CORV

Audio and video transcription tool for research data

HI!

Who?

- Hugo Hueber
- DCSR Research engineer dedicated for FDCA
- Maintainer for Corv

What?

- **DCSR** (Division Calcul et Soutien à la Recherche): **Scientific Computing and Research Support Unit**
- **FDCA** (Faculté de droit, des sciences criminelles et d'administration publique): **Faculty of Law, Criminal Sciences and Public Administration**

UNIL | Université de Lausanne

Where do crows go to get educated?

– Caw-lege

WHAT IS CORV?

TRANSCRIPTION?

- Transformation of audio/video into text.
- Numerous use cases.
  - Interviews, discussions, subtitling, …
- Useful for multiple researchers at UNIL.
  - Identified cases in social sciences, forensics, economics, …

```
3
00:00:05,597 --> 00:00:09,618
[SPEAKER_01]: Hello, I'm Susan Thompson, resource manager here.

4
00:00:09,638 --> 00:00:12,879
[SPEAKER_00]: Hi, I'm Mary Hansen and I'm applying for one of your kitchen jobs.

5
00:00:13,240 --> 00:00:13,700
[SPEAKER_00]: Great.

6
00:00:13,840 --> 00:00:15,300
[SPEAKER_00]: Here's a copy of my resume.

7
00:00:15,981 --> 00:00:17,121
[SPEAKER_01]: Great, have a seat, Mary.

8
00:00:17,161 --> 00:00:17,821
[SPEAKER_00]: Thank you.
```

AN EFFORT THAT CAN BE AUTOMATED

- For a long time, the work was done by hand.
  - Listening to the audio multiple times and typing whatever one heard.
- Nowadays, numerous tools are available.

A FEW PROBLEMS

- Quality of transcription.
  - Some softwares are not that good.
- Cost of commercial solutions.
  - Licensing is expensive.
- Difficulties in using certain open-source solutions.
  - The technological slope is too complicated for a non-technical user.
- Data protection.
  - Where is the data going exactly?

```
(env)                                                whisperx
 sample.mp3 --model small --compute_type int8 -
-output_dir sample_results --verbose True --dia
rize --hf_token hf_ceciestuntokendexemplehahabi
enessayé
```

00:00:09 Présentateur 2

Marie Hansen, and I am applique for One of your Kitchen job. Here's a copy of my resine.

00:00:23 Présentateur 1

No, but I want.

00:00:26 Présentateur 2

To learn, I work **** and a lot at home.

00:00:31 Présentateur 1

Ok, Tell me about yourself.

Etat: 8 septembre 2020

Etat de la protection des données dans le monde

DCSR'S PROPOSAL

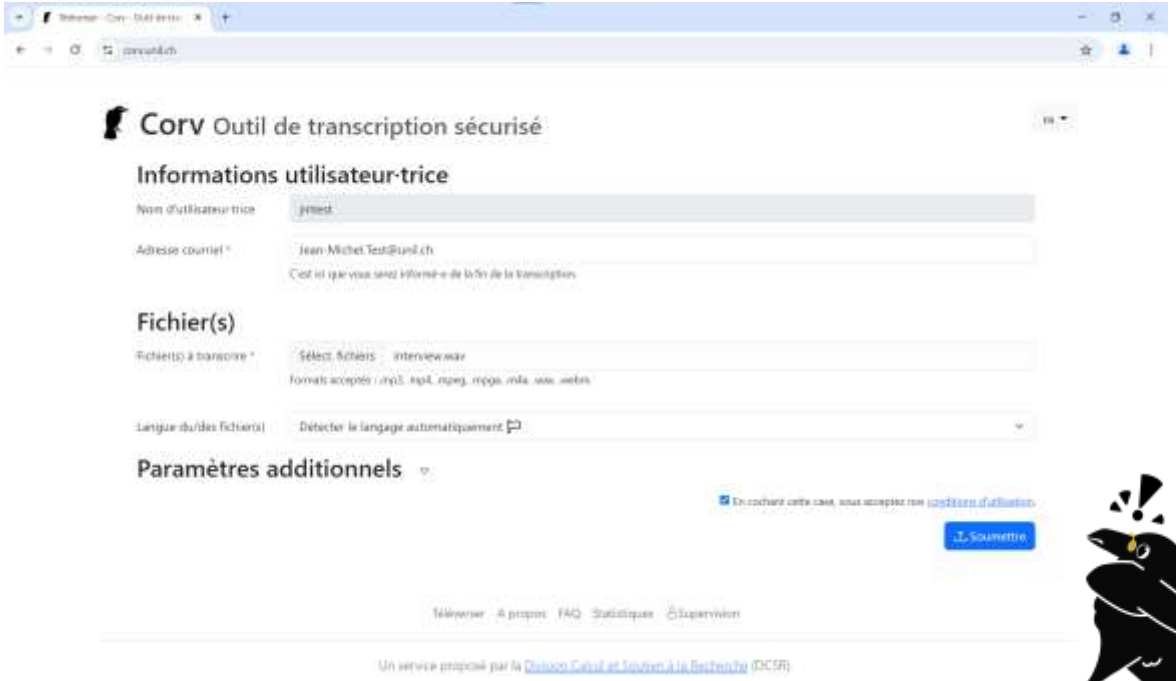
- Corv ("Crow" in romanche).
- Developed by the DCSR, for research.
- Open source, on-premise at UNIL, easy to use.

SOME TECHNICAL INFORMATION

https://corv.unil.ch/about

- DCSR's VM cluster for sensitive personal data.
- Backend: WhisperX with OpenAI transcription model and various Hugging Face alignment models
- Frontend: Python Flask and Bootstrap.
- Access only on UNIL network or via VPN.
- All data is processed locally, no data is transferred over the Internet.
- Data is deleted after processing, some metadata is retained.

~~DEMO EFFECT~~ DEMONSTRATION!



What did the PhD crow say to Prof. Scarecrow?

– “You’re outstanding in your field!”

FIELD FEEDBACK

DATA PROTECTION OFFICE ARC

- From a technical point of view, Corv was already secure – but we wanted to be sure it was compliant with standards.
  - June 2024 – November 2024, six months to get it right.
- Guide du préposé fédéral mon amour.

le DPO de l'Université et la DCSR. L'analyse a conclu que les mesures mises en place lors du développement du logiciel sont suffisantes pour garantir la sécurité des données personnelles, même sensibles, pendant leur traitement dans l'outil.

Objectifs

- L'outil propose un service de transcription (transformation audio/vidéo vers texte) automatique sécurisé à la communauté des chercheuses et chercheurs de l'UNIL.
- L'outil peut être utilisé pour les données normales, pour les données personnelles, pour les données sensibles, au sens de LPD Art. 5 et au sens de LPrD Art. 4, et pour les données personnelles liées à la santé, au sens de LRH Art. 3.
- L'outil propose des mesures de sécurité appropriées, en conformité avec TOM 8.6 "Niveau de sécurité et protection" (page 43), dont les détails sont donnés ci-après.

Mesures physiques

- Le serveur utilisé pour l'outil est dans la ferme de machines virtuelles pour données personnelles sensibles de la DCSR, dont les machines physiques sont situées à l'UNIL.
- Le serveur est chiffré "at-rest" avec la technologie VMWare NKP (Native Key Provider), et chiffré "in-transit" avec SSL/TLS, en conformité avec TOM 8.6 "Niveau de sécurité et protection" (page 43, "Chiffrer").
- Détails techniques du serveur :

OS	Red Hat Enterprise Linux 8.10
CPU	2 vCPU (AMD EPYC 7742 64-Core Processor)
Mémoire	16GB RAM
GPU	16GB VRAM Carte nVidia Tesla T4
Réseau	10 Gbps Ethernet

- L'accès au serveur peut être effectué par les administrateur-trice-s du service et de l'infrastructure.

Accès à l'outil

- L'accès à l'outil est réservé à la communauté UNIL.
- L'outil ne peut être accédé que depuis le réseau local ou le VPN de l'UNIL.
- L'accès à l'outil est protégé, en conformité avec TOM 8.6 "Niveau de sécurité et protection" (page 43, "Protéger") ; pour l'utiliser, il faut se connecter à l'aide de son identifiant Switch.
  - Les informations suivantes sont récoltées automatiquement :

STRESS TEST AND TESTERS

- Before going to production, we asked people to test the solution.
- Various people, from various fields.
  - Some of them never responded… But are still using the tool to this day. I guess it's useful then (;.

num of jobs	total files size	total files length
39	3.1 GB	42 heures 17 minutes 8 secondes

- We also stress-tested the solution ourself.
- Corrupt files, big files, multiple jobs at the same time, …
  - Lots and lots and lots of bugs discovered!

WHAT PEOPLE LOVE

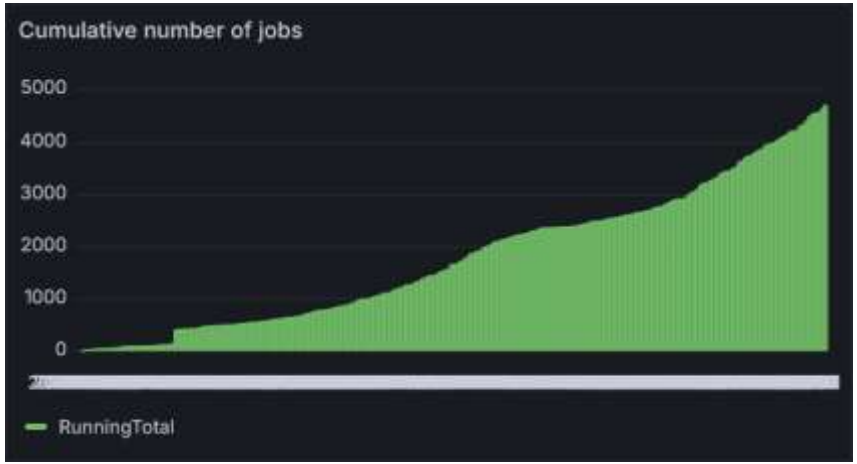
- It’s at UNIL.
  - Local processing, no data on the Internet.
- It’s secure.
  - Both from a technical and a legal point of view (LPD, LPrD, LRH, ORH, …).
- It’s free as in free beer.
  - You do not have to pay a cent.
- It’s free as in freedom.
  - Free and open source.
- It’s easy to use.
  - Two clicks!
- It can be used for both research and non-research data.
  - Useful for meetings.

WHAT PEOPLE WANT

- Most requested features:
  - Integrated editor.
  - Speech disfluency.
  - Various Microsoft Word formats modifiers.
- Most encountered problems:
  - “I cannot connect!” → Forgot to use the VPN.
  - “My ZIP file is empty!” → Problem when Corv does not recognize a language.
- Most asked questions:
  - Whatever people did not read in the FAQ.

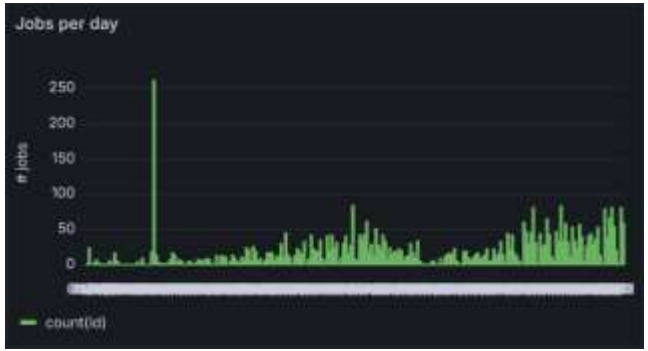
SOME STATISTICS

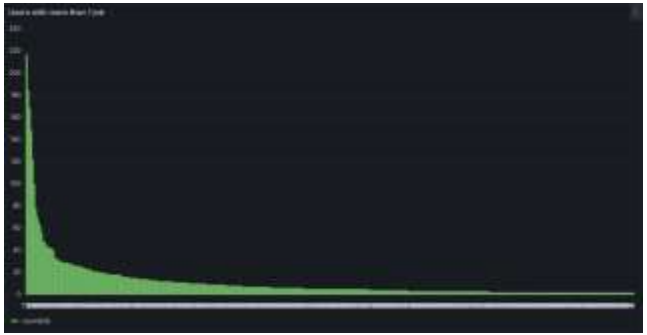
- Since about September 2024:
  - ~1’400 unique users.
  - ~10’000 jobs, ~1.4 TB
  - ~8’000 hours of audio/video.
    - That’s almost fifteen times every Star Trek shows and movies together (source).



MORE STATISTICS

- We lately have on average 50 jobs a day.
- Most users have less than 20 jobs, a couple of them have more than 100.





SCHENANIGANS

- Since the start of the year, three of us have been working on Corv, theoretically at a maximum of 20%.
  - Answering tickets and requests for help, development, representation (*hiii* 👋).
- The software is mostly stable and running well, but we're still adding features.
  - ~20 PRs since January 2025.

What do you call a bunch of crows gathering money?

– A crowd funding

COST COMPARED TO A PAID SOLUTION

COST OF CORV

- 60 minutes of transcription costs around CHF -.53, i.e., around CHF 4’700 per year.
  - Assuming full operation of the current infrastructure.
- Corv's development can be estimated at around CHF 10’000, i.e., around CHF 12’000 per year.
  - Assuming 10 months of development, one to three people, around 20% per week.
- Corv therefore costs roughly CHF 17’000 per year.
  - Warning: this is a *rough* estimate (not taking into account hidden costs, ecological impact, …).

COMPARATIVE TABLE OF OFFERS AVAILABLE ONLINE

Prix par an par personne	Pour 1'000 personnes	Réduction de 50%	Prix effectif par an	Nombre fichiers par an	Taille fichiers (Mo)	Durée fichiers (minutes)	Minutes transcription par an
- CHF	- CHF	- CHF	17 000,00 CHF	∞	1000	∞	525960
- CHF	- CHF	- CHF	/	/	1000	60	720
125,40 CHF	125 400,00 CHF	62 700,00 CHF	/	/	10000	600	72000
191,40 CHF	191 400,00 CHF	95 700,00 CHF	/	/	10000	600	72000
- CHF	- CHF	- CHF	/	36	/	/	/
105,49 CHF	105 494,40 CHF	52 747,20 CHF	/	∞	/	/	/
73,81 CHF	73 814,40 CHF	36 907,20 CHF	/	∞	/	/	/
- CHF	- CHF	- CHF	/	1095	/	30	/
105,60 CHF	105 600,00 CHF	52 800,00 CHF	/	∞	/	1800	/
87,99 CHF	87 991,20 CHF	43 995,60 CHF	/	/	/	/	28800
211,20 CHF	211 200,00 CHF	105 600,00 CHF	/	/	/	/	36000
/	/	/	/	/	/	/	/
211,09 CHF	211 094,40 CHF	105 547,20 CHF	/	/	/	/	3840
422,29 CHF	422 294,40 CHF	211 147,20 CHF	/	/	/	/	9600
950,29 CHF	950 294,40 CHF	475 147,20 CHF	/	/	/	/	24000
547,20 CHF	547 200,00 CHF	273 600,00 CHF	/	84	/	/	/
592,80 CHF	592 800,00 CHF	296 400,00 CHF	/	∞	/	/	/
/	/	/	/	∞	/	/	/
167,20 CHF	167 200,00 CHF	83 600,00 CHF	/	∞	/	/	/
- CHF	- CHF	- CHF	/	3	/	30	3600
87,96 CHF	87 964,80 CHF	43 982,40 CHF	/	120	/	90	14400
211,20 CHF	211 200,00 CHF	105 600,00 CHF	/	∞	/	240	72000
/	/	/	/	∞	/	240	72000
126,72 CHF	126 720,00 CHF	63 360,00 CHF	/	/	/	/	7200
253,44 CHF	253 440,00 CHF	126 720,00 CHF	/	/	/	/	21600
580,80 CHF	580 800,00 CHF	290 400,00 CHF	/	/	/	/	28800
- CHF	- CHF	- CHF	/	/	/	/	720
/	/	/	/	/	/	/	/

UNIL | Université de Lausanne

RESULTS

- On average:
  - Cost of an online solution for about 1’000 people, approximately CHF 220’000.-/year.
    - Even with 50% reduction, around CHF 110,000/year.
  - For only around 30’000 minutes of transcription per year.
- Best offers:
  - Cost of around CHF 74,000/year.
  - In principle, for unlimited transcriptions.
- Corv costs four times less.

What kind of crows always stick together?

– Vel-crow-es

SHARING THE FUN: OPEN SOURCE

WE WOULD LIKE TO SHARE CORV

MORE EYES!

- More users can spot bugs.
- More developers can help develop more features.
- More nerds can help better the product.

> In software development, **Linus's law** is the assertion that "given enough eyeballs, all bugs are shallow". The law was formulated by Eric S. Raymond in his essay and book *The Cathedral and the Bazaar* (1999), and was named in honor of Linus Torvalds.[1][2]

SAVE MONEY!

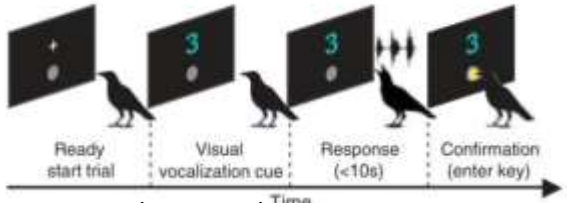
- No need to reinvent the wheel each time.
- Local solutions are less expensive in the mid- and long-term.
- A product that ‘belongs’ to the institutions means no licensing problems later on.
  - E.g. an increase in the cost of a commercial solution on which you are dependent.

COMMUNITY!

- It's important to be able to build on common interests.
- In particular, sharing knowledge, tools and methods will help the entire research community in general.
- Developers, legal people, end users, researchers, staff, … Diversity is our strength.

INTERESTED?

- We have the green light to publish Corv under an AGPL licence.
- The installation is documented and has been tested.
- We want to completely refactor it and provide even better code documentation.
- We can offer you one hour of *pro bono* installation support, or we can perform a paid installation over the course of a day.
  - Call me (;



doi/10.1126/science.adl0984

What do you call a group of really loud crows?

– A caw-caw-phony

LESSONS LEARNED

LISTENING TO PEOPLE IS USEFUL

- IT specialists often propose IT solutions.
  - That's not necessarily what end users want.
- It's a long and complex job listening to all the people involved (tech, law, end users, etc.) but it's worth it for the final product.

KEEP IT SIMPLE, SILLY

- People do not want extra fancy complex features.
- Corv's clean, uncluttered interface has been a big plus point for many people.
- We do not want to add to many extra features, like translation or automatic summary.

SECURITY IS HARD, RESPECTING THE LAW IS HARDER

- We have tried to ensure that the tool and the server are properly secured.
- Above all, we made sure that we followed all possible recommendations.
  - «Je ne suis pas juriste, mais...»
  - We followed the guide issued by the Federal Commissioner, spoke to legal experts, discussed the matter with the DPO, etc.
  - Teamwork!

Please don't hesitate to contact me if you have any questions, suggestions, problems, comments or improvements...

Hugo.Hueber at UNIL.ch.

THANK YOU FOR YOUR ATTENTION!