how to arrange all the services that we offer so that users can find what they need. Throughout the school, as much as the content presented was very important, so was the way in which it was presented (see below). Also, because the school was a wonderful wholesome overview of the field of Data Stewardship and Data Science it made me aware of all the little pockets of information and tools that are available and that need to be shared. It also made me aware of the context in which people will use these tools and how to best present them. It was a view you find on top a mountain. Throughout the two weeks, but more especially during the data steward stream, next to my notes I was marking down which little phrases, links, tools that would be good to include on our website to help others. I think the overall approach on one of the days of teaching RDM was great which was expressed something like: "this is what RDM is, in all detail, but you will not tell it to researchers ALL of this." The implication being that there need to be approaches to communicate this knowledge in a way that is easy to digest. And for me, even though there was not a specific amount of focus on this, it became important to think how it translates to the written guides and the digital material. There is a wealth of open science info out there, but part of what the Data Steward should do is to distill it or adjust it to suit the audience that they are working with. In this case of my context, and am sure many others, that would include researchers, librarians and students (as evidenced by our role-playing exercises). So with the website I hope to contribute to my team's writing to find the right tone and approach with our audience and also make sure that the most appropriate content is immediately accessible. I also want to integrate the Rmarkdown skills we obtained to work on our quickguides for certain of our services and tools, allowing us to make quicker changes to that material.

2. Develop Suitable Teaching/Training Formats on Data Stewardship

a. Catering for a number of audiences in terms of spreading the practices of data stewardship is quite a challenge. The material needs to be presented in different ways within diverse forums and various audiences. What was really wonderful at the data school was the number of lecturers and having a chance to engage with everybody's teaching approach in terms of connecting with the audience and how they structured their content. It was possible to take a step back and observe which approaches suited which people and content better. Is it better to allow

students to ask questions, or do you work with green and pink stickers? How much content do you put on your slides, and how much do you deliver in words only? Do you start at the beginning or at the “end” to hook people in? Do you give a lot of theory or make people do things immediately to get the experience? Of course there is no one right approach, but the opportunity to see different ones in action was eye-opening and then to cross-match them with the content highlighted how much a data steward - in terms of advocacy - should be flexible in their approach. I think more than anything a data steward should be able to switch between the theory aspect and an interactive workshop environment seamlessly to keep the flow of advocacy. It might sound obvious but simply showing a tool and what it does - without providing the larger motivations for using the tool - might fail to connect with the audience. The other important thing is choosing how much information to share before it might become an overload, thus knowing when not to cross that line that might stop people from engaging. It made me rethink and make sure that I plan different strategies when doing outreach and advocacy programmes.

3. Increase Participation within Data Steward Community

- a. I think the big part of the Data Steward community is that it needs to keep growing to the point where everybody is a Data Steward because they are managing their data and engaging in solid digital scholarship practices. We do quite a bit of outreach already in terms of hosting data steward meet-ups and going to present at various faculties and departments. So obviously the aim to increase the participation is only by making people aware of what they are already doing. In a strange way I found the session on linked data was useful for this, because developing a directory of people with related interests is very useful for this task, and being able to identify points of intersection between people through linked data is a digital scholarship project on its own. As requested by my manager of our team, we need to develop such a graphic that “maps” out our community of data stewards. Wanting to be part of this map could encourage others to join the community of data stewards. What we could do with this map is also connected to the common repositories, metadata standards, queries that each department gets (something we at DLS should keep track of and catalogue), so that new users can simply navigate the map to get some of their initial questions answered - either by looking at connections or by contacting one of the people along the connection list.

2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

1. A more interactive website that improves efficiency of the researchers and puts us more in contact with the research community. If we can make clear progress on how we communicate the information and the kind of information we put out - we might find that

researchers are using our resources without needing direction from us or even having to intervene on their behalf. While not eliminating the need for a personal connection and hands-on, face to face outreach, having a single resource that can assist you in multiple ways without getting you lost would improve our status within the research community as a trusted resource. I suppose if we see more traffic on our website that would be a good start.

2. On top of this, I see a bigger outcome in having more digital scholarship projects arise and take place within the university. This is more of a very long term wish but it is related to the experience gained at the school. For me the Digital Scholarship within my position encompasses all of the data steward work plus a digital humanities angle that seeks to encourage and support research projects which use digital tools and technologies. I think partly the potential of engaging with R markdown, and the semantic web has made me realize how the awareness of these tools could save a lot of future work for researchers, but also encourage them to do some interesting projects by being aware of these tools in the first place.
3. Overall the aim to grow the data stewards community in three ways - in the library among staff, in the faculty among lecturers and researchers, and in the student body. Having a number of different data stewards in different fields and disciplines would ensure that it is not necessary for one data steward to serve the needs of an entire multifaceted research body. Instead the data steward might be a first port of call who can direct research to those who are more equipped to deal with the query and give the most appropriate workflow, tool or procedure. Building a data culture through the website, advocacy, a visual registry would support all of this. I don't know if the aim is to have people within DLS do less advocacy, but actually have the time to drive further investigations and discover the next steps in data stewardship while the initial steps are done with the data stewards who are students, researchers and librarians. If we empower them, we can seek further vantage points.

Sothearath Seang's data stewardship action plan:

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?

- Report back about the summer school and discuss about research data policy with my university in the beginning of 2019-2020 academic year:
 - Provide information about the programme, what it covered and how the materials and knowledge can be used in the context of the university
 - Develop a first version of research data policy
 - Call for consultation from all university members
 - Make the policy publicly available on the website
- Get researchers and other university or laboratory personnel such as librarians and IT people to be engaged in the process together:
 - Organise events such as workshops, seminars and conferences with local experts on the subject
 - List out the benefits of collaboration between the different actors in the research sphere of the university and of having a concrete research data management plan
 - Find and summarise regional and national projects and initiatives for references and as a state of the art to encourage the stakeholders to take action
- Create a working group specifically oriented towards research data:
 - Within the newly formed Open Science working group, a sub-working group should be created
 - Co-develop training courses with the regional unit of training in addition to the existing ones in Open Science general curriculum
 - Actively approach researchers and other stakeholders for surveys, discussions and help concerning any issues they are facing with research data
 - Engage PhD candidates in the topic through our local association of junior researchers

2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

- As a first objective, it is essential that the research community will become more aware of the opportunities/challenges and what are at stake for research data
- Because this university has never put their focus on this topic, the outcomes are a bit difficult to predict but I believe that making things clearer for all the people involved in the first place, it can help the community to grow and develop in the right direction

- By putting in place this action plan, I expect that more positions will be created to fill the current void of research data experts in my university and
- This should also foster more collaborations between actors locally and regionally as well as more Open Science and Open Data projects

Frans Huigen's Data Stewardship action plan

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?

From the 2019 Summer School, the most striking notion for me was that of the Responsible Science Citizenship and the responsibilities inherent to it. At DANS, I will continue to invest in being such a citizen. I will do so through the following three activities (chronologically ordered):

1. Being responsible is being open and transparent. I thus have written two pieces of reporting of the Summer School. One for our DANS-website (which I will translate later on), and one for the e-data&research journal. Its topic is Data Stewardship in the ever changing world of digital research data.
2. Together with experts in the Horizon2020 project FAIRsFAIR, I will work on what's called the FAIR Terminology. This terminology is derived from the FAIR4S framework (written in the context of the EOSCpilot project). Concrete action is to further define skills necessary for 21st century Data Stewards in a structured way.
3. I became part of the DANS Training group, specialized in giving workshops, training, lectures, about Open Science and related topics. Giving a good workshop is not a walk in the park. Thus, I will take up training to give training. On December 4th, and on January 15th, I will receive - with colleagues - a training in how to give a solid workshop.

2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

For the first two: yes, concrete outcomes, namely:

1. Reporting of the event in Trieste, referring further to useful website and publications on the subject. This will create digital traffic and ensure attention is given to the important topic.
2. An iterative process of understanding, and growing of a structured list of FAIR skills necessary for Data Stewards. Through it, the FAIR principles will be concretized, operationalized; ready to be implemented in necessary curricula.

For the third, the outcome will be less concrete but nevertheless crucial. In following workshops and investing in my professional growth, I will become an expert in the field and improve my soft skills.

Lina Sitz action plan:

1. Please tell us what you plan to put into practice when you get back to your home institution as a result of taking part in this training. Specifically, what are three things you would like to achieve within the next six months?

After taking part in this training I realised that in my institution, even if there are a lot of actions devoted to promote open science and data sharing, we don't have a common way of dealing with our research data.

So I would like to achieve the following activities:

- 1) Try to help the scientist and IT of our pilot project to fulfill of "good practices" the use case. So this use case can become an "example" to help other scientist at the ICTP in the process of sharing their research data.
- 2) Use the RISE tool, in collaboration with scientists and directives, in order to identify the "real" gaps in my institution in RDM offering. I believe that in the case of my institute the chart will be close to a dot :), but I'm planning to use the options of any section in order to understand which are the real objectives and needs in my institute.
- 3) Make a list of actions needed in order to develop a RDM roadmap.

2. Are there any specific outcomes that you anticipate from these actions? If so, can you please describe them?

- 1) A FAIR shared dataset and a "real example" (with real issues and the possible solutions, in order to take "FAIR decisions") to help other scientists from and visiting our centre in the process of sharing their research data.
- 2) A frame to define the actions to take in order to provide a RDM service in our institute.