IJDC | *Conference Pre-print*

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
DANS

Hugh Shanahan
Royal Holloway, University of London

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 is provided to the origin of 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 by the time of the IDCC meeting will have been held in Trieste in August 2019. The paper will discuss the overall curriculum for the pilot school, how it matches with the FAIR4S framework, and plans for getting feedback from the students.

Finally, the paper discuss future plans for the school, in particular how to deepen the integration between the Data Stewardship strand with the Early Career Researcher strand.

.

Submitted 8 July 2019 ~ Accepted 19 February 2020

Correspondence should be addressed to Hugh Shanahan, Postal Address Department of Computer Science, Royal Holloway, University of London, TW20 0EX, UK. Email: Hugh.Shanahan@rhul.ac.uk

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

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).

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 (Shanahan, 1993)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[1] (Whyte, A, 2019; Whyte, Angus et al., n.d.). 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).

[1] https://www.eoscpilot.eu/sites/default/files/fair4s_eoscpilot_skills_framework.pdf

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 Science (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, Author Carpentry (Caltech Library, 2017), Data Visualisation, Machine Learning and Computational Infrastructures. Some schools also delivered material on Information Security and in December 2019 the school in San José, Costa Rica will use Python as the main language. By February 2020, nine schools will have been run on three continents to students from over 40 countries.

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 will be run in parallel with the CODATA-RDA school at the International Centre for Theoretical Physics. A cohort of five students with a Data Stewarding background will be taught in the pilot school. The programme is summarised in table 1. In the first week the Data Stewards attend the same modules as the ECRs. For approximately 80% of the second week they attend their own modules with some key overlap with the ECRs.

Table 1. Programme for Pilot school. Week 1 is shared with ECRs and Data Stewards. Courses marked with an asterix in week are also shared. 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

Week	Topics and matching FAIR4S terms
1	Open Science (AE, GA); Introduction to the Command Line (IA, CP, PD); R (IA, CP, PD, ED); Git (PR, ED); Author Carpentry (CP, AP, PR); Introduction to Research Data Management (PD, AP, ED, GA)
2	Finding and Reusing Data (CP, AP); Introduction to Information Security* (GA); Data Management Planning (PD, AP, PR, ED, GA); Data sharing and findability (PD, AP, PR, ED, GA); Data steward practice (PD, SR, AE); Train the trainers (AE); Introduction to Repositories (PD, AP, PR, SR,); Computational Infrastructures* (IA, SR); Machine Learning* (IA); Linked Data (PD,ED)

Assessment

It is essential that the pilot school gets as much assessment as possible from the students to improve its content for future events with larger numbers of data steward students. During each CODATA-RDA school, feedback is gathered from its students at the end of each day, through the use of post its and online surveys. The curriculum is reviewed on a regular basis as iterative improvement is a powerful technique (Wolf, 2007). This will also be carried out for the students of the Data Steward pilot. Students will receive a more formal questionnaire at the end of the school to get insights on how to improve the curriculum. The students will also be asked to complete questionnaires on a longitudinal basis to determine the overall impact that the school has had on their own practice as a Data Steward. The initial results of this will be discussed at the IDCC conference.

Future Plans

The planned school for August 2019 represents a first step; focusing on ensuring the draft curriculum is apposite for the Data Stewards. ECR and Data Steward students will have an opportunity to meet and work with each other during the first week of the school and learn more about how they can work together to support reproducible research. While this pilot will provide integration points between the ECR and Data Stewardship strands, the groups will not have the opportunity to fully explore the complementary roles that they have. The pilot will help to identify optimal points for integration and shared activities to be delivered during future iterations of the school. This can 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.

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 - Press release - G20 Leaders' Communique Hangzhou Summit. (2016, September 5). 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

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

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. (n.d.). *EOSCpilot D7.5 Strategy for Sustainable Development of Skills and Capabilities*. Retrieved from https://eoscpilot.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

Wolf, P. (2007). A model for facilitating curriculum development in higher education: A faculty-driven, data-informed, and educational developer–supported approach. *New Directions for Teaching and Learning*, *2007*(112), 15–20. https://doi.org/10.1002/tl.294