FAIRsFAIR

Fostering Fair Data Practices in Europe

Project Title	Fostering FAIR Data Practices in Europe
Project Acronym	FAIRsFAIR
Grant Agreement No	831558
Instrument	H2020-INFRAEOSC-2018-4
Topic	INFRAEOSC-05-2018-2019 Support to the EOSC Governance
Start Date of Project	1st March 2019
Duration of Project	36 months
Project Website	www.fairsfair.eu

D6.5 Report on Three Annual schools in core data skills for researchers

Work Package	WP6
Lead Author (Org)	Louise Bezuidenhout (DANS)
Contributing Author(s) (Org)	Hugh Shanahan (RHUL)
Due Date	28.02.2022
Date	18.02.2022
Version	1.0
DOI	10.5281/zenodo.6074588

Dissemination Level

☒ PU: Public

□ PP: Restricted to other programme participants (including the Commission)

□ RE: Restricted to a group specified by the consortium (including the Commission)

☐ CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	21.11.2021	Louise Bezuidenhout	Initial draft
0.2	17.12.2021	Hugh Shanahan	Formatted to match report standard. Added Executive summary, introduction and conclusions.
0.9	20.12.2021	Hugh Shanahan	Final corrections. Ready for internal review.
	21.01.2022	Claudia Engledhardt(UGOE)	Internal Review
	10.02.2022	Perdro Principe (UMinho)	Internal Review
1.0	18.02.2022	Hugh Shanahan	Content Ready

Disclaimer

FAIRsFAIR has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no. 831558 The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Abbreviations and Acronyms

CODATA	The Committee on Data for Science and Technology
ECR	Early Career Researcher
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
HEI	Higher Education Institution
ICTP	International Centre for Theoretical Physics
RDA	Research Data Alliance
RI	Research Institution
VLE	Virtual Learning Environment

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

Executive Summary

This report provides a summary of the three schools run for early career researchers (ECRs) during the FAIRsFAIR project. The curriculum provides a broad but shallow introduction to the necessary technical and social skills in Data Science for Early Career Researchers. In particular there are modules in Open and Responsible Research, Software Carpentry (the Unix command line, R or Python, git), Research Data Management, Visualisation, Information Security, Machine Learning, Author Carpentry and Computational Infrastructures. The schools were originally designed to be an intensive two week school run on a face to face basis.

Over the three year period there was a transition (because of the necessary pivot to online teaching) from

- a face to face school (2019) which also ran a pilot Data Steward Instructor training event in parallel covering the above topics over two weeks.
- A transitional school (2020) for alumni with an emphasis on encouraging the alumni to teach what they have learnt, which was carried out virtually over two weeks with some video materials and discussion sessions.
- A fully virtual school (2021), with video materials, exercises, forums for questions and live sessions that ran over 11 weeks. This has an equivalent number of learning hours as the original face to face school.

All of these were hosted at the ICTP in Trieste with the first being physically hosted there and the latter years being virtually hosted through the services that they provided. A coherent set of teaching materials for online delivery is now available as a result of this including an image of the Moodle VLE of the final taught course with videos.

The schools activities are valuable and should go on in the future. Hence it is necessary to consider the sustainability of the schools. This includes the development of an Advisory Board, the development of a variety of different business models and an external assessment of the value of the schools.

Table of contents

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

1. Introduction

These schools provide early career researchers with the core skills that will allow them to be more effective and efficient in the work they do, to analyse the data at their disposal and to take advantage of the emerging culture of Open Science.

Contemporary research – particularly when addressing the most significant, transdisciplinary research challenges – cannot effectively be done without a range of skills relating to data. This includes the principles and practice of Open Science and research data management and curation, the use of a range of data platforms and infrastructures, large scale analysis, statistics, visualisation and modelling techniques, software development and data annotation. This is defined here 'Research Data Science' as the ensemble of these skills.

Research Data Science skills are common to all disciplines and training in 'Research Data Science' needs to take this into account. For example, all disciplines need to ensure that research is reproducible and that provenance is documented reliably and this requires a transformation in practice and the promotion of the necessary culture, practice and skills.

The CODATA-RDA Schools for Research Data Science have a long-standing relationship with the International Centre for Theoretical Physics (ICTP). The first school was held at the ICTP Trieste campus in 2016, and further schools were held annually between 2017 and 2019. These schools were residential, making use of the ICTP computer labs and lecture theatres, as well as accommodation and catering. Each school lasted two weeks and involved around 50 students and a pool of roughly 20 instructors and helpers. The schools covered a diverse "broad and shallow" curriculum of data science using open training materials developed by the volunteer instructors (Shanahan et al., 2019).

In partnership with FAIRsFAIR, three of these schools were scheduled to be run during the project.In addition to these schools, the possibility of franchising the schools, where they are run by separate was to be explored - the findings of that is discussed elsewhere (Shanahan *et al.*, 2022). Largely, the plan was to run the schools with little change in the delivery or curriculum. The COVID-19 pandemic overturned and required a radical rethink of the delivery of the schools. Therefore, at the end of the FAIRsFAIR project an entirely new delivery of the schools was developed.

1.1 The curriculum

The initial curriculum for the schools was an official output of the RDA (Shanahan *et al.* 2019). The final version of the curriculum is summarised in figure 1. Overall the school provides a broad introduction to a variety of topics that cover the technical and social aspects of Data Science skills for researchers. Open and responsible research is based on seminars and discussion groups of the students to determine how much of open research principles can be taken up in their labs. All the

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

other modules have an ethical exercise associated with it to bring it back to this specific topic (Bezuidenhout *et al.* 2020). Software Carpentry is based on the modules provided by the Carpentries on the Unix command line, the programming languages R or Python and the tool git. For the rest of the school the same programming language is used. Research Data Management provides a necessary introduction to the relevant topics in this area for researchers, including a discussion of FAIR. Visualisation provides an introduction to data visualisation using a specific programmatic library such as ggplot2 or seaborn. Information Security focuses on describing possible Information Security issues associated with opening up data and software and practical steps to address that. Author Carpentry describes authorship in the 21st Century with an introduction to tools such as curl and markdown. FInally Computational Infrastructures introduces computing beyond purely local computing a researcher may have, such as cloud computing platforms and tools such as singularity for porting their environments onto such platforms.

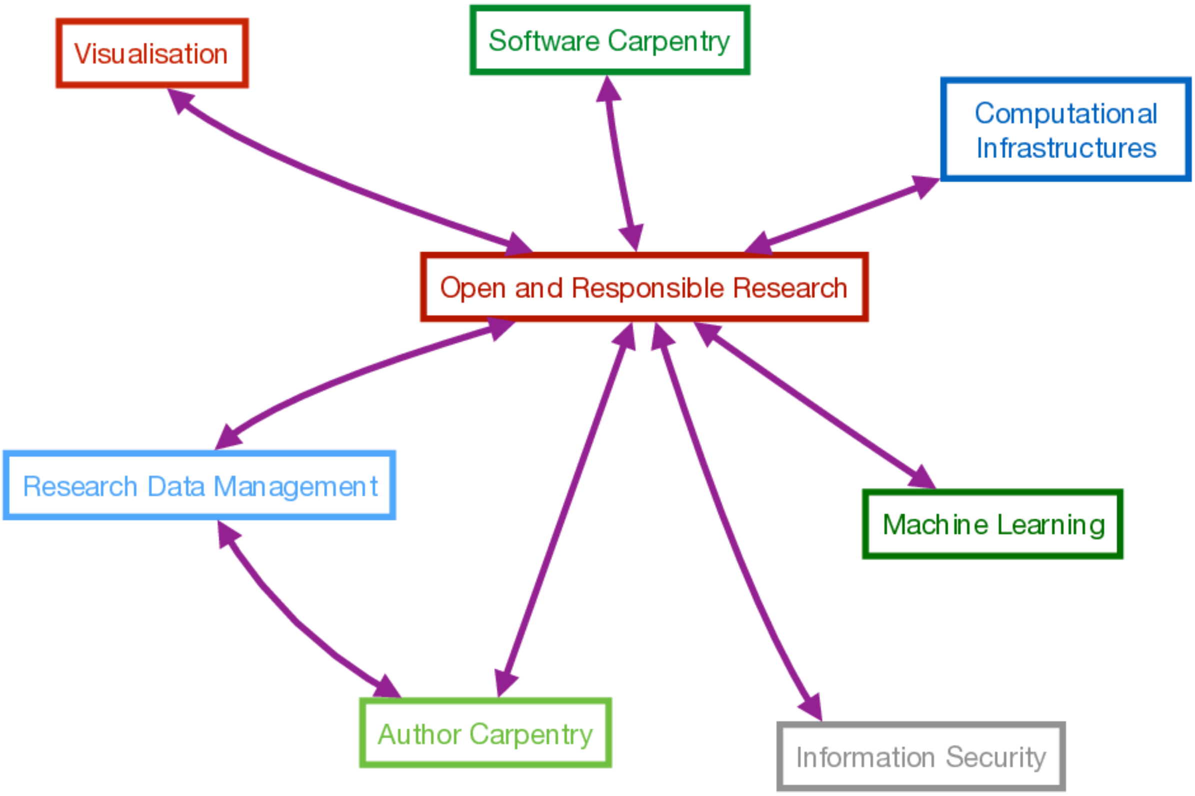

Figure 1 diagrammatic representation of schools curriculum expressed in terms of individual modules. Arrows indicate the connections between the individual modules.

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

2 The schools

2.1 2019 school - physical school

This school ran along the same lines as the previous schools as outlined above. The programme is listed in Annex 1. It was run from Aug 5 - Aug 16th 2019 at the ICTP in Trieste, Italy. In parallel with the school, the first Data Steward Instructor training was also run, where the attendees did the first week of the school and then did a separate week of activities. Further information on the Data Steward Instructor training is provided in the report D6.6 Data Steward Instructor Training (Davidson *et al.* 2022). An attendee would have had 72 contact hours in total during the school.

2.2 2020 school - intermediate virtual school

The COVID-19 pandemic made it impossible to run in-person schools after 2019, and the co-chairs took the decision to transition to an online format for the annual Trieste schools. In 2020, an abridged version of the school was run online as a train-the-trainer event running from Aug 31 - Sep 11 2020. The programme can be found in Annexe 2. The school was still hosted by the ICTP in Trieste (e.g. using their zoom service). This school was open to alumni and included curriculum refresher lectures and discussions on training on the different modules. Three short videos on each module were provided that attendees were expected to watch prior to the discussion sessions. One gave a recap of the material covered in the module, another described a specific challenge associated with teaching that module and finally a description of what would be include in a more advanced version of the module. This was accompanied by two live one hour sessions to discuss these videos. The school ran over ten weeks with five hours of video material and 20 hours of meetings.

2.3 2021 school - full virtual school

In 2021, the first fully online version of the annual school was run. It ran from Sep 6 - Nov Nov 19 2021. The programme can be found in Annexe 3. This was hosted by the ICTP using their Moodle platform and Zoom (this latter point was particularly important as Zoom is normally subject to constraints on usage from a number of countries - the ICTP have arranged a concession from Zoom to provide access). Being supported by FAIRsFAIR, this was open not only to LMIC attendees but also

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

European ECRs.   The structure of this school followed a previous online school that had been hosted earlier in the year in conjunction with the University of Pretoria in South Africa. The online school ran over 11 weeks, with each week focusing on a different module within the curriculum. Students followed the curriculum using pre-recorded videos and self-paced exercises. Support was provided through live “helper sessions”, two weekly discussion sessions with the instructors and a forum for the attendees to raise questions. The materials were hosted on a moodle VLE run by the ICTP service. The image of the moodle VLE can be found here. This can be uploaded onto any other VLE running moodle. Each attendee had in total 72 learning hours, consisting of materials for attendees to watch, live sessions to attend and time for exercises that the students needed to do.

3 Conclusions

The ongoing benefits of the schools demonstrate that its activities should carry on beyond the lifetime of the FAIRsFAIR project. With this in mind an Advisory Board was set up which will advise the schools on how to become financially sustainable. The Terms of Reference for the Advisory Board are listed in Annexe 4. In addition an additional meeting was held to consider sustainability. Two points were raised there - the schools should consider a pathway to becoming its own non-profit organisation and should seek funding for an external appraisal of the value it brings to make the case for future support.

The CODATA-RDA Schools for Research Data Science model has been well-received by the international data science community. For example, the schools were used in a case study in OECD report on digital workforce capacity (OECD, 2020). The curriculum has recently been recognised as an official output of the Research Data Alliance (RDA) (Shanahan et al., 2019). A number of partner institutions who have previously co-hosted regional schools, such as the University of Pretoria, are committed to hosting annual schools. In addition, the success of the school curriculum and both models of delivery has led to it being considered as a formal training partner in the roll-out of the African Open Science Platform.

In table 1, we see a summary of registered attendees of the schools. We note the large number of attendees in 2020 is due to the school being open to all of the alumni.

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

Table 1 - summary of registered attendees of the schools

	Number of attendees	Number of Countries Represented
2019 (in person)	38	18
2020 (online)	123	42
2021 (online)	76 (8 EU, 5 UK)	32 (6 EU countries and UK)
Total	**243**	**56** (unique countries)

Despite the enormous challenges, there is now a complementary set of online materials that can be used for similar training events.

A longitudinal study of attendees has demonstrated take up of the materials and that the attendees stay in touch with each other and continue to learn with each other (Bezuidenhout *et al.,* 2021).

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

4. Bibliography

Bezuidenhout, L., Quick, R., Shanahan, H. 2020, Software and Engineering Ethics, 26 (4), 2189-2213.

Bezuidenhout, L., Drummond-Curtis, S., Walker, B., Shanahan, H., Alfaro-Córdoba, M., 2021, A School *and* a Network: CODATA-RDA Data Science Summer Schools Alumni Survey, Data Science Journal, 20 (1)

Davidson, J., Bezuidenhout, L., Newbold, E., Shanahan, H., Walker, B., Yates, K., 2022, Data Steward Instructor Training, https://doi.org/10.5281/zenodo.6074458

OECD, 2020, Building digital workforce capacity and skills for data-intensive science, OECD Science, Technology and Industry Policy Papers, No. 90, OECD Publishing, Paris, https://doi.org/10.1787/e08aa3bb-en

Shanahan, H., Newbold, E., Davidson, J., 2022, Report on schools run through franchising with local organisers, https://doi.org/10.5281/zenodo.6043906

Shanahan, H., Quick, R., Córdoba, M.A., Clement, G., Bezuidenhout, L., Shanmugasundaram, V., Jones, S., Ashley, K., Diggs, S., Gillespie, C., El Jadid, S., Sorokina, M., Barlow, R., Okorafor, E., Sipos, G., Constantini, A., Short, H., 2019. Submission of curriculum specification of CODATA-RDA Research Data Science schools as an RDA output. https://doi.org/10.5281/zenodo.3478590

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

Annex 1 - Programme for 2019

Summary		
	Common for ECR and Data Steward attendees	
	For ECR attendees only	
	For Data Steward attendees only	
Open Science & RCR Overview 1	August 5	Day 1 am
Shell		Day 1 pm
Ethics exercise for Shell		Day 1 pm
Git	August 6	Day 2 am
Ethics exercise for Git	August 6	Day 2 am
R		Day 2 pm
R	August 7	Day 3 am
R		Day 3 pm
Ethics exercise for R		Day 3 pm
RDM and OS and DMPs 1	August 8	Day 4 am
RDM and OS and DMPs 2		Day 4 am
Author Carpentry: Reproducible Reportng (using RStudio)		Day 4 pm
How to get DOI for your report and deposit it into repositories		Day 4 pm
Predatory publishing (optional - remote call)		Day 4 pm
Author Carpentry Copyright and Data Licensing around disseminating your report ("A Dramatic Interpretation in 3 parts")	August 9	Day 5 am
Open Science 2		Day 5 am
RDM lab		Day 5 pm
RDM lab		Day 5 pm
Visualisation with R	August 12	Day 8 am/pm
Ethics exercise for visualisation		
Introduction to Data Stewardship; Finding and reusing data; Data management planning		Day 8 am

Data sharing and findability		Day 8 pm
Information Security		Day 8 pm
Information Security - Advanced Topics (Optional)		Day 8 pm
Machine Learning Overview - Recommendation	August13	Day 9 am
Recommender Systems		Day 9 am/pm
Soft skills for Data Stewards		Day 9 am
Data discovery, automated DMP's		Day 9 pm
Repositories		Day 9 pm
PID's and the Digital Object architecture (optional)		Day 9 pm
Repositories	August 14	Day 10 am
DMP's		Day 10 pm
Train the trainer game		
Artificial Neural Networks	August 14	Day 10 am/pm
Ethics exercise for Artifiical Neural Networks		
RDA Outputs (Optional)		Day 10 pm
Other Machine-learning systems	August 15	Day 11 am
Research Computational Infrastructure		Day 11 am/pm
Research Computational Infrastructure	August 16	Day 12 am
Ethics exercise for Research Computational Infrastructure		
Linked Data; Sparql queries		Day 12 am
Closing ceremony		Day 12 am

Annex 2 - Programme for 2020

	session number	Contents
Monday August 31	1	RDM
Tuesday September 1	2	Open and Responsible Research
Wednesday September 2	3	Author Carpentry
Thursday September 3	4	Visualisation
Friday September 4	5	Carpentries
Monday September 7	6	Machine Learning
Tuesday September 8	7	ANN
Wednesday September 9	8	Info. Sec.
Thursday September 10	9	Comp. Infr.
Friday September 11	10	Planning (how to run sessions)

Sessions run 07:00-08:00 UTC and 14:00-15:00 UTC

Annex 3 - Programme for 2021

	Date from	Date to	Modules
1	6-Sept	10-Sept	OS / Working with spreadsheets/ Open Refine
2	13-Sept	17-Sept	R
3	20-Sept	24-Sept	R
4	27-Sept	1-Oct	Visualisation / Being Open and Responsible at home
5	4-Oct	8-Oct	Machine Learning / Neural Networks
6	11-Oct	15-Oct	Neural Networks
7	18-Oct	22-Oct	RDM
8	25-Oct	29-Oct	Data/Information security
9	1-Nov	5-Nov	Unix / Git
10	8-Nov	12-Nov	National Infrastructure / International Infrastructure
11	15-Nov	19-Nov	Responsible authorship/ Reproducible writing / Reputation Management

Live sessions run 14:00-15:00 UTC on Wednesdays and Fridays

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

Annexe 4

CODATA – RDA Data Schools Advisory Board
Terms of Reference

0. Context

The CODATA – RDA Data Schools have been active since 2016 and have delivered 12 schools in 5 different continents, to over 500 students, delivering volunteer teaching and training for the equivalent of over half a million US dollars. All teaching to date has been 100% volunteer.
“Data Schools” refers to the activities of the CODATA- RDA schools of Research Data Science, including the early careers schools, the data steward instructor training, and other activities as defined on the website datascienceschools.org
“Founding organisations” refers to CODATA (the Committee on Data of the International Science Council (ISC)) and RDA (the Research Data Alliance).
“Data Schools Co-chairs” refers to the co-chairs of the CODATA-RDA Schools of Research Data Science as listed at https://www.datascienceschools.org/members/

These Terms of Reference (ToR) set out the working arrangements for the CODATA – RDA Data Schools Advisory Board (AB) and list information about the AB, its purpose, chair and membership, meeting schedule, level of administrative support, and dispute resolution processes.

1. Role/Purpose

The role of CODATA – RDA Data Schools AB is to:

- Assist and advise the Data Schools and the ‘founding organisations’ with planning and implementing a sustainability and expansion model.
- Advise and oversee strategic development of the schools and help ensure good and transparent governance.
- Advise on the development of a business plan and sustainability model.
- In due course, advise on any recruitment which may need to be taken (understanding that the specific appointment may be made by a partner institution)

2. Term & Renewal

Advisory Board members are invited to serve for a two year renewable term. Members have been nominated by the founding organisations – CODATA and RDA – and by the co-chairs of the Data Schools. A member may serve for a maximum of two consecutive two-year terms. At two year intervals 50% of the board is expected to step down or come to the end of their maximum terms.

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

3. Membership & Chair

The AB is composed of a minimum nine and a maximum of fifteen members. Where possible, the Advisory Board will be composed of members that ensure gender, geographical, organizational and knowledge / expertise balance.

Each year, the AB will select one chair to serve a one-year term, at the first AB meeting. A chair may only serve one term.

CODATA and RDA Secretariats will each have one *ex officio* member. At least two Data Schools Co-chairs will attend meetings in a non-voting role.

4. Roles, Responsibilities and Expectations

The advisory board will provide support and advice on specific areas which include:

- insight and advice on the definition of priority areas and targets
- supporting the definition of 5-year Data Schools roadmap
- helping maintain the focus and activities of the Data Schools on the agreed scope, outcomes and benefits
- Identifying factors outside the data school's control that are critical to its success.
- fostering collaboration
- advice to enable successful delivery, adoption and franchising
- identification of potential funders and supporters

The members of the advisory board commit to:

- attending all scheduled Advisory Board meetings and if necessary nominate a proxy
- wholeheartedly championing the data schools within and outside of work areas
- sharing all communications and information across all Advisory Board members
- making timely decisions and taking action so as to not impede progress
- notifying members of the Advisory Board, as soon as practical, if any matter arises which may be deemed to affect the development of the Data Schools

Members of the advisory board will expect:

- that each member will be provided with complete, accurate and meaningful information in a timely manner
- to be given reasonable time to make key decisions
- to be alerted to potential risks and issues that could impact the Data Schools, as they arise
- consensus driven discussions and decisions
- open and honest discussions, without resort to any misleading assertions
- ongoing 'health checks' to verify the overall status and 'health' of the data schools.

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

5. Meetings

All meetings will be chaired by the chair or co-chairs to be appointed.
A meeting quorum will be 50% of the named members of the advisory board
Decisions made by consensus (i.e. members are satisfied with the decision even though it may not be their first choice). If not possible, advisory board chair makes final decision
Meeting agendas and minutes will be provided by the CODATA and RDA Secretariat representatives in rotation, this includes:

- preparing agendas and supporting papers
- preparing meeting notes and information.

Meetings will be held every six months and may take place in-person or on-line (virtual).
If required subgroup meetings will be arranged outside of these times at a time convenient to subgroup members.

6. Amendment, Modification or Variation

These Terms of Reference may be amended, varied or modified in writing after consultation and agreement by Advisory Board members.

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