Notes on subtitles

by Maria Dimou / Academic Training & IT e-learning
Thanks are due to Vint Cerf, Ken Harrenstien, Dimitri Kanevsky (Google), Alex Manzoni (Red Cross), Andreas Hoecker, Manuella Vincter, Thomas Baron, José Benito Gonzalez, Pete Jones, Jean-Yves LeMeur, René Fernandez, Lorys Lopez, Ruben Gaspar (CERN), Matthew Goodman (EPFL).

Latest news

Click on or scroll down to the Action's Log (https://codimd.web.cern.ch/s/BJDNA-oiB#Actions%E2%80%99-log) to find them.

Background

We have colleagues who can't hear. Elementary diversity awareness requires that we equip all CERN-made videos with subtitles. This note is about an investigation of:

1. A good transcription software.
2. A scalable way to display subtitles in CDS.

Executive summary of conclusions

Scroll down for the latest news at the end of this page.

About the transcription software - December 2019 status

The selected product should be used for *all* events, live and recorded (also the ones of the past). An Open and free-of-charge solution would be ideal, if quality corresponds.
The one used by *YouTube* seems to be the best quality-wise from the ones we've seen.
Concerns were raised by ATLAS and IT/CDA-DR management about potential legal aspects, the text having transited, via the API, through YouTube. The concern refers data ownership in general and data confidentianlity in particular of *restricted* videos from important meetings.

A formal proposal can be presented to the HR *Diversity Office* requesting (co-)funding of the total audiovisual webcasting/recording process review, with systematic subtitles' inclusion in ALL cases. Manual review of the subtitles' quality is *indispensable* if we opt for quality results. This requires a permanent human effort provision that we may not afford. Hence, it has to be a priori agreed how much error margin we can accept in the output.
See Appendix for details.

About displaying from CDS with subtitles - December 2019 status

We need a new *player* software to display simultaneously *camera* and *slides* .mp4 **with** subtitles. Lorys found the Open Source player paella (https://paellaplayer.upv.es/) being now evaluated in the team of audiovisual experts in IT/CDA-IC. If proven as good as it looks, it will replace the current Theo player (https://www.theoplayer.com/). *paella does* display the CERN Document Server (https://cds.cern.ch/) (CDS) and videos.cern.ch required **.vtt** format.
Moreover, if the filename has the right format, *paella* displays the subtitles with no manual

configuration needed (see here today's required manual configuration (https://it-e-learning.docs.cern.ch/administration/#how-do-i-include-the-relevant-subtitlesvtt-file-in-a-published-video-record)).

A requirement of the audiovisual service experts is to find new/long-lasting operational solutions for the whole stack of webcast/recording/transcription equipment and software. We are in a good path with the opencast (https://opencast.org/) evaluation concerning this matter, as *Opencast* seems able to replace both *Sorenson* (https://en.wikipedia.org/wiki/Sorenson_Media) for transcoding and *Micala* (http://micala.sourceforge.net/) for archiving.

For the few videos which contain subtitles at present, CDS displays well the e-learning collection (https://cds.cern.ch/collection/E-learning%20modules?ln=en) because it *does* invoke the player. For records containing several files .mp4 (e.g. camera and slides) and .vtt (i.e. even if subtitles are already present) CDS currently only offers to download, hence the subtitles are not merged.

Details about the transcription software

1. **amara** For the IT e-learning short videos we do the subtitles by hand using *amara* (https://amara.org/en/). This affordable because the videos are <=5 minutes-long & the script is pre-written. Still the process making the subtitles_en.vtt file and configuring CDS is entirely manual and very time consuming (>3 hours for a 5 mins' video). All relevant links here:

    - Video library sorted by topic (https://twiki.cern.ch/Edutech/VideoLibrary)
    - which are sorted views of the relevant CDS collection (https://cds.cern.ch/collection/E-learning%20modules?ln=en).
    - Documentation on using amara (https://it-e-learning.docs.cern.ch/video/faq/#q5-how-do-i-introduce-subtitles)
    - Documentation on configuring subtitles' display in CDS (https://it-e-learning.docs.cern.ch/administration/#how-do-i-include-the-relevant-subtitlesvtt-file-in-a-published-video-record)
    - How we did mass subtitling of existing videos (https://it-student-projects.web.cern.ch/projects/e-learning-it-collaboration-devices-applications-insert-subtitles-video-tutorials).

2. **YouTube** For selected videos from the Academic Training recordings in CDS (https://cds.cern.ch/collection/Academic%20Training%20Lectures?ln=en)we rely on *YouTube* transcription with impressively good quality results. Nevertheless, family names and particle names are not always right), examples in the CERN Lectures YouTube channel (https://www.youtube.com/channel/UCwXkOx0EuKBR5m_OOiaZRUA). Input from Jean-Yves on the quality: *Dommage, lorsque le discours devient technique, c'est quand même approximatif. Par exemple, regarde https://www.youtube.com/watch?v=8HIse5Y5Tho (https://www.youtube.com/watch?v=8HIse5Y5Tho) autour de 42mn10s. At 42mn59s, "neutrino" is transcribed into "student Reno's" :-).*
    The current process (aveditor) (https://it-e-learning.docs.cern.ch/aveditor/#the-academic-training-use-case) to publish a CDS record in YouTube doesn't scale (manual, very time consuming and not covering recordings with access restricted). Notes on tests done by Pete and Alex:

    - Some help for manual download (https://www.4kdownload.com/howto/howto-

download-youtube-subtitles). Manual and temporary if YouTube policy changes.
    - For Linux users, the youtube-dl script (https://www.ostechnix.com/download-youtube-videos-with-subtitles-using-youtube-dl).
    - Other downloader (https://youtubedownload.video/en1/).
    - Transcription process in our e-learning collection (https://twiki.cern.ch/Edutech/TranscribeYourVideo).

1. **S2T** The Digital Memory project evaluated *S2T* tool by WIPO (https://www.wipo.int/s2t/) (World Intellectual Property Organisation): It has been trained with 60 hours of training data from HEP so far (after we signed a use agreement). First comparative report (https://cds.cern.ch/record/2304470). It makes reference to the Google Cloud Speech API. It was written in 2017. Conclusion was that human review was indispensable. Examples:
    - Reviewed by human here (http://digital-memory.web.cern.ch/digital-memory/media-archive/video/open/subtitles/script/displaySubtitle.php?videoid=Video-2293460-a).
    - Not reviewed here (http://digital-memory.web.cern.ch/digital-memory/media-archive/video/open/subtitles/script/displaySubtitle.php?videoid=Video-423243)
2. **ai-media** Australian commercial company Ai-Media (https://www.ai-media.tv/), also in the UK. Was hired by Thomas to do some live captioning of Vidyo meetings. Contact James.Ward@ai-media.tv. The service involves paying money. Input by Jean-Yves: *I think the only service tested with AI-Media was live captioning, which turned out to be below the expectations of the user (Thomas can provide the details as I think it was on a very specialized topic).*
3. **3playmedia** USA commercial company 3playmedia (https://www.3playmedia.com/) was listed as candidate, evaluation not yet done. The service involves paying money.
4. **MLLP service (https://ttp.mllp.upv.es/index.php?page=faq)** from the Universitat Politècnica de València (UPV) (http://www.upv.es/), a "service" by their automatic transcription/translation/interpretation research unit, sponsored by the EU for the EMMA project (http://project.europeanmoocs.eu/about/).
5. **EPFL input** Maria contacted MOOC makers (Matthew Goodman)at EPFL. They mostly use *amara* for the MOOCs, like us for the e-learning. For all-day events like the EPFL Open Science day (https://www.epfl.ch/campus/events/celebration-en/open-science-day/) they just rely on *YouTube*.

SNOW tickets during the investigation

Investigation started in June 2018 and is managed by SNOW tickets. On-going as per the Executive summary above.

- 2018 analysis and conclusion (https://cern.service-now.com/service-portal/view-request.do?n=RQF1041572)
- November 2019 investigation (https://cern.service-now.com/service-portal/view-request.do?n=RQF1460345) of possible new products and workflow.
- April 2020 extract of Indico (webcast)and CDS/videos recordings (https://cern.service-now.com/service-portal?id=ticket&n=RQF1562305) See files attached to the ticket.

Appendix

Maria contacted Vint Cerf, Internet father, now at Google, who also has a hearing impairment. Replies below led to the recommendation in the Executive Summary above. Comments by Ken Harrenstien (Google): Human review always required. One advantage of YouTube is that you can solicit volunteers world-wide to help fix and translate caption tracks.
Suggested APIs:

1. https://developers.google.com/youtube/v3 allows to upload a private video, wait a while until the ASR track is created, download that track, then delete the video. Free.
2. Cloud Speech to Text API (https://cloud.google.com/speech-to-text/) almost the same as what YouTube uses.
3. Google Live Transcribe code now Open Source (https://github.com/google/live-transcribe-speech-engine). It can be tried via Android devices here (https://www.android.com/accessibility/live-transcribe/).

Subtitles' quality review

For reviewing the transcription we could see if the CERN host-states' office can advise us on contracts like:

- Place d'apprentissage (https://junior.gateway.one/apprentissages/lieu-geneve?region=fr-CH) which involves small amount of money contributed by the state or
- The swiss army civil service (https://www.zivi.admin.ch/zivi/fr/home.html)

Nevertheless such contract solutions can't be available without interruption. ATLAS doesn't mind the imperfections of the automatic transcription, as the community is aware of the expected terms. In any case, it is important to have an agreement on the quality of the expected transcription result from the beginning of the project.

Actions' log

Most recent first:

2021/01/11 CDA Internal meeting

Present: Jean-Yves, Thomas, Maria D., Ruben, Tim, José, Nicola.

Meeting agenda (https://indico.cern.ch/event/958294/)

A lot of work was done by all of the CERN IT/CDA-IC section participants in this meeting, during these last few months/weeks. The MLLP Pilot is now over and we have to go to a Call for Tender. During the pilot, with the help of student Amine Hadjiat on internship from the University of Geneva (see project [a]), we post-processed MLLP transcripted and translated lectures from the LHCP2020 conference to train the system (see notes in [b]). The MLLP software capabilities and the professionalism of the MLLP experts were impressive. The collaboration was very satisfactory. The final report from the MLLP pilot will be linked from [c]. Still, we can't say anything for the long run, before the Tender exercise goes through. The MLLP advantage is its good integration with the rest of the workflow of Weblectures' publishing tools and the player (paella). It is a Universitat Politècnica de València (UPV) spin-off, with 10 years of experience and quite promising for survival. High-profile webcasts by famous speakers, would have good live-transcriptions with MLLP, as the system can be trained by existing papers by the speaker fed into the system in advance, for "teaching" the system with the technical jargon.
For a limited-time contract, we can always fall back to another co-tender, if MLLP has not the

bright future, we hope it will have.
If we create a backlog of manually-prepared subtitles, we can always use it to train any new system, if we need to change.

The Microsoft (MS) Translator or the Google Meet automatic transcription have very good quality.
We should **find out if the new "MS Campus" agreement includes such services** (Thomas).
Our strategy allows us to go for a GAFAM solution, if we find nothing better. Still, we hope for a european solution, one that ensures control over our own data. In this spirit, cloud solutions are not great.
The WIPO product gives us a free license for offline transcription but not for live transcriptions, nor for support. The community now requires that as well.

With the most recent CERN-Zoom contract, captions can appear automatically for every event. Still, we keep investigating an independent, ideally open, good quality solution instead (in additon). The Zoom hosts can enable live transcription, save the video and publish the Zoom subtitles (in .vtt or .dfxp format), although the user community finds their quality not great. In this way, the tool that will win the tender will not have to take the burden of *all* meetings and lectures organised at CERN.

First priority for now is to **assemble the *Specifications for the Call for Tender*** (Thomas, Ruben, Maria, with existing material from Jean-Yves' experience with the WIPO product).

Colleagues from ATLAS, University of California, Rio de Janeiro and BNL collaborating in the *Diversity, Inclusion, and Equity topical group of the Snowmass process* [c] are very interested in our work and willing to help, in the framework of their study *accessibility for all participants* [d]. The Snowmass process is akin to the European strategy and one of the goals is to ensure that our communities work together to a great extent. Maria will ask their **input for the *Specifications***.

The date of the next meeting will be agreed in email, based on the *Tender* progress.

[a] https://it-student-projects.web.cern.ch/projects/correct-automatically-transcribed-videos-input-mllp
[b] https://codimd.web.cern.ch/kkvPgm8kQfisHkREE2szFA#
[c] https://codimd.web.cern.ch/CkA_VyauS_CYqXZrqPzPQg#
[d] https://snowmass21.org/community/diversity
[e] https://docs.google.com/document/d/1a_XqcI7r76Vo_Mjirxdzd_1fPoyHWqyV2m8wCdU38YA/edit#

2020/09/21 CDA Internal meeting

Present: Jean-Yves, Thomas, René, Maria D., Ruben.
Apologies: Tim, José, Nicola.

Meeting agenda with Ruben's slides (https://indico.cern.ch/event/923007).

Conclusions

- **On otter**:Thomas purchased *one* otter.ai license at 240 US$, currently used in the section with a collaborator who has a hearing impediment. One can see events https://cds.cern.ch/record/2719117 and https://cds.cern.ch/record/2719020 (also linked from the agenda), where otter was used to introduce subtitles. It use requires linking

between the otter and Zoom accounts. This is why use by more people would require purchase of more licenses. A second one will probably be purchased by IT-CDA-IC as a new request for *live captioning* just came in. Further licenses, if purchased, would have to be paid by the requesting departments.
- **On amberscript**: Abandoned. Compared to otter showed too bad results to be retained.
- **On Google Docs dictation**: Ruben said that it can be proposed as a "do-it-yourself" option. Users should remember that quality is non good enough and that the audio should be re-directed to googledocs, so they have to only *listen* and not speak at the Zoom meeting, as the microphone is no more active. Not bad for listening to webinars, for instance.

Anyway, all the above are interim solutions. The MLLP deployment is the one that will give a rich service, including Online & Offline transcription, voice customisation for known speakers (e.g. the DG) and (later) translation.
Action: Ruben will post in the Discourse of the Zoom@CERN pilot (https://videoconference.web.cern.ch/) the current options and their limitations, including otter.
DONE https://videoconference.web.cern.ch/t/diy-automatic-transcription-online-events/172

On MLLP: Ruben presented these slides (https://cernbox.cern.ch/index.php/s/LGvN9jBprGi54KM), linked from the agenda, also present in cernbox. The slides contain the links to the DAI and NDA (Non Disclosure Agreement), which are now done, as well as links to internal work notes and MLLP Guidelines (https://cernbox.cern.ch/index.php/s/fGADW4qK610Ykyd).

Suggestions:

1. Jean-Yves, surprised that the videos from the archive were rejected for transcription quality, sent this link to videos which have been human-checked (http://cds.cern.ch/search?lcc=CERN+Moving+Images&sc=1&p=8564_i%3A%22Video+with+Subtitles%22&actior
2. Maria asked if extracting the YouTube-made subtitles *and then* manually correcting the errors could be useful input to MLLP system training. Ruben said "yes". This would be from the Academic Training lectures' subset (https://www.youtube.com/channel/UCwXkOx0EuKBR5m_OOiaZRUA) about 35 hours of physics and computing lectures. **Action** Maria with student Mira Buzanszky.

A.O.B.

- Miguel-Angel joined IT-CDA-IC to work on *Opencast* (https://opencast.org/), the Open Source replacement of Sorensen and Micala.
- Next meeting will be on **Monday December 7th at 2PM.** Agenda (https://indico.cern.ch/event/958294). This meeting had to be moved to January 11th 2021, due to the CERN ARCHIVER project kick-off that took place on December 7th 2020.

2020/05/25 CDA Internal meeting

Present: Tim, José, Nicola, Jean-Yves (JY), Thomas, René, Maria D.
Absent: Ruben.

Meeting agenda (https://indico.cern.ch/event/905172/).

Conclusions

1. The MLLP service cost exceeds the limit that would spare us a call for tender. This is why Thomas and Ruben re-scoped the MLLP pilot project into a consulting action from the Technical University of Valencia, to help us establish adequate specifications for a future call for tender, including extensive quality indicators from a trained ASR system.
2. News from the tender process will only be known in the autumn. This is why our next meeting will be on **September 21st at 2PM**. Agenda here (https://indico.cern.ch/event/923007/).
3. The MLLP service seems to offer all the features that we need (automated transcription, translation, training, cloud service and extensive API, it's important to confirm this and understand what quality levels we can expect from such a service in order to run an efficient tender afterwards.
4. ATLAS would like to see evaluation between multiple products. This was done and is still being done as explained below.
5. Google Docs dictation (https://www.pcworld.com/article/3038200/how-to-use-voice-dictation-in-google-docs.html) and Zoom Closed Captions (https://support.zoom.us/hc/en-us/articles/207279736-Using-closed-captioning) were evaluated and showed too many errors to be retained.

    1. The Zoom Cloud recording service, comes with an automatic transcription.Thomas will check if the format is .vtt.
    2. If the format is .srt by default, the CDS team will accept them or convert them.

6. The TECH student (see job description here) (https://it-student-projects.web.cern.ch/projects/deploy-subtitles-service-cern-videos) was dropped, although approved, because we won't have a *service* ready to tune, in the given timeframe.
7. For displaying the videos in CDS the *Theo player* is still used for most videos, e.g. the DFS directory *Weblectures*. Nevertheless, the plan is to expand the use of the *paella player*, which is already in use for *live webcasts*.
8. The WIPO S2T product might not be able to participate in the tender, given that *Cloud service provision* is part of the tender specification.

    1. We can be covered for *Privacy issues* with a Cloud-based service, by explicit clauses in the contract.

9. Comparison between Open Source solutions and MLLP was done and presented by Ruben last time (slides below). None of them met the requirements.
10. Requiring an on-prem (https://en.wikipedia.org/wiki/On-premises_software) service is not a good idea because of many hidden costs.
11. For *live captioning*, Thomas purchased a otter.ai (https://otter.ai/login) license (20 CHF/person for 6K minutes/month) for use from within Zoom for evaluation. To limit the need for too many licenses, the CERN Audiovisual service, will be doing the set-up when requested in a SNOW ticket. The quality of automatic live transcription is poor, despite the 800-words' vocabulary that can be fed into the product, to avoid gross errors.
12. For *offline transcription* Amberscript (https://www.amberscript.com/en/) and otter.ai will be used to transcribe the Zoom recordings of the LHCP Online conference this week. One or the other will be chosen after comparing the quality of their output. CDS embedding should be trivial as both have either an .srt or .vtt export. See the method here (https://it-e-learning.docs.cern.ch/administration/#how-do-i-include-the-relevant-subtitlesvtt-file-in-a-published-video-record).

2020/03/31 CDA Internal meeting

Present: Tim, José, Nicola, Jean-Yves (JY), Thomas, René, Maria D., Ruben.

Meeting agenda (https://indico.cern.ch/event/890353/).

Ruben's slides that drove the discussion (https://indico.cern.ch/event/890353/attachments/1989051/3363027/Ruben-OnlineOffinetranscripts.pdf) containing all the numbers that lead to the conclusions below.

Conclusions

1. The scope is now larger, as the requirements will become pressing anyway and our transcoding and transcription stack cannot be changed very often. The tool of our choice should be able to offer offline AND online transcription, selectively translation and voice synthesis, the whole also for the material of the Digital Memory Project (https://cds.cern.ch/record/2200146/files/CERN%20Digital%20Memory%20Specs.pdf?).
2. After Ruben's investigation of various products, including some in the original list above (https://codimd.web.cern.ch/rAX3vM6XTi657uCYsMyZXQ#Details-about-the-transcription-software), the MLLP and WIPO S2T remained the most serious ones, as per our last meeting.
3. We can not test these products now because we don't have the software on-site.
4. The choice is in favour of MLLP, because we need an API integrated with the CERN SSO. WIPO S2T doesn't provide an API.
5. When evaluated, the WIPO S2T ran at CERN, so our data didn't leave CERN, which was good. Unfortunately, only the backend engine is provided by WIPO. The rest would have to be developed by us, an effort that we cannot provide.
6. There will be a **Pilot** with MLLP (https://ttp.mllp.upv.es/index.php?page=faq) from from the Universitat Politècnica de València (UPV). Reasons:
    - The product integrates well with the new technical stack of the transcoding service, e.g. the paella player (https://paellaplayer.upv.es/).
    - The MLLP topic annotation feature is very useful.
    - The pilot has modest pricing - see here slides 9 & 12 (https://indico.cern.ch/event/890353/attachments/1989051/3363027/Ruben-OnlineOffinetranscripts.pdf).
    - A PJAS will come from UPV, in July 2020, to work mainly on Opencast.
    - The pilot will lead, if we are satisfied, to a 3-year *service* contract 55KCHF/yr, which will be a **collaboration CERN-UPV**, with support by an engineer.
    - We have to make sure the pilot engages with all of the use cases. The contract signing should be delayed until the end of the pilot.
7. Tim emphasised that if we make an investment for the coming pilot+3 years, it would be wise to have a set-up that allows us to move to an Open Source solution *without losing this work*. It is good to get a service. Still, at the end of the contractual period, transcriptions should belong to us. We'd like to also have full ownership of our transcription data and algorithms. Ruben replied that we don't have control over the MLLP or WIPO engines. The Mozila Deepspeech *is* Open Source, indeed, but the WER (Word Error Rate) is bad.
8. José asks *what is behind the products?* Ruben says that MLLP, being a research group , although closed source, it is a home-made software. As they are service providers they can be reliable for the respect of data privacy and the quality of service they will provide.
9. Nicola needs APIs for the interface between CDS and the CDA/IC transcription service. MLLP provides that.

10. Maria regularly meets with the ATLAS Deputy Spokespersons Manuela Vinkter and Andreas Hoecker. Their position is that it will be *a great prestige for CERN to ensure subtitles for our videos*. Giordon Holsberg Stark - ATLAS member with hearing impediment - offered to help during the product evaluation, comparison, subtitles' correction, glossary making etc. JY has already collaborated with Giordon during the WIPO S2T evaluation in the past. Both JY and Ruben think his offer to help is **not** needed at the moment.
11. Maria suggested a mini-project for the identification of WER of the subtitled CERN Lectures in YouTube (https://www.youtube.com/channel/UCwXkOx0EuKBR5m_OOiaZRUA) by a *CERN Child* or *Intern*, in order to build a *Glossary of HEP terms*. Ruben and JY said this is not necessary. MLLP works with ML algorithms, no way to "input a glossary".
12. Maria suggests to obtain from the Indico and CDS experts all the events that requested webcastin 2019 the recording backlog, in order to estimate the effort needed to get all pre-recorded lectures transcribed and live events subtitled. Done - see the results attached to this SNOW request (https://cern.service-now.com/service-portal/view-request.do?n=RQF1562305).
13. Maria asked what will happen with this TECH student proposal (https://it-student-projects.web.cern.ch/projects/deploy-subtitles-service-cern-videos) she wrote after the January meeting. The proposal is submitted indeed. Ruben will be the supervisor. She had called Luisa Carvalho from Diversity for the funding but a f2f meeting was postponed after COVID-19.

Next meeting 2020/05/25

2020/03/03

- Maria met ATLAS deputy spokespsersons Manuella Vincter and Andreas Hoecker.
- Ruben wrote a note from MLLP and WIPO S2T meetings and evaluations in February here (https://codimd.web.cern.ch/8LsIM46XSsK1hhuJZtRQOQ#)

2020/01/09 CDA Internal meeting

Present: Tim, Jose, Nicola, Jean-Yves, Thomas, Rene, Maria D., Ruben.

Meeting agenda (https://indico.cern.ch/event/868274/). One can see, on this event, 2 lectures with subtitles made while Ruben tested the MLLP service (https://ttp.mllp.upv.es/index.php?page=faq) from the UPV University, a "service" by their automatic transcription/translation/interpretation research unit, sponsored by the EU for the EMMA project (http://project.europeanmoocs.eu/about/).

Summary and Conclusions

Transcoding tools

Ruben reminds us of the legacy components: CES (Central Encoding System), *Micala* (http://micala.sourceforge.net/) and *Sorenson* (https://en.wikipedia.org/wiki/Sorenson_Media) (company no more exists). Affected services are CDS, Indico and SNOW.

Transcoding The, originally Zurich University, Open Source product Opencast (https://opencast.org/) is chosen, as *Opencast* is well maintained by a large community and covers both our transcoding and archiving needs.

Player The new Open Source paella (https://paellaplayer.upv.es/) is chosen. It comes with *Opencast*, displays simulataneously *camera* and *slides* .mp4 **with** subtitles in .vtt format. Any user can edit the paella subtitles as a file and send it back to us (maintainers in IC) for validation and integration in the subtitles. It is validated also in webcast and mobile environment. It will replace the current Theo player (https://www.theoplayer.com/).

- Timeline for Opencast: Collaboration with Universitat Politècnica de València (UPV) signed. PJAS comes in July 2020. Consultancy contract can be signed earlier. Ruben to communicate his optimal plan with the order of services affected. Get agreement from DR section.
- About the actual videos' filesystem: Opencast has NFS as default. We shall use CEPH (ask the service managers to mirror the current mediaarchive now in DFS). Opencast's access to storage is done via a web service.
- About restricted videos: Nicola finds this is an opportunity to improve the current workflow, where the user has to login twice, once for the restricted event and one for viewing the video itself.

Running of the services Responsibility remains with Ruben, will need DR participation on the CDS integration part of the new solutions.

Transcription tools

List of products reminded and enhanced with the MLLP service (https://ttp.mllp.upv.es/index.php?page=faq) from the Universitat Politècnica de València (UPV) (http://www.upv.es/), by their automatic transcription/translation/interpretation research unit, sponsored by the EU for the EMMA project (http://project.europeanmoocs.eu/about/).

Scope
We should provide subtitles for *all pre-recorded* videos, not live. Subtitles should appear by default, unless the info is important and the event manager asks not to include the automatically-produced ones but check/make them manually, to guarantee transcription quality. When we reach operation status, the one who checks the automatic transcription's quality should be the **speaker** (Tim).

Preferred products

- MLLP (https://ttp.mllp.upv.es/index.php?page=faq) can take slides and article/script/notes accompanying a video and use them via their AI algorithms to fix the vocabulary of the automatic trascription. They also provide a solution for live speech. We could offer to run the service here as they lack computing power. This solves the problem of "black box in Valencia".
- WIPO S2T (https://www.wipo.int/s2t/) is also free-for-CERN and Open Source. It requires GPUs. We have an evaluation license. We shall have to go through the CERN Legal service if we wish to run it as a service.

After the meeting Jean-Yves circulated the requirements' list (https://codimd.web.cern.ch/s/Hk1edZjlU#) he has for the Digital Memory project. Ruben and Jean-Yves will extend the table to select one of the 2 above products best suiting the criteria best, also in terms of functionality, OS and infrastructure. Thomas suggests to add the optional requirement for the selected *automated transcription* product for our offline base of videos, to *also* do live transcription. Maria wrote this TECH student project proposal (https://it-student-projects.web.cern.ch/projects/deploy-subtitles-service-cern-videos) to request *Diversity Office* funding. The student will set-up the *service* for the pre-selected product.

A.O.B.

- There was a question of impact of the transcoding infrastruction change to the timescale of videos' move from cds.cern.ch to videos.cern.ch. This is not obvious and should be discussed at the next checkpoint meeting.
- Suggested date **Thursday March 19th @ 4PM** (Agenda (https://indico.cern.ch/event/890353/))
- Discussion items include:
    - Storage issues, i.e. EOS vs CEPH.
    - Where to display the final video.

Appendix

1. CERN Academic Training - New lecturing method (https://codimd.web.cern.ch/s/B13U3qjiL#).