Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D4.2

Periodic Training Report and Planning for the next Period 1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData-PlusPlus
PROJECT TITLE	SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics
STARTING DATE	01/01/2020 (60 months)
ENDING DATE	31/12/2024
PROJECT WEBSITE	http://www.sobigdata.eu
TOPIC	INFRAIA-01-2018-2019 Integrating Activities for Advanced Communities
GRANT AGREEMENT N.	871042

DELIVERABLE INFORMATION	
WORK PACKAGE	WP4 NA3 Training
WORK PACKAGE LEADER	KCL
WORK PACKAGE PARTICIPANTS	CNR, USFD, UNIPI, FRH, UT, IMT, LUH, SNS, ETH Zürich, UNIROMA1, CNRS, URV, KTH, SSSA
DELIVERABLE NUMBER	D4.2
DELIVERABLE TITLE	Periodic Training Report and Planning for the next Period 1
AUTHOR(S)	Mark Coté (KCL), Marco Braghieri (KCL), Beatrice Rapisarda (CNR)
CONTRIBUTOR(S)	Giulio Rossetti (CNR), Fabrizio Lillo (SNS), Tommaso Venturini (CNRS), Axel Meunier (CNRS)
EDITOR(S)	Beatrice Rapisarda (CNR)
REVIEWER(S)	Jesús A. Manjón Paniagua (URV), Valerio Grossi (CNR)
CONTRACTUAL DELIVERY DATE	31/12/2021
ACTUAL DELIVERY DATE	22/12/2021
VERSION	1.0
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	48
KEYWORDS	Online Learning, In-person events, Covid-19, Online training materials, summer schools, datathons, diversity

EXECUTIVE SUMMARY

This deliverable provides an overview on training activities performed between M6 and M24 of the SoBigData++ project and planning for the forthcoming reporting period (M24-M48). Moreover, it includes an assessment of the impact of the Covid-19 pandemic on the training within the SoBigData++ community. Each of the Work Package's four tasks is divided into a 'reporting' and 'planning' section in order to facilitate the assessment of performed and planned activities.

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042.

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance on such ambitious tasks thanks to SoBigData, the predecessor project that started this construction in 2015. Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments.

This document contains information on SoBigData++ core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData++ Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The content of this publication is the sole responsibility of the SoBigData++ Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

GLOSSARY

EU	European Union
EC	European Commission
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
Covid-19	Sars2 Coronavirus
WHO	World Health Organisation
MOOC	Massive Open Online Courses
STS	Science and Technology Studies
XAI	eXplainable AI
MS	Master of Science
VRE	Virtual Research Environment
AI	Artificial Intelligence
XDMS	eXplainable Decision-Support Making

TABLE OF CONTENTS

1 Relevance to SoBigData++

Work Package 4, Training, aims to establish a joint training and education resource on big social data promoting the education of the next generation of data science researchers. The Work Package explores and develops both conventional and unconventional training experiences for master students, PhD students and early career post-doctoral researchers as well as an academically interested general public. Likewise, Work Package 4 proposes campaigns aimed promoting interest and participation of under-represented communities in data science with special emphasis on gender issues.

1.1 Purpose of this document

This document provides an overview of the planned activities for the forthcoming reporting period and offers an overview of the activities that have already taken place, organised, and performed by Work Package 4 between M6 and M24.

In order to provide further comprehension on the impact of the Covid-19 pandemic on training, we have included a questionnaire that aims to provide insight into how the SoBigData++ community has been impacted by the ongoing pandemic in relation to training activities (See Section 2).

The document provides an overview of performed and planned activities during the reporting period, following the Work Package's task structure (See Section 1.2) which include T4.1 (Online Training Modules), T4.2 (Summer Schools), T4.3 (Datathons) and T4.4 (Cultivating Diversity in Data Science Through Training).

1.2 Relevance to project objectives

The training activity within the SoBigData++ project is developing a unique, joint training and education resource centre on big social data. Building on the experience of the first iteration of the SoBigData project, Work Package 4 explores and develops conventional and unconventional training experience for master and PhD students and post-doctoral trainees. These experiences include the organisation of a number of different events and the development of the e-learning Area which has been created and integrated into the SoBigData Research Infrastructure. Among events, project-oriented summer schools and datathons have been planned and organised in order to match research (and industrial) needs and people skills. Moreover, activities will aim to address gender and diversity issues in data science through training.

This Work Package is organised around four different tasks.

- Task T4.1, 'Online Training Modules' is centred on creating open-source training materials that are integrated into the SoBigData Research Infrastructure within the e-Learning Area. This part of the Research Infrastructure was designed, created, and integrated into the SoBigData catalogue during the first iteration of the SoBigData project.
- Task 4.2 is centred on the organisation of a yearly SoBigData Summer School, introducing participants to techniques and methodologies for analysing big data, in order to provide them with a solid

background in the computational and mathematical theories behind algorithmic tools for empowering their future research. The summer schools will be strongly interdisciplinary and include experts across arts and sciences.

- Task T4.3 is centred around the organisation of Datathons, with a minimum of one per year, whose aim is to bring together young and bright minds in smaller dedicated groups, providing complementary theoretical and practical skills to visualise and analyse social big data questions addressing important societal problems. All datathons will be supported by the Operational Ethics and Law Board (operated by Work Package 2) in order to include Ethical and Law aspects in the Datathon activities.

Task 4.4 Computer Science and Data Science currently fail to adequately embody staff equality and diversity issues. For instance, not only females but also minority groups, etc are still woefully underrepresented in data science. The aim is to leverage existing networks in order to raise awareness regarding the opportunities provided by employment in the field of data science. SoBigData++ will support specific events and provide travel grants for young female and minority group researchers, continuing an experience started in the first iteration of the SoBigData project.

1.3 Relation to other work packages

The SoBigData++ project is organised around work packages (*Fig.1*) which are combined in order to follow three main axes:

- Community building (including innovation and networking activities)
- Social mining research infrastructure building
- User accessibility (granted by virtual and trans-national access)

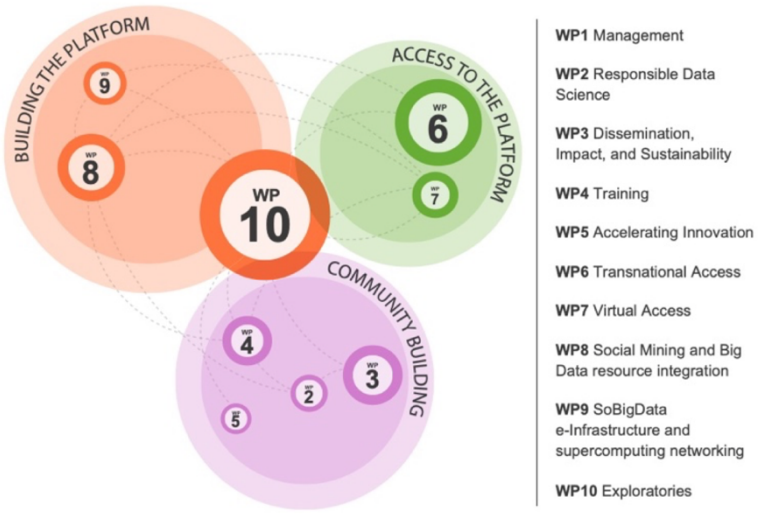

Figure 1 *SoBigData++ Work Package organization*

Among all work packages, WP2 (Responsible Data Science), WP3 (Dissemination, Impact and Sustainability), WP4 (Training) and WP5 (Accelerating Innovation) are aimed at community building between excellence centres, other academic and industrial users, and trainee data scientists. Thus, Work Package 4 works closely with:

- WP2 – Responsible Data Science – This work package is mainly tasked with operationalising a legal and ethical framework for the whole SoBigData++ Research Infrastructure.
- WP3 – Dissemination, Impact and Sustainability – This work package is mainly tasked with developing dissemination and impact strategies for the entire SoBigData++ project.
- WP5 – Accelerating Innovation – This work package is tasked with widening the project's impact through innovation activities aimed at industry and other stakeholders, such as government bodies, non-profit organisations, funders, and policy makers.

Aside from these work packages, WP4 will also work in collaboration with WP7 (Virtual Access) in order to design and integrate training modules into the SoBigData++ Research Infrastructure. Moreover, WP4 will work alongside WP9 (JRA2 - E-Infrastructure and Supercomputing Network) to create operation manuals for facilitating platform exploitation in all the aspects will be made accessible through a specialised operation portal dedicated to developers, ICT managers, and service providers. Finally, WP4 will also work alongside WP10 (JRA3 – Exploratories).

1.4 Structure of the document

This document (D4.2) entitled 'Periodic Training Report and Planning for the next Period' is divided in seven sections and features two Appendixes. Section 2 regards the effect of the Covid-19 pandemic on Training. Since the reporting period covered by the document ranges from M6 to M24, it covers entirely the period since the Sars-Cov2 virus has been declared a pandemic by the World Health Organisation. In order to assess the impact of the Covid-19 pandemic, a questionnaire was administered to the SoBigData++ community at large, to better understand the impact on teaching and training, workload, and effectiveness of distance learning. Section 3 regards 'Online Training Modules' which is Task 4.1 of WP4. This section is divided into two sub-sections, one reporting on what has been done in the reporting period and another listing actions to be undertaken in the forthcoming reporting period (M24 – M48). The same structure is used to report on Task 4.2 (Summer Schools) in Section 4, on Task 4.3 (Datathons) in Section 5 and on Task 4.4 (Cultivating Diversity in Data Science Through Training) in section 6. The report features a further section (Section 7) which reports on other forms of training events that have taken place during the reporting period (M6 – M24). Finally, Appendix A features the full questionnaire administered to the SoBigData++ community and Appendix B features the full text of a semi-structured interview that was performed within Task 4.4 activities.

2 Assessing the Effect of the Covid-19 Pandemic on Training

The current Covid-19 outbreak was declared a pandemic by the World Health Organisation on 11 March 2020 (WHO, 2020). In the reporting period covered by this deliverable (June 2020 – December 2021), a variety of different containment measures have been adopted worldwide. Many activities have been disrupted, others have moved to virtual spaces (i.e., 'work from home', 'distance learning' etc.). The impact on academic activities and academics has been severe, as 'confinement policies enacted by most countries have implied a sudden switch to home-working, a transition to online teaching and mentoring, and an adjustment of research activities' (Corbera *et al.*, 2020).

The SoBigData++ project has continued to operate under these circumstances, which have impacted many different aspects, including training activities. Work Package 4, which is devoted to training, bases its tasks on activities that traditionally take place in face-to-face settings, such as summer schools and datathons. Hence, WP4 distributed a questionnaire between 14 and 24 November 2021 in order to assess the impact of Covid-19 on training activities within the SoBigData++ community. The questionnaire is based on close-ended questions (Reja *et al.*, 2003) and includes a comment space at the end (See Appendix A). Respondents were purposedly not identified and the questionnaire was administered to 263 contacts which are part of the SoBigData++ contact list. The questionnaire totalled 21 respondents, which in 70% of cases answered all 9 close-ended questions, while the open-ended question had 7 replies.

2.1 Questionnaire Results

Our respondents were representative of all three stages of academic career, as 38,1% were Early Career Researchers, 42,9% Senior Lecturers \ Full Professors and 19% were Associate Professors \ Lecturers [Q1]. All respondents agreed that the Covid-19 pandemic has impacted their institution [Q2]. Our respondent pool developed different virtual training experiences, with synchronous lectures performed by all respondents, whereas asynchronous lectures and hands-on training materials were performed by respectively 33,3% and 23,8% of respondents [Q3].

The vast majority of our respondents (85,7%) created new content for virtual lectures and training, most being slides (88,9%), followed by Interactive Training Materials such as Python Notebooks (50%), Training Materials including wikis, MOOCs[1] (33,3%) and other types of materials (22,2%) [Q4] (these results were possible because this was a multiple-choice question). Over half of respondents plan to keep on using to some extent these materials when resuming in person teaching (55,6%), while over a third plan to keep on using them entirely (38,9%). Almost 6% of respondents do not plan to keep on using materials developed during virtual teaching \ training during the Covid-19 pandemic [Q5].

[1] MOOC stands for Massive Open Online Course

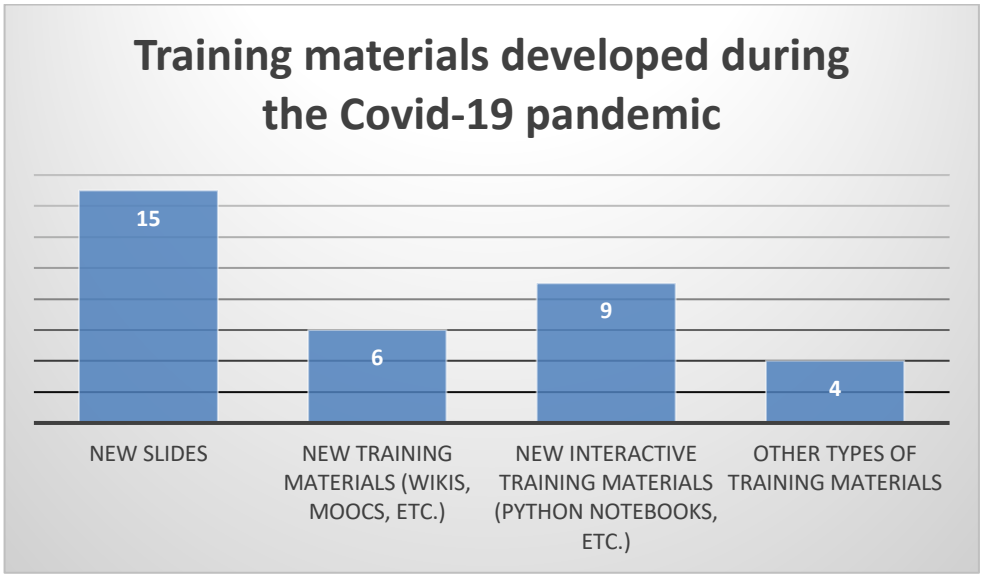

Figure 2 *Answers to Q5 of the Questionnaire 'If you have created new content, it has been (multiple answers possible)'*

As we can see in *Figure 2*, respondents have engaged in the creation of new materials and, more broadly, have perceived their role in virtual teaching and training as being more labour intensive (75%), while the rest replied that it was the same as before [Q6]. Moreover, the vast majority of respondents perceived the students and trainees to be less engaged by distance teaching training (66,7%), whereas just over a third of replies reported students to be well responding (33,3%) [Fig. 3 Answers to Q7].

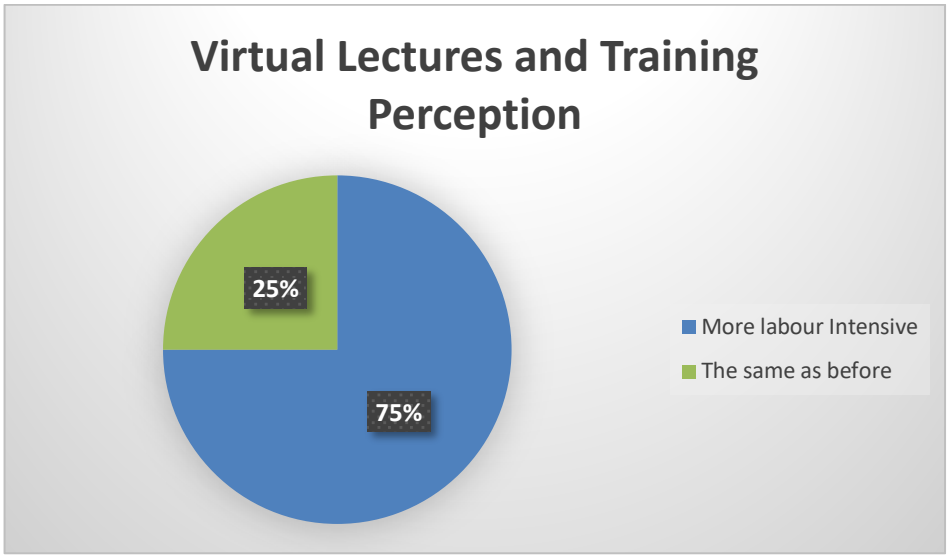

Figure 3 *Answers to Q7 of the Questionnaire: 'If you resorted to virtual lectures and training, have they been'*

As of the end of November 2021, over half of respondents have resumed in-person teaching albeit for a small number of individuals (52,4%), whereas 28,6% have not yet done so. A little under a fifth (19%) of respondents is back on a pre-pandemic schedule regarding in-person activities [Q8]. Finally, attendance to conferences and workshops was gauged by the last question: over half (57,1%) of respondents have attended

events but only in a virtual setting, 38,1% have attended both virtual and in-person events, whereas 4,8% has not attended any type of event since March 2020 [Q9].

The answers to the last, open-ended question [Q10] ('If you wish, please add a description of your teaching\training experience during the Covid-19 Pandemic'), helped assess the situation faced by academic staff. Respondent 2 (R2), described teaching \ training as 'exhausting and less effective', a perspective shared by R6 who underlined how 'For both teacher and students, the lessons are much less engaging than usual... Therefore, the lessons are much less effective than usual'. However, there were also respondents who highlighted the training material they created: 'in this past year I made approximately 30 videos, most of which I am able to reuse' (R4) and how 'teaching on-line takes more work but it allows a better teaching job than face-to-face teaching in a classroom with a physical blackboard. Slides and the electronic whiteboard are more integrated, you can show students resources on the internet, you can keep the electronic whiteboards from previous classes and show them again to the students, etc. The scores I got from students in 2021 have been among the highest I ever got. Nonetheless, students participate less in an on-line class than in a face-to-face class' (R7).

2.1.1 Questionnaire Analysis

While our respondents pool was fairly balanced among different stages of academic life between early career researchers, lecturers and full professors, results indicate that the impact of the Covid-19 pandemic has been sizeable. While there has been a strong engagement in adjusting to the virtual setting, both by resorting to the creation of new materials and adapting to virtual lectures, results also indicate that this has put added to the workload on respondents. Their answers indicate that this period has been more labour intensive and yet students seem to be – in the vast majority – less engaged than before. As of November 2021 (at the time when the questionnaire was administered), some activities have started to take place within an in-person setting, this usually regards small groups of students. Likewise, events have registered participation in a virtual setting, but are yet to resume in full in-person setting.

In broader terms, respondents' answers seem to underline the increased workload that virtual teaching has demanded, along with difficulties in participation from students \ trainees. As WP4 regards training, questionnaire results aid in assessing to what extent the current Covid-19 pandemic has impacted the SoBigData++ community.

3 Task 4.1 – Online Training Modules

Task leader: KCL

Participants: LUH, UT, ETHZ, USFD, CNR, UNIPI, SNS, UNIROMA1, URV, SSSA

During the reporting period, WP4 has been involved in the creation of the novel 'Data Literacy' Section within the SoBigData++ Research Infrastructure in collaboration with WP2, WP7 and WP9. This has allowed the harmonisation between training materials and a working environment based on a curated collection of literature of interest for the SoBigData++ Community. The collection consists of a catalogue service enacting authorized members to publish literature of interest and organize the selected contents to facilitate discovery and access. By using the SoBigData Catalogue the integration between these two entities has been seamless and research can be carried out either by word search, thematic cluster, and a multi-level tagging system.

At present, there are 41 training materials, as shown in *Table 1*, that have been uploaded into the SoBigData++ Research Infrastructure. Section 3.1 is devoted to the description of the new training materials that have been uploaded in the reporting period.

Training Materials (in alphabetical order)		
1.	Archive Crawling	https://ckan-sobigdata.d4science.org/dataset/archive_crawling
2.	Archive Spark	https://ckan-sobigdata.d4science.org/dataset/archive_spark
3.	Automated Methods of Urban Green Analysis	https://ckan-sobigdata.d4science.org/dataset/automated_methods_of_urban_green_analysis
4.	Business Data Analytics Course	https://ckan-sobigdata.d4science.org/dataset/business_data_analytics_course
5.	Can Big Data Bridge Gaps in Migration Statistics?	https://ckan-sobigdata.d4science.org/dataset/can_big_data_bridge_gaps_in_migration_statistics
6.	Complex Network Analysis Lecture	https://ckan-sobigdata.d4science.org/dataset/complex_network_analysis_lecture
7.	Compressed and Learned Data Structures Seminar	https://ckan-sobigdata.d4science.org/dataset/compressed_and_learned_data_structures_seminar
8.	Data Inquiries Initiative	https://ckan-sobigdata.d4science.org/dataset/data_inquiries
9.	Data Journalism and Story Telling	https://ckan-sobigdata.d4science.org/dataset/data_journalism_and_story_telling
10.	Data Management for Business Intelligence Module	https://ckan-sobigdata.d4science.org/dataset/data_management_for_business_intelligence_module
11.	Data Mining and Machine Learning Module	https://ckan-sobigdata.d4science.org/dataset/data_mining_and_machine_learning

12.	Data Mining and Machine Learning for Social Science	https://ckan-sobigdata.d4science.org/dataset/data_mining_and_machine_learning_for_social_science
13.	Data Visualisation and Visual Analytics Module	https://ckan-sobigdata.d4science.org/dataset/data_visualisation_and_visual_analytics_module
14.	Data in soccer: an athletic trainer's point of view	https://ckan-sobigdata.d4science.org/dataset/data_in_soccer_an_athletic_trainers_point_of_view
15.	Database Module	https://ckan-sobigdata.d4science.org/dataset/database_module
16.	Deep Learning Course	https://ckan-sobigdata.d4science.org/dataset/deep_learning_course
17.	Efficiency - Effectiveness Trade-offs in Learning to Rank	https://ckan-sobigdata.d4science.org/dataset/efficiency_-_effectiveness_trade-offs_in_learning_to_rank
18.	Ego network analysis, information-driven social links, and impact on information diffusion	https://ckan-sobigdata.d4science.org/dataset/ego_network_analysis_information-driven_social_links_and_impact_on_information_diffusion
19.	Epidemics and city. How mobility and well-being changed with COVID19 era	https://ckan-sobigdata.d4science.org/dataset/epidemics_and_city_how_mobility_and_well_being_changed_with_covid19
20.	Evaluating the significance of network observables with a maximum entropy-based approach	https://ckan-sobigdata.d4science.org/dataset/evaluating_the_significance_of_network_observables_with_a_maximum_entropy-based_approach
21.	Explaining Explanation Methods	https://ckan-sobigdata.d4science.org/dataset/explaining_explanation_methods
22.	First Awareness Panel SoBigData++	https://ckan-sobigdata.d4science.org/dataset/first_awareness_panel_sobigdata_plus_plus
23.	GATE Course	https://ckan-sobigdata.d4science.org/dataset/gate_course
24.	High Performance and Scalable Analytics Module	https://ckan-sobigdata.d4science.org/dataset/high_performance_and_scalable_analytics_module
25.	Information Retrieval Module	https://ckan-sobigdata.d4science.org/dataset/information_retrieval_module
26.	Interactive Learning Environments	https://ckan-sobigdata.d4science.org/dataset/interactive-learning-environments
27.	Introduction to Data Curation	https://ckan-sobigdata.d4science.org/dataset/introduction_to_data_curation
28.	Introduction to Data Science for Social Scientists	https://ckan-sobigdata.d4science.org/dataset/introduction_to_data_science_for_social_scientists
29.	Jupyter Notebooks	https://ckan-sobigdata.d4science.org/dataset/jupyter_notebooks
30.	Legal Materials as Big Data: (algo)Rithms Support Legal Interpretation. A Dialogue with Data Scientists	https://ckan-sobigdata.d4science.org/dataset/legal_materials_as_big_data_algo_rithms_support_legal_interpretation_a_dialogue_with_data_scientists

31.	Medical Device Regulation and Digital Health: Problems and Perspectives	https://ckan-sobigdata.d4science.org/dataset/tmedical_device_regulation_and_digital_health_problems_and_perspecti
32.	Mobility data sharing: application potential and ethical issues webinar	https://ckan-sobigdata.d4science.org/dataset/mobility_data_sharing_application_potential_and_ethical_issues_webinar
33.	SOS Online Abuse of Politicians	https://ckan-sobigdata.d4science.org/dataset/sos_online_abuse_of_politicians
34.	Platforms Data Protection & IP Issues part 1 and 2	https://sobigdata.d4science.org/catalogue-sobigdata?path=/dataset/second_sobigdata_plus_plus_awareness_panel_r_i_platforms_data https://ckan-sobigdata.d4science.org/dataset/second_sobigdata_plus_plus_awareness_panel_r_i_platforms_data_protection_and_ip_issues_2_-_part_2
35.	SoBigData++ e-infrastructure webinar	https://ckan-sobigdata.d4science.org/dataset/sobigdata_plus_plus_e-infrastructure
36.	Social Network Analysis	https://ckan-sobigdata.d4science.org/dataset/social_network_analysis
37.	Social Network Analysis @MasterBigData2021	https://ckan-sobigdata.d4science.org/dataset/social_network_analysis_masterbigdata2021
38.	Social Network Analysis with Python	https://ckan-sobigdata.d4science.org/dataset/network_diffusion_library_tutorial
39.	Text Analytics and Opinion Mining Module	https://ckan-sobigdata.d4science.org/dataset/text_analysis_and_opinion_mining
40.	Tutorial on Learning to Rank	https://ckan-sobigdata.d4science.org/dataset/research_and_software_on_learning_to_rank
41.	Visual Analytics for Data Scientists	https://ckan-sobigdata.d4science.org/dataset/visual_analytics_for_data_scientists

Table 1 ***A list of all the Training Materials currently uploaded onto the SoBigData++ Research Infrastructure***

These training materials which comprise slides, Jupyter Notebooks, videos, hands-on material now cover most of the communities that SoBigData++ aims to serve. A synergic approach has been adopted towards the SoBigData++ community, issuing periodic calls using the contacts@sobigdata.eu mailing list, in order to allow partners to contact WP4 once new training materials are ready to be uploaded on the SoBigData++ Research Infrastructure.

3.1 User Data regarding e-Learning VRE

During the reporting period, the SoBigData++ Research Infrastructure has registered the following data regarding user accesses and registrations (*Fig.4* and *Fig.5*):

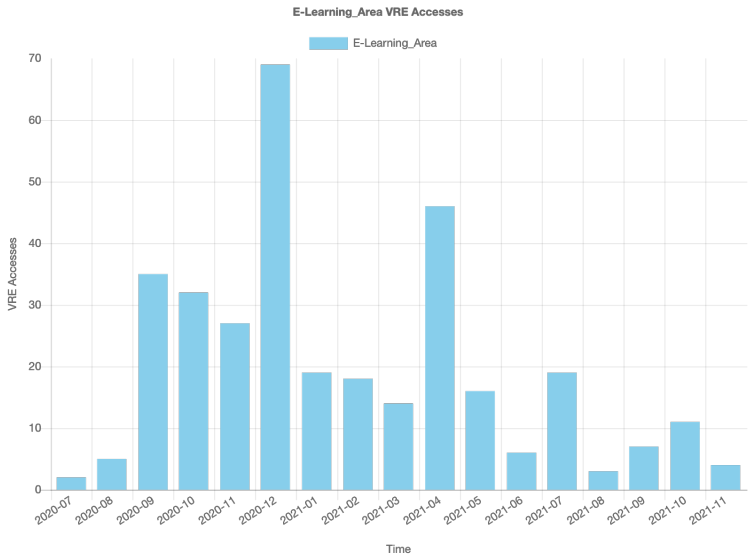

Figure 4 ***User accesses to the SoBigData++ RI e-Learning VRE between July 2020 (M7) and November 2021 (M17)***

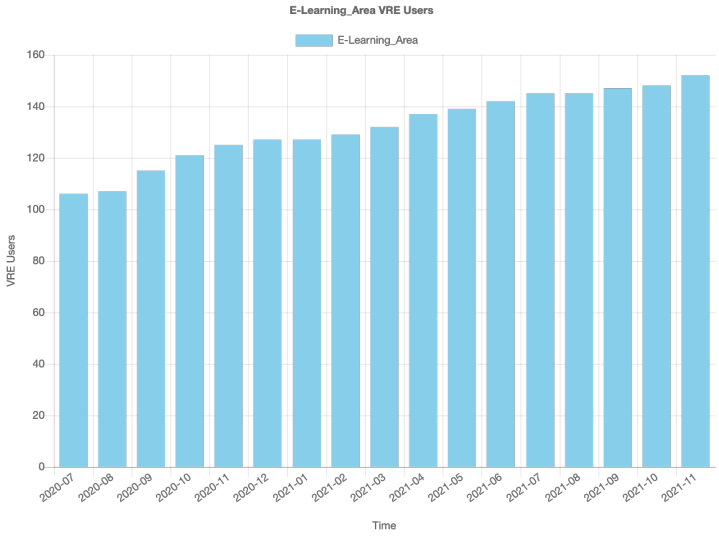

Figure 5 ***Registered users to the e-Learning VRE of the SoBigData++ Research Infrastructure between July 2020 (M7) and November 2021 (M17)***

3.2 Reporting

This section describes each of the training materials that have been uploaded onto the SoBigData++ Research Infrastructure during the reporting period. They comprise a webinar series, which was developed by various Work Packages within the SoBigData community and online training materials such as lectures, courses, and tutorials.

3.2.1 Webinar Series

Due to the ongoing Covid-19 pandemic, the SoBigData++ project organised a series of virtual webinars in order to overcome the impossibility of in-person meetings. This series of webinars, which was organised as a collective effort by the project, involved many different Work Packages. The series has been integrated into the SoBigData Research Infrastructure as a training material, providing an opportunity for users to access webinars in an asynchronous manner, aside from the live event. Webinars are featured both in the SoBigData++ Research Infrastructure and on the video platform YouTube, in order to broaden their reach as much as possible.

3.2.1.1 SOBIGDATA++ E-INFRASTRUCTURE WEBINAR

This webinar, devoted to the explanation of the SoBigData++ Research Infrastructure, was held in M5 (21 May 2020) which is outside the remit of this deliverable. However, it has been integrated into an online training material which features both a link to the recording of the webinar and all the slides featured during the event, which explore in detail every aspect of the Research Infrastructure. Slides regard a review of the SoBigData++ Infrastructure, a presentation of the website, gateway and catalogue search and a presentation of the exploratories. Moreover, the training material includes a series of guides on how to execute an experiment within the SoBigData++ RI which focuses on method engine, execution, and exploitation; how to integrate a new method and on how to integrate a new dataset using the workspace, catalogue DB and storage devices.

3.2.1.2 EPIDEMICS AND THE CITY: HOW HUMAN MOBILITY AND WELL-BEING CHANGED DURING THE COVID-19 ERA

This webinar, which took place on 3 July 2020, was aimed at exploring the impact of the Covid-19 pandemic on mobility, individual well-being, and virus transmissibility from a data science and environmental epidemiology perspective. The webinar was moderated by Angelo Facchini (IMT) and Luca Pappalardo (CNR) and featured two speakers: Professor Dino Pedreschi (UNIPI), Full Professor of Computer Science at the University of Pisa where he co-leads the KDD Lab and is Coordinator of the Data Science PhD and Paolo Vineis, Chair of Environmental Epidemiology, and leader of the Exposome and Health theme of the MRC-PHE Centre for Environment and Health at Imperial College, London.

3.2.1.3 1ST SOBIGDATA++ AWARENESS PANEL

This webinar, which took place on 22 July 2020, was aimed at exploring the theme of data protection for research and statistical purposes. Named 'Towards Legally Attentive Datathons, it featured three speakers and a question-and-answer session. The first speaker was Professor Gianni Comandé (SSSA) who focused on Data Processing for Scientific Research and Statistics and the SoBigData++ Framework. The second speaker was Dr. Giulia Schneider (SSSA), who focused on risk and opportunities connected to datathons. The third speaker was Dr. Denise Amram (SSSA) who focused on legal aspects concerning datathons.

3.2.1.4 CAN BIG DATA BRIDGE GAPS IN MIGRATION STATISTICS?

This webinar, which took place on 29 September 2020, was organised by the Migration Studies Exploratory in WP10 and hosted Professor. Tuba Bircan, Professor at the Department of Sociology and research coordinator of the Interface Demography research group at Vrije Universiteit Brussel (VUB), Belgium. The webinar focused on the possible role of Big Data in bridging the gaps in migration statistics, as traditional statistical data on international migration suffer from the problems (gaps) of inconsistency in definitions, differences in geographical coverages, absence of reasons for migration, timeliness, and limitations in demographic characteristics. New sources of data and Big Data (i.e., mobile phone records, social media data, etc.) exist and could be used to fill some of the gaps of the traditional data.

3.2.1.5 2ND SOBIGDATA++ AWARENESS PANEL

This webinar, which took place on 10 November 2020, was aimed at exploring the theme of data protection and intellectual property issues in platforms. The first speaker was Professor Gianni Comandé (SSSA) who focused on the relationship between data protection and research, specifically on the relationship between market innovation and the protection of data subjects' fundamental rights. The second speaker was Dr. Giulia Schneider (SSSA), who focused on GDPR's research exception and the incidence of data protection principles on data sharing. The third speaker was Professor Caterina Sganga (SNS) who provided an overview of intellectual property relevant aspects regarding research infrastructure and online practices. The fourth speaker was Dr. Giulia Priora (SSSA) who focused on the European developments concerning copyright exceptions and national Open Access policies.

3.2.1.6 EXPLAINING EXPLANATION METHODS

This webinar, which took place on 30 November 2020, focused on existing explanation problems in the field of Artificial Intelligence, the main strategies adopted to solve them, and the most common types of explanations are illustrated with references to state-of-the-art explanation methods able to retrieve them. The speaker was Riccardo Guidotti, researcher at UNIPI and member of the KDD Lab at CNR. In this webinar, he aimed to investigate the lack of transparency on how AI systems make decisions – a clear limitation in their adoption in safety-critical and socially sensitive contexts. Moreover, the webinar was devoted to assessing current research in eXplainable AI (XAI) field, which has recently caught much attention, with specific distinct requirements for different types of explanations for different users.

3.2.1.7 EVALUATING THE SIGNIFICANCE OF NETWORK OBSERVABLES WITH A MAXIMUM ENTROPY-BASED APPROACH

This webinar, which took place on 12 December 2020, hosted Enrico Maiorino, researcher at Bringham and Women's Hospital of Harvard University in Boston, USA. The webinar revolves around how to statistically evaluate observed results in Networks. In regard to human complex diseases, biological networks are powerful resources for discovering and understanding their underlining mechanisms. In this seminar, Maiorino introduces basic concepts around maximum entropy modelling and the proposed methodology and presents a Python package implementing the methodology which he is developing, called 'claude', showing several use cases and examples.

3.2.1.8 DATA IN SOCCER: AN ATHLETIC TRAINER'S POINT OF VIEW

This webinar, which took place on 12 February 2021, was organised as part of the activities of the Sports Data Science exploratory in WP10. The webinar hosted Cristoforo Filetti, an athletic trainer from football team Paris Saint-Germain, who described how soccer clubs collect data and how these are processed and used for players' assessment and training schedule. Dr. Filetti is a sports scientist with a master's degree in Sports Science obtained at the University of Tor Vergata, Roma. During his PhD, Cristoforo focused his research on soccer and in particular on the study of training workload effects in collaboration with AS Roma and during his experience in Qatar. After that, he started his career as an athletic trainer at US Salernitana and is now part of Paris Saint-Germain.

3.2.1.9 3RD SOBIGDATA++ AWARENESS PANEL

This webinar, which took place on 15 February 2021, was aimed at exploring the issue of Medical Device Regulation and Digital Heath. The webinar hosted five speakers: Professor Dr. Paul Quinn from Vrije Universiteit Brussel how presented a talk entitled 'Medical Devices (or not) in the Age of Human Enhancement?'; Dr. Giulia Schneider, from SoBigData++ partner SSSA, who presented a talk entitled 'Data Protection for Digital Health/ Artificial Intelligence Devices: Search for Standards and Certifications'; Dr. Andrea Parziale, from SoBigData++ partner SSSA, who presented a talk entitled 'How Medical Devices Product Liability is Poised to Develop With the Rise of Digital Heath and Artificial Intelligence'; Dr. Alessandro Blassimme, from SoBigData++ Partner ETH Zurich, who presented a talk entitled 'Advancing Digital Health Governance: Ethical and Policy Aspects'; and Professor Dr. Giovanni Comandé, from SoBigData++ partner SSSA, who presented a talk named 'Issues in Translational Biases in the Digital Health Sector'.

3.2.1.10 4TH SOBIGDATA++ AWARENESS PANEL

This webinar, which took place on 10 June 2021, was aimed at exploring the application of potential and ethical issues regarding mobility data sharing. The webinar hosted five speakers: Josep Domingo-Ferrer (SoBigData++ partner Universitat Rovira i Virgili, Catalonia) who gave a talk on 'Decentralized anonymization of mobility data'; Frederick Richter, (Foundation for Data Protection, Germany) who spoke about 'Data Sovereignty in the connected vehicle'; Thierry Chevallier (AKKA Technologies, France) who presented 'MobiDataLab', another Horizon2020 project founded by the EU; Dr. René Peralta (National Institute of Standards and Technology, United States Department of Commerce), who presented a talk entitled 'Studying population dynamics without compromising people's privacy' and Professor Giovanni Comandé (SoBigData++ partner SSSA) who gave a talk entitled 'Mobility data reidentification opportunities and… risks'.

3.2.1.11 5TH SOBIGDATA++ AWARENESS PANEL

This webinar, which took place on 6 July 2021, focused on the interplay between legal data and data science. The webinar, entitled 'Legal Materials as Big Data: (algo)Rithms to Support Legal Interpretation. A Dialogue with Data Scientists hosted over ten speakers and was a joint awareness panel between SoBigData++ and Legality Attentive Data Scientists (LeADS), another Horizon2020 project founded by the EU. The webinar included speakers from both projects and presented a mixed expertise between legal and data science fields, allowing for an interdisciplinary discussion on the subject. Speakers included Professor Francesca

Chiaromonte, Professor Gaetana Morgante, Professor Francesca Romano, Dr. Daniele Licari and Dr. Denise Amram from SoBigData++ partner SSSA; Vern R. Walker, Professor Emeritus of Law at Hofstra University; Professor Shel Mellouli, Laval University; Professor Monica Palmirani, from University of Bologna; Professor Paolo Ferragina from SoBigData++ partner University of Pisa. The moderator was Professor Gianni Comandé from SoBigData++ partner SSSA.

3.2.1.12 AUTOMATED METHODS OF URBAN GREEN ANALYSIS

This webinar, which took place on 13 July 2021, aimed at providing a preliminary definition of the state-of-the-art upon automatic methods for systematized urban green data collection. The webinar focussed on methods and tools that are currently available for the analysis of urban green, considering their degree of accuracy (i.e., location, size, aboveground volume, canopy cover, leaf area, species identity) in relation to the development of the urban green infrastructures. The webinar hosted Giorgio Vacchinano, researcher, and associate professor of Forest Management and Planning at the Department of Agricultural and Environmental Sciences - Production, Landscape, Agroenergy (DISAA) at the University of Milan, Italy. Author of numerous scientific publications, he was named by the journal Nature among the eleven best emerging scientists in the world in 2018. He is a member of the Italian Society of Silviculture and Forest Ecology, Ecological Society of America, and Pro Silva Italia.

3.2.2 Deep Learning Course

This training material is a complete course regarding Deep Learning, created by SoBigData++ partner Barcelona Supercomputing Center and Universitat Politècnica de Catalunya. The course provides an applied approach to Deep Learning, presenting an overview of methods and approaches, and is organised in three types of sessions: Theory (most content is provided by the lecturer, often through slides); Guided Laboratory (content provided by lecturer to be used by trainees); Autonomous Laboratory (work by to be performed by trainees). Most of the course (CNNs, RNNs, Transfer Learning and Transformers) is taught by Dario Garcia-Gasulla. The High-Performance Computing (HPC) part is taught by Marc Casas.

3.2.3 Ego Network Analysis, Information-driven Social Links and Impact on Information Diffusion

This training material comprises a lecture, entitled 'Ego network analysis, information-driven social links and impact on information diffusion' held by Dr. Chiara Boldrini of SoBigData++ partner CNR. The lecture focuses on four different aspects: 'The social brain and its constraints'; 'Dunbar's model in the online world'; 'Ego networks and information diffusion' and 'The Egonetwork package'.

3.2.4 Complex Networks Analysis Lecture

This deck of slides presents a lecture, entitled 'Network Analysis' held by Vaiva Vasiliauskaite, professorship of Computational Science at ETH Zurich. The lecture is part of the Data Science in Techno-Socio-Economic Systems course, and its goal is to introduce network science tools that can be used for the analysis of

networked data and to present network modelling that can provide further understanding of data of interest. The lecture covers network science concepts, network models, community detection and discusses the limitation of network science, ending with a presentation of simplicial complexes as a generalisation of networks.

3.2.5 Compressed and Learned Data Structures Seminar

In this seminar cycle, trainees are guided in the direct usage of a powerful C++ library implementing many state-of-the-art compressed data structures for big data. Other than providing a walkthrough of the API of this library, the seminar encourages students to write, execute and play with some small example programs based on such API (these examples are shown slide decks). The instructions and the code to set up an environment to experiment with this library are available in a GitHub .zip repository file. Finally, trainees are asked to test what they have learned by implementing a toy program that stores and searches through a dictionary of words. The efficiency of their implementation was then tested on a dataset of 2.8 million words. Both the dataset and a solution to the exercise are available in the GitHub zip repository. This seminar was originally prepared by Giorgio Vinciguerra and attended by the Algorithm Engineering students course held by Prof. Paolo Ferragina for the master's degree in Computer Science at the SoBigData++ partner University of Pisa.

3.2.6 Tutorial on Learning to Rank

This tutorial was created and presented by Claudio Lucchese and Franco Maria Nardini at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) in 2018. The tutorial comprises slide decks, Jupyter file notebooks and an instruction file. In the last years, learning to rank (LtR) had a significant influence on several data mining tasks and in particular in the Information Retrieval field, with large research efforts coming both from the academia and the industry. Indeed, efficiency requirements must be fulfilled in order to make an effective research product deployable within an industrial environment. The evaluation of a model can be too expensive due to its size, the features used and several other factors. The tutorial discusses the recent solutions that allow to build an effective ranking model that satisfies temporal budget constrains at evaluation time.

3.2.7 Social Network Analysis @ Master in Big Data

This course, part of the 2021 MS in Big Data at the University of Pisa, introduces students to the theories, concepts, and measures of Social Network Analysis (SNA), which is aimed at characterizing the structure of large-scale Online Social Networks (OSNs). The course presents both classroom teaching to introduce theoretical concepts, and hands-on computer work to apply the theory on real large-scale datasets obtained from Online Social Networks like Facebook and Twitter. The course aims to discuss in particular how the structural properties of social networks can be analysed through Social Network Analysis techniques, and how these properties can be used to characterize social phenomena arising in society.

3.2.8 Data Inquiries Initiative

The Data Inquiries Initiative training material comprises four elements: a Syllabus Example, a Workshop Example, a Project Example and a Project clearinghouse \ greenhouse. In order to facilitate comprehension of this novel approach, introduction and instruction files are also present. The four elements are described in detail in section 5.1.20.3 of this document. The aim of Data Inquiries is to emphasise the social life of data on two levels. The first is the conceptual level, as Data Inquiries promotes a connection between practice of data research with interdisciplinary literature regarding data studies. The second is a practical level, which suggests situating data research within actual social situations, in order to allow synergies with other and diverse societal actors such as civil society groups et alii.

3.3 Planning

Training modules made available within the SoBigData Research Infrastructure so far are predominantly pre-existing teaching materials designed as standalone supports for university courses (i.e., slides, exemplificative notebooks, etc.). Due to their nature, these materials are loosely integrated within the RI, where they can be searched and accessed only as items within the resources catalogue. To enhance their visibility and appeal to the end users WP4 aims to provide a more integrated solution able to better deliver existing and future contents (e.g., through the adoption of MOOC technologies, Video courses, wikis, etc.).

The goal of the forthcoming reporting period is to provide the first example of a prototype course specifically designed for the SoBigData++ end-users. These novel resources will provide a more coherent and embedded access to all other RI functionalities: they will be designed to allow active learning (i.e., introducing student self-evaluation quizzes and other activities), video lectures fruition, and online coding exercises (leveraging the SoBigData JupyterHub).

3.3.1 Prototypical Interactive Course

WP4 has engaged in discussions with SoBigData++ partner University of Amsterdam (UvA) in order to conceive an interactive training material prototype. University of Amsterdam has an extensive experience in the creation of training materials that bridge between top-down settings and hands-on research, as demonstrated by their Digital Methods Initiative (https://wiki.digitalmethods.net/Dmi/DmiAbout), which 'designs methods and tools for repurposing online devices and platforms (such as Twitter, Facebook and Google) for research into social and political issues' according to their own definition.

Thus, WP4 and University of Amsterdam have begun discussing a possible prototype which will be based on training materials developed by UvA. The working title for this training material is 'Visual Analytics for Social Media Research'. The theoretical background is provided by an article authored by Professor Richard Rogers (UvA) published on *Big Data & Society* (Rogers, 2021).

The inception of this structurally novel training material, built in order to exploit interactive tools and methods, will be based on a forthcoming publication 'How to use Visual Analytics for Social Media Research' to be published by SAGE research Methods: Doing Research Online Collection. Moreover, it will include video

case-studies and tutorials, more specifically one case-study and two tutorials, both also to be published by SAGE research Methods: Doing Research Online Collection.

These elements, together with the collaboration with other WPs directly involved in the management of the SoBigData++ Research Infrastructure, will allow the creation of a prototype training material which will encompass not only static and catalogue-based entities but also leverage the most recent additions to the SoBigData++ Research Infrastructure, such as the SoBigData JupyterHub, in order to finalise the creation of an immersive, interactive, and fully developed online training material.

3.3.2 Integration of 'Discovering and Attesting Digital Discrimination' into the SoBigData++ RI

As per work packaged description in the Reporting Period 1 document, WP4, 'together with other work packages, such as WP7 and WP9, has been working to integrate as an application into the SoBigData Research Infrastructure an AHRC founded project, named "Discovering and Attesting Digital Discrimination" (https://dadd-project.github.io/index.html)', a project that adopts a interdisciplinary perspective encompassing technical, legal and social dimensions of the problem. Hence, WP4 will interact with WP7 (Virtual Access) and WP9 (E-Infrastructure and Supercomputing Network), to submit a micro-project proposal regarding the integration of 'Discovering and Attesting Digital Discrimination' as an application. WP4 believes that, thanks to the preliminary meetings held with other project WPs, it will be possible to integrate DADD as an application, which would ease fruition by users.

3.3.3 Jupiter Notebooks for Data Anonymization and Discrimination Discovery

In the forthcoming period, SoBigData++ partner Universitat Rovira i Virgili (URV) plans to create two Jupyter notebooks leveraging libraries for data anonymisation and discrimination discovery that have been developed in partnership with Work Package 8 (Social Mining and Big Data Resource Integration). URV also intends establish this as a working pattern within the SoBigData++ project, developing training materials for every module, library and method which will be developed in the future.

3.3.4 Techno-socio-economic System Course

In the forthcoming period, SoBigData++ partner ETH Zurich plans to expand their contribution to online training modules (See section 3.1.4) with further lecture materials for the data science course in Techno-Socio-Economic systems. Topics will include issues related to the question of how data science and complexity can be used hand in hand.

3.3.5 Tutorial and Code to Create Graphs that Simulate Epidemics

In the forthcoming reporting period, SoBigData++ partner ETH Zurich plans to create a tutorial for the use of a code that will be submitted to Task 8.4 (Social Network Analysis) of WP8 (Social Mining and Big Data Resource Integration). The code is based on the shortest path kinetic Monte Carlo simulation - a numerical

method of solving mathematical problems through random sampling (Sobol, 2017) - of epidemics on graphs, as described in a paper authored by three authors part of the SoBigData++ project at ETH Zurich (Vasiliauskaite, Antulov-Fantulin and Helbing, 2021).

4 Task 4.2 – Summer Schools

Task leader: SNS

Participants: USFD, LUH, UNIROMA1

Summer schools – which are per Grant Agreement the core objective of Task 4.2 – have been severely impacted by the Covid-19 pandemic. Travel restrictions and outbreaks have limited the overall number of events taking place and travelling possibilities. Moreover, planning for events has become increasingly difficult as the Covid-19 pandemic-imposed restrictions have been partially lifted and enforced in cycles across the last 18 months. During the period covered by this deliverable (M6 – M18), one summer school was organised. However, all four project partners involved in this task plan to organise a summer school in 2022.

4.1 Reporting

4.1.1 Summer School on Machine Learning of Dynamic Processes and Time Series Analysis

The Summer School on Machine Learning of Dynamic Processes and Time Series Analysis took place on 26 and 27 November 2020 and was organised by SoBigData++ partner Scuola Normale Superiore (SNS), based in Pisa, Italy. The event was based on four plenary lectures (each of 2.5 hours) by Professor Josef Teichmann of ETH Zurich, Switzerland (a SoBigData++ partner); Professor Christa Cuchiero of Vienna University, Austria; Jun. Professor Lyudmila Grigoryeva of University of Konstanz, Germany and Professor Juan-Pablo Ortega Lahuerta of University of St. Gallen, Switzerland. The summer school also featured seven contributed talks (each of 20 minutes of length). The total number of registered participants was 268 (228 males) coming from academic institutions, research institutes and industry. The event took place virtually and the average number of participants for each session was 120. *Fig.6* shows the flyer of the Summer School.

Figure 6 ***The flyer of the Summer School***

4.2 Planning

Each of the four SoBigData++ partners contributing to this task plans to organise a summer school in 2022. However, due to the current state of the Covid-19 pandemic, organisers are not able to determine if these will be in-person, hybrid, or virtual events.

4.2.1 Summer School on Misinformation Analysis

SoBigData++ partner USFD (UK) plans to organise a summer school between September and October 2022. However, due to sick leave of one of the organisers, details will be defined at a later stage.

4.2.2 Summer School on Data Science

SoBigData++ partners UNIROMA1 and UNIPI (Italy) will organise a summer school on Data Science in the summer of 2022.

4.2.3 Summer School on Explainable AI

SoBigData++ partner LUH (Germany) will organise a summer school on Explainable Artificial Intelligence between the end of July 2022 and August 2022. The expected number of participants, if the event takes place in-person, will be around 50 individuals.

4.2.4 Summer school on Machine Learning of Dynamic Processes and Time Series Analysis – 2nd Edition

SoBigData++ partner SNS (Italy) will organize the second edition of the summer school on Machine Learning of Dynamic Processes and Time Series Analysis in May or June 2022. The format will follow the one employed in the first edition (described in section 4.1.5).

5 Task 4.3 - Datathons

Task leader: CNRS

Participants: CNR, IMT, KTH, ETHZ

Due to the ongoing Covid-19 pandemic, organising in-person datathons was not possible during the reporting period. Datathons, per Grant Agreement convey together "smaller dedicated groups, providing complementary theoretical and practical skills to visualise and analyse social big data questions addressing important societal problems". Hence, aside from reporting the datathon which has taken place during the reporting period (See Section 5.1.19), this section focuses especially on the development of the Data Inquiries initiative by task leader CNRS. This initiative is an on-going, project-wide reflection on the datathon format and includes a deep theoretical framework that has been developed in the reporting period. Moreover, the Data Inquiries initiative has also led to the creation of ad-hoc training materials that have been integrated into the SoBigData++ Research Infrastructure (See Section 3.1.8).

5.1 Reporting

During the reporting period, while one virtual datathon was organised (See 5.1.22), task leader CNRS devoted a consistent effort in a profound reassessment of the datathon format. This reassessment led to the proposal of the Data Inquiries Initiative by CNRS, which is also been integrated as an online training material.

5.1.1 HACK@EO L'Aquila2021

This virtual datathon started on 1 May 2021 and ended on 31 July 2021. A total of 30 individuals divided in 10 teams took part in the datathon, 69% males. The datathon was built around four challenges on building three indexes: one on air quality, one regarding road walkability and the third regarding access to services in L'Aquila territory in Italy (which in 2009 was struck by an earthquake). The fourth challenge regarded the construction of a digital dashboard to access information regarding the three indexes plus the creation of a well-being index that combined the three indexes. All the data yielded during the datathon is available on a dedicated VRE

(https://territoriaperti.d4science.org/).

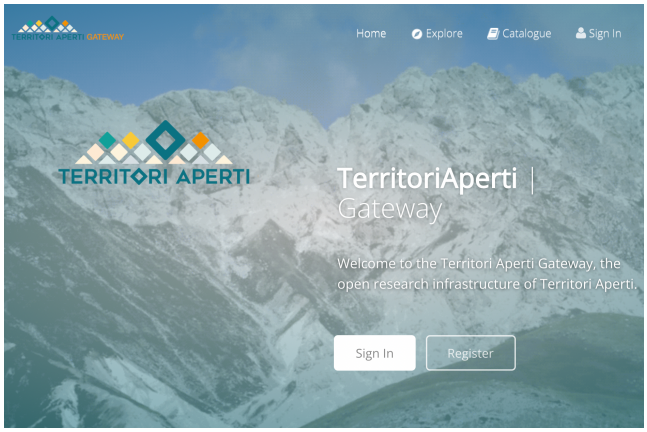

Figure 7 ***The VRE Gateway for the Aquila Hackathon***

During the hackathon's preliminary phase, in order to create the ground truth of machine learning approaches, 30 citizens were involved. They evaluated the walkability of the roads and the accessibility to services following the indications provided. In particular, the recruitment followed a perfect gender balance

(50% female and 50% male) and the evaluators recruited were 70% citizens and 30% students. As per participants, there were no constraints, except the contemporary presence of stem and humanities experts and a minimum quota of females in the team. *Fig.7* shows the VRE Gateway for the Aquila Hackathon.

5.1.2 Data Inquiries Initiative

In previous projects carried out in Science and Technology Studies, SoBigData++ partner and task leader CNRS realised that datathons and workshops constitute the best settings to gather participants from different backgrounds and develop interventions capable of putting digital data and digital technologies at the service of societal debates (Venturini *et al.*, 2015). Building on this, CNRS developed a workshopping format called "data sprint" (Venturini, Munk and Meunier, 2018), (Munk, Venturini and Meunier, 2019), (Munk, Madsen and Jacomy, 2019), (Berry *et al.*, 2015) along with an informal research network dedicated to this type of intervention – the Public Data Lab (www.publicdatalab.org) – which gathers researchers from various European institutions (including SoBigData++ partners – i.e., UvA and KCL).

5.1.2.1 ASSESSED CHALLENGES WITHIN THE PRACTICE OF SOCIAL BIG DATA

The inception of Data Inquiries began with assessing the role of three major aspects of social big data practices.

The first was the role of disciplinary reflexes in data science. The first is the extremely rapid pace of publication demanded in this field. More than other scholars, computer scientists are expected to produce a steady flux of publication in journals and conferences, while monographs and other types of publication are rarer. CNRS undertook an evaluation of the data challenge format, which was used during the first datathon held under SoBigData++, the Epidemic Datathon organised by ETH Zurich. This format consists in asking several teams to compete in answering a well-defined research question through a dataset made available by the organiser. Data challenges are designed to focus the attention of the participant on experimenting and comparing different analytic techniques. In this specific case, the datathon participants were tasked with the prediction of the pandemic.

The second aspect was the role of 'moral panic', as defined by Marshall McLuhan, a situation in which 'a condition, episode, person or group of persons emerges to become defined as a threat to societal values and interests… in a stylized and stereotypical fashion' (Cohen, 2011). One example that CNRS focused on is the issue of misinformation and the role of so-called 'fake news', which according to the Council of Europe is a term which is both 'insufficient and dangerous to use because it has been appropriated by politicians around the world to describe news organisations whose coverage, they find to be problematic. The term 'fake news' is being used as a mechanism for clamping down on the free press, and serves to undermine trust in media institutions, hoping to create a situation whereby those in power can circumvent the press and reach supporters directly through social media' (Wardle and Derakhshan, 2017). While data scientists are aware that the problem of online disinformation cannot be reduced to a true\false opposition, richer and sharper conceptualisations of misinformation have been proposed in journalism studies and media studies. However, these conceptualisations are generally difficult to operationalise with computational methods. Thus, CNRS worked in collaboration with SoBigData++ partner Barcelona Supercomputing Centre, to develop a definition

of disinformation that is richer and more nuanced, but that is still operationalizable through standard data science techniques. As a proxy of the quality of public debate, CNRS decided to use the temporal profile of its attention cycles.

The third aspect regarded the impact of the Covid-19 pandemic, which shifted research activities from in-person to online settings. In order for interdisciplinary collaborations to be successful in an online setting, informal spaces have proven of great value albeit difficulty replicable. This led CNRS to reconsider the organisation of a single, relatively large datathon and instead fraction it into smaller events. Within CNRS's contribution to WP10, and more specifically to the Societal Debates and Misinformation Analysis exploratory, a first workshop in this series took place in January 2021, during the 2021 Winter School of the Digital Methods Initiative and focussed on the attention rhythms of political streamers (wiki.digitalmethods.net/Dmi/WinterSchool2021ConspiracyFolklore).

5.1.2.2 AIM OF DATA INQUIRIES

This avenue of research has led CNRS to believe that there is the need for a constant reflection on formats such as datathons and has identified four main areas of assessment:

1. the nature of advanced computational techniques (and in particular social big data and artificial intelligence) and how they influence collective phenomena at the second-degree (i.e., not only mediating them, but exploiting the records of such mediation to influence their development).
2. the development of social big data which shouldn't be oriented exclusively to the development of technical solutions but should aim at facilitating the emergence of richer and more complex collective solutions that involve technology but are not limited to it.
3. an exploration on how advanced computational techniques and social big data can contribute to collective life by increasing its inclusivity – i.e., the capacity to consider the voices of a wider spectrum of social actors.
4. the preparation of events that objectives are both socially relevant and technologically challenging, that the participants come from different disciplines and backgrounds; that the data have been properly vetted for the bias and power relations that come with them; that the logistics of the workshop is adequate to support its purposes.

5.1.2.3 STRUCTURE OF DATA INQUIRIES

Data Inquiries draws attention to the social life of data both conceptually and practically:

- Conceptually, it proposes to connect the practice and teaching of data research to the transdisciplinary literature on critical data studies and its description of the implicit assumptions and side effects of data infrastructures.
- Practically, it suggests experimenting and teaching data research not in abstract challenges, but in actual social situations, that is in collaboration with civil society groups using data in their projects.

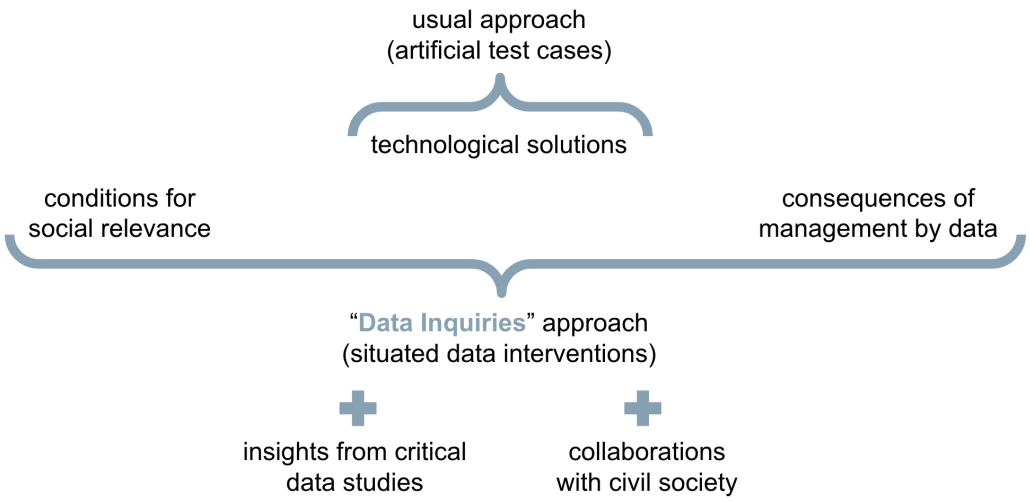

Figure 8 ***Differences between the Data Inquiries approach and common approach***

Data Inquiries, as featured in the SoBigData++ Research Infrastructure Training Material section (See Section 3.1.10) and on the website http://datainquiries.publicdatalab.org/, comprises four main elements:

1- PROJECT

The first is a project example. In particular, the example regards the Fake News Field Guide, an initiative built on a series of data sprints involving journalists and fact-checkers to co-design a set of mapmaking recipes that would help develop a richer and more complex understanding of digital misinformation. The result was the *Field Guide to Fake News and Other Information Disorders*, a digital book gathering a series of methodological recipes to study various aspects of online misinformation, i.e., its circulation; its monetization strategies; the use of Internet memes and images; the connection with trolling, etc.

2- DATA SPRINTS

A crucial element of the Data Inquiry approach is the collaboration between the expert/apprentice data scientists and the actors engaged in actual societal situations. This collaboration helps the data scientists to consider not only the technical dimension of their intervention, but also the context in which it takes place and the conditions that can make it more than a simple exercise of technical skill. Civil society groups, for their part, will find not only help in the collection and treatment of data, but more importantly a fresh perspective on how their action can exploit digital records. Data sprints import two things from its open-source predecessors:

a. **The 'quick and dirty' (or 'design to cost') approach.** The short and intensive nature of data sprint shields these events from the dream of exhaustivity sometimes associated with 'big data'. Participants know that they will only be able to treat a limited quantity of records and that they will only achieve imperfect results, but they accept such constraints more as a challenge than as a flaw. Making the most out of light infrastructures, simple logistics and agile organizations, participants are aware that their work should reuse code and data gathered in earlier projects and that their outcomes should become the basis for further ventures.

b. **The heterogeneity of the actors involved.** The need to achieve deliverable results by the end of the event requires the gathering of all competences required both as in terms of technical skills and in terms of the knowledge of the social situation at stake (hence the importance of convening all the participants during the sprint).

Unlike hackathons and bar-camps, however, data sprints are always preceded by extensive preparation. Because the time available during a data sprint is limited, it is crucial to carry out some activities before the data sprint:

• **Posing the intervention objectives.** While identification of the specific goals of the intervention should be carried out before the sprint itself, it is important that these objectives are not just imposed by the civil society group initiating the intervention but discussed and co-defined among all the participants of the sprint.

• **Operationalizing the intervention objectives into feasible projects.** CNRS found that an excellent way of doing this initial vetting is to have collective discussion about which existing datasets could be accessed and exploited for the intervention. This provides a chance for all participants to attune to what the data project can and cannot achieve.

• **Procuring and preparing datasets**. If relevant datasets are identified beforehand, their harvesting and cleaning should be carried out before the sprint, as these are generally time-consuming operations. If no suitable dataset exists, then the objective of the sprint can become precisely such a primitive collection.

A careful preparation allows to dedicate as much as possible of the time allotted to the workshop for the activities that can only be carried out during the data sprint:

• **Writing and adapting scripts.** Since the focus of the sprint is the preparation of an actual societal intervention, designing new technical solutions is less important than adapting existing code to the goals of the interventions.

• **Designing data visualizations and interfaces.** One of the driving forces of data sprints is that they deliver tangible outcomes. These outcomes may have different forms, but they always share the characteristic of being usable by societal actors. Often, this results in the civil society groups leaving the sprints with tangible results that they immediately mobilize in their actions. More generally, this means that specific efforts should be invested to design the outcomes of the sprint so that they are relevant not only for the data scientists, but also for their potential users.

• **Considering and managing on societal implications.** This last mode of engagement is necessary to create a space for researchers and actors to experiment with new collective arrangements – or hybrid forums as defined above. The six activities described above should all be arranged so that this type of engagement might be achieved. If sprints fail in creating a common space for the co-produce knowledge between social scientists and social actors, they fail in all other respects as well.

Finally, a greater follow-up than hackathons and bar camps is necessary after the data sprint. The 'quick and dirty' approach that characterizes the sprinting days should be complemented by an extensive work of refinement and documentation, in order to make sure that the work of the sprint actually bears fruit and generates the desired societal outcomes. Besides following-up on the specific objectives of the spring, efforts should be invested in making datasets, scripts, and visualizations reusable beyond their original projects. Sprints should remain faithful to their open-source roots and ensure that all the data, code and content produced are freely available through open licenses.

3- SYLLABUS

The syllabus explores the consequences of the digital traceability of collective phenomena with a critical and empirical approach. Considering a variety of computational methods, it offers first-hand experience of digital quantification, examining its potential, but also its shortcomings and biases. Trainees will learn techniques of data collection, corpus cleaning, exploratory analysis, network analysis, natural language processing, artificial intelligence, and information visualization. Working in groups, they will apply these techniques to actual societal situations. Through this experience, they will be led to consider reflexively the insights of the transdisciplinary field of critical data studies.

The course is divided in three parts:

- The **theoretical part** can be developed considering the impacts of media and digital infrastructures to encourage critical reflection on subjects related to *Gender Studies* and *Political and Cultural Geography*, discussing the social and cultural consequences of digital media, particularly with regard to constructions of gender, class, ethnicity, etc.; *Political Economy, Security Studies* and *Political Sciences*, by extending the reflections on the effects of quantification in the government of modern societies at the national and international levels.

- The **methodological part** can be developed for curricula in *Digital Humanities*, *Journalism*, *Media Studies and Social Research and Intervention*, by introducing more advanced numerical and computational methods and discussing the advantages and disadvantages compared to traditional qualitative and quantitative methods, as well as the possible integration with them.

- The **practical part** can be developed and focussed on case studies around subjects directly relevant for programs in *Public Management*, *Sustainable Development*, *Science and Technology Studies*, *Innovation Management*, etc.

4- PROJECT CLEARINGHOUSE \ GREENHOUSE

The final element of Data Inquiry is the clearinghouse \ greenhouse function. The clearinghouse is a space where societal actors can describe the objectives of their data projects and define the kind of contribution they seek and where students and experts can find occasion for practicing data research in real-life situations. The greenhouse is where new projects and collaborations are nurtured.

5.1.2.4 AUDIT OF SOBIGDATA++ MICRO-PROJECT CANDIDATES FOR DATA INQUIRIES

Aside from fulfilling an advisory role of other SoBigData++ partners for the organisation of datathons, CNRS decided to elicit responses from project partners on the Data Inquiries approach. CNRS authored an article on Issue n°6 of the SoBigData Magazine (https://data.d4science.net/BYMa) and contacted five SoBigData++ partners to propose a collaboration. The labs were contacted on the basis of an audit of the microprojects running at the time. Out of all micro-projects, five were selected (*Table 2*) because they had the potential to involve multiple stakeholders and to be a site of socio-technical controversies.

TITLE	DESCRIPTION	POTENTIAL FOR DATA INQUIRY
Urban green dataset for Sustainable cities for citizens in Pisa, Italy	The Micro-Project 'Urban green dataset for Sustainable cities for citizens in Pisa, Italy' aims at developing an exportable dataset for urban green data collection and analysis. The micro-project will be developed on the case study based in Pisa (Italy). Specifically, the following activities are foreseen: -Contact with the Municipality of Pisa for the raw data acquisition. -Data cleaning and supervision of the dataset. -Methods for dataset integration.	- Collaboration with a municipality on urban nature, a stakeholder setup CNRS has had previous experience with (NATURPRADI project on Nature and Digital Practices in Paris, with the city of Paris as partner) - It questions what phenomena 'urban green data' make visible and how - has relevance for policies regarding urban sustainability
Estimating countries' peace index with GDELT and machine learning techniques	Peacefulness is a principal dimension of well-being, and its measurement has lately drawn the attention of researchers and policymakers. Exploit information extracted from GDELT to capture peacefulness through the Global Peace Index (GPI) with machine learning techniques. News media attention, sentiment, and social stability from GDELT can be used as proxies for measuring GPI at a monthly level.	- GPI (from https://www.visionofhumanity.org/maps/#/: 'The GPI comprises 23 indicators of the absence of violence or fear of violence. The indicators were originally selected with the assistance of the expert panel in 2007 and have been reviewed by the expert panel on an annual basis. All scores for each indicator are normalised on a scale of 1-5, whereby qualitative indicators are banded into five groupings and quantitative ones are scored from 1 to 5, to the third decimal point'. - how to measure peacefulness must certainly be controversial and disagreed upon, especially if the project aims at proposing an alternative definition of the GPI based on social data rather than the rating of experts. - a particular point of attention could be: could the proposition lead to a less 'Western-oriented' definition of peace?

TITLE	DESCRIPTION	POTENTIAL FOR DATA INQUIRY
Difference between men's and women's in soccer	The aim of this study is to analyse the spatiotemporal events during matches in the last World Cups to compare male and female teams based on their technical performance.	- gender discrimination in sport is an important topic, and the lesser interest in women's soccer events is sometimes implicitly 'justified' in the discourse of sport enthusiasts by a lesser 'technical performance'. - Hence why what data enter into the characterization 'technical performance' should be a controversial issue. - potential for involvement of various types of expertise to produce qualitative descriptions of the 'spatiotemporal events' in question.
Validation on home location detection algorithms on ground truth	Home detection, assigning a phone device to its home antenna, is a ubiquitous part of most studies in the literature on mobile phone data. Despite its widespread use, home detection relies on a few assumptions that are difficult to check without ground truth, i.e., where the individual that owns the device resides. In this microproject, aim is to provide an unprecedented evaluation of the accuracy of home detection algorithms on a group of sixty-five participants for whom exact home address is known and the antennas that might serve them.	The idea to is help the gathering of statistical data about people's addresses through mobile phone data (See https://arxiv.org/pdf/1809.07567.pdf). - The modification of technologies for official statistics, in particular the use of citizen-generated data, should bring its share of interesting problems with regard to format, quality of data etc. - The project can possibly involve stakeholders, from homeowners (planned in the project) to governmental authorities, that has relevance for 'ground-truthing'.
Detecting Content That Triggers Polarization in Social Networks	During the last decade it has been popular to study polarization in social networks. There have been many works that have focussed on understanding polarizing interactions between different users. However, little attention has been paid to the content that triggers these interactions. For example, when a controversial article is shared in the social network, this might cause a heated discussion which eventually leads to more polarization among the users. In this project, our goal is to obtain a better understanding of such phenomena. Algorithm development which locates content (such as articles or tweets) that triggers polarizing user interactions in the social network.	- One interesting issue could around the definition of 'controversial article', which might depend on the content but also the context in which it is shared. - There is place for qualitative assessment of the polarizing content that would involve various stakeholders and perspectives. - CNRS has previous experience with the topic they are addressing, as well as methods to suggest.

Table 2 *Audit of SoBigData++ Micro-projects*

5.2 Planning

Within the timeframe encompassed by Deliverable D4.1 and Deliverable D4.2, two datathons have taken place within the SoBigData++ Project: The Epidemic Datathon and Hack@EO L'Aquila 2021. Both have been virtual and not in-person events due to the ongoing Covid-19 pandemic. At the present time, WP4 is not able to predict if the planning section that follows will take place in an in-person setting or in an online setting.

5.2.1 Planned Datathons

In the next reporting period, CNRS plans to organise two datathons. The first will be a workshop on participatory practices during a project's General Meeting to:

- Reflect critically upon data challenges as a participatory practice
- Render visible and share other, less visible or overlooked, mundane participatory practices
- Rework the design strategies published in this deliverable so they can become useful for the SoBigData++ community
- Help out partners to design their own datathon or collective event

Following this first event, CNRS hopes that at least one SoBigData++ partner will be able to develop a 'data inquiry' event (aided by CNRS). Hence, organise another datathon.

5.2.2 Planned Related Activities

The Data Inquiries reflection within the datathon framework should be understood as an articulation between methods from different and diverse disciplines, such as Sociology, Computer Science and Design. For the next reporting period, CNRS will:

1. Populate training materials related to these three disciplines within the Data Literacy framework of which Online Training Materials are part of.
2. Submission of an academic article that follows, complements and completes the already published literature on 'data sprints'. Likewise, CNRS envisions a possible academic contribution regarding the critical analysis of data challenges based on results stemming from a workshop (See section 5.2.13) with other SoBigData++ partners.

6 Task 4.4 – Cultivating Diversity in Data Science Through Training

Task leader: KCL

Participants: ALL

6.1 Reporting

Within this task, WP4 traditionally moved along two lines of action. The first being dedicated training events aimed at raising awareness regarding the opportunities provided by employment in the field of data science. Jointly, SoBigData++, in its first iteration, began the practice of funding bursaries dedicated females and under-represented minorities to allow attendance at conferences and events with the disciplines of interest, such as data science, machine learning, network analysis and beyond. However, due to the Covid-19 pandemic, it has not been possible to undertake efforts in this strand of the task. Jointly, WP4 has begun tracking best-practices of tracking diversity participation in conferences and publications within the SoBigData++ fields of interest. The work package has identified two main experiences, one being BiasWatchNeuro (http://www.biaswatchneuro.com), which tracks diversity in the field of neuroscience and the other being DivinAI (http://divinai.org), which tracks diversity in the field of Artificial Intelligence. In order to progress the WP's knowledge on this experience, WP4 interviewed Ana Freire, Dr. Ana Freire of Universitat Pompeu Fabra (SoBigData++ partner) and research fellow of the EU Joint Research Centre Humanint project (Human Behaviour and Machine Intelligence).

6.1.1 Investigating best-practices

WP4 decided to do a semi-structured interview (Galletta, 2013) with Dr. Ana Freire of Universitat Pompeu Fabra (SoBigData++ partner) and leader of DivinAI. According to its website (*Fig.9*), DivinAI aiming 'to research and develop a set of diversity indicators, related to Artificial Intelligence developments, with special focus on gender balance, geographical representation and presence of academia vs companies'.

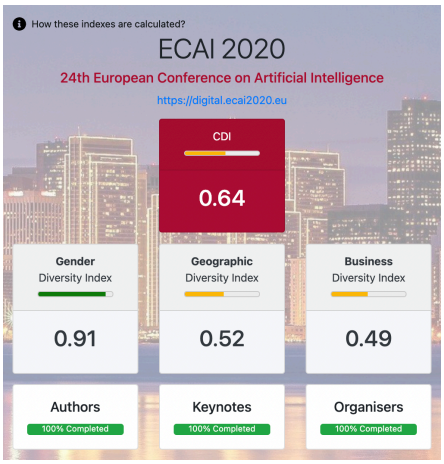

Figure 9 *DivinAI index calculation on the 24th European Conference on Artificial Intelligence*

DivinAI currently tracks four different indexes, which directly regard the issue of under-representation (gender diversity index, geographic diversity index and business diversity index together with a joint conference diversity index). *Table 3* details how these indexes are calculated.

Gender diversity index We consider three different species (S = 3) in the gender dimension: "male", "female" and "other". We compute Shannon evenness by means of the Pielou diversity index. For calculating the Gender Diversity Index, we consider three different communities: keynotes (k), authors (a) and organisers (o). Our final GDI performs a weighted average among the Pielou index in each community with the following weights: 1/3 for keynotes, 1/3 for authors and 1/3 for organizers.
Geographic diversity index As we have multiple species (S = number of countries present by community), we want to measure the richness together with the evenness, so we apply the weighted average of the Shannon Index community. We compute the Shannon Index for each of the following communities: keynotes (k), authors (a) and organizers (o). We equally weight each community using the following weights: 1/3 for keynotes, 1/3 for authors and 1/3 for organizers.
Business diversity index We consider three different species (S=3) in the business dimension: "academia", "industry" and "research centre". We compute Shannon evenness by means of the Pielou diversity index. For calculating the Business Diversity Index, we consider three different communities: keynotes (k), authors (a) and organisers (o). Our final BDI performs a weighted average among the Pielou index in each community with the following weights: 1/3 for keynotes, 1/3 for authors and 1/3 for organizers.
Conference diversity index This index is computed by the combination of gender, geographic and business indexes using the following formula: CDI = 1/3 *(GDI + GeoDI/2 + BDI)

Table 3 ***DivinAI different index calculation (from http://divinai.org)***

6.1.2 Interview with Dr. Ana Freire (DivinAI project)

WP4 performed a semi-structured interview (Galletta, 2013) with Dr. Ana Freire, head of the Tech Academic Unit, Senior Lecturer and Researcher at Universitat Pompeu Fabra (SoBigData++ partner) Barcelona School of Management. Freire leads DivinAI (Diversity in Artificial Intelligence), an initiative of the HUMAINT project at Joint Research Center (EC) and the ICT Department at Universitat Pompeu Fabra, Barcelona. For the full text of the interview, see Appendix B.

In the interview, Freire underlines how the goal of DivinAI is to 'define a methodology to monitor diversity', focusing on Artificial Intelligence conferences 'as they are the most relevant outcome now for AI research dissemination'. DivinAi monitoring is divided in four indexes (See Table 2) and keeps track of each conference's evolution over time.

Currently, the DivinAI dataset comprises 35 conferences. As to trends, Freire says that 'we couldn't find a common positive pattern, but we could see, for instance, that there is an effort in selecting female keynotes for invited talks. However, organisers and authors are mostly "male" and women do organize more than authorise'. Moreover, 'most of the keynotes come from North America or Europe and very few researchers from Africa were found among authors and organisers.' Freire also underlined how where conferences take place directly impacts the presence of minorities, especially regarding their country of origin. While data sharing from conferences does raise privacy issues, some conference organisers, according to Freire, 'have shown interest on knowing these data and even showed them to the audience during the conference'.

DivinAI's methodology is applicable to other fields as it (Freire, Porcaro and Gómez, 2021) 'has been published […] and anyone can replicate it with their own data. Also, our website offers the possibility of registering as an editor to input data and calculate the index automatically'. Moreover, 'a very similar analysis in the field of "affective computing" in the *9th International Conference on Affective Computing and Intelligent Interaction (ACII).'*

Finally, Freire underlined the. Major challenges faced by index trackers such as DivinAI, since "the data publicly available from AI conferences is restricted to the name, surname and affiliation of authors, keynotes and organisers. This leads to some limitations when computing the proposed indicators. For instance, no data is provided about gender, so it needs to be inferred based on the name and surname, which introduces some errors and oversimplification to binary labels". Moreover, Freire added 'another limitation aspects the way in which we compute the geographical diversity index. On one side, having information just about the affiliation and not about the nationality, makes ethnic-based analysis extremely difficult to be performed'.

Hence, Freire's interview provided insight into how an index such as DivinAI works, and what are the challenges it faces, especially regarding data collection and how this data is the computed to calculate indexes. While DivinAI uses data deriving from publicly available sources, 'limitations can be solved if the conferences' organisers collect more data at registration time, for instance, and share them using a safe procedure', according to Freire. Scalability and applicability of DivinAI's indexes to other fields of study is possible, as Freire underlined, since the methodology has been published and thus 'anyone can replicate it with their own data'. In light of generating an index which covers SoBigData++ fields of interest, DivinAI certainly provides an open, composite, and scalable model.

6.2 Planning

WP4 plans, together with all SoBigData++ partners, to further initiatives in this task by pursuing two lines of action. The first will be the creation and integration of a diversity tracker within the SoBigData++ Research Infrastructure, in the manner that will be rendered possible by work in collaboration with WP7 and WP9. Moreover, by integrating 'Discovering and Attesting Digital Discrimination' as an application within the SoBigData++ Research Infrastructure, WP4 plans to provide a further tool to achieve the goal of cultivating data science through training. The second line of action will be – whenever it will be possible – to resume the bursary activity for under-represented individuals in data science to attend conferences.

6.2.1 Integration of diversity tracker in the SoBigData++ RI

With the interview with Dr. Ana Freire of Universitat Pompeu Fabra (SoBigData++ partner) and research fellow of the EU Joint Research Centre Humanint project (Human Behaviour and Machine Intelligence), WP4 has concluded the first phase of its investigation into best practices regarding diversity participation trackers within SoBigData++ fields of interest. The long-term goal of this activity is to explore the possibility to build and host within the SoBigData++ Research Infrastructure of a similar tool that will allow tracking of participants to events related to the project's field of inquiry. Due to the nature of the SoBigData++ Research Infrastructure, the WP's work will focus on collaborating with other Work Packages in order to explore the

possibility to create and integrate a diversity tracker within the Research Infrastructure. WP4 believes that this might become a useful tool to overcome biases which are present – among other fields – in STS (Science and Technology Studies). These biases which may regard gender and ethnicity are not necessarily explicit or voluntary (Raymond, 2013) and thus such a tool might aid awareness diffusion in the organisation of events, publications, and other scientific activity.

6.2.2 Integration of 'Discovering and Attesting Digital Discrimination' into the SoBigData++ RI

The integration of 'Discovering and Attesting Digital Discrimination as a running application within the SoBigData++ Research Infrastructure (See Section 3.2.10) will aid the WP in addressing T4.4 as part of DADD is the 'Language Bias Visualiser' which is a tool which interactively compares men and women stereotypes inherent in large textual datasets taken from the internet, as captured by Word Embeddings models.

6.2.3 Travel bursaries dedicated to under-represented individuals in SoBigData++ related events

SoBigData++ partner Universitat Rovira i Virgili (URV) plans to offer grants to the Privacy in Statistical Databases conference dedicated to females and under-represented scholars. URV plans to do the same for the successive edition of the conference, which is held every two years.

7 Other Training Events

7.1 Reporting

What follows is a list of training events that took place virtually during the reporting period. They include tutorials that were hosted during conferences and workshops.

7.1.1 XAI Tutorial at AAAI 2020

The XAI tutorial was a virtual event which took place on 5 January 2020 organised by Università di Pisa, Italy. This tutorial was part of the 34th Conference on Artificial Intelligence sponsored by the Association for the Advancement of Artificial Intelligence (AAAI) and took place in New York, USA between 7 and 12 February 2020. The tutorial was entitled: 'Explainable AI: Foundations, Industrial Applications, Practical Challenges, and Lessons Learned'. The tutorial focused on providing a snapshot on the work of Explainable Artificial Intelligence to date, and survey the work achieved by the AI community with a focus on machine learning and symbolic AI related approaches. The goal of the tutorial was to provide answers to the following questions: what is XAI? Why shall we care? Where it is critical? How does it work? What did we learn? What is next? The attendance of the tutorial was of 100 individuals (50 males).

7.1.2 XDMS Tutorial at DSAA 2020

The XDMS (eXplainable Decision-Support Making) tutorial is a virtual event organised by the Università of Pisa, Italy on 9 October 2020. This tutorial was one of four scheduled in the 7th IEEE International Conference on Data Science and Advanced Analytics organised in Sidney, Australia. One hundred people took part in the tutorial (50 males). The tutorial's purpose was to illustrate state-of-the-art approaches for explainable data mining and interpretable machine learning. Moreover, the tutorial also focused on the current problems, issues, and challenges in the field in order to encourage principled research in order to promote explainable, transparent, ethical, and fair data mining and machine learning.

7.1.3 Incontra Informatica Workshop

Incontra Informatica was a virtual event which took place on 16 April 2021 and was organised by Università di Pisa, Italy. The workshop was targeted to high school students, who were shown how to read, pre-process and analyse soccer match event data, allowing them to develop some basic statistics to describe single player and team performance. A total of 27 students attended the event (17 males). Organisers underline how engagement by students was considerable. This workshop was one of many events designed for students in their final high school year to promote informatics as a MS.

7.1.4 SoBigData.eu Tutorial at DSAA 2021

This tutorial led by Consiglio Nazionale delle Ricerche (CNR) was a virtual event which took place on 9 October 2021 during the 8th IEEE International Conference on Data Science and Advanced Analytics organised in Porto, Portugal. The tutorial's goal was to showcase the services provided by the SoBigData++ Research Infrastructure and focus on the resources which are available to users (*Fig. 10*). Examples of usage of the SoBigData libraries and its method engine were presented, and users were able to follow and repeat the experience on a dedicated Virtual Research Environment built for DSAA 2021. A total of 16 participants attended (12 males) and organisers report that during the tutorial there were a lot of questions and interactions between the speakers and the attendees, albeit limited by the virtual nature of the event.

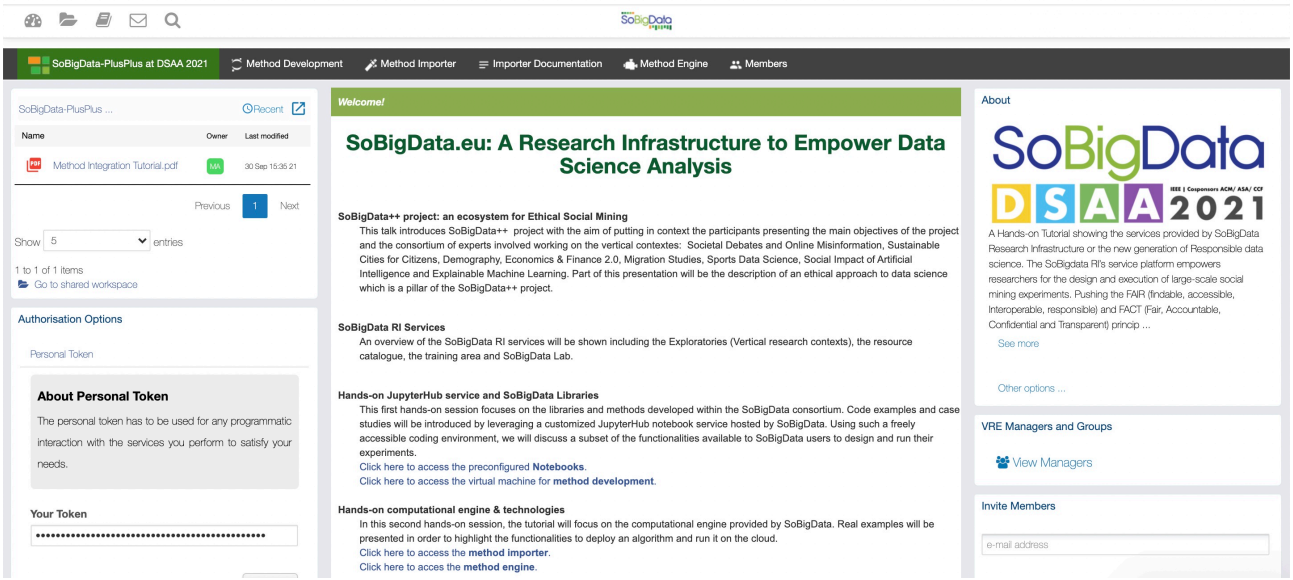

Figure 10 ***Bespoke VRE developed for the DSAA 2021 tutorial in the SoBigData++ Research Infrastructure***

Appendix A. Training during the Covid-19 Pandemic Questionnaire

A.1 Questionnaire design

The questionnaire was created via Google Forms and was administered to the 263 contacts of the SoBigData++ contact list. The questionnaire is completely anonymous and is mainly based on close-ended questions, except for a comment box at the end, to allow participants to express possible further comments outside of the close-ended questions. The total number of respondents was 20.

A.2 Questionnaire

Q1 – Are you a:

a) Early Career Researcher
b) Lecturer – Associate Professor
c) Senior Lecturer – Full Professor

Q2 – Has the Covid-19 Pandemic impacted in-person teaching \ training in your institution?

a) Yes
b) No

Q3 – Since March 2020, have you resorted to virtual teaching \ training? (Multiple answers possible)

a) Yes, synchronous lectures
b) Yes, asynchronous lectures
c) Yes, hands-on training materials
d) No, all teaching and training activities have continued as before

Q4 – Did you create new content for your virtual lectures \ training?

a) Yes
b) No

Q5 – If you have created new content, it has been (Multiple answers possible)

a) New Slides
b) New Training Materials (Wikis, MOOCs, etc.)
c) New Interactive Training Materials (Python Notebooks, etc.)
d) Other types of materials

Q6 – Do you plan to keep using the new content you created during in-person teaching?

a) Yes, all of it
b) Yes, some of it
c) No, none of it

Q7 – If you resorted to virtual lectures and training, have they been:

a) More labour intensive
b) Less labour intensive
c) The same

Q8 – How have your students \ trainees responded to virtual teaching \ training?

a) Well, they seemed engaged
b) Not so well, they seemed less engaged
c) Just about the same

Q9 – Have you resumed in-person teaching \ training?

a) Not yet
b) Yes, but only for small numbers of individuals
c) Yes, we are back on a pre-pandemic schedule with all activities in person

Q9 – Since March 2020, have you attended any events (i.e., conferences, workshops, etc.)?

a) Yes, but only virtually
b) Yes, both virtually and in-person
c) Yes, but only in-person
d) No

Q10 – If you wish, please add a description of your teaching \ training experience during the Covid-19 pandemic

(Open-ended question)

Appendix B. Interview on the DivinAI project

B.1 Interview with Dr. Ana Freire (DivinAI project)

The following is a full transcript of a semi-structured interview (Galletta, 2013) with Dr. Ana Freire of Universitat Pompeu Fabra (SoBigData++ partner) and research fellow of the EU Joint Research Centre Humanint project (Human Behaviour and Machine Intelligence) and one of the promoters of DivinAI.

B.2 Full Interview transcript

Q1. DivinAI has developed four different indexes: Gender Diversity, Geographic Diversity, Business Diversity and Conference Diversity. What led to the creation of these four indexes? What issues does each index address? How do they work?

It is well recognized that Artificial Intelligence (AI) eld is facing a diversity crisis, and that the lack of diversity contributes to perpetuate historical biases and power imbalance. As a consequence, the research community has established different initiatives for increasing diversity such as mentoring programs, visibility efforts, travel grants, committee diversity chairs and special workshops. However, there is no mechanism to measure and monitor the diversity of a scientific community and be able to assess the impact of these different initiatives and policies.

Our goal is to define a methodology to monitor the diversity of the scientific community. We focus on AI conferences as they are the most relevant outcome now for AI research dissemination. We measure not only diversity of a conference, but its evolution.

We consider diversity in terms of gender, geographical location, and business (understood as the type of institution, i.e., academia vs industry). Gender diversity is the focus of programs such as Women in Machine Learning; Geographic diversity is linked to the presence of different countries and cultures in AI research. Finally, the business index provides a way to assess the type of institutions contributing to AI research. We think these are three key socio-economic aspects of AI communities.

We measure diversity considering different roles:

- Authors: the collective who is improving the state of art of AI.
- Keynotes: the most visible part of AI, representing the experts in the field.
- Organisers: who are working in dissemination.

We calculate these indexes through widely used diversity Indexes: shannon and simpson indexes.

Q2. How many conferences does your dataset comprise at this time? Has conference performance shown changes over time?

More than 35 conferences. Unfortunately, we couldn't find a common positive pattern, but we could see, for instance, that there is an effort in selecting female keynotes for invited talks. However, organisers and authors are mostly "male" and women do organize more than authorise. Those conferences also balancing the gender among organisers got the highest GDI.

Another important issue we could find was that most of the keynotes come from North America or Europe and very few researchers from Africa were found among authors and organisers. We would also like to note the importance of the conference location for promoting the presence of minorities. We highlight the case of RecSys 2020, located in Brazil, with a high representation of organisers (13 out of 38) and even one keynote (out of 3) from South America, a continent with almost no representation in the rest of the events.

Finally, we could also conclude that in general, most of the conferences are academic, having low representation from industry or research centres.

Q3. Have you been in contact with conference organisers regarding the issues your work raises? Have they acknowledged the issue of bias? In your article, "Measuring Diversity of Artificial Intelligence Conferences" (Freire, Porcaro and Gómez, 2021), you write "the research community has established different initiatives for increasing diversity such as mentoring programs, visibility efforts, travel grants, committee diversity chairs and special workshops". Do you think the tracking proposed by DivinAI can strengthen the awareness for further initiatives?

Yes, some conference organisers shown some interest on knowing these data and even showed them to the audience during the conference. Our project can be very useful to raise awareness on some hidden issues, such as the underrepresentation of some countries/continents. It would be great if conference organisers could find a way of facilitating us their data to make these indexes more accurate, but this is still difficult due to privacy issues.

Q4. Your website and article explain in detail the inception of the four indexes that DivinAI has developed. Do you think these methods can be applied to other fields of academic study? Are they available for individuals or groups to implement in their field of study? Are you aware of similar tracking experiences in other fields?

Sure, indeed, we have recently published a very similar analysis in the field of "affective computing" in the *9th International Conference on Affective Computing and Intelligent Interaction (ACII).* Our methodology has been published on the 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI) and anyone can replicate it with their own data. Also, our website offers the possibility of registering as an editor to input data and calculate the index automatically.

Q5. Is there any issue that you'd like to address that wasn't featured in the questions above?

The data publicly available from AI conferences is restricted to the name, surname and affiliation of authors, keynotes and organisers. This leads to some limitations when computing the proposed indicators. For instance, no data is provided about gender, so it needs to be inferred based on the name and surname, which introduces some errors and oversimplification to binary labels. Another limitation aspects the way in which we compute the geographical diversity index. On one side, having information just about the affiliation and not about the nationality, makes ethnic-based analysis extremely difficult to be performed. On the other side, an index based just on the number of countries might hide a lack of diversity regarding, for instance, the presence of researchers from least developed countries. Thus, we also computed the number of developing countries present (following the United Nations classification). We couldn't find any representation of any of the countries included in this list containing 46 countries. Thus, we also grouped the affiliations data in order to report the presence of continents and explore the variability of the index in considering these major geographical divisions. We could see that, in general, the indexes computed for the continents are very similar to those related to the countries, and they have a value below 0.5. We consider 7 continents (Africa, Antarctica, Asia, Europe, North America, South America, and Oceania), in order to avoid hiding lower representation of Latin American countries. In most of the conferences explored, there are few "species" (usually 3 -North America, Europe and Asia – and rarely 4 – including Oceania). However, these limitations can be solved if the conferences' organisers collect more data at registration time, for instance, and share them using a safe procedure.

Appendix C. References

Berry, D.M. *et al.* (2015) 'The data sprint approach: exploring the field of Digital Humanities through Amazon's application programming interface', *Digital Humanities Quarterly*, 9(4). Available at: http://www.digitalhumanities.org/dhq/ (Accessed: 26 November 2021).

Cohen, S. (2011) *Folk Devils and Moral Panics*. London: Routledge. doi:10.4324/9780203828250.

Corbera, E. *et al.* (2020) 'Academia in the Time of COVID-19: Towards an Ethics of Care', *Planning Theory & Practice*, 21(2), pp. 191–199. doi:10.1080/14649357.2020.1757891.

Freire, A., Porcaro, L. and Gómez, E. (2021) 'Measuring Diversity of Artificial Intelligence Conferences', in *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI). Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, PMLR, pp. 39–50. Available at: https://proceedings.mlr.press/v142/freire21a.html (Accessed: 2 December 2021).

Galletta, A. (2013) *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication*. New York, New York: NYU Press.

Hupont, I. *et al.* (2021) 'How diverse is the ACII community? Analysing gender, geographical and business diversity of Affective Computing research', *arXiv:2109.07907 [cs]* [Preprint]. Available at: http://arxiv.org/abs/2109.07907 (Accessed: 2 December 2021).

Munk, A.K., Madsen, A.K. and Jacomy, M. (2019) 'Thinking Through The Databody: Sprints as Experimental Situations', in Mäkitalo, Å., Nicewonger, T., and Elam, M. (eds) *Designs for Experimentation and Inquiry*. London: Routledge, pp. 110–128. doi:10.4324/9780429489839.

Munk, A.K., Venturini, T. and Meunier, A. (2019) 'Data Sprints: A Collaborative Format in Digital Controversy Mapping', in Vertesi, J. and Ribes, D. (eds) *Digital STS*. Princeton, New Jersey, USA: Princeton University Press, pp. 472–496. doi:10.2307/j.ctvc77mp9.34.

Raymond, J. (2013) 'Most of us are biased', *Nature*, 495(7439), pp. 33–34. doi:10.1038/495033a.

Reja, U. *et al.* (2003) 'Open-ended vs. Close-ended Questions in Web Questionnaires', *Developments in applied statistics*, 19(1) ((1)), pp. 159–177.

Rogers, R. (2021) 'Visual media analysis for Instagram and other online platforms', *Big Data & Society*, 8(1), p. 20539517211022370. doi:10.1177/20539517211022370.

Sobol', I.M. (2017) *A Primer for the Monte Carlo Method*. Boca Raton: CRC Press. doi:10.1201/9781315136448.

Vasiliauskaite, V., Antulov-Fantulin, N. and Helbing, D. (2021) 'Some Challenges in Monitoring Epidemics', *arXiv:2105.08384 [cond-mat, physics: physics]* [Preprint]. Available at: http://arxiv.org/abs/2105.08384 (Accessed: 28 November 2021).

Venturini, T. *et al.* (2015) 'Designing Controversies and Their Publics', *Design Issues*, 31(3), pp. 74–87. doi:10.1162/DESI_a_00340.

Venturini, T., Munk, A. and Meunier, A. (2018) 'Data-Sprinting: A Public Approach to Digital Research', in Lury, C., Fensham, R., and Heller-Nicholas, A. (eds) *Routledge handbook of interdisciplinary research methods*. Routledge (Routledge international handbooks). doi:10.4324/9781315714523-24.

Wardle, C. and Derakhshan, H. (2017) 'Information disorder: Toward an interdisciplinary framework for research and policy making', *Council of Europe*, 27.

WHO (2020) *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (Accessed: 23 November 2021).