The Covid Information Commons & Columbia University Libraries – using translation & transcription to increase accessibility to NSF-funded research

by Lauren Close, Lylybell Teran, and Esther Jackson, with editorial support from Florence Hudson, Macy Moujabber, Isabella Graham-Martinez, and Jeremiah Mercurio

With thanks to Lara Azar, Elia Bregman, Brian Buckley, Cora Lee Cole, Victoria Horrocks, Saanya Subasinghe, Rhyley Vaughan, and Kathryn Pope.

As our scholarly communication ecosystem becomes increasingly reliant on digital media, it is important to keep diversity and accessibility at the center of our digital designs and infrastructure. Accessible design means that all people can share, perceive, navigate, and interact with electronic information regardless of their ability/disability status. It is equally important to share materials in multiple media formats and languages to reach diverse audiences around the world.

In collaboration with the Columbia University Libraries, the Northeast Big Data Innovation Hub (NEBDHub) began a project in 2021 to make the resources associated with the NSF-funded COVID Information Commons (CIC) virtual events accessible to the broader public. The NEBDHub team of staff and students, with support from the Columbia University Libraries' Academic Commons, created English transcriptions and captions and Spanish translations for each of the CIC webinar research lightning talks.These COVID-19 resources were designed in alignment with the NEBDHub's website accessibility principles and Diversity, Equity, Inclusion, and Accessibility policy guidelines. The materials will be maintained by the Academic Commons in perpetuity, ensuring their availability to students, faculty, staff, and members of the public.

In 2019, 41.4 million Americans (12.7% of the U.S. population) self-identified as being individuals with disabilities, whether that disability signified cognitive, ambulatory, hearing, or vision difficulties.[1] Given the sweeping significance of the COVID-19 pandemic, an event which has had an unprecedented impact on global health and economic outcomes, it is the position of the CIC and the Libraries that it is the responsibility of the academic and scientific community to ensure these research and educational materials are accessible to the broadest possible audience. When resources such as these are developed with web accessibility in mind, it means that all members of society have equal access to crucial information about pandemic prevention and preparedness measures.

Similarly, the CIC team believes in the importance of translating our web materials into languages other than American English. After English, the second most commonly used language in the U.S. is Spanish. In 2021, approximately 41.7 million Americans (13.5% of the U.S. population), spoke Spanish at home.[2] To reach this sizable audience, the CIC team

---

[1] https://www.census.gov/newsroom/facts-for-features/2021/disabilities-act.html

[2] https://www.weforum.org/agenda/2021/12/spoken-language-united-states-america-english-spanish-mandarin/

committed to translating the valuable Lightning Talk presentations into Universal Spanish. We have also spearheaded initiatives to partner with Hispanic Serving Institutions (HSIs) at the undergraduate and graduate level to ensure the broad reach of these materials.

Below, the NEBDHub and Academic Commons teams share the processes developed for this project in the hope that it will be of use to other teams interested in digitizing their materials in an accessibility-friendly manner and in alignment with broader DEIA efforts.

Background

The COVID Information Commons (CIC) is funded by the National Science Foundation (NSF) through grants #[2028999](#) and #[2139391](#). The CIC is an open resource for exploring research on the COVID-19 pandemic and offers an open portal of over 9,400 NSF and NIH-funded research projects and community events to enable researcher collaboration.

In July 2020, the COVID Information Commons (CIC) began hosting monthly webinars for NSF and NIH-funded researchers to present their COVID-19 research to a community of interested professionals and students from around the world. The researchers have shared their insights on all aspects of the COVID-19 pandemic, ranging from epidemiology to education impacts and healthcare outcomes. Presentations are formatted as short lightning talks and followed by open Q&A sessions with audience members. Through January 2023, the CIC hosted 118 presentations in 24 webinars, reaching over 9,900 audience members via live events and the [NEBDHub's YouTube Channel](#). [Individual Lightning Talks](#) and recordings of [the webinar sessions](#) can be found on the CIC website and in [Academic Commons](#).

In 2021, the CIC Project Team began the process of transcribing the presentation videos and providing written summaries of the events. In line with the NEBDHub's Diversity, Equity, Inclusion and Accessibility (DEIA) goals, we took further steps to make the content broadly available. First, we transcribed the CIC lightning talk videos into written English to enable the hard of hearing to better access the content. Second, we leveraged Adobe Acrobat functions to modify the transcripts to meet accessibility standards suitable for individuals who require screen readers. To further extend our NEBDHub and CIC community reach, students and staff translated the CIC English transcripts into written Spanish, as there are hundreds of Spanish-speaking members of our community. The Spanish transcripts were likewise brought into alignment with Adobe Acrobat's accessibility standards so as to be suitable for individuals who require screen readers.

The Columbia University Libraries Digital Scholarship team partnered with the CIC Project Team to advise on certain aspects of the project (e.g., suggesting that translation start with the time-encoded caption files, and then be reflected in narrative transcript documents) and further highlight the outcomes of the transcription and translation project. Academic Commons offers CIC researchers the opportunity to generate a unique DOI (Digital Object Identifier) for their presentation so that their work can be cited and referenced in academic publications in perpetuity. It is also indexed by scholarly research aggregators such as Google Scholar and OpenAlex.

Process

The CIC Project Team is pleased to share the details of the processes used for this initiative.

After each CIC virtual webinar, the CIC team spliced the 60-90 minute webinar into the separate COVID researcher lightning talks, each at approximately 10-15 minutes in length. Each lightning talk was then added to the Project Tracker spreadsheet (Sample Project Tracker) and the initial transcription and translation responsibilities were assigned to CIC team members. The English transcribers began working directly in the settings of the NEBDHub YouTube account. When the English captions were correctly updated in YouTube, the text was transferred to a shareable word document (Sample Word Document Template) for formatting. When complete, edits were solicited from a second team member, who was also responsible for generating the Adobe Acrobat version of the document with built-in accessibility features. The final English transcription was published to the CIC website for public consumption. All documentation was then passed to the Spanish translation team, who used a combination of digital tools (such as Reverso) to establish a baseline translation for the talk. The team further refined the text with particular focus on accurate translations of the technical terms used in the talks. Once completed, the Spanish translations were also posted to the CIC website and the Spanish text was manually added to YouTube as captions.

An overview of the process is shared here, including a timeline for securing permissions from webinar speakers.

After the final videos and translations were posted to the CIC website, Northeast Big Data Hub staff worked with Libraries staff to deliver the materials on a schedule, in a structured format, for inclusion in Academic Commons. This process included requesting signed Academic Author agreements from all speakers. English and Spanish records for each talk (for which the author has given approval) were then published to Academic Commons on a rolling basis, and these records include the lightning talk videos, transcripts, and captions.

Results

As a result of the CIC Project Team's efforts, all 118 Lightning Talks from July 2020 through January 2023 have been transcribed into written English and translated into Universal Spanish. Over 190 unique DOIs have been generated from the resulting documentation and shared with the PIs for grant reporting purposes, further dissemination through their ORCID (Open Researcher and Contributor IDentifier), and other mechanisms. In 2023, 24% of all video views on the NEBDHub's YouTube channel made use of English and Spanish subtitles.

The 194 talks[3] in Academic Commons have been downloaded 9,905 times as of April 2023, and their records have been viewed 9,494 times.

Resources
Suggested Process for Event Transcription & Translation

---

[3] English originals and Spanish translations are counted separately, so this represents 97 original talks.

Sample Transcription Template
Template Transcription Project Tracking Document
Columbia - Accessibility

Four student employees of the NEBDHub, one student employee in the Libraries, and five REAL Volunteers have contributed to this initiative, providing transcription edits and translation support. We would like to thank Lara Azar, Elia Bregman, Brian Buckley, Cora Lee Cole, Victoria Horrocks, Isabella Graham-Martinez, Macy Moujabber, Saanya Subasinghe, Lylybell Teran, and Rhyley Vaughan for their project support. This has been a truly global initiative and the benefits of our accessibility and diversity-forward efforts are already being realized.