Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Deliverable 6.15 Report on Training Webinars

Dissemination Level	PU
Due Date of Deliverable	30/04/2022 (M40)
Actual Submission Date	14/03/2022
Work Package	WP6 - Fostering Communities, Empowering Users, & Building Expertise
Task	Task 6.5 Coordinating Targeted Training in the Social Sciences and Humanities
Type	Report
Approval Status	Approved by EC - 04 May 2022
Version	V1.0
Number of Pages	p.1 – p.69

Abstract:

This report provides detailed information about nine training webinars offered by SSHOC between March 2019 and December 2021. The aim of these webinars was to offer targeted training to the SSH community. The webinars offered both generic and highly topic-specific content covering six thematic clusters, namely Data Science for the SSH, Data Science for Heritage Science, Data Protection and the GDPR, Data Stewardship and RDM, Data Citation, and Text Mining for the SSH. The webinars reached a total of 1765 participants worldwide: 638 participants of livestreams and 1127 views of webinar recordings. The focus target audience of the webinars included, but was not limited to, researchers, librarians and archivists, and representatives of research institutions and infrastructures.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.1	04/02/2022	First draft	Kristina Pahor de Maiti (CLARIN/UL-FF) Darja Fišer (CLARIN/UL-FF)
0.2	11/02/2022	First review received	Tatsiana Yankelevich (LIBER)
0.3	14/02/2022	Implementation of changes required by first review	Kristina Pahor de Maiti (CLARIN/UL-FF) Darja Fišer (CLARIN/UL-FF)
0.4	22/02/2022	Second review received	Judith Wehmeyer (CESSDA/GESIS)
1.0	28/02/2022	Implementation of changes required by second review and submission of the final version	Kristina Pahor de Maiti (CLARIN/UL-FF) Darja Fišer (CLARIN/UL-FF)

Author List

Organisation	Name	Contact Information
CLARIN/UL-FF	Kristina Pahor de Maiti	Kristina.Pahordemaiti@ff.uni-lj.si
CLARIN/UL-FF	Darja Fišer	Darja.Fiser@ff.uni-lj.si

Editor List

Organisation	Name	Contact Information
LIBER	Tatsiana Yankelevich	Tatsiana.Yankelevich@libereurope.org
CESSDA/GESIS	Judith Wehmeyer	Judith.Wehmeyer@gesis.org

Executive Summary

The aim of T6.5 activities was to provide targeted training events to the SSH community in the form of workshops and webinars. The aim of the events was to maximize the uptake of SSHOC resources and to promote data-driven and cross-disciplinary research directions. Targeted training webinars aimed to complement training workshops. This report concerns training webinars which were conceived as thematically narrower and shorter online events that would precede or follow the workshops, while the workshops were conceptualized as comprehensive and immersive training sessions. Because of their online format, the webinars were highly inclusive since the attendance was possible for a great variety of individuals who for different reasons would be unable to attend the in-person events, and consequently, greatly expanded the outreach of SSHOC outcomes. Given the training design in the form of workshop-webinar pairs and the fact that most of the workshops were, like webinars, delivered online, the organisation of the events largely overlapped. In order to provide a concise report on the work done in SSHOC Task 6.5, this deliverable gives a detailed account of the webinars delivered, while for all other aspects that concern training workshops, please refer to Deliverable 6.14 *Report on Training Workshops* (Pahor de Maiti & Fišer, 2022).

The nine targeted training webinars listed below were organised from March 2019 to December 2021 and followed six thematic clusters:

1. **Data Science for the Social Sciences and Humanities**
   1.1. Hands-on Tutorial on Transcribing Interview Data (03/2019)
   1.2. Sharing Datasets of Pathological Speech (10/2020)
2. **Data Science for Heritage Science**
   2.1. Use and Re-Use of Scientific Data in Archaeology and Heritage (04/2020)
3. **Data Protection and the General Data Protection Regulation**
   3.1. GDPR and the DARIAH ELDAH Consent Form Wizard (10/2020)
4. **Data Stewardship and Research Data Management**
   4.1. Tools and Resources for FAIR Data (05/2020)
   4.2. Introducing the Newly Launched Ethnic and Migrant Minority Survey Registry (10/2020)
5. **Data Citation**
   5.1. FAIR SSH Data citation: Practical Guide (12/2021)
6. **Text Mining for the Social Sciences and Humanities**
   6.1. Quanlify With Ease: Combining Quantitative and Qualitative Corpus Analysis (04/2020)
   6.2. SSHOC'ing Drama in the Cloud: the Added Value of SSHOC/CLARIN Services (06/2021)

The webinars usually lasted for an hour and consisted of presentations and a moderated questions and answers sessions at the end. Special care was taken in the preparation of the program and moderation of the live stream in order to ensure an engaging experience for the participants. The webinars were attended live by 638 people, that is 70 participants per webinar on average. Additional outreach was

gained through playbacks of the recordings, published on the SSHOC YouTube channel,[1] which have so far accounted for another 1127 views (data obtained on 28/02/2022). In total, the webinars reached 1765 people. Thanks to the online format, the webinars attracted a very diverse audience with regard to participants' geographical location (on average 25 countries were represented at each webinar). Training webinars successfully reached its key target groups which included researchers, research performing institutions and research libraries, but the webinars were also followed by other stakeholders identified as relevant for SSHOC (e.g. research infrastructures, private sector, or civil society). The webinars were followed by a blogpost, furthermore presentations slides and recordings were uploaded to SSHOC channels for future use. When relevant, lessons learned about the organisation of webinars were informally shared with other SSHOC members in order to contribute to knowledge sharing about impactful online training events.

Training webinars proved to be a cost- and time-efficient training format which was warmly welcomed by user communities. The attendance numbers and the feedback show that the webinars addressed current topics, successfully engaged with diverse user communities and established collaborations between speakers that extend beyond the SSHOC project. Due to the online format, webinars were highly accessible to a great variety of individuals regardless of their background, geographical location, family situation, work duties, and career level. Despite the *virtual fatigue* due to the COVID-19 pandemic, these training webinars managed to attract a high number of participants and covered a wide variety of topics which were intertwined with other efforts realised within SSHOC. The webinars thus crucially contributed to one of the main SSHOC objectives—empowering individuals to maximise data re-use through Open Science and FAIR principles.

Abbreviations and Acronyms

FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
Q&A	Questions and Answers
RDM	Research Data Management
SSH	Social Sciences and Humanities
SSHOC	Social Sciences & Humanities Open Cloud
WP	Work Package

[1] SSHOC YouTube Channel: https://www.youtube.com/channel/UCw-mY8v84yeHW2z4KG3ZLtA/featured

Table of contents

1. Introduction

This report concerns the targeted training webinars organised by SSHOC WP6, Task 6.5. The aim of the webinars was to maximize the impact of SSHOC among data users, data producers and data experts. This was done by providing them with relevant training content that helps transfer the knowledge needed to foster cross-disciplinary cooperation in the SSH and fully leverage SSHOC services, tools and data in order to facilitate and promote data-driven research in the SSH.

Targeted training webinars were part of training activities provided by Task 6.5 and formed a methodological training pair with training workshops. The webinars functioned as teasers or follow-ups to workshops and provided complementary information on selected topics. Because the training webinars and workshops were conceptualized as complementary events, several organisational aspects overlapped. This overlap was further amplified by the fact that both events, although initially planned as online (webinars) and onsite (workshops) events, they were delivered online due to pandemic-related restrictions. For this reason, this report describes webinar-specific aspects, while it refers to Deliverable 6.14 *Report on Training Workshops* (Pahor de Maiti & Fišer, 2022) in all other aspects shared with training workshops. For additional information about the rationale for the events and their focus, please see Section 1 of Deliverable 6.14 *Report on Training Workshops* (ibid.).

This report is organised as follows: Section 2 of this report outlines the webinar-specific aspects of the process followed in the organisation of targeted training events, Section 3 provides a summarized list of the webinars, and Section 4 gives an overview of the outcomes. Full reports for all webinars are annexed under Section 6.

2. Organisation and Structure of the Webinars

In general, the webinars, that are online events, lasted for an hour and consisted of presentations given by up to three speakers and a Q&A session at the end. Each webinar included a short presentation of the SSHOC project and its connection to the project's goals and/or outcomes. The webinars were moderated by a host who opened the event, introduced presenters and moderated the Q&A session, while technical support was provided by a separate person in the background. Due to time limitations, the webinars mainly pursued an informative and/or demonstrative function, but when relevant, short interactive exchanges were encouraged (e.g., in break-out rooms, or through live polling options). Interactive elements were more thoroughly addressed by training workshops (cf. Pahor de Maiti & Fišer, 2022).

For more details about the organisational process, the general challenges faced and mitigation measures used, see Section 2 of the Deliverable 6.14 *Report on Training Workshops* (ibid.).

3. Summary of the Webinars

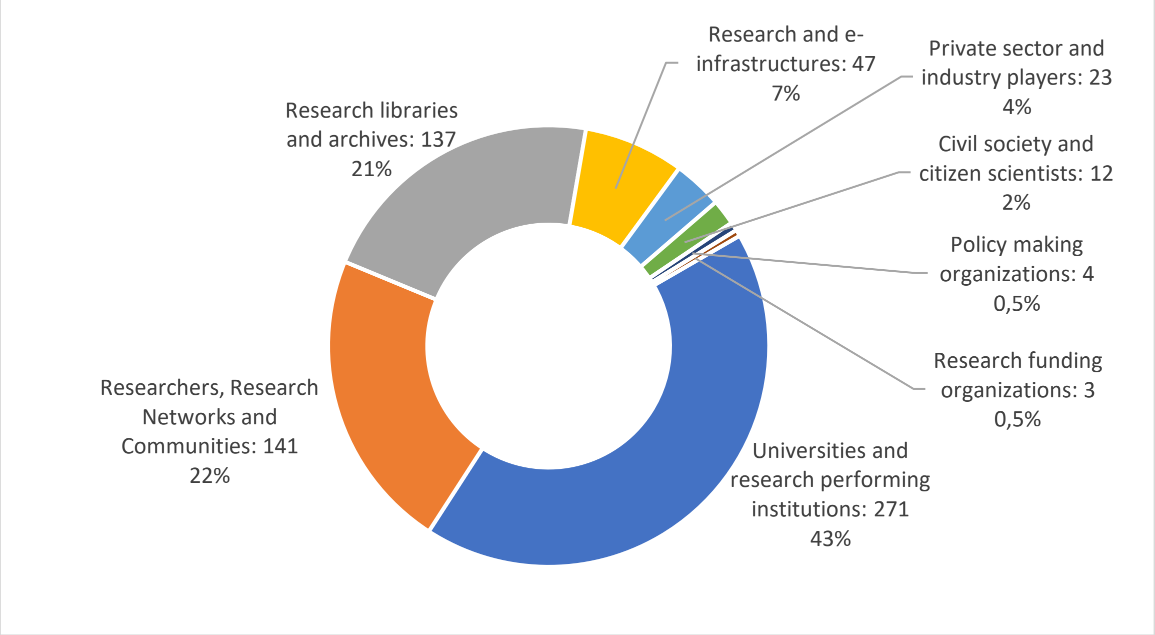
Overall, nine webinars were organised from March 2019 to December 2021: six were required as per the SSHOC Grant Agreement and three were organised in addition (see Table 1).

THEMATIC CLUSTER	WEBINARS
Data Science for the SSH	2
Data Science for Heritage Science	1
Data Protection and the GDPR	1
Data Stewardship and RDM	2
Data Citation	1
Text Mining for the SSH	2

Table 1: The number of webinars per thematic cluster

The webinars were attended live by **638 participants**, that means on average by 70 participants per webinar. However, the overall reach is higher thanks to recording playbacks. These account for another 1127 views (on 28/02/2022) which gives a total reach of 1765 individuals. The geographical coverage of the webinars was broad, since webinar participants represented on average 25 different countries per webinar covering Europe (EU and non-EU member states), Africa, North and South America, Asia, Australia, and the Middle East.

The distribution within the stakeholder categories is represented in Graph 1. A great majority (85%) is taken up by the target groups which were identified as key target audiences for training webinars, i.e., *Researchers, research networks and communities; Research libraries and archives; and Universities and research performing institutions*. Other stakeholder groups were also represented at the webinars, but with a smaller share.



Graph 1: The number and percentage of participants per stakeholder group

For a comparative analysis of workshops and webinars, see Section 3 and Section 4 of the Deliverable 6.14 *Report on Training Workshops* (Pahor de Maiti & Fišer, 2022).

The following sections provide a list of the webinars according to the predefined topics in the Grant Agreement together with a brief overview of the target audience, organisers and the links to published materials. Each webinar is described in further detail (e.g., the aim of the webinar, content of the presentations, participants' feedback) in the particular webinar report annexed at the end (see Section 6. *Annexes*).

3.1 Data Science for the SSH

3.1.1 Hands-on Tutorial on Transcribing Interview Data

Date and Venue	03/2019; online
Links to materials	
Announcement	https://www.sshopencloud.eu/hands-on-tutorial-transcribing-interview-data
Blogpost	https://sshopencloud.eu/news/sshoc-webinar-clarin-hands-tutorial-transcribing-interview-data
Presentation	https://doi.org/10.5281/zenodo.3694223
Recording	https://www.youtube.com/watch?v=X6bFGJpMjVQ

Audience	
Total participants	86
Recording playbacks[2]	346
Targeted audience	Researchers and others working with audio data
By stakeholder category	University and/or research performing institution: 64 Research library and archive: 9 Private sector and industry player: 4 Research and e-infrastructure: 4 Researcher: 2 Civil society and citizen scientist: 1 Policy making organisation: 1 Research funding organisation: 1
Organisers	T6.5 in cooperation with CLARIN ERIC and SSHOC T4.4
Workshop report	see Annex 1: Webinar report: Hands-on Tutorial on Transcribing Interview Data

3.1.2 Sharing Datasets of Pathological Speech

Date and Venue	10/2020; online
Links to materials	
Announcement	https://sshopencloud.eu/events/sshoc-webinar-sharing-datasets-pathological-speech
Blogpost	https://www.sshopencloud.eu/news/webinar-notes-sharing-datasets-pathological-speech
Presentation	https://doi.org/10.5281/zenodo.4081578
Recording	https://www.youtube.com/watch?v=qjTJ4ZxzfvI
Audience	
Total participants	73

[2] For all events, the numbers of recording playbacks have been collected on 28/02/2022.

Recording playbacks	52
Targeted audience	Researchers and others working with corpora of speech disorders
By stakeholder category	University and/or research performing organisation: 29 Researcher: 26 Research library and/or archive: 9 Research and e-infrastructure: 4 Civil society and/or citizen scientist: 3 Private sector and/or industry player: 2
Organisers	T6.5 in cooperation with CLARIN ERIC and SSHOC T5.4
Workshop report	see Annex 2: Webinar report: Sharing Datasets of Pathological Speech

3.2 Data Science for Heritage Science

3.2.1 Use and Re-Use of Scientific Data in Archaeology and Heritage

Date and Venue	04/2020; online
Links to materials	
Announcement	https://www.sshopencloud.eu/events/use-and-re-use-scientific-data-archaeology-and-heritage
Blogpost	https://www.sshopencloud.eu/news/sshoc-webinar-use-and-re-use-scientific-data-archaeology-and-heritage
Presentation	https://doi.org/10.5281/zenodo.3783568 https://doi.org/10.5281/zenodo.3783583 https://doi.org/10.5281/zenodo.3783586 https://doi.org/10.5281/zenodo.3783589
Recording	https://www.youtube.com/watch?v=h6V9KDtfpYQ https://www.youtube.com/watch?v=z_YV-EnbQL0
Audience	
Total participants	32
Recording playbacks	Part 1: 92 Part 2: 47
Targeted audience	Researchers and others interested in novel policy and best practice around use and re-use of scientific data within heritage science and archaeology

By stakeholder category	University and/or research performing institution: 17 Research & e-infrastructure: 7 Policy making organisation: 3 Research library and archive: 2 Private sector and industry player: 1 Researcher: 1 Research funding organisation: 1
Organisers	T6.5 in cooperation with UCL
Workshop report	see Annex 3: Webinar report: Use and Re-use of Scientific Data in Archaeology and Heritage

3.3 Data Protection and the GDPR

3.3.1 GDPR and the DARIAH ELDAH Consent Form Wizard

Date and Venue	10/2020; online
Links to materials	
Announcement	https://www.sshopencloud.eu/putting-data-protection-practice-gdpr-and-dariah-eldah-consent-form-wizard-0
Blogpost	https://www.sshopencloud.eu/news/webinar-notes-gdpr-and-dariah-eldah-consent-form-wizard
Presentation	https://doi.org/10.5281/zenodo.4090351
Recording	https://www.youtube.com/watch?v=eAKhI0qde2w
Audience	
Total participants	41
Recording playbacks	107
Targeted audience	Researchers interested in ensuring GDPR compliance of their work
By stakeholder category	University and/or research performing organisation: 15 Research library and/or archive: 13 Researchers, research networks and communities: 9 Research and e-infrastructure: 2

	Civil society and/or citizen scientist: 1 Private sector and/or industry player: 1
Organisers	T6.5 in cooperation with DARIAH and CLARIN ERIC
Workshop report	see Annex 4: Webinar report: GDPR and the DARIAH ELDAH Consent Form Wizard

3.4 Data Stewardship and RDM

3.4.1 Tools and Resources for FAIR Data

Date and Venue	05/2020; online
Links to materials	
Announcement	https://www.sshopencloud.eu/sshoc-webinar-emm-survey-data-fair
Blogpost	https://www.sshopencloud.eu/news/tools-and-resources-fair-data-sshoc-webinar-notes
Presentation	https://doi.org/10.5281/zenodo.3831892
Recording	https://www.youtube.com/watch?v=QxV7Nwii3GU
Audience	
Total participants	136
Recording playbacks	180
Targeted audience	Researchers and others interested in ensuring FAIR principles in the entire data lifecycle
By stakeholder category	Research library and/or archive: 50 University and/or research performing organisation: 47 Researcher: 23 Research and e-infrastructure: 9 Private sector and/or industry player: 5 Civil society and/or citizen scientist: 2
Organisers	T6.5 in cooperation with CESSDA/UKDS
Workshop report	see Annex 5: Webinar report: Tools and Resources for FAIR Data

3.4.2 Introducing the Newly Launched EMM Survey Registry

Date and Venue	10/2020; online
Links to materials	
Announcement	https://www.sshopencloud.eu/events/sshoc-webinar-introducing-newly-launched-ethnic-and-migrant-minorities-emm-survey-registry
Blogpost	https://sshopencloud.eu/news/webinar-notes-introducing-newly-launched-emm-survey-registry
Presentation	https://doi.org/10.5281/zenodo.4134060
Recording	https://www.youtube.com/watch?v=UFRj6Lz0v_w
Audience	
Total participants	107
Recording playbacks	130
Targeted audience	Researchers and others interested in working with or learning more about quantitative survey research on Ethnic and Migrant Minorities' (EMM) integration
By stakeholder category	University and/or research performing organisation: 50 Researcher: 44 Private sector and/or industry player: 4 Research library and/or archive: 4 Research and e-infrastructure: 3 Civil society and/or citizen scientist: 2
Organisers	T6.5 in cooperation with Sciences-Po and SSHOC T9.2
Workshop report	see Annex 6: Webinar report: Introducing the Newly Launched Ethnic and Migrant Minorities (EMM) Survey Registry

3.5 Data Citation

3.5.1 FAIR SSH Data Citation: Practical Guide

Date and Venue	12/2021; online
Links to materials	
Announcement	https://www.sshopencloud.eu/events/fair-ssh-data-citation-practical-guide

Blogpost	https://www.sshopencloud.eu/news/sshoc-workshop-notes-fair-ssh-data-citation-practical-guide
Presentation	https://doi.org/10.5281/zenodo.5751880
Recording	https://www.youtube.com/watch?v=dRMAnuxvY88
Audience	
Total participants	41
Recording playbacks	48
Targeted audience	Researchers and repository managers as well as anyone else interested in data citation in the SSH domain
By stakeholder category	University and/or research performing organisation: 11 Research and e-infrastructure: 10 Research library and/or archive: 10 Researcher: 7 Civil society and/or citizen scientist: 2 Research funding organisation: 1
Organisers	T6.5 in cooperation with CNRS (Huma-Num) and CLARIN ERIC; WP3
Workshop report	see Annex 7: Webinar report: FAIR SSH Data citation: Practical Guide

3.6 Text Mining for the SSH

3.6.1 Quanlify with Ease: Combining Quantitative and Qualitative Corpus Analysis

Date and Venue	04/2020; online
Links to materials	
Announcement	https://www.sshopencloud.eu/3rd-sshoc-webinar-quanlify-combining-quantitative-qualitative-corpus-analysis
Blogpost	https://www.sshopencloud.eu/news/quanlify-ease-combining-quantitative-and-qualitative-corpus-analysis-sshoc-webinar-notes
Presentation	https://doi.org/10.5281/zenodo.3754060
Recording	https://www.youtube.com/watch?v=SkXchcHJk6I
Audience	

Total participants	75
Recording playbacks	93
Targeted audience	Researchers and others interested in analysing large text collections by combining quantitative and qualitative methodology
By stakeholder category	University and/or research performing organisation: 26 Researcher: 25 Research library and/or archive: 10 Research and e-infrastructure: 7 Private sector and/or industry player: 6 Civil society and/or citizen scientist: 1
Organisers	T6.5 in cooperation with CLARIN ERIC
Workshop report	see Annex 8: Webinar report: Quanlify with Ease: Combining Quantitative and Qualitative Corpus Analysis

3.6.2 SHOC'ing Drama in the Cloud: the Added Value of SSHOC/CLARIN Services

Date and Venue	06/2021; online
Links to materials	
Announcement	https://www.sshopencloud.eu/events/sshoc%E2%80%99ing-drama-cloud
Blogpost	https://sshopencloud.eu/news/sshoc-webinar-notes-sshocing-drama-cloud-added-value-sshocclarin-services-post-event-report
Presentation	https://doi.org/10.5281/zenodo.5082521
Recording	https://www.youtube.com/watch?v=KJmI3C20KiE
Audience	
Total participants	47
Recording playbacks	32
Targeted audience	Librarians and others interested in services offered by the SSH research infrastructures

By stakeholder category	Research library and/or archive: 30 University and/or research performing organisation: 12 Researcher: 4 Research and e-infrastructure: 1
Organisers	T6.5 in cooperation with CLARIN ERIC
Workshop report	see Annex 9: Webinar report: SSHOC'ing Drama in the Cloud – Encoding Theatrical Text Collections and the Added Value of SSHOC & CLARIN Services

4. Outcomes and Conclusions

Training webinars represent a crucial contribution to the efforts made by SSHOC members to promote the achievements of the project and transfer invaluable skills and knowledge to the SSH communities. Targeted training activities ensure sustainable use of SSHOC outcomes in the future and encourage data-driven interdisciplinary research approaches. The initial set of six webinars which covered a broad array of topics was expanded with three additional webinars addressing specific training needs identified by the SSHOC community and/or showcasing tools and services developed in the SSHOC project. Such was, for example, the webinar *Introducing the newly launched EMM Survey Registry* (see Section 3.4.2 Introducing the Newly Launched EMM Survey Registry for more) which was the second most attended webinar with over 100 live participants. The webinars were successful in establishing connections inside the SSHOC community and beyond by inviting speakers from other projects, initiatives, and institutions. A great example of in-house cooperation was the webinar *FAIR SSH Data citation: practical guide* delivered by experts from WP3 and WP6 (see Section 3.5.1 FAIR SSH Data Citation: Practical Guide for more). External collaboration proved valuable, for example, for the webinar *Hands-on Tutorial on Transcribing Interview Data* (see Section 3.1.1 Hands-on Tutorial on Transcribing Interview Data for more) which brought together experts from SSHOC and the *Speech Data & Technology Working Group*.[3]

High attendance at each webinar (i.e., 70 participants/event on average; in total 1765 people (638 participants of live events and 1127 views of recordings)) confirms that the topics and speakers were highly relevant Lively Q&A sessions at the end of the webinars as well as participants' feedback prove that the presentations were valuable and thought provoking. The organisers' experience shows that webinars are a cost- and time-effective form of training that is also environmentally friendly due to its online format. This is valuable for both generic and specific training activities when the content is well

[3] Speech Data and Technology: https://speechandtech.eu/home/who-we-are

adapted to an online format and approximately one-hour time frame. Examples of such trainings with high attendance rates are, on one hand, the topic generic webinar *Tools and Resources for FAIR Data* which was also the most well attended webinar attracting almost 140 live participants (see Section 3.4.1 for more). On the other hand, the highly specific webinar *Sharing Datasets of Pathological Speech* (see Section 3.1.2 for more) which was attended by slightly more than 70 participants. These numbers demonstrate the importance of training events for highly specialized research communities. Furthermore, the webinars captured a highly diverse audience. At each webinar, the participants came from 25 different countries on average and represented six stakeholder categories —among which the most prominent ones were *Researchers, research networks and communities; Research libraries and archives; and Universities and research performing institutions* (accounting for 85% of the audience).

The attendance report data confirms that webinars were successfully organised and delivered since all Key Performance Indicators (Torma et al., 2019, p. 21) were reached or exceeded: the attendance numbers were always above the set minimum of 20 per webinar (minimum: 32, maximum: 136, average: 70); the distribution in terms of geographical location and stakeholder groups was very diverse; the three key target groups (researchers, research performing institutions, and research libraries) represented a great majority of the audiences; the webinars were equally distributed across the project's timeframe; and the final webinar count (9 webinars) exceeded the required minimum (6 webinars).

Lastly, training activities in the form of targeted training webinars resulted in a rich set of online materials in the form of blogposts, presentation slides and webinar recordings. They extend the positive impact of the webinars beyond the actual live stream and continue to promote the vision of the SSHOC project after its formal conclusion. The organisational process also had a positive side effect: the experiences, best practices and recommendations were informally shared with the SSHOC community on several occasions in order to support the efforts put into the organisation of future high-impact training events.

In conclusion, the SSHOC T.5 organisers' experience shows that training webinars are an invaluable training opportunity highly sought after in the research community. The format offers a cost- and time-efficient way of transferring skills, knowledge and information about new developments and findings in the SSH community. Webinars are also a highly inclusive format of training events since they can be attended by a great variety of individuals regardless of their personal background or career situation. Despite much lower networking potential in comparison to more extensive face-to-face workshops, webinars, if planned correctly, can nonetheless be engaging for the audience and can encourage an active exchange of ideas and contacts. Based on the T6.5 members' results, it can be concluded that the webinars importantly contributed to the success of targeted training offered by Task 6.5 in the period struck by the pandemic. Training webinars and workshops thus helped promote maximal reuse of tools, services, and data in line with the Open Science and FAIR principles by offering trainings on current topics with high quality presentations. The event series leaves behind a set of valuable and openly accessible materials for future use, such as presentations and recordings.

5. Reference list

Pahor de Maiti, K. & Fišer, D. (2022). SSHOC D6.14 Report on Training Workshops. Zenodo.

Schwabe, A., Ausserhofer, J., Marino, L., Willems, M., Kalaitzi, V., Vipavc Brvar, I., Smith, E., & Muscella, S. (2019). *SSHOC D2.1 Overall Communication and Outreach Plan (approved 18 Nov 2019)*. Zenodo. https://doi.org/10.5281/zenodo.3595936

Torma, M., Kalaitzi, V., Dijk, E., Wittenberg, M., Willems, M., Fišer, D., Pahor de Maiti, K., Durco, M. & Vipavc Brvar, I. (2019). SSHOC D6.2 Building Expertise Strategy (v1.0). Zenodo. https://doi.org/10.5281/zenodo.4558294

6. Annexes

Annex 1: Webinar report: Hands-on Tutorial on Transcribing Interview Data

Background

The webinar — CLARIN Hands-on Tutorial on Transcribing Interview Data — was held on 3 March 2020 and was organised as a follow-up to the SSHOC workshop — *The Case of Interview Data – A Multidisciplinary Approach to the Use of Technology in Research* — which was held on 6 July 2019 at the Digital Humanities 2019 conference in Utrecht. It was organised by T4.4 in cooperation with T6.5 of the SSHOC project on the pre-defined topic of Data Science for the Social Science and Humanities.

As a follow-up to the event, the webinar was also focused on the data-creation phase – the first phase of the data lifecycle – and thus highly important for the following steps. Providing means and guidelines for effective data management lifecycle is one of the major goals of the SSHOC project, while Work Package 4 (WP4) specifically addresses the data-creation phase.

Webinar Overview & Format

Aim. Spoken audio data, such as interview data, is a scientific instrument used by researchers in various disciplines. Despite different scientific methods of analysis used by these researchers, core processing methods of this kind of data are cross-disciplinary. Creating transcriptions with an appropriate level of detail is one of the initial and most important steps in the spoken audio data analysis, but this step can also be very time-consuming. This is why researchers can greatly benefit from at least partial automation of the transcription process. However, choosing high-quality tools and learning how to use them is not always a straightforward process, and researchers can quickly lose their enthusiasm for automation for the fear of that the automation process might be too complex or non-transparent.

The webinar focused on the presentation of the central part of the transcription workflow (developed by the Oral History & Technology research group) which is integrated into the OH-portal. This portal provides researchers with high-quality and easy-to-use tools that help them move from a digital audio signal to a time-aligned transcript.

Speakers. The webinar was delivered by 2 speakers who are part of the Oral History & Technology research group:

- Henk van den Heuvel (Radboud University, Nijmegen), and
- Christoph Draxler (Ludwig Maximilian University Munich).

Organisers. The webinar was organised in cooperation with the partners in T6.5 and members of the Oral History & Technology research group, participating in T4.4 of the SSHOC project: Christoph Draxler (Ludwig Maximilian University Munich) and Henk van den Heuvel (Radboud University Nijmegen).

Participants. There were 172 viewers of the webinar.[4] The majority of participants came from the EU countries, but the webinar was also followed by some participants from countries outside Europe (i.e. the USA, several African countries, China, etc.). The audience of the livestream included all stakeholder categories identified in D6.1. The great majority (approx. 70%), belonged to the categories: "Researchers, Research Networks and Communities"; "Universities and Research Performing Institutions". These two categories were followed by: "Research Libraries and Archives"; "Research and E-infrastructures"; "Private Sector and Industry Players"; their representation accounted to approximately 20% of the entire audience. The remaining categories: "Policy Making Organisations"; "Research Funding Organisations"; "Civil Society and Citizen Scientists" were represented only by a few participants (approx. 10%).

Brief summary of the event structure. The webinar lasted for a full hour and was divided into 4 parts. After the introduction of the house rules and webinar etiquette, Henk van den Heuvel shortly introduced SSHOC and CLARIN ERIC. This short session was followed by the main part, which was divided between the two speakers. Henk van den Heuvel first presented the cross-disciplinary appeal of automated workflows for the analysis of spoken audio data together with the context of the initiative for the transcription portal. He then handed the microphone over to Christoph Draxler, who first introduced the participants to the theoretical basis for good transcription outputs and demonstrated the technology available through the OH-portal for transcribing spoken language. The webinar concluded with an informative Q&A session.

Presentations & Discussions: Key Points

First Session. Presentation of the SSHOC project and CLARIN ERIC infrastructure
Speakers. Henk van den Heuvel
Main points. The speaker presented the goals and the expected impact of the SSHOC project and placed the webinar in relation to them. Furthermore, he briefly presented the CLARIN ERIC infrastructure, which

[4] The number consists of viewers of the livestream (86) and of viewers of the webinar recording (86). It should also be noted, that the number of the recording views was extracted on 2 April 2020 and is subject to change.

is directly linked to the portal demonstrated in this webinar. CLARIN ERIC proves to be a relevant partner in the SSHOC project by providing language resources and technologies, and by boosting cross-disciplinary cooperation between researchers that use language data in their work.

Links to materials.

- Presentation slides (pp: 5–9)
- Webinar recording (min: 4:23–7:44)

Second Session. The initiative for the transcription portal

Speaker. Henk van den Heuvel

Main points. Interviews are a central research instrument for oral historians, but are also widely used in many other disciplines (Law, Social Economy, Medicine, etc.). The initiative for a transcription portal, which is supported by CLARIN ERIC, started from a series of discussions on the topic of oral history. Tools and services for data types used in oral history are not yet as well-established as in some other disciplines. The initiative's taskforce developed a transcription chain and integrated it into the OH-portal. The portal combines different services needed to process spoken audio data. One of the main steps in the transcription workflow is automatic speech recognition (ASR), however, other natural language processing (NLP) tools push the automation process further and provide enriched final results.

Links to materials.

- Presentation slides (pp: 10–15)
- Webinar recording (min: 7:46–13:00)

Third Session. ASR and the OH-portal

Speaker. Christoph Draxler

Main points. A high-quality orthographic transcript is the basis for all types of analyses of spoken language data. Since transcribing speech manually is a time-consuming task, ASR tools have been recognized as important developments in the last years. The OH-portal enables access to external providers of ASR services for a number of languages (English, German, Dutch, Italian, and Czech in preparation). In addition, the portal has an editor for a manual correction of the transcripts, a tool for automatic word-time alignment and an interface to look into phonetic details. Files can be exported in various formats, which comes in handy for further analyses with different NLP tools.

Transcripts can differ in the level of details they provide, so it is important to know in advance what purposes the transcript will serve. For example, it is never a good idea to dispose of the audio signal even when a detailed transcript has been produced, indeed detailed transcripts might not fit all kinds of analysis, and it might be easier to repeat the transcribing process than to clean an existing transcript of redundant information.

The OH-portal includes services needed to address different stages of the transcription chain. During the webinar, the speaker guided the participants step-by-step through the process of creating a transcript with the help of multiple examples. He also added some useful tips, which can save the users some frustration when using the portal (i.e., the importance of good audio signal, technical restrictions of the portal, privacy issues, etc.). The attendees got a clear demonstration of the features of the portal and learned where and to what extent manual intervention into data is necessary.

Links to materials.

- Presentation slides (pp: 16–45)
- Webinar recording (min: 13:08–50:33)

Fourth Session. Q&A session
Speaker. Henk van den Heuvel and Christoph Draxler
Main points. The participants were interested in ASR performance on data from speakers with special linguistic accents. The best way to find this out is currently by trial. The use of ASR for special topics, such as ASR for dialects or for speakers with language disorders are currently at the centre of captivating research. Other questions addressed technical aspects regarding the files and the portal: for example, the input and output file formats; the languages available for processing; the file sizes that can be processed by the services integrated into the portal; the need to downsample the signal; the use of special characters in file names; the recording equipment, etc. Participants also asked for some tips on what to do when one signal encompasses multiple speakers. Unfortunately, the only solution for this kind of situation in order for the ASR to work is to have either as many microphones as the speakers, or alternatively, to have very disciplined speakers. Usually, these kinds of situations require manual intervention. Participants were also curious about the GDPR compliance of the incorporated services and about storing policy of the processed files. Christoph Draxler explained that the user can check the privacy policy of the ASR providers and select a service accordingly (some services reserve the right to perform additional analysis of the audio files or to keep the files, others don't). Files are kept for a maximum of 24 hours in order to allow the users to resume their work in case they needed to interrupt the process.

Links to materials.

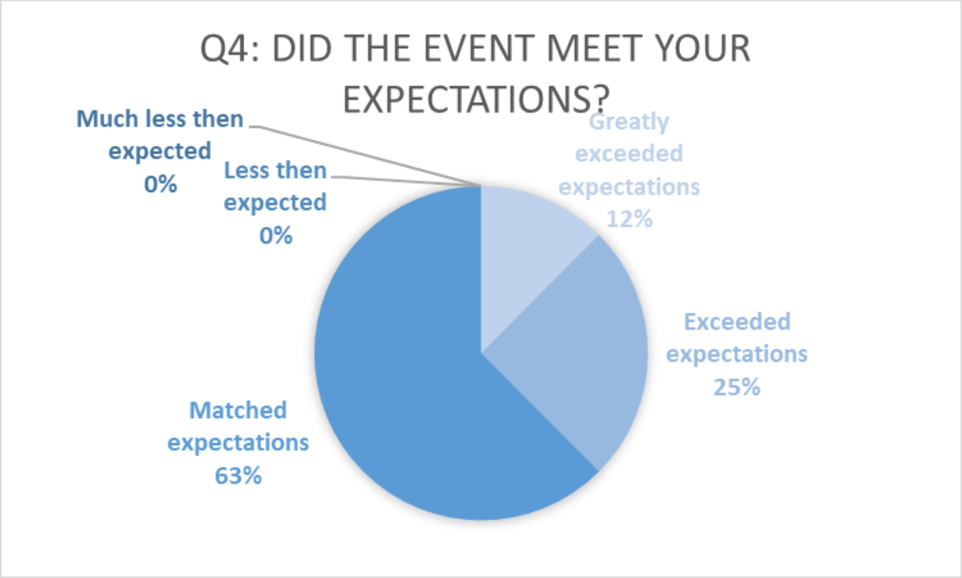
- Presentation slides (pp: 46–47)
- Webinar recording (min: 50:35–1:00:52)

Outcomes & Feedback

By following the webinar, the participants were able to discover the transcription workflow and learned how to use the services in the OH-portal. They were also given some practical advice on spoken audio data as a data type and about the operation of the OH-portal. By aggregating state-of-the-art services and offering an easy-to-use workflow, the portal encourages cross-disciplinary cooperation and represents an important contribution to the processing of less well-established data (i.e. the spoken audio data). The portal also promotes the objectives of the SSHOC project. For this reason and also due to its accessibility and careful management of the data in terms of privacy, the OH-portal is an important tool for researchers that use spoken audio data and seek to partially automate their work.

Participant satisfaction. Participants rated the webinar as either 'Very good' (56%) or 'Excellent' (44%) (cf. Q1). The majority stated that the webinar "Matched their expectations' (63%), while other participants felt that the webinar 'Exceeded their expectations' (25%) or even 'Greatly exceeded expectations' (12 %) (cf. Q4).

> "I have found an optimum amount of information for an hour." (Respondent no.7, Q7)



Participants really appreciated the structure of the webinar, demonstration of the portal and practical advice (cf. Q6). Several of them also mentioned that they liked the online format. The positive aspects of the online format could be summarized in the following statement:

> "/…/ it saves time and money for travel and in the same time I can ask or comment." (Respondent no. 11, Q6)

When asked about possible improvements to the content of the webinar, the participants mentioned that they would appreciate having more time to try out their own audio clips, getting some information about the sustainability of the portal, as well as having pointers to scientific work referencing the transcription chain (cf. Q7).

The big majority stated that the webinar will have a direct positive impact on their work: either by using the portal in their own research or by being able to direct interested individuals to the portal (cf. Q5). Participant no.8 (Q5) sees the advantage of using the OH-portal in the fact that privacy issues linked to the use of spoken audio data were not overlooked in the design of the portal.

Feedback regarding the Organisation. Participants generally felt the webinar was 'Excellently' or 'Well organised' (cf. Q9). The only Organisational remark concerning possible improvements referred to the fact that the participants were not informed about the sign-in requirements of the OH-portal in advance as articulated by Respondent no.16 in Q7:

> "/.../ it would have been nice to know in advance that we had to have an institutional/CLARIN login to access the tool ourselves."

Regarding the promotion of the webinar, we still see that more could be done through SSHOC channels (e.g., SSHOC newsletter) as none of the respondents learnt about the event through dedicated SSHOC channels (cf. Q2). Nonetheless, the good response in terms of the number of participants shows that the promotion of the event has been successful.

Future work. Based on the respondents' answers (Q3&5) we can claim that the topic in focus for this webinar on the pre-defined topic of "data science for social sciences and humanities" was chosen well. Furthermore, answers like:

> "/other webinars on this topic or similar would be interesting" (Respondent no.5, Q8),

show that knowledge about this topic is in demand. This is not surprising given that spoken audio data is used by researchers from various scientific disciplines. The T6.5 organisers of this webinar do not plan any future events on this topic at the moment. However, in case there will be great interest from the community and available means to organise additional events, T6.5 team will consider organising events that could expand on the topic of spoken audio data processing. Nonetheless, in the framework of SSHOC, future events on this topic are planned, e.g. SSHOC Workshop - Linking Social Survey and Linguistic Infrastructures through EOSC.

ANNEX 2: WEBINAR REPORT: SHARING DATASETS OF PATHOLOGICAL SPEECH

Background

The webinar — Sharing Datasets of Pathological Speech — was held on 14 October 2020 and was organised in cooperation between the DELA initiative and SSHOC T5.4 and T6.5 as an additional SSHOC T6.5 training activity with the aim to promote the best practices for sharing datasets of pathological speech including innovative approaches developed in the framework of SSHOC. The webinar is part of T6.5 activities that fall under the broad topic of "Data Science for the Social Science and Humanities".

In general, the webinar focused on data sharing and open access which are two crucial aspects in the process of ensuring Open Science, but more specifically, the webinar dealt with sensitive data which requires especially careful treatment. Therefore, the webinar lent itself also to the promotion of T5.4 efforts since T5.4 specifically works on enhancing and extending the infrastructure for secure remote access to sensitive research data while respecting all relevant legislation and the best interest of data owners. Moreover, the webinar provided much needed targeted content for the research community dealing with the niche language resources, such as datasets of pathological speech, and addressed real-life problems, offered solutions through use cases, provided support and sketched future steps in the development of providing solutions to process sensitive research data.

Webinar Overview & Format

Aim. Corpora and datasets of pathological speech also called speech corpora of individuals with communication disorders (CSD) are hard to get because they are costly to obtain, but mainly because they are hard to share. In order to promote best practices and raise awareness regarding processing of such data, experts from DELAD initiative and SSHOC Tasks 5.4 and 6.5 organised a webinar where a group of speakers presented several alternatives for obtaining, processing and sharing CSD.

The webinar addressed several related topics:

- Progress achieved by the DELAD initiative for sharing corpora of speech disorders (CSD) and the role of the CLARIN Knowledge Centre on Atypical Communication Expertise;
- GDPR and the ethics of special category data relevant for collecting and sharing CSD;

- How storing and sharing CSD is arranged in a GDPR compliant way at The Language Archive of the Max Plank Institute for Psycholinguistics and the collaboration with the Talkbank;
- The CAVA audio-visual human communication archive project - a digital video repository to support the work of the international human communication research community. This resource enhances the discoverability and re-usability of expensively-created, specialist video content;
- Infrastructure requirements for secure remote access to sensitive research data with diverse legal (e.g. social media terms of service), ethical (e.g. children as subjects), and technical (e.g. audio and video) challenges, and assessment of several existing platforms;
- The curation and disclosure of pathological speech corpora: how CSD can be found through one Organisation and made accessible through another - includes a demonstration using the example of the Polish Cued Speech Corpus of Hearing-Impaired Children.

Speakers. The webinar was delivered by five speakers who participate in the DELAD initiative and the SSHOC project or independently cooperate with these teams:

- Henk van den Heuvel (CLST, Radboud University, DELAD and SSHOC T5.4)
- Nicola Bessell (Department of Speech and Hearing Sciences, University College Cork, DELAD)
- Paul Trilsbeek (The Language Archive, Max Planck Institute for Psycholinguistics)
- Libby Bishop (GESIS - Leibniz Institute for Social Sciences, SSHOC T5.4)
- Katarzyna Klessa (Adam Mickiewicz University, DELAD)

Organisers. The webinar was organised in cooperation among the DELAD initiative and SSHOC T5.4 and T6.5.

Participants. There were 90 viewers of the webinar.[5] The majority of viewers (80%) came from the EU countries, 12% from non-EU countries and the remaining share from countries outside Europe (Bahrain, Canada, etc.). The audience of the livestream included 6 stakeholder categories identified in D6.1. Almost half of all the viewers (76%) represented the categories: “Researchers, Research Networks and Communities” and “Universities and Research Performing Institutions”. These two categories were followed by “Research Libraries and Archives” (12%), while the representation of the remaining three categories: “Research and E-infrastructures”; “Civil Society and Citizen Scientists”, “Private Sector and Industry Players” account to approximately 12% of the entire audience.

Brief summary of the event structure. The webinar lasted for a full hour and was divided into four main parts with an introduction to the SSHOC project, the DELAD initiative and the CLARIN infrastructure at the beginning, and a Q&A session at the end. After the introduction of the webinar etiquette and programme, Henk van den Heuvel shortly introduced SSHOC, the DELAD initiative and CLARIN ERIC. This was followed by Nicola Bessell’s talk about the GDPR provisions and the Ethics concerning special

[5] The number consists of viewers of the livestream (73) and of viewers of the webinar recording (17). It should also be noted, that the number of the recording views was extracted on 19 November 2020 and is subject to change.

category data, Paul Trilsbeek's talk on storing and sharing CSD at The Language Archive, Libby Bishop's presentation of the CAVA project and the remote access to sensitive data, and finally by Katarzyna Klessa's presentation of a use case about a corpus of speech produced by hearing-impaired children.

Presentations & Discussions: Key Points

First Session. Presentation of the SSHOC project, the DELAD initiative and the CLARIN ERIC infrastructure
Speakers. Henk van den Heuvel
Main points. The speaker presented the goals and the expected impact of the SSHOC project and placed the webinar in relation to them.

He then presented the DELAD (meaning *shared* in Swedish) which is an initiative that works towards facilitating the exchange and investigation of CSD (corpora of speech of individuals with communication disorders) in compliance with the GDPR. DELAD strives to connect to existing research infrastructures, and therefore cooperates closely with the CLARIN Knowledge Centre for Atypical Communication Expertise (ACE). The speaker briefly introduced the work of CLARIN ERIC, an important partner also in the SSHOC project, and underlined the points of cooperation between the DELAD initiative and the CLARIN infrastructure. The ACE centre collaborates with the CLARIN Data Centre at the Max Planck Institute for Psycholinguistics (The Language Archive), and with Talkbank for storage of sensitive data. DELAD also organises annual workshops covering topics related to ethical, legal and technical aspects of working with CSD. These cover everything from collecting, formatting, processing and sharing CSD, to ensuring access to such data by collaborating with existent research infrastructures and providing a quality inventory of relevant datasets.

Links to materials.

- Presentation slides
- Webinar recording (min: 0:00–8:33)

Second Session. GDPR & Ethics of special category data
Speaker. Nicola Bessell
Main points. Nicola Bessell first highlighted the ethical and GDPR considerations when collecting corpora of speech disorders. She underlined that the GDPR stipulates that processing of health-related data is only allowed for research purposes, while archiving of such data in order to be legal must be in the public interest.

In order to ensure GDPR-compliant use of CSD data, the researchers and other users of such data must obtain consent from data owners. This can be done via consent forms which need to address the following aspects:

- **Use of data**: The data user must obtain explicit consent to use the data for the intended purpose. In addition, the consent form used must outline how the confidentiality of data owners will be protected.
- **Dissemination of data:** The consent form must also list the terms that will govern the dissemination of data, and should state what future use is envisaged for research purposes.
- **Archiving of data:** It is recommended that the consent form specifies the archival period.

Links to materials.

- Presentation slides
- Webinar recording (min: 8:34–15:44)

Third Session. Storing and sharing CSD at The Language Archive
Speaker. Paul Trilsbeek
Main points. Paul Trilsbeek presented the GDPR-compliant way in which data is stored and made accessible at The Language Archive (TLA). He put special emphasis on the issues regarding the anonymization process. This process can often invalidate the data for many research purposes. Furthermore, he stressed the necessary legal agreements for archiving and sharing personal data which are in essence of two types: deposit/processing agreements and data use agreements/licenses, and underlined the need for thorough examination of licenses used since many existing licenses are "perpetual" and may therefore be in conflict with the GDPR under certain conditions.
Paul Trilsbeek also elaborated on the technical and logistic requirements needed and implemented at TLA in order to ensure "data protection by design and by default" as stated in the GDPR. This includes up-to-date systems and software, secure transport of data (HTTPS) and an elaborate system of access policies and authorisation. At the TLA, all archived copies reside within the EU at trusted data centres within the Max Planck Society, which is another important aspect for ensuring data security.

Links to materials.

- Presentation slides
- Webinar recording (min: 15:45–24:35)

Fourth Session. Presentation of the CAVA project
Speaker. Libby Bishop
Main points. Libby Bishop first shared some insights into a decade old data collection project called CAVA (Human Communication Audio-Video Archive) which includes data covering a wide range of disorders and is hosted at UCL. The project aimed at establishing a digital video repository for human communication sciences, cataloguing those videos and provide transcripts and supporting materials for them, as well as developing procedures for managing access and ensuring the sustainability of the repository. After shortly presenting the content of the repository, she addressed the legal and technical issues related to sustaining and possibly expanding such a collection. She went also into some detail

about post-project issues which mainly have to do with ensuring a sustainable financial model and choosing a platform/administrator that would ensure repository's longevity.

Links to materials.

- Presentation slides
- Webinar recording (min: 24:35–31:56)

Fifth Session. Presentation of the remote access to sensitive data
Speaker. Libby Bishop
Main points. Libby Bishop presented an innovative access method to CSD that is now explored in the SSHOC project, i.e. the remote secure access. This type of data access does not bring the data to the user, but the user to the data. The data resides at the local server and the user can perform analyses by using the tools available at the remote end. In this way only aggregated analysis results can be downloaded by the user but not the data which ensures higher security of the data but nonetheless enables easy reuse of the data for the researchers. The main concern raised was that there is currently no reliable path to a sustainable infrastructure. Open cloud-based solutions, such as those (that will be) provided by SSHOC/EOSC offer a promising way forward, but only time will tell how successful this approach really is.

Links to materials.

- Presentation slides
- Webinar recording (min: 31:57–37:58)

Sixth Session. The case of the Polish Cued Speech Corpus of Hearing-Impaired Children
Speaker. Katarzyna Klessa
Main points. Katarzyna Klessa presented a very recent curation project which includes legacy data from Polish children with hearing impairment. She specifically highlighted the legal basis for sharing the data, and issues of interoperability when it comes to obsolete data formats. The CLARIN Knowledge Centre for Atypical Communication Expertise helped make this data accessible via a new and unique sharing model: all metadata and information on the dataset can be found at the Talkbank, whereas the audio data is stored on European servers only, more specifically at The Language Archive. This is a novel and promising example for data storage and access that opens up new possibilities for European researchers, since it uses a well-established data centre in the USA for hosting the landing page and part of the CSD, whilst keeping the most sensitive part of the data on European servers.

Links to materials.

- Presentation slides
- Webinar recording (min: 37:59–50:56)

Sixth Session. Q&A session
Speaker. Henk van den Heuvel, Nicola Besselli, Libby Bishop, Paul Trilsbeek
Main points. The webinar was followed by a very informative Q&A sessions during which many practical questions have been brought up and the answers showed that often, especially regarding the GDPR, there are currently no definitive solutions. However, the speakers indicated possible paths one could take in order to avoid illegal action.

Among others, the discussion brought a useful remark regarding the use of clinical data for research purposes. Since such use is not classified as repurposing by the GDPR, it is fully allowed. In any case, Nicola Besselli suggests being as explicit as possible in the consent forms, especially when future use of the data might involve data sharing. With regard to consent forms, Libby Bishop underlined that it is good practice to store a blank consent form. This is especially important since it is a general consensus that consents are not stored in repositories, as is the case with the TLA, because this would represent an additional risk of personal data disclosure due to a possible security breach into the repository. This is, as Paul Trilsbeek put it, problematic if at a certain moment one needs to prove that consent actually exists, but then the practice is to fall back on the data depositors who, hopefully, store the original consents. Given that sometimes it is not possible to acquire written consents, one viewer asked about the possibility to use video recorded consents. The answer to this was that this option is possible, but not recommended.

Links to materials.

- Webinar recording (min: 50:57–1:01)

Outcomes & Feedback

Speech corpora of individuals with communication disorders (CSD) are highly valuable and rare language resources. They are costly to obtain and due to their personal properties hard to share among researchers. Therefore, researchers often collect the data themselves but much time could be saved and deeper research insights gained if existing data could be (re-) used more easily. The GDPR is often seen as a big hurdle in this pursuit, but as we could see in the webinar, it does not need to be if targeted support is given to researchers. The DELAD initiative was brought to life in order to help researchers with exactly that: sharing datasets of pathological speech. A part of the DELAD team is also involved in SSHOC and the project provides an excellent framework for them in their efforts to develop a sustainable, GDPR-compliant and attractive approach to the access and processing of CSD. By following the webinar, the viewers were able to raise awareness about possible issues and solutions when dealing with CSD data, get acquainted with best practices as well as raise questions and participate in a lively debate.

Participant satisfaction[6]**.** The two respondents to the post-event survey rated the event as "Excellent" or "Very good" and indicated that it "Greatly exceeded" or "Matched expectations" (cf. Q1&4). They joined the webinar because they work with the type of data in focus and/or work on corpus collection (cf. Q3). They praised the clarity of the information delivered and appreciated the exchange of views among researchers from different continents (cf. Q5&6).

Feedback regarding the Organisation. The respondents rated the organisation of the webinar as excellently or well organised (cf. Q9). They stressed, however, that the webinar could be longer and that it would be helpful if the presenters delivered their speech in a more slowly pace.

Future work. Since this webinar was targeted at a very specific community and dedicated to a presentation of issues and solutions related to the sharing of CSD, including the advances made in the framework of SSHOC, as well as being an additional activity for the T6.5 team, the T6.5 team currently does not plan to organise any follow-up events on the topic. However, provided there will be requests for related training activities, we will try to support their Organisation.

[6] It should be noted that despite the same procedure as used for our other events (including sending reminders), the response rate to the post-event survey was very low (i.e. only 2 respondents).

ANNEX 3: WEBINAR REPORT: USE AND RE-USE OF SCIENTIFIC DATA IN ARCHAEOLOGY AND HERITAGE

Background

The webinar — Use and Re-use of Scientific Data in Archaeology and Heritage — was held on 2 April 2020 and was organised as a teaser to the SSHOC T6.5 workshop. The workshop will be organised in spring 2021 and will bring additional opportunities for knowledge transfer in the Heritage Science domain. The webinar was organised by T6.5 of the SSHOC project in cooperation with the Saving European Archaeology from the Digital Dark Age COST Action (SEADDA) and the European Research Infrastructure for Heritage Science (E-RIHS) on the topic of Data Science for the Heritage Science, and in the framework of the SEADDA Exploratory Workshop on Use and Reuse of Archaeological Data.

The webinar contributed to one of the principal SSHOC goals to maximise the efficiency and effectiveness of data re-use which is a topic applicable to all SSH domains. By addressing the precise needs of the Archaeology and Heritage Science community, the webinar also fulfilled the SSHOC's aim to empower specific research communities by ensuring tailored knowledge and skills transfer.

Webinar Overview & Format

Aim. The heritage and archaeology sectors are producing increasing volumes of very diverse data through extensive use of digital technologies and data analysis. However, processes for effective use and re-use of such data are still neither clear nor universal. This webinar brought together a group of experts to provide insight into the latest guidelines and best practice for key aspects of archaeological and heritage data management, with lessons learnt extending to other fields in the SSH domain.

Heritage science and archaeology have been identified by SSHOC as a target area for training in response to their characteristic challenges. Data in this area are often derived from non-repeatable interventions, given the processes and the uniqueness of the objects and sites studied. It is also highly multidisciplinary and combines a wide range of methodologies and forms of data, often adapting novel technology, which can lead to the data being particularly fragile and subject to obsolescence. Under these conditions, management of data is complex. The purpose of the webinar was to present the latest guidance and best practice through the lens of use and reuse of data, offering principles and discoveries that are not only relevant to the sector, but can be adapted to other science and humanities contexts.

Speakers. The webinar was delivered by 5 speakers:

- Julian Richards (University of York)
- Alejandra Albuerne (UCL)
- Holly Wright (University of York)
- Jessica Hendy (University of York)
- Scott Orr (UCL)

Organisers. The webinar was organised in cooperation with partners in T6.5 of the SSHOC project and members of the Saving European Archaeology from the Digital Dark Age COST Action (SEADDA) and the European Research Infrastructure for Heritage Science (E-RIHS). These are the three leading actors in developing and promoting good-practice for the full data cycle for archaeological and heritage science data.

Participation. There were 65 viewers of the webinar.[7] The majority of viewers came from the EU countries, with some participation from countries outside Europe (e.g., the USA). The audience of the livestream included all stakeholder categories identified in D6.1, except for "Civil Society and Citizen Scientists". A majority of viewers (over 50%), belonged to the category "Universities and Research Performing Institutions". The second largest viewers pool belonged to the "Research and e-infrastructures" (over 20% of viewers) category. "Policy Making Organisations" were well represented, amounting to nearly 10% of viewers. "Research Libraries and Archives" represented 6% of participation. The remaining categories: "Researchers, Research Networks and Communities"; "Research Funding Organisations" and "Private Sector and Industry Players"; were represented only by a few viewers (approx. 10% all together).

Brief summary of the event structure. Due to the Covid-19 pandemic, the webinar, which was initially planned as a face-to-face workshop, had to be adapted for the online environment. For this reason, it lasted for three hours instead of the usual one-hour webinar format. The extension of the webinar duration allowed for more time to insightfully exchange knowledge and skills. The majority of the audience was happy to attend a longer online session due to their previous commitment to the in-person workshop. It comprised 4 presentations, each followed by a lively Q&A session that helped contextualise these initiatives and present ways for the audience to engage with them. Introductory remarks were presented by Julian Richards and followed by the first – theoretical –part, which included introductions to SSHOC, SEADDA and E-RIHS Data Curation policy (delivered by Alejandra Albuerne and Holly Wright). The second – practical – part focused on presentation of best practices (delivered by Jessica Hendy and Scott Orr).

[7] The number consists of viewers of the livestream (32) and of viewers of the webinar recording (33). It should also be noted, that the number of the recording views was extracted on 1 July 2020 and is subject to change.

Presentations & Discussions: Key Points

First Session. Introduction to SSHOC
Speakers. Alejandra Albuerne
Main points. The webinar was given in the framework of the SSHOC project. The speaker presented the goals and the expected impacts of the SSHOC project and situated the webinar in relation to them.
Links to materials.

- Presentation slides
- Webinar recording (min: 0:40–10:55)

Second Session. Introduction to SEADDA, and E-RIHS Heritage Science Data Curation Policy
Speaker. Holly Wright
Main points. FAIR principles (Findable, Accessible, Interoperable and Reusable) are broadly accepted for research data, but the question remains how to effectively implement them. Holly Wright presented the data curation policy framework recently developed by E-RIHS for heritage science data, which provides guidelines for meeting the FAIR principles. Making data open is not enough to ensure its use and re-use: it is also necessary to understand how data are being reused from both, a qualitative and quantitative perspective, which has been a key objective of E-RIHS.
The guidelines are organised around the four components of FAIR. The use of persistent identifiers for datasets and researchers is one of the key recommendations to make data findable, as well as the development of relevant metadata schemas and standards for heritage science by the relevant communities.
Accessibility can be enhanced by making datasets open wherever possible. It was recognised during the session that there are numerous instances in the sector where data privacy prevails, in which cases it is recommended that metadata is made open to make the data discoverable. Repositories with well-defined access conditions are also a must.
Interoperability requires standards that are both human- and machine-readable and use non-proprietary file formats. E-RIHS intends to offer metadata models as port of the resources in their DIGILAB.
Re-usability relies on producing data that are ready for future research processing. This requires systematic documentation, version control and even file naming, as well as sufficiently rich and consistent metadata that informs about provenance, methodology and equipment, among other things. Licensing is greatly important and needs to be clear. CC-O licensing is recommended for metadata and CC-BY for datasets wherever possible.
The presenter also encouraged the viewers to share their case-studies with the E-RHIS team to help them better understand how different methodologies and workflows for data are being used in the Heritage sector.
Links to materials.

- Presentation slides
- Webinar recording (min: 30:18–51:32)

Third Session. Challenges in practice: using molecular biology as a lens
Speaker. Jessica Hendy
Main points. One of the key strategies for enabling the re-use of data is effective data archiving. Jessica Hendy reflected on her work on molecular biology in archaeological science to offer her perspective on the practice or data archiving, discussing the possibilities and challenges it brings.

In her area of work, ancient DNA (aDNA) and ancient proteins, it is common practice to make data available in publications and to deposit data in international dedicated repositories, with up to 97% of research data being deposited for aDNA. The reasons for this large uptake include the use of readily available archive platforms in the field of molecular biology and the regular sharing of data, probably as a result of historical concerns about data authentication. The benefits of archiving data are many: it allows for a thorough analysis of the interpretation and for quality assurance, it enables the replication of data analysis strategies in peer review and after, and it provides a means of long-term data storage that is non institute-specific. Data archiving is key for allowing future research to reanalyse existing data when new strategies are developed.

There is an increasing demand for transparency on how data were generated, both in the lab and computationally, which is promoting the recording of laboratory protocols using protocols.io and computational processing using GitHub or Bitbucket:

There are nonetheless challenges in archiving aDNA:

- Datasets can be massive, and this can mean that only institutions with sufficient computational support can analyse available data. We need to be aware of the risk of certain institutions dominating the field.
- Data can be highly specialised, requiring specialised knowledge to critically interpret it.
- There remains lack of awareness and communication of what data are produced and how it is stored.

The presenter suggested strategies for addressing these challenges. First, data exploration can be made more easily accessible and not exclusively of interest to just a few research groups, which can be done, for example, through online processing capacity. Second, it is necessary to raise awareness among research partners and collaborators that it is considered standard to share data and communicate how it was acquired.

Links to materials.

- Presentation slides
- Webinar recording (min: 0:40–22:00)

Fourth Session. Best practice and tools for use and re-use
Speaker. Scott Orr

Main points. The last session of the webinar, delivered by Scott Orr, had the purpose of offering hands-on advice in the form of best practice and tools to plan for the re-use of scientific data in heritage and archaeological contexts. The session was geared towards data users and was suitable for all levels, including those new to digital data production and management.

Scientific data in the context of heritage and archaeology can take many formats and explore a diversity of issues. It ranges from scientific imaging to environmental monitoring or collection records. Potential uses include conservation and management of artefacts and sites, interpretation and engagement. The key to all this data when it comes to managing it for re-use is to focus on comprehension. The speaker's suggestion when it comes to planning the management of your data is to think of your future self: if you come back to the data in a few years' time, what will you want to know about it in order to interrogate it again?

Many points and questions were raised in this session. Highlights include:

- Metadata and paradata become integral for the re-use of dataset, offering comprehensive information about the data and how it was obtained.
- File formats need to be considered from the start. Consider the use of lossless files vs lossy files (e.g., RAW, TIFF or PNG instead of JPG or GIF). When proprietary data are obtained, consider whether it can be exported to an open format, but pay attention to potential loss of information in the process. If any information is going to be lost in going from proprietary to open format, why not save the data in both formats?
- 5-star open data is a deployment scheme for open data that goes from the most basic form of sharing data to the most comprehensive and linked way of making data available for re-use on the web. The main skills jump is in going from 3- to 4-star: from making your data available in non-proprietary open format to doing so using Uniform Resource Indicators that identify common elements that can be searched across several databases.
- It is important to think about re-use when it comes to analysis, in particular code and algorithms, which are part of the procedures used for processing and interpreting the data. Documentation and annotations, a clear format and version control are invaluable for re-use.

Links to materials.

- Presentation slides
- Webinar recording (min: 42:00–1:16:44)

Fifth Session. Q&A sessions (after each presentation)

Speakers. Interactive discussion chaired by Julian Richards (University of York)

Main points. There was lively discussion and knowledge share in the Q&A time that followed each presentation. Attendees from different disciplines participated generously with their questions., Their insights and contributions were very valuable. Contributions included ample discussion of CARE principles for Indigenous Data Governance and how they can be generalised for other context where privacy needs consideration. Other topics of discussion included the role of DOIs and URIs in linking data to sites and objects, and the challenge with speed of data publication and data embargoes in different fields of heritage and archaeological science. Ron Dekker, SSHOC coordinator and member of EOSC

Executive Board, elaborated on the active role of SSHOC in supporting the SSH community and the specific need for SSH vocabularies.

Links to materials.

- Webinar recording (after each presentation)

Outcomes & Feedback

The webinar offered the viewers a range of complimentary perspectives and resources to support the re-use of scientific data in the heritage and archaeology remits. The webinar explored what it means to re-use data, which is essential in order to plan for it from multiple perspectives. Viewers were equipped with novel guidelines for implementing FAIR principles, which is an effective data management strategy to enable future re-use of data. They were also provided with practical advice on planning for and implementing data re-use from diverse research perspectives in archaeology and heritage science. In these highly interdisciplinary and diverse fields, it is important to share best practice and key messages derived from experience to facilitate re-use. Relevant key issues for the field were discussed and disseminated, such as the benefits of open data and how to plan for it, the role of data archiving or the relevance of specialised knowledge to interpret and analyse datasets.

Viewers' satisfaction. The great majority of the viewers rated the event as Excellent or Very good (approx. 90%), the rest thought it was good (cf. Q1). For half of the respondents the webinar exceeded expectations, while the second half answered that it matched their expectations (cf. Q4).

The viewers felt that the event will positively influence their future work since it "broadened /their/ perspective of the work being done in the region" (Respondent No. 9, Q5) and provided "useful contacts" (Respondent No. 7, Q5). They especially liked the discussion that followed each presentation (cf. Q6).

Feedback regarding the Organisation. The viewers felt that the webinar was Well to Excellently organised (cf. Q9) and that the "online /edition/ made it easier to attend" (Respondent No. 6, Q3). Many also stressed that the ample discussion time added value to the event (cf. Q6).

Certain viewers enjoyed the online format as expressed below:

> "The format allowed users less familiar with the topic to raise questions / points through an online form whereas they probably would not engage in an open meeting" (Respondent No. 1, Q6),

while others missed "the face-to-face networking". (Respondent No. 8, Q7).

The data also show that the majority of respondents learned about the webinar through the SEADDA network which is not surprising given the focus of the webinar on heritage science and archaeology research communities.

Future work. Survey data show that the presentations were of high quality and on a topic that research community finds useful (cf. Q6). In addition, viewers expressed their wish for follow-up discussions (cf. Q7) which proves an active interest for data (re-)use issues. For this reason, a follow-up workshop will be organised in 2021 in the framework of SSHOC (T6.5). SSHOC will again join forces with key actors in heritage science data to prepare a workshop where the leading experts from the field will be discussing more aspects of the management of heritage data.

Annex 4: Webinar report: GDPR and the DARIAH ELDAH Consent Form Wizard

Background

The webinar — Putting Data Protection into Practice: GDPR and the DARIAH ELDAH Consent Form Wizard — was held on 13 October 2020 and was organised in cooperation between the DARIAH ELDAH working group and T6.5 with the aim to discuss GDPR-compliant ways of processing personal data in research, education and cultural heritage, and to demonstrate a tool that helps create consent forms. The webinar was part of T6.5 activities that fall under the broad topic of "Data Protection and the GDPR".

The webinar focused on data protection which is an important aspect in ensuring legally compliant data that can be safely and productively re-used without posing any danger to the well-being of data subjects. By providing targeted theoretical and hands-on content for a broad audience of researchers in the Humanities and Social Sciences, the webinar successfully contributed to the SSHOC efforts of ensuring high data and knowledge re-use through Open Science and FAIR principles.

Webinar overview and format

Aim. The General Data Protection Regulation (GDPR) strongly impacts the work of SSH researchers and research infrastructures as well as other research/educational bodies. However, it is sometimes hard to know exactly what the areas impacted by the GDPR are, and how researchers and research infrastructures can work together in order to maximize the positive aspects of GDPR. The first aim of this webinar was to offer an overview of the principles of GDPR and identify the main situations in which management of personal data plays a role in the daily activities of SSH researchers. The second aim was to demonstrate the Consent Form Wizard developed by the DARIAH ELDAH Working Group – a tool that helps researchers to obtain GDPR compliant data.

Speakers. The webinar was delivered by four speakers:

- Koraljka Kuzman Šlogar (University of Zagreb/DARIAH-HR)
- Vanessa Hannesschläger (Austrian Academy of Sciences/CLARIAH-AT)
- Walter Scholger (University of Graz/CLARIAH-AT)
- Laure Barbot (DARIAH-EU)

Organisers. The webinar was organised in cooperation between T6.5 and the DARIAH-ELDAH group.

Participants. There were 61 viewers of the webinar.[8] The great majority of viewers came from European countries (EU 75%, non-EU 12%), while the others represented countries outside Europe (Canada, South Africa, the USA, etc.). The audience of the livestream included 6 stakeholder categories identified in D6.1. The great majority of the viewers (95%) represented three categories: “Researchers, Research Networks and Communities”, “Universities and Research Performing Institutions” and “Research Libraries and Archives”. These were followed by the category “Research and E-infrastructures” (5%). The representation of the categories “Civil Society and Citizen Scientists” and “Private Sector and Industry Players” comprised 2% of the entire audience.

Brief summary of the event structure. The webinar lasted for almost an hour and a half and was divided into three main parts. The first one, preceded by an introduction of the SSHOC project, included the three presentations by the speakers and covered (1) the presentation of the DARIAH-ELDAH Working group and the important ethical aspects when processing personal data, (2) a condensed overview of the GDPR provisions relevant for the research community, and (3) the development and functionalities of the Consent Form Wizard. The second part comprised three breakout rooms on two topics: (1) Gather Data/Consent for Communication and Hosting Academic Events, and (2) Gather data from and/or about living people for research purposes where the viewers were able to test the Consent Form Wizard for the specified purpose and discuss its functionalities. The third part was dedicated to a summary of important insights gathered during the breakout room sessions and to Q&A.

Presentations & Discussions: Key Points

First Session. Presentation of the SSHOC project and the DARIAH-ELDAH working group
Speakers. Laure Barbot
Main points. The speaker presented the goals and the expected impact of the SSHOC project and placed the webinar in relation to them.
Links to materials.

- Presentation slides
- Webinar recording (min: 0:00–8:30)

Second Session. General overview of the importance of data protection and ethics in research and a short introduction of ELDAH WG

[8] The number consists of viewers of the livestream (41) and of viewers of the webinar recording (20). It should also be noted that the number of the recording views was extracted on 20 November 2020 and is subject to change.

Speaker. Koraljka Kuzman Šlogar
Main points. Koraljka Kuzman Šlogar presented the DARIAH ELDAH Working Group, its mission and activities, and highlighted how the changing technical environment and digital information landscape was creating new ethical questions and new regulations in terms of data protection & ethics in research. The development of the Consent Form Wizard by the ELDAH WG can be seen as a tool to support researchers in this transition by providing standardized consent form templates for obtaining legal consent from human participants.
Links to materials.

- Presentation slides
- Webinar recording (min: 8:31–18:13)

Third Session. Introduction to GDPR principles and research exceptions
Speaker. Walter Scholger
Main points. Walter Scholger uncovered the important aspects of the *European legal framework for data privacy and processing personal data*, in short the GDPR, its principles and definition, explaining that this regulation should be seen as a set of guidelines and not a directive. He went into more detail regarding the applicability of the GDPR (e.g., processing personal data outside of the digital), its limitations and exceptions (e.g. that it does not apply to processing of data of deceased persons, or to anonymized data) and the different roles described in the regulation (i.e., Data Controller, Data Processor and Data Subject) which need to be identified in a GDPR-compliant personal data processing scenario. Finally, he also addressed the principles of data processing as defined in the GDPR, and touched upon the process of defining and obtaining consent for personal data processing which might be the one of the most frequent cases where SSH researchers are faced with GDPR requirements.
Links to materials.

- Presentation slides
- Webinar recording (min: 18:14–37:50)

Fourth Session. Introduction to prevalent personal data processing scenarios in research context and the Consent Form Wizard
Speaker. Vanessa Hannesschläger
Main points. To conclude the round of presentations, Vanessa Hannesschläger highlighted the prevalence of personal data processing scenarios in the SSH research context and explained how these scenarios were used to build the DARIAH-EU ELDAH Consent Form Wizard. Launched last September, the *Consent Form Wizard* supports humanities researchers within the EU in obtaining valid consent for data processing in the context of their specific professional activity. During the webinar, three scenarios were considered that would require a consent form in order to comply with the GDPR:

- communicating through electronic media (e.g. mailing lists);
- organising academic events;
- collecting data for research purposes.

Vanessa first explained some principal aspects of the creation of GDPR-compliant forms tailored to specific purposes and data categories via the *Consent Form Wizard,* which were identified through surveys at international Digital Humanities conferences and workshops among the DARIAH-EU community, she then provided more details about the purpose of the tool. Vanessa also underlined that no formal legal advice can be derived from the suggestion provided by the tool. Despite the fact that legal experts were involved in the creation of the *Consent Form Wizard*, the templates produced via the tool cannot replace the expertise of a dedicated data protection lawyer and cannot provide cover for formal legal liability. In fact, in particular cases, the tool will even explicitly recommend obtaining additional legal advice. Within the scope of the offered scenarios, however, the templates provided can be considered as *best practice* recommendations.

Links to materials.

- Presentation slides
- Webinar recording (min: 37:51–54:50)

Fifth Session. Discussion in small groups and Q&A time

Speaker. All speakers

Main points. The presentations were followed by a hands-on session in small groups where participants were invited to test the *Consent Form Wizard* and comment on the tool's functionalities or ask questions. While one group focused on gathering data/consent for communication and hosting an academic event, two others were discussing the GDPR challenges that can arise while gathering data from and/or about living people for research. One group discussed the limits of the very concept of consent: how can we, as researchers, organise ourselves when the data subjects are people with mental disability, or children, for example? In another group, participants' questions and presentation of individual cases were an opportunity to discuss how the forms produced by the *Consent Form Wizard* (i.e., downloadable in a raw format) can be adjusted to an institutional template or integrated in different communication channels.
The conclusion allowed Vanessa, Koraljka and Walter to highlight what the main goals of the DARIAH ELDAH Working Group are: contributing to the education of those who collect the consents to ensure that data privacy in research is better understood and that practices are GDPR-compliant.

Links to materials.

- Presentation slides
- Webinar recording (min: 1:01:31–1:17:43)

Outcomes & feedback

By following the webinar, participants were able to learn about the GDPR principles applied to SSH research and to identify the main situations in which they have to be careful about data protection. They learnt how to use the Consent Form Wizard and will be able to reuse it in order to comply with GDPR rules in their daily research activities.

Participant satisfaction. The majority of the respondents to the post-event survey rated the event as "Excellent" or "Very good" and indicated that it "Exceeded" or "Matched expectations" (cf. Q1&4). They learned about the event through various channels (personal recommendation, social media, newsletters, SSHOC website, meetings), but mostly through the mailing lists (cf. Q2). The respondents joined the webinar mainly because their work is highly impacted by the GDPR and wanted to get additional information about the implications, but also about the usefulness of the tool presented (cf. Q3). They stated that more detailed information about GDPR could be useful because many other events already covered the basic aspects of it (cf. Q7), but that overall, the event will positively impact their work since they became more aware of legal and ethical aspects regarding the research work, and also of useful tools and initiatives in this field (cf. Q5). This goes especially for the consent form wizard which is seen as "a needed and very useful tool" (Respondent No. 17, Q5).

Feedback regarding the Organisation. The respondents indicated that the webinar was excellently or well organised (cf. Q9). They especially praised the webinar's format, in particular the topic-related breakout groups, "the possibility to actually try out the consent wizard form" (Respondent No. 7, Q6) and the Q&A session at the end (cf. Q6). One participant suggested that it would be useful to have an additional session where other researchers would share their real-life experience with the tool (Respondent No. 3, Q7).

Future work. Survey answers show that the viewers would be interested in a "follow-up webinar if the consent form wizard is further developed" (Respondent No.11) and/or in "a series of lectures covering different disciplines/methodologies, e.g. language acquisition, ethnography, etc." (Respondent No. 16) since different aspects needs to be emphasised for different disciplines with regard to data protection (cf. Q8). This feedback is valuable and will be taken into account for a follow-up workshop in the context of T6.5 in 2021. We will strive to organise a face-to-face event if travel and health restrictions due to the COVID-19 pandemic will allow equal participation for individuals from different geographical areas. If this is not possible, we will first postpone the workshop, and then as a final resort, opt for an online virtual event.

ANNEX 5: WEBINAR REPORT: TOOLS AND RESOURCES FOR FAIR DATA

Background

The webinar — Tools and Resources for FAIR Data — was held on 18 May 2020 and was organised as a follow-up to the SSHOC T6.5 workshop Caring for Sharing – Data Management and FAIRness of Migration Data. The workshop was held on 9 March 2020 at the COST Action 16111 - ETHMIGSURVEYDATA Work group (WG) plenary meetings and 2nd Annual Policy Dialogue Conference in Brussels, Belgium. The webinar was organised by T6.5 of the SSHOC project on the topic of "Data Stewardship and RDM in theory and practice".

As a follow-up, the webinar's principal focus, just like that of the workshop, was on data management, a crucial aspect in the process of ensuring Open Science. As such, the webinar contributed to the main SSHOC goals by transferring skills and knowledge to the broad SSH community with which researchers can ensure maximised re-use of research data.

Webinar Overview & Format

Aim. Applying FAIR principles (Findable, Accessible, Interoperable and Reusable) to research data is crucial for assuring effective and efficient data re-use. However, ensuring FAIRness of research data can be a challenging and tedious task. This can discourage researchers from publishing their data under such constraints, representing a great loss for the scientific community since the already acquired data will not be available for replication and reuse purposes in the future. Therefore, this webinar focused on tools and resources that help researchers improve the quality of their data in terms of FAIR principles with less effort. Its aim was to convey the importance of respecting the FAIR principles throughout the research process and to equip the viewers with knowledge about available tools and resources that can be used to simplify the task of making data FAIR as well as showing them in detail how to use one of such resources.

Speakers. The webinar was delivered by Anca Vlad who is a data repository administrator at UK Data Service. She manages data deposits through UK Data Service self-deposit repository (ReShare), advises on data management, data deposit and ethical considerations when archiving data, and conducts data disclosure checks on deposited data.

Organisers. The webinar was organised in cooperation with partners in T6.5 of the SSHOC project.

Participation. There were 201 viewers of the webinar.[9] The majority of viewers came from European countries (EU 47%, non-EU 34%), but the webinar was also followed by viewers from other continents (the USA, Canada, Australia, Benin, Brazil, Israel, etc.; 19%). The audience included six stakeholder categories identified in D6.1. The great majority of viewers (72%) belonged to the categories "Research Libraries and Archives" and "Universities and Research Performing Institutions". "Researchers, Research Networks and Communities" category accounted for 17% of the entire audience. The remaining categories "Research and e-infrastructures", "Private Sector and Industry Players", and "Civil Society and Citizen Scientists" represented 12% of the viewers.

Brief summary of the event structure. The webinar lasted for an hour and was divided into three parts. After the introduction of the house rules and webinar etiquette, Kristina Pahor de Maiti, a Research Assistant at the University of Ljubljana and the moderator of the webinar, introduced the SSHOC project. This short presentation was followed by two main parts. Anca Vlad first presented the FAIR principles and an overview of existing tools and resources that can be used as an aid in ensuring FAIRness of the data. Then she proceeded with a demonstration of the QAMyData tool. The webinar concluded with a Q&A session.

Presentations & Discussions: Key Points

First Session. Introduction to SSHOC
Speakers. Kristina Pahor de Maiti
Main points. The webinar was given in the framework of the SSHOC project. The speaker presented the goals and the expected impacts of the SSHOC project and situated the webinar in relation to them.
Links to materials.

- Presentation slides
- Webinar recording (min: 00:00–5:43)

Second Session. Tools and Resources for FAIR Data
Speaker. Anca Vlad
Main points. The speaker first presented the four principles of FAIR data. (1) **Findable:** To ensure that data is findable, the researcher must check whether it has a globally unique and persistent digital identifier (such as a DOI in the citation). If data is published with an accredited archive or data repository (preferably CoreTrustSeal accredited), it will be assigned this unique identifier before publishing. (2)

[9] The number consists of viewers of the livestream (136) and of viewers of the webinar recording (65). It should also be noted, that the number of the recording views was extracted on 1 July 2020 and is subject to change.

Accessible: To be accessible, (meta)data should be retrievable using an open, universally applicable and standardized protocol. Data does not need to be open to ensure its accessibility; what is essential is that the process of obtaining data is made clear and described in plain language for each dataset. (3) **Interoperable**: To be interoperable, (meta)data needs to be in specific formats, and use community agreed standards and vocabularies/ontologies, such as the DDI Schema. Specific formats should be respected because digital data is very software dependent and as such prone to corruption/loss due to the obsolescence of software/hardware. The EMM Survey Registry, a tool developed in SSHOC, is a good example as it uses a multitude of machine readable metadata fields to describe its datasets (such as scope, region(s), start date, end date of survey, target population, etc.). (4) **Reusable**: And finally, to be reusable, the quality of supporting documentation, metadata and of course data itself, as well as the applicable license, are important. The license should allow the data to be available to the widest possible audience with the widest possible range of uses.

Next, the speaker presented some tools and resources that can simplify the process of ensuring FAIRness of data: SSH Open Marketplace, CESSDA Data Management Guide, DMP Online, Go FAIR starter kit, a handbook on Managing and Sharing Research data, A guide to Good Practice.

Finally, the speaker showcased the QAMyData which is a free open source tool that provides a ´health check´ for numeric data. The tool was designed to be easy to use by employing automated methods to detect and report some of the most common problems in survey data: missingness, duplication, outliers and direct identifiers (information that can point to one person in particular, such as name, address etc.). Outliers and direct identifiers are of particular concern when sharing data as researchers need to uphold confidentiality agreements. The tool offers a number of configurable tests, categorized by type: file, metadata, data integrity, and identifiers. It can run popular file formats, including SPSS, Stata, SAS and CSV.

Links to materials.

- Presentation slides
- Webinar recording (min: 5:43–39:15)

Third Session. Q&A session
Speaker. Anca Vlad
Main points. The viewers were interested in the development process of the QAMyData tool, especially to what extent the researchers as the primary users were involved in the creation process, and how intensely the tool is used by the UK Data Service data repository administrators. The speaker briefly explained the development stages of the tool as well as her experience of using it in her work. The viewers also sought further information about pros and cons of different repositories and were interested in ways of finding good repositories for specific data, such as oral history interviews. The main advice was to focus on the list of CoreTrustSeal accredited repositories.

Links to materials.

- Webinar recording (min: 39:16–47:37)

Outcomes & Feedback

FAIR principles represent one of the pillars of modern data management. However, they are often not well understood by the researchers or disregarded because it is often not simple to ensure their implementation. Viewers of this webinar first received a concise overview of what the principles are all about and what tools and resources they can use in the data management process. Apart from getting a quick overview of various "FAIRness-ensuring" tools, they were able to follow a detailed demonstration of one such tool, namely the QAMyData. Since it is a free tool, researchers can directly use the knowledge acquired during the webinar in their research projects.

Viewer's satisfaction. Viewers rated the webinar either as Very good or Excellent (cf. Q1). The respondents to the post-event survey stated that the webinar matched their expectations (cf. Q4). Most of them said that they will use the QAMyData tool in their work and will also recommend the presented tools and resources to their colleagues (cf. Q5). The viewers liked the concise presentation of FAIR principles, e.g. "Great intro to FAIR" (Respondent No. 6, Q6), as well as the useful QAMyData demo (cf. Q6). However, some noted that additional details would be welcome (cf. Q7).

Feedback regarding the Organisation. The viewers generally felt that the webinar was Excellently or Well organised (cf. Q9). The comments about possible improvements included a suggestion for incorporating more interactive activities into the webinar format, and to provide clearer promotional material since it was not absolutely clear that there will be a specific tool demonstration (cf. Q8 & 9). On the other side, the SSHOC promotion activities seemed to be efficient, since the majority learned about the webinar through the SSHOC channels.

Future work. Since this was a follow-up to a previous workshop, the T6.5 team does not plan to organise directly related events in the future. However, provided there will be requests for related training activities, we will try to support their organisation.

ANNEX 6: WEBINAR REPORT: INTRODUCING THE NEWLY LAUNCHED ETHNIC AND MIGRANT MINORITIES (EMM) SURVEY REGISTRY

Background

The webinar — Introducing the newly launched Ethnic and Migrant Minorities (EMM) Survey Registry — was held on 26 October 2020 and was organised in cooperation between T9.2, T6.5 and the COST Action ETHMIGSURVEYDATA as an additional SSHOC T6.5 activity with the aim to promote the outcomes of WP9 in the form of the EMM Survey Registry. The webinar is part of the T6.5 activities that fall under the broad topic of "Data Stewardship and RDM in theory and practice". The webinar is associated with the Deliverable 9.4 Database with the metadata of surveys to EMMs across Europe.

The webinar's principal focus was on data management and data sharing, a crucial aspect in the process of ensuring Open Science. The webinar served as a platform to present a new tool developed inside SSHOC that promotes sharing and re-use of data within the ethnic and migration studies domain in accordance to Open Science and FAIR principles thus supporting one of the primary goals of SSHOC.

Webinar Overview & Format

Aim. In spring 2020, SSHOC T9.2 and the COST Action ETHMIGSURVEYDATA launched the beta version of the EMM Survey Registry which is a free online tool that will, once finalized, display survey-level metadata for over 800 surveys on ethnic and migrant minorities' (EMMs') integration from over 30 European countries. The aim of the webinar was twofold: (1) to discuss the developments of the EMM Survey Registry with a focus on its FAIR-compliant[10] and sustainable design and the diverse pool of possible target users; and (2) to offer a tutorial of the currently available beta version.[11]

Speakers. The webinar was delivered by two presenters and a host:

- Laura Morales, Professor in Political Science/Comparative Politics at Sciences Po (Paris, France), affiliated with CEE and LIEPP, and the Chair of the COST Action "International Ethnic and Immigrant Minorities' Survey Data Network". She specialises in the political dynamics and

[10] FAIR principles: findable, Accessible, Interoperable, Reusable

[11] The EMM Survey Registry is already fully functional, but only a certain part of it is currently available to general public, i.e. the front-end with approximately half of the database entries. This is what we refer to as the beta version.

consequences of immigration, in the civic and political inclusion of migrant-origin minorities, and in survey research on migrant-origin populations.

- Ami Saji, a junior researcher based at Sciences Po (Paris, France). She supports the SSHOC project in making quantitative data on ethnic and migrant minorities FAIR. Prior to this role, she served as the Network Coordinator for ETHMIGSURVEYDATA and worked in the NGO sector, specialising in refugee resettlement, migrant integration, and workforce development.
- Darja Fišer, Vice-Executive Director responsible for User Involvement at CLARIN ERIC and is leading task 6.5 in the SSHOC project. She is also Associate Professor at the Faculty of Arts, University of Ljubljana, and Senior Research Fellow at the Department of Knowledge Technologies, Jožef Stefan Institute.

Organisers. The webinar was organised in cooperation between T9.2 and T6.5 of the SSHOC project, as well as the COST Action ETHMIGSURVEYDATA.

Participation. There were 168 viewers of the webinar.[12] The majority of viewers came from European countries (EU 65%, non-EU 15%), but the webinar was also followed by viewers from countries outside Europe (Algeria, Brazil, Canada, China, Japan, the USA, etc.; 20%). The audience included six stakeholder categories identified in D6.1. The great majority of viewers (89%) belonged to the categories "Universities and Research Performing Institutions" and "Researchers, Research Networks and Communities". Other categories ("Research Libraries and Archives", "Research and e-infrastructures", "Private Sector and Industry Players", "Civil Society and Citizen Scientists") were less well represented with a 3% share on average.

Brief summary of the event structure. The webinar lasted for an hour and was divided into five parts. After a short introduction about the webinar etiquette and presentation of the SSHOC project delivered by the webinar's host, Darja Fišer, the first part of the webinar was dedicated to a general introduction about the goals of the webinar and the communities involved and was delivered by Ami Saji. Afterwards, Laura Morales presented the rationale behind the tool, its FAIR-compliant design and the EMM Survey Registry's front-end. This was followed by a presentation of the tool's back-end by Ami Saji which is also planned to be accessible for the public after the release of the tool in its finalized version. The webinar concluded with a Q&A session.

Presentations & Discussions: Key Points

First Session. Introduction to SSHOC
Speakers. Darja Fišer
Main points. The webinar was given in the framework of the SSHOC project. The speaker presented the goals and the expected impacts of the SSHOC project and situated the webinar in relation to them.
Links to materials.

- Presentation slides
- Webinar recording (min: 00:00–4:15)

[12] The number consists of viewers of the livestream (107) and of viewers of the webinar recording (61). It should also be noted, that the number of the recording views was extracted on 19 November 2020 and is subject to change.

Second Session. The ethnic and migration studies data community
Speaker. Ami Saji
Main points. The ethnic and migration studies data community brings together various stakeholders of quantitative survey research undertaken with EMM populations. For the purposes of executing its scientific work, the data community is organised into three distinct groups: the T9.2 team of SSHOC, COST Action ETHMIGSURVEYDATA and FAIRETHMIGQUANT (an Open Science project funded by the French *Agence National de la Recherche*). As part of SSHOC, the T9.2 team and COST Action ETHMIGSUVREYDATA have jointly produced and delivered the EMM Survey Registry (D9.4 of T9.2).

Links to materials.

- Presentation slides
- Webinar recording (min: 4:16–39:15)

Third Session. The EMM Survey Registry: A FAIR tool for EMM survey data & The front-end of the EMM Survey Registry: The key functionalities
Speaker. Laura Morales
Main points. At its core, the EMM Survey Registry is intended to be a user-friendly and user-centric tool. It is also envisioned to be a single access point to metadata about quantitative surveys undertaken (primarily) in Europe since January 2000 with EMM respondents. Given these objectives, the FAIR principles naturally became the framework for the EMM Survey Registry design. The EMM Survey Registry promotes the FAIR principles in the following ways:

- **Findable:** Existing EMM surveys conducted in Europe are easier to locate, because the EMM Survey Registry acts as a living 'census'. Furthermore, the metadata are easy to navigate and understand due to the systematic documentation of the metadata and the user-friendly interface of the tool.
- **Accessible:** The EMM Survey Registry makes all of its metadata publicly available online. For each survey, the metadata include detailed information about how to access the data set(s), technical documentation, questionnaire(s) and other relevant publications whenever possible.
- **Interoperable:** The metadata are set up in both human- and machine-readable formats.
- **Reusable:** The metadata are detailed, informative, organised and structured. In addition, the metadata itself can be reused for research. In fact, reuse of the survey is promoted and facilitated, since the metadata include information about how to access the relevant original sources (e.g. data set(s), technical documentation) whenever possible.

The beta version of the EMM Survey Registry, which was officially launched in spring 2020, has a fully functional front-end and currently displays metadata for almost 480 surveys from 14 different countries. Users can freely explore the front-end of the EMM Survey Registry and learn about the already included surveys through the various front-end functionalities among which are the following:

- **Keyword search**: Users are able to look for specific surveys using a keyword search based on Boolean logic.
- **Simple filtering**: Users are able to refine their search results using 15 different filters (i.e. metadata variables from the EMM Survey Registry's metadata schema).

- **Advanced filtering**: Users can refine their search with up to 28 filters.
- **Summary of applied filters**: After a user applies all of their desired filters, a summary of their selection is displayed above the list of surveys.
- **Sorting**: Users are able to sort the list of surveys based on country, scope (national or subnational/local level), region, start/end dates, EMM target population and sample size (achieved).
- **Abbreviated view** of the metadata compiled for a survey: When viewing the list of surveys, users are presented with an abbreviated view of a survey's metadata that provides users with the critical information about a survey, so they can quickly decide whether or not the survey is of interest to them.
- **Full view** of the metadata compiled for a survey: Each survey captured on the EMM Survey Registry is set up with its own page that provides the full view of the compiled metadata. As the full view contains an immense amount of information, it is structured into sections and offers a navigation menu.
- **XML files** of the metadata: For each survey, the compiled metadata can be downloaded as an XML file by clicking on the "XML" icon. The XML file that is generated provides the actual metadata to conform with DDI Codebook.

Links to materials.

- Presentation slides
- Webinar recording (min: 11:18–34:31)

Fourth Session. The back-end of the EMM Survey Registry: the key to sustainability
Speaker. Ami Saji
Main points. The back-end of the EMM Survey Registry, which is where all the metadata is managed, is also fully functional. Currently, its access is limited to the administrators of the EMM Survey Registry, because the protocol of granting access has not yet been established; nonetheless, its functionalities were showcased in the webinar. In the coming months, the back-end will be opened up to external users, so that they can contribute their own metadata to the EMM Survey Registry.
To ensure the sustainability of the EMM Survey Registry, Ami Saji also called upon data producers and data users to support the development of the tool through providing (meta)data for the registry and the feedback regarding functioning of the tool.

Links to materials.

- Presentation slides
- Webinar recording (min: 34:32–49:21)

Fifth Session. Q&A session
Speaker.
Main points: Audience members asked questions to better understand the surveys being in the EMM Survey Registry and to learn more about the process for external users (i.e., individuals outside of the ethnic and migration studies data community) to contribute their own metadata to the registry. Ami Saji and Laura Morales provided responses to the questions and shared resources (e.g., documents created

by the ethnic and migration studies data community) that provide detailed information about how metadata was collected and compiled).

Links to materials.

- Webinar recording (min: 49:22–59:28)

Outcomes & Feedback

Quantitative surveys undertaken with EMM (ethnic and migrant minority) populations can provide valuable insights about the integration experience and process that EMMs go through. Yet, EMM survey data are often underutilised for research and in policymaking, because this data are often difficult to find, access and re-use. To help untap this unmet potential of EMM survey data being produced in Europe and beyond, the ethnic and migration studies data community (T9.2), in partnership with the COST Action ETHMIGSURVEYDATA, developed the EMM Survey Registry. Most importantly, the tool is FAIR-compliant, has a sustainable management design and is free to use. Ensuring FAIRness and sustainability of the tool were at the core of the creation process, since these are the crucial aspects of modern data management. With this taken into consideration and in addition, ensuring free and open access, we believe that the tool has all that is needed for producing far-reaching impact on the EMM research community.

Viewer's satisfaction. All respondents but one rated the webinar either as Very good or Excellent (cf. Q1). One stated that he/she expected more, however all others were more satisfied since half of them indicated that the webinar matched their expectations and the other half that it exceeded their expectations (cf. Q4). The respondents mainly joined the webinar because they have been actively involved with the EMM community either by doing research or by working directly with those individuals. Therefore, they wanted to know how the EMM Survey Registry could help advance their research, while others wanted to participate as data producers (cf. Q3). The respondents indicated that the tool will positively impact their work, but that they would wish to see more countries included (cf. Q5). In general, they praised the delivery (including the demonstrations of the tool's functionalities) which was "clear and efficient" (Respondent No. 6, Q6).

Feedback regarding the Organisation. The viewers generally felt that the webinar was Excellently or Well organised, except for one who indicated that the webinar was Fairly organised (cf. Q9). The only suggestion was that it would be good if the viewers received the post-event survey form during the event or immediately after it, instead of receiving it together with the webinar recording a day or two after the live streaming (Respondent No.2, Q7). This suggestion will be taken into consideration for our next events. Regarding the pre-event promotion, the answers show that the information about the webinar reached the viewers through different channels (personal invitation, newsletters, SSHOC/EOSC website, newsletters, CLARIN conference) which is an encouraging finding showing that SSHOC events are disseminated through non-SSHOC channels as well (cf. Q2).

Future work. Since this webinar was dedicated to a presentation of a newly launched tool in the framework of SSHOC and since this was an additional activity on the part of T6.5, the T6.5 team currently does not plan to organise any follow-up events on the topic. However, provided there will be requests for related training activities, we will try to support their Organisation.

ANNEX 7: WEBINAR REPORT: FAIR SSH DATA CITATION: PRACTICAL GUIDE

Background

This online webinar — FAIR SSH Data citation: practical guide — was held on 3rd December 2021 and was organised by SSHOC T6.5 in cooperation with SSHOC WP3 as a follow-up to a workshop on a similar topic delivered on 15th June. By focusing on practical aspects of citation practice in Social Sciences and Humanities, the webinar offered an opportunity to showcase the work done in SSHOC T3.4 and more broadly, to underline the importance of citation protocols for Open Science while simultaneously outlining best tools and practices to cite SSH data.

Webinar Overview & Format

Aim. Open Science relies heavily on FAIR principles and consequently on FAIR data, tools and services. Proper data citation is a crucial element needed to ensure reproducibility and transparency of research, for instance by giving credit to the creator of data. It also heavily impacts the visibility of the research and their authors as well as encourages data reuse. However, approaches to data citation in SSH are very diverse which limits the positive effects of cited items to a great extent. The aim of the webinar was therefore to raise awareness of the importance of data citation and to show how to build robust data citation outputs. The webinar focused on the value/necessity of data citation, use of the *FAIR SSH Citation* prototype and other existing tools, practical advice on citing SSH data and ideas for data-based scholarship. The webinar was especially tailored to the needs of SSH researchers and repository managers.

Speakers. The webinar was delivered by 3 speakers and a host:

- Nicolas Larrousse (Huma-Num/CNRS, SSHOC T3.4 leader) is the Vice Director of TGIR Huma-Num and head of its Coordination of National and International User Communities group. He is an IT-specialist particularly interested in interoperability issues and long term preservation. He is involved in several European infrastructures and projects.
- Edward J. Gray (Huma-Num/CNRS, SSHOC WP3 member) is the Research Infrastructure Coordinator at the TGIR Huma-Num (CNRS) and the Officer for National Coordination at DARIAH ERIC, the European Research Infrastructure for the Digital Arts and Humanities. He earned his doctorate in history from Purdue University, after a dissertation on early modern French familial politics. While a doctorant invité at the École nationale des chartes in Paris, he earned a master's

degree in Technologies numériques appliquées à l'histoire (TNAH), where he is also chargé de cours.

- Cesare Concordia (ISTI-CNR, SSHOC WP3 member) is a full-time researcher at ISTI-CNR, where he works on topics related to distributed Information Systems and Digital Libraries. He is a member of the AI for Media and Humanities (AIMH) laboratory of ISTI. His research interests include also: semi-structured data, Service Oriented Architectures and Semantic Web frameworks.
- Daan Broeder (CLARIN, SSHOC WP3 leader) as the host and the SSHOC representative who moderated the webinar. He has a long career working on research infrastructure, working in different capacities at different CLARIN centres and was managing tasks in several European and national projects.

Organisers. The webinar was organised as a stand-alone online event in cooperation with partners from T6.5 and WP3 of the SSHOC project.

Participation. There were 41 viewers of the webinar.[13] The great majority of the viewers (93%) came from European countries, both EU (85%) and non-EU (8%), while the remaining participants came from Brazil, Tunisia and Turkey. The audience of the livestream covered six stakeholder categories identified in D6.1. Almost half of the audience was represented by "Researchers, Research Networks and Communities" (17%) and "Universities and Research Performing Institutions" (27%). The other two groups of considerable size included participants from "Research and e-infrastructures" (24%) and "Research Libraries and Archives" (24%). The remaining participants identified as belonging to "Civil Society and Citizen Scientists" (5%) and "Research Funding Organisations" (3%). Most of the participants are working across fields, so 46% of the participants stated they work with topics relating to SSH in general, Digital humanities, Open Science and Information Science. Approximately a fifth of the participants (22%) work in Social Sciences, another fifth (17%) in Humanities (History, Linguistics, Ethnology), while the remaining participants (15%) work in Natural Sciences.

Brief summary of the event structure. The webinar was a one-hour online event which consisted of a short introduction of the SSHOC project, the main part of three topic-related presentations and an engaging Q&A session at the end.

[13] The number represents only the viewers of the livestream (excluding the organisers and presenters), while the registrations were much higher, amounting to 95 individuals. It should be noted that the overall reach is higher than 41 viewers since this number does not include the views of the webinar recording which is available on the SSHOC YouTube channel and promoted by SSHOC and its partners.

Presentations & Discussions: Key Points

First Session. Introduction to SSHOC

Speakers. Daan Broeder

Main points. The webinar was given in the framework of the SSHOC project. The speaker presented the goals, the expected impacts as well as some results of the SSHOC project and situated the webinar in relation to them.

Links to materials.

- Presentation slides (pp.: 1–8)
- Webinar recording (min: 00:00–04:54)

Second Session. Data Citation in SHS

Speaker. Nicolas Larrousse

Main points. Nicolas Larrousse (Huma-num/CNRS) first presented the current state of data citation practices in SSH. Given the many different disciplines within SSH, the current data citation practices are very diverse. As Nicolas Larrousse pointed out, SSH lack a common approach to data citation even though that communities of practices exist. Nevertheless, there are a number of recommendations already available (DASISH Project, ICPSR, CESSDA, SHARE, W3C's Web Annotation Data Model, RDA Data Citation of Evolving Data, etc.) that allow researchers to cite the data in such a way that it can be shared and reused across different hardware and software platforms. But those recommendations are not fully adapted for SSH, which is why specific recommendations were needed and developed in SSHOC.

Links to materials.

- Presentation slides (pp.: 9–13)
- Webinar recording (min: 04:55–12:36)

Third Session. Data Citation Recommendations and Survey of Repositories

Speaker. Edward J. Gray

Main points. In the second presentation, Edward Gray (Huma-Num/CNRS) presented data citation recommendations developed in the SSHOC project. He noted that a few years ago it might have been enough to cite one source as a simple string of bibliographic information that include author and title in Chicago style, APA or MLA, but today this is no longer sufficient. In order to make data citation machine-actionable, a set of recommendations based on FORCE11 has been developed in the framework of SSHOC. The recommendations have been further developed and adapted to the needs specific to SSH, mainly based on peer review and the feedback from the Round Table of Experts.
Recommendations for FAIR Data Citation in the Social Sciences and Humanities. They are published in a document that describes each recommendation separately. More specifically, it lists eight recommendations – Importance, Credit and Attribution, Evidence, Unique Identification, Access,

Persistence, Specificity and Verifiability, and Interoperability and Flexibility, and each of them is described from three different perspectives: first, the general, societal and technical challenge with specific data citation principle is explained, then practical recommendations to address this problem are given, and finally, the document specifies the expected outcome for each of the principles.

These recommendations are aimed at all stakeholders in SSH, from researchers and engineers to funders and research infrastructures. However, the *Recommendations for FAIR Data Citation in the Social Sciences and Humanities* also list a number of use cases that can help any reader identify whether the recommendations might be beneficialin their case. The use cases that are further elaborated in the document, include:

- I am a Researcher and/or an Engineer working for a project
- I am a Research software engineer working for a project and/or research infrastructure.
- I am a Manager of a data repository.
- I am a Data Librarian, a Data Steward or an Open Science Officer.
- I am a Research Funder.
- I am a Researcher who is conceiving a research project.
- I am a member of the public who wishes to re-use data.

Links to materials.

- Presentation slides (pp.: 14–23)
- Webinar recording (min: 12:37–28:45)

Fourth Session. Data Citation Prototype
Speaker. Cesare Concordia
Main points. The *FAIR SSH Citation prototype* presented by Cesare Concordia (ISTI-CNR) is a software tool designed for harvesting metadata and presenting the information collected from various sources in a standardized and unified way.
The main functionalities of the prototype are:

- exploring metadata on datasets,
    - retrieving metadata from landing pages via PIDs (DOIs, handles and others) or URLs,
    - retrieving information from other sources like APIs knowledge graphs etc.,
- providing facilities for curation and semantic annotation of citations,
- visualising and exploiting citation metadata, and
- disseminating metadata.

The prototype can be tested by anyone interested in the tool through the following services:

- Citation Service API
- The Citation Metadata Viewer

Additionally, citations from the abstracts of all (ADHO) DH conferences 2015–2020 and from DHQ journal articles can be checked on this link.

Links to materials.

- Presentation slides (pp.: 24–40)
- Webinar recording (min: 28:46–44:12)

Fifth Session. Conclusion and discussion

Speaker. All speakers

Main points.

Nicolas Larrousse concluded the webinar by emphasising that good data citation practice requires a complete ecosystem that not only includes cited data according to broadly accepted norms and standards, but also trusted repositories and dissemination tools. He also shortly addressed the future developments in the field of data citation, e.g. using citation-based information to associate tools and data. This was followed by an elaborate discussion in which the speakers addressed questions from the audience. Questions, for example, included issues related to data journals in general and their lack in SSH, as well as knowledge graphs, quality metadata and interoperability that enrich the research process. Additional recommendations were also discussed in the Q & A session, including questions such as who to cite as authors of datasets. The discussion ended with the speakers pointing out areas for future work, such as widening the scope of existing recommendations, and commenting on possible improvements of specific repositories.

Links to materials.

- Presentation slides (pp.: 41–43)
- Webinar recording (min: 44:13–1:05:00)

Outcomes & Feedback

Data citation in Social Sciences and Humanities can be a rather complicated task. In particular, when it comes to one crucial step, this is making it machine actionable. The fact that data are not available at the same place and described in the same manner with the same attributes, significantly lowers the scope of benefits that come from well recorded items of available knowledge. In order to address the incompleteness of existing citation methods and complexity of the technical landscape, SSHOC T3.4 performed an inventory of citation practices. The project further developed recommendations and software to make SSH datasets citable; visualise and exploit citations; and provide facilities for curation and semantic annotation of these resources. By participating in this workshop, the attendees learned about most recent developments in the field of data citation in SSH, acquired some practical advice on how to best prepare their citation entries, e.g. starting with planning the data lifecycle of their research projects carefully and preparing a precise description of their data, and heard about future plans and ideas for data-based scholarship.

Viewers' satisfaction.[14] The respondents indicated that the webinar matched their expectations and that it was very well organised. The overall feedback was very positive and the participants underlined that they appreciated up-to-date information as well as high information density of the webinar. In addition, the respondents also commented that the webinar was especially valuable since many supporting materials and references have been provided throughout the event. One respondent suggested that it would be valuable to have recommendations especially developed for particular communities/types of users, and that the webinar could be followed by a demonstration of the prototype. Based on the feedback and active engagement of participants in the final discussion, it can be claimed that the webinar addressed a timely topic with high quality content, and considerably contributed to the promotion of SSHOC results.

[14] The post-event survey was filled out by 3 participants. The report includes also feedback given in the chat box during the webinar.

ANNEX 8: WEBINAR REPORT: QUANLIFY WITH EASE: COMBINING QUANTITATIVE AND QUALITATIVE CORPUS ANALYSIS

Background

The webinar — Quanlify with ease: Combining quantitative and qualitative corpus analysis — was held on 16 April 2020 and was organised as a follow-up to the SSHOC masterclass Using Corpora for Implementing Validation. Workflows that combine quantity and quality which was held on 30 September 2019 at the CLARIN Annual Conference 2019 in Leipzig. It was organised by T6.5 of the SSHOC project on topic of Data Science for the Social Science and Humanities.

As a follow-up to the masterclass, the webinar fell under the SSHOC objective focusing on developing directly applicable tools for researchers from the SSH domain and supporting them in their use of these tools.

Webinar Overview & Format

Aim. Researchers that rely on large-scale corpora are confronted with particular challenges to validate their findings. The validation process is far from a trivial task since there are numerous requirements and technical limits of the existing software solutions. This webinar focused on the practical implementation of *quanlification* (i.e. the combination of quantitative and qualitative approaches to corpus analysis) while at the same time setting the theoretical background for this methodological approach. In the webinar, Andreas Blätte, head of the PolMine project and developer of the polmineR package from the University of Duisburg-Essen, shared tools and experiences of researchers implementing *quanlification* in practice. He also presented the polmineR analysis environment.

Speaker. The masterclass was taught by Andreas Blätte, head of the PolMine project and developer of the polmineR package from the University of Duisburg-Essen. Andreas Blätte is also a professor of Public Policy and Regional Politics at the University of Duisburg-Essen. He is a political scientist by training, but has developed a strong interest in data and analytical tools for using corpora in the Social Sciences and Humanities. He has published three R packages (polmineR, RcppCWB and cwbtools) that are available via CRAN. His prime substantial research interest is the discursive dimension of migration and integration policy.

Organisers. The webinar was organised by the T6.5 of the SSHOC project.

Participation. There were 108 viewers of the webinar.[15] The majority of viewers came from European countries (EU 84%, non-EU 9 %), but the webinar was also followed by some viewers from countries outside Europe (e.g., the USA (5%), South Africa (1%), Indonesia and other Asian countries (2%)). The audience included six stakeholder categories identified in D6.1.-The great majority (67%) belonged to the categories of "Researchers, Research Networks and Communities" and "Universities and Research Performing Institutions". The second cluster of viewers represented "Research Libraries and Archives" (14%); "Research and e-infrastructures" (10%), and "Private Sector and Industry Players" (8%); making up 33% of the entire audience. The remaining audience belonged to the "Civil Society and Citizen Scientists" category, and were represented by only a few viewers (approx. 1%).

Brief summary of the event structure. The webinar lasted for an hour and was divided into 3 parts. After the introduction of the webinar etiquette delivered by Christoph Leonhardt (a research associate at the Institute of Political Science of the University of Duisburg-Essen), Andreas Blätte continued with the main part. He introduced the theoretical background of combining qualitative and quantitative techniques and described the technical design behind the tools that enable applying these techniques to research results while showcasing examples from the polmineR analysis environment. The webinar concluded with an informative Q&A session.

Presentations & Discussions: Key Points

Main Session. Combining quantitative and qualitative corpus analysis with polmineR

Speaker. Andreas Blätte

Main points. The central assumption is that the validity of research results obtained from large-scale corpora depends on researcher's ability to combine the quantitative and qualitative analysis of textual data. Drawing from the ideas of distant reading and processing texts as data, Andreas Blätte proposed that the findings of corpus analysis are based on both, the text itself, as well as its quantitative numerical representation. While this supposed juxtaposition suggests a methodological divide of qualitative and quantitative approaches which might necessitate separate means of validation, Blätte argued that despite this divide, there is a common challenge in research practice: while every qualitative finding requires quantitative support in order not to be deemed simply anecdotal, every quantitative analysis

[15] The number consists of viewers of the livestream (75) and of viewers of the webinar recording (33). It should also be noted, that the number of the recording views was extracted on 30th June 2020 and is subject to change.

needs to rely on qualitative confirmation to avoid misleading interpretation of patterns. Therefore, Blätte concluded that the necessity to combine qualitative and quantitative approaches to text is conceptually undisputed. Both perspectives should always be applied in tandem. Distant reading and close reading should be blended in order to validate the findings of the research. It is, however, not clear how a research process allowing an easy interplay between quantitative and qualitative methods should be implemented, since existing software solutions do not yet support this requirement. Although there are various not interlinked tools, setting up a truly *quanlitative* project remains expensive and difficult to implement.

In the webinar, Blätte elaborated on design decisions that are a prerequisite to create tools which facilitate *quanlitative* research. He showcased examples from one such tool, namely the polmineR analysis environment which is written in the statistical programming language R. The tool enables validation of the results by displaying the full context which allows contextualizing the findings. This can be done through different modules that were presented in the webinar and is applicable to the results which are either based on numerical approaches (e.g., co-occurrences and topic models) or represent only a part of the original text (e.g., concordances and subcorpora). The viewers of the webinar also got to know different possibilities to visualize the results of statistical analyses to gather the semantic sense of the returned numerical values. Other examples for the *quanlitative* approach discussed in the webinar included the annotation of sentiment weights in a keyword-in-context representation, and the visualization of co-occurrences in a three-dimensional co-occurrences graph enriched with the underlying concordances. The package also includes features to annotate any kinds of tables which can be used to both facilitate intersubjectivity and to generate training data for machine learning approaches. In addition, the presenter showcased some early implementations of the approach which enables applying *quanlitative* methods to textual data at every point of the research project.

Links to materials.

- Presentation slides
- Webinar recording (min. 00:01–42:52)

Fourth Session. Q&A session

Speaker. Andreas Blätte

Main points. The questions were varied and encompassed the technical aspects of the polmineR environment as well as its pedagogical potential and methodological questions related to analysing political debates. Andreas Blätte explained that there is currently a limitation in corpus size that can be used inside the polmineR, but before the tools become more powerful, an easy solution can also be to simply split up the corpus. He also mentioned that the polminR code is constantly improving in order to become more efficient and invited everybody to join the community of developers on polmineR Github channel. He further noted the fruitful collaboration with CLARIN ERIC through ParlaClarin workshop. In addition, the speaker informed the viewers about access possibilities to different corpora that are made available through polmineR environment, specifics of data formats and the available packages for linguistic annotation of corpora from scratch (e.g., cwbtools). There was a discussion of the pedagogical materials for social scientists and possible methodological approaches to analysing governmental measures taken during the Covid-19 pandemic in different countries.

Links to materials.

- Webinar recording (min: 43:00–58:37)

Outcomes & Feedback

Corpus analysis techniques are well established in the Humanities and are becoming more and more popular in the Social Sciences, but a full and simple integration of quantitative and qualitative analysis approaches into tools remains an unfulfilled promise. However, the viewers of the webinar were able to learn about tools which make the task of combining these analysis techniques much more manageable and, therefore, directly applicable in research projects. On top of that, the webinar primarily targeted Social Scientists and thus promoted the corpus analysis techniques among research communities where such approaches are not yet heavily used.

Viewers' satisfaction.[16] Respondents to the post-event survey felt that the event was Excellent or Very good and that it exceeded their expectations (cf. Q1 & Q4). The viewers liked the "precision and informative nature" (Respondent No. 2, Q6) of the webinar and think that the content will directly impact their work (cf. Q5 & Q6).

Feedback regarding the Organisation. According to the responses, the webinar was excellently organised and provided a well-timed, concise and enjoyable learning experience (cf. Q6 &Q9).

Future work. The feedback shows that the webinar is addressing the needs of researchers who would appreciate further support in using the tools:

> In a longer event, a workshop involving participants actually trying procedures would be instructive. (Respondent No. 2, Q7)

Given that this was a follow-up to the masterclass on similar topic, the T6.5 team does not plan to organise related events in the future. However, provided there will be requests for related training activities, we will try to support their organisation. In addition, we will make sure to boost the visibility of the webinar recording through SSHOC and partners' channels.

[16] The results are based on the post-event survey. Despite multiple reminders, we were only able to obtain three replies.

ANNEX 9: WEBINAR REPORT: SSHOC'ING DRAMA IN THE CLOUD – ENCODING THEATRICAL TEXT COLLECTIONS AND THE ADDED VALUE OF SSHOC & CLARIN SERVICES

Background

The webinar — SSHOC'ing Drama in the Cloud – encoding theatrical text collections and the added value of SSHOC & CLARIN services — was held on 23th June 2021 as part of the LIBER annual conference and was organised by SSHOC T6.5 in cooperation with SSHOC WP3. The webinar gave a general overview how SSH researchers can benefit from the resources and services which are curated in SSH research infrastructures. The content of the workshop was especially designed for librarians, but the workshop was open to other interested individuals as well. In this way, the participants of the workshop not only extended their general knowledge about the offer of research infrastructures, but also learned how to guide other potential users towards a successful and efficient use of various tools and resources that are available in research infrastructures and can be discovered through SSH Open Marketplace. The webinar, therefore, contributed to the objectives of SSHOC, specifically the promotion of FAIR and Open Science principles.

Webinar Overview & Format

Aim. The objective of this webinar was to equip the participating librarians with some general knowledge on how researchers in the field of Social Sciences and Humanities (SSH) can benefit from the resources and services offered by SSH research infrastructures for producing and exploiting highly encoded historical textual data. The webinar wished to train the participants, so that they are able to successfully guide and advise SSH researchers (with a particular focus on literature studies) in their choice amongst existing resources and tools, based on their research question.

This objective was achieved by:

- Familiarising the participants with the Text Encoding Initiative (TEI) format that is widely adopted in SSH for the XML-based mark-up of textual documents and demonstrating the potential benefits;

- Teaching the participants how to explore and visualise TEI collections with the help of tools and services offered by CLARIN and SSHOC;
- Showing the participants how to optimize research workflows with the help of SSH Open Marketplace (SSHOC).

The workshop use case was based on ongoing work carried out within the SSHOC project (WP3) on a corpus of theatrical play texts from the 17th and 18th century covering examples in three languages (English, French, and Spanish).

Speakers. The workshop was delivered by 3 speakers:

- Francesca Frontini (CLARIN ERIC, ILC CNR, SSHOC T3.1) is currently a member of the Board of Directors of CLARIN ERIC and Research Scientist at ILC-CNR. She focusses on further developing and steering the CLARIN Ambassadors programme, and is interested in the development and use of language resources, named entity recognition and textual analysis. In particular, she has worked on NLP methods for the analysis of literary texts and literary criticism. In addition, she has published extensively on issues relating to language resource documentation, preservation and standardisation.
- Maria Eskevich (CLARIN ERIC, SSHOC T3.3) is CLARIN ERIC Central Office Coordinator. Since joining CLARIN in 2018, she has been involved in a number of H2020 Projects, such as PARTHENOS, SSHOC, TRIPLE, as well as in CLARIN collaboration with Europeana. She has a strong background in language and speech technologies, information retrieval and evaluation, digital humanities. Her research interests span from multimedia retrieval, evaluation of multimedia and podcast search, use of crowdsourcing technologies and social media, text and data mining (TDM), archiving, and use of those technologies in the context of digital humanities. Among other things, she is currently co-organising Multilingual Semantic Search task at Community Evaluation Effort for MultiLingual Information Access on COVID-19 (MLIA).
- Iulianna van der Lek-Ciudin (CLARIN ERIC, SSHOC T3.1) performed the role of the workshop moderator. Iulianna is a Training and Education Officer at CLARIN ERIC and is working closely with the CLARIN community to ensure that the training and resource activities relevant to the humanities and social sciences disciplines are developed according to the FAIR principles for scientific data management.

Organisers. The workshop was co-located with the LIBER Annual Conference and was organised in cooperation with partners from SSHOC T6.5 and WP3.

Participation. 47 people participanted in the workshop.[17,18] The participants came from EU (90%) or non-EU (10%) countries. Given the target audience of the LIBER conference, it is not surprising that the great majority of the participants represented "Research Libraries and Archives". A solid share of 25% was represented by those who belonged to the "Universities and Research Performing Institutions", while around 10% of the audience represented "Researchers, Research Networks and Communities" and "Research and e-infrastructures". "Private Sector and Industry Players", "Civil Society and Citizen Scientists", "Policy Making Organisations" and "Research Funding Organisations" were not represented in the audience.

Brief summary of the event structure.
Due to the COVID-19 pandemic, the LIBER annual conference, which co-hosted this workshop, was moved online and consequently, the workshop itself was organised in a virtual environment. The workshop lasted for one hour and twenty-five minutes and was divided into three main parts. The workshop opened with a presentation of the CLARIN and the SSHOC project. This was followed by a longer session dedicated to the presentation on one possible scenario of use and the motivation for this workshop, as well as the presentation of the Text Encoding Initiative (TEI). Finally, the largest portion was dedicated to a hands-on session, where the participants were able to test offered methods, resources and tools. The main speakers answered questions throughout the event.

Presentations & Discussions: Key Points

First Session. Introduction to CLARIN and SSHOC.
Speakers. Francesca Frontini and Maria Eskevich

Main points. The workshop was given in the framework of the CLARIN and SSHOC project. Francesca Frontini first presented CLARIN and focused on the tools and services offered by this research infrastructure. Maria Eskevich continued with a presentation of the goals and some outcomes of the SSHOC project and situated the workshop in relation to them.

Links to materials.

- Presentation slides
- Workshop recording (min: 9:40–23:35)

[17] It should be noted that the overall reach is higher since this number does not include the views of the workshop recording which is available on the SSHOC and LIBER YouTube channels and promoted by SSHOC and its partners.

[18] The data about stakeholder categories and country of origin/work was not directly collected via registration form. This information is, therefore, based on the *affiliation/company* and *position* field.

Second Session. Scenario of use and motivation.
Speaker. Francesca Frontini
Main points. Francesca Frontini (CLARIN ERIC) presented a basic scenario that motivated the webinar: for example, a researcher is passionate about a research question dealing with theatrical characters, but this same researcher has very limited knowledge of digital sources and methods (e.g. TEI).

In order to help and guide the researcher in this case, the librarian can use the CLARIN Virtual Language Observatory (VLO) to:

- find appropriate data,
- get access to the source material,
- process the text with the Language Resource Switchboard which tools allow you to analyse the text and
- visually explore the text.

Links to materials.

- Presentation slides
- Workshop recording (min: 23:40–29:35)

Third Session. TEI in Details.
Speaker. Maria Eskevich
Main points. Maria Eskevich (CLARIN ERIC) introduced the Text Encoding Initiative (TEI) and showed where librarians can find tools, resources, services, and various teaching materials. The TEI default structure consists of a header, body and textual components. It is recommended that the *header* consists of, at a minimum, bibliographical information (e.g. author, distributor, publisher) and that the *body* consists of annotations such as names, dates, people and places.

Alternatively, there already exist some SSHOC workflows that explain in detail how to create a TEI-based corpus. In addition, a researcher can use corpora and collections in different archives (e.g. VLO, DraCor - Drama Corpora Project and OBVIL - corpus Molière) that are already annotated and available in TEI format.

Links to materials.

- Presentation slides
- Workshop recording (min: 29:40–36:15)

Fourth Session. SSHOC/CLARIN Use Case.
Speaker. Maria Eskevich

Main points. The largest part of the workshop was dedicated to the hands-on session, where the participants were able to test offered methods, resources and tools. The use case – *Intertextuality phenomena in European drama history* – was an interesting research problem because it necessitated the analysis of the literary language of individual dramas of the respective historical language level as well as a comparative literary analysis. The main challenges in performing this study are the sheer volume of the material which cannot be processed manually, multilingualism and the absence of any annotation for parts of the collections.

The sample data which is based on a corpus of theatrical play texts from the 17th and 18th century, is available in three different languages (English, French, and Spanish). One of the issues encountered is the inconsistency of available formats: the documents can be available as TEI-XML, but do not follow any valid schema, they can be encoded with proprietary formats or are only available as plain text files. In order to fix these issues and normalise the corpora format, XSLT (Extensible Stylesheet Language Transformations) and Python scripts are used to clean up different parts of the corpus.

The participants learned how to extract the spoken text of two literary characters (the master and the servant) from the sample data in the single plain text format. The extraction steps included:

- The first step is to find annotated data via an aggregator (e.g. VLO, SSH Open Marketplace) and then download it from the original source of the data collection.
- The second step is to find workflows with scripts, processing examples and compatible tools (via, for example, SSH Open Marketplace, CLARIN LRS). The tedious work is predominantly done by the scripts, so it just takes different actions/commands to complete the process. In this workflow, there are 16 steps in total, and the accompanying documentation includes all the details to successfully run the elements of the workflow.
- The last step is data processing. This can be done offline (by following the instructions provided in a workflow) or online (e.g. via the Language Resource Switchboard).

Links to materials.

- Presentation slides
- Workshop recording (min: 42:20–1:02:50)

Outcomes & Feedback

In this webinar the participants first learned about CLARIN ERIC and the SSHOC project and then observed how encoded documents can be searched for in the CLARIN Virtual Language Observatory (VLO). Then, they were introduced to the basics of the XML-TEI encoding, in particular to those elements that concern theatre plays, characters and their respective lines. Through concrete examples, the participants were shown how simple scripts can be used to generate separate sub-corpora containing the speech for each character or a group of characters. In sum, the participants acquired a general overview of the possibilities offered by CLARIN and SSHOC-related services, and learned how to accompany researchers in their search of useful resources and services.

Participants' satisfaction.[19] The respondents felt that the webinar was excellent or very good (cf. Q1) and that it matched or exceeded their expectations. All of them stated that the webinar will have a positive impact for their work and that they especially liked the combination of general presentations, use cases and hand-on exercises (cf. Q5 & Q6). They report being able to get a broad overview of services that can be further explored, and to learn important details that will help them better understand the needs of researchers and the solutions that can be offered in response (cf. Q5 and Q6).

Feedback regarding the Organisation. Most of the respondents felt that the event was excellently organised (cf. Q9) and that they joined for the topic of the webinar as well as because they were interested in the outcomes of the SSHOC projects (cf. Q3). Suggestions for possible improvements included using a live demonstration of a tool rather than showing the process on the slides, and making sure the slides are unshared during the discussion since seeing other participants rather than slides encourages interaction (cf. Q7 and Q8).

Future work. There is no direct follow-up event envisaged as part of T6.5, but there will be other training and awareness-raising events showcasing SSHOC-related outcomes targeting various stakeholders.

[19] The post-event survey was filled out by 4 respondents.