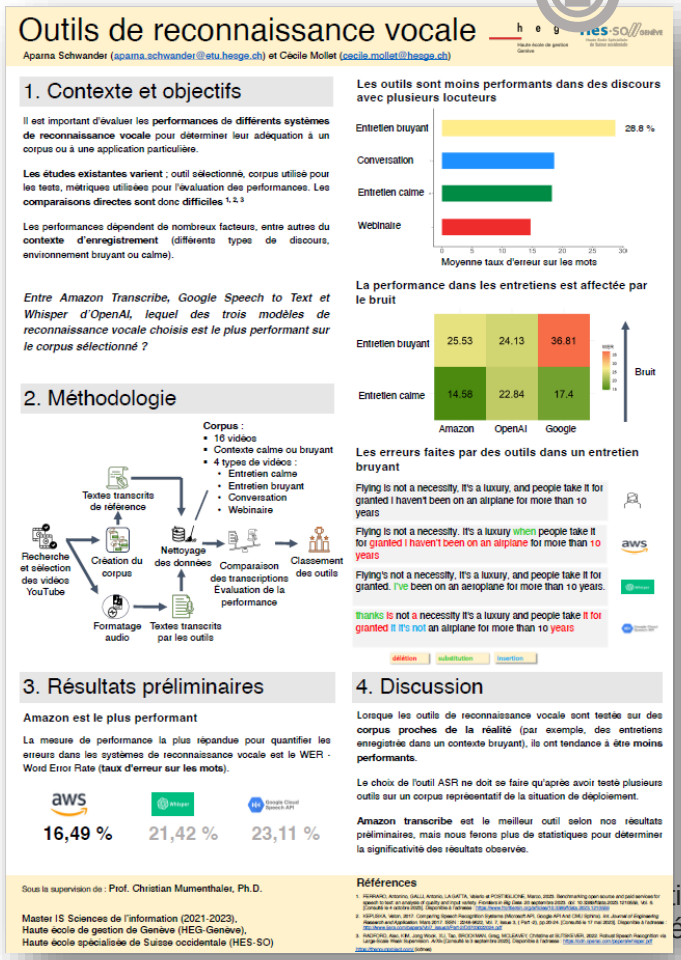
Posters – Top 4

https://www.hesge.ch/heg/formations/masters/sciences-information



Outils de reconnaissance vocale

Aparna Schwander (aparna.schwander@etu.hesge.ch) et Cécile Mollet (cecile.mollet@hesge.ch)

1. Contexte et objectifs

Il est important d'évaluer les **performances de différents systèmes de reconnaissance vocale** pour déterminer leur adéquation à un corpus ou à une application particulière.

Les études existantes varient ; outil sélectionné, corpus utilisé pour les tests, métriques utilisées pour l'évaluation des performances. **Les comparaisons directes sont donc difficiles** [1, 2, 3]

Les performances dépendent de nombreux facteurs, entre autres du **contexte d'enregistrement** (différents types de discours, environnement bruyant ou calme).

Entre Amazon Transcribe, Google Speech to Text et Whisper d'OpenAI, lequel des trois modèles de reconnaissance vocale choisis est le plus performant sur le corpus sélectionné ?

2. Méthodologie

Corpus :
- 16 vidéos
- Contexte calme ou bruyant
- 4 types de vidéos :
  - Entretien calme
  - Entretien bruyant
  - Conversation
  - Webinaire

Recherche et sélection des vidéos YouTube → Création du corpus → Textes transcrits de référence / Nettoyage des données → Formatage audio → Textes transcrits par les outils → Comparaison des transcriptions / Evaluation de la performance → Classement des outils

3. Résultats préliminaires

Amazon est le plus performant

La mesure de performance la plus répandue pour quantifier les erreurs dans les systèmes de reconnaissance vocale est le WER - Word Error Rate (**taux d'erreur sur les mots**).

aws	Whisper	Google Cloud Speech API
16,49 %	21,42 %	23,11 %

Sous la supervision de : Prof. Christian Mumenthaler, Ph.D.

Master IS Sciences de l'information (2021-2023),
Haute école de gestion de Genève (HEG-Genève),
Haute école spécialisée de Suisse occidentale (HES-SO)

Les outils sont moins performants dans des discours avec plusieurs locuteurs

La performance dans les entretiens est affectée par le bruit

	Amazon	OpenAI	Google
Entretien bruyant	25.53	24.13	36.81
Entretien calme	14.58	22.84	17.4

Bruit

Les erreurs faites par des outils dans un entretien bruyant

Flying is not a necessity, It's a luxury, and people take it for granted I haven't been on an airplane for more than 10 years

Flying is not a necessity, It's a luxury when people take it for granted I haven't been on an airplane for more than 10 years

Flying's not a necessity, It's a luxury, and people take it for granted. I've been on an aeroplane for more than 10 years.

thanks is not a necessity It's a luxury and people take it for granted It it's not an airplane for more than 10 years

délétion substitution insertion

4. Discussion

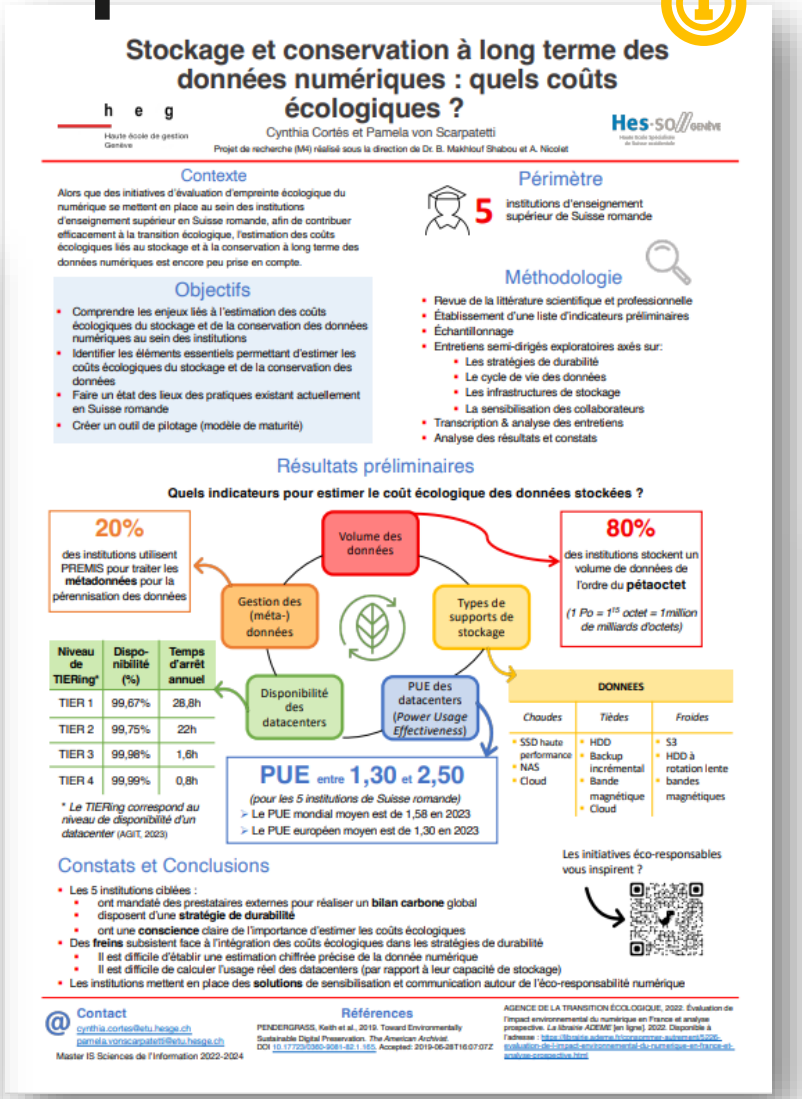
Lorsque les outils de reconnaissance vocale sont testés sur des **corpus proches de la réalité** (par exemple, des entretiens enregistrés dans un contexte bruyant), ils ont tendance à être **moins performants**.

Le choix de l'outil ASR ne doit se faire qu'après avoir testé plusieurs outils sur un corpus représentatif de la situation de déploiement.

Amazon transcribe est le meilleur outil selon nos résultats préliminaires, mais nous ferons plus de statistiques pour déterminer la significativité des résultats observée.

Références

1. FERRARO, Antonio; GALILI, Antonio; LA GATTA, Valerio et POSTILLIONE, Marco. 2023. Benchmarking open source and paid services for speech to text an analysis of quality and input variety. Fondazione Bi Rep. Date 20 septembre 2023. doi: 10.5281/Rama.2023.121.0836. Vol. 6. [Consulté le 4 octobre 2023]. Disponible à l'adresse : https://www.fondazionebirep.com/article/2023/121/0836
2. KSHIRSAGAR, Vaibhav. 2017. Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). en. Journal of Engineering Research and Application www.ijera.com ISSN : 2248-9622, vol. 7, iss. 3 (part -5), p. 20-24. [Consulté le 17 mai 2023]. Disponible à l'adresse : http://www.ijera.com/papers/vol7_issue3/Part-5/D0703052024.pdf
3. RADFORD, Alec; KIM, Jong Wook; XU, Tao; BROCKMAN, Greg; MCLEAVEY, Christine et SUTSKEVER, Ilya. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. Arxiv [Consulté le 5 septembre 2023]. Disponible à l'adresse : http://arxiv.org/abs/2212.04356 [cs, eess]. https://theeg.pocket.pub (Infinite)



Stockage et conservation à long terme des données numériques : quels coûts écologiques ?

Cynthia Cortés et Pamela von Scarpatetti
Projet de recherche (M4) réalisé sous la direction de Dr. B. Makhlouf Shabou et A. Nicolet

Contexte

Alors que des initiatives d'évaluation d'empreinte écologique du numérique se mettent en place au sein des institutions d'enseignement supérieur en Suisse romande, afin de contribuer efficacement à la transition écologique, l'estimation des coûts écologiques liés au stockage et à la conservation à long terme des données numériques est encore peu prise en compte.

Périmètre

5 institutions d'enseignement supérieur de Suisse romande

Objectifs

- Comprendre les enjeux liés à l'estimation des coûts écologiques du stockage et de la conservation des données numériques au sein des institutions
- Identifier les éléments essentiels permettant d'estimer les coûts écologiques du stockage et de la conservation des données
- Faire un état des lieux des pratiques existant actuellement en Suisse romande
- Créer un outil de pilotage (modèle de maturité)

Méthodologie

- Revue de la littérature scientifique et professionnelle
- Établissement d'une liste d'indicateurs préliminaires
- Échantillonnage
- Entretiens semi-dirigés exploratoires axés sur:
  - Les stratégies de durabilité
  - Le cycle de vie des données
  - Les infrastructures de stockage
  - La sensibilisation des collaborateurs
- Transcription & analyse des entretiens
- Analyse des résultats et constats

Résultats préliminaires

Quels indicateurs pour estimer le coût écologique des données stockées ?

20% des institutions utilisent PREMIS pour traiter les **métadonnées** pour la pérennisation des données

80% des institutions stockent un volume de données de l'ordre du **pétaoctet** (*1 Po = 1^{15} octet = 1 million de milliards d'octets*)

Volume des données — Gestion des (méta-) données — Types de supports de stockage — Disponibilité des datacenters — PUE des datacenters (*Power Usage Effectiveness*)

Niveau de TIERing*	Dispo-nibilité (%)	Temps d'arrêt annuel
TIER 1	99,67%	28,8h
TIER 2	99,75%	22h
TIER 3	99,98%	1,6h
TIER 4	99,99%	0,8h

** Le TIERing correspond au niveau de disponibilité d'un datacenter* (AGIT, 2023)

PUE entre 1,30 et 2,50
(pour les 5 institutions de Suisse romande)
- Le PUE mondial moyen est de 1,58 en 2023
- Le PUE européen moyen est de 1,30 en 2023

DONNEES

Chaudes	Tièdes	Froides
SSD haute performance, NAS, Cloud	HDD, Backup incrémental, Bande magnétique, Cloud	S3, HDD à rotation lente, bandes magnétiques

Constats et Conclusions

- Les 5 institutions ciblées :
  - ont mandaté des prestataires externes pour réaliser un **bilan carbone** global
  - disposent d'une **stratégie de durabilité**
  - ont une **conscience** claire de l'importance d'estimer les coûts écologiques
- Des **freins** subsistent face à l'intégration des coûts écologiques dans les stratégies de durabilité
  - Il est difficile d'établir une estimation chiffrée précise de la donnée numérique
  - Il est difficile de calculer l'usage réel des datacenters (par rapport à leur capacité de stockage)
- Les institutions mettent en place des **solutions** de sensibilisation et communication autour de l'éco-responsabilité numérique

Les initiatives éco-responsables vous inspirent ?

Contact
cynthia.cortes@etu.hesge.ch
pamela.vonscarpatetti@etu.hesge.ch
Master IS Sciences de l'Information 2022-2024

Références

PENDERGRASS, Keith et al., 2019. Toward Environmentally Sustainable Digital Preservation. The American Archivist. DOI 10.17723/0360-9081-82.1.165. Accepted: 2019-06-28T16:07:07Z

AGENCE DE LA TRANSITION ÉCOLOGIQUE, 2022. Évaluation de l'impact environnemental du numérique en France et analyse prospective. La librairie ADEME [en ligne]. 2022. Disponible à l'adresse : https://librairie.ademe.fr/cadre-vie-batiment/5206-evaluation-de-l-impact-environnemental-du-numerique-en-france-et-analyse-prospective.html

Indexer les sexualités

Alexandra Andrade et Laura Lorenzo

Quels termes utiliseriez-vous pour décrire cet ouvrage ?
a. Pratique sexuelle, Société, Milieu du 19e
b. Perversion sexuelle, Masturbation, Érotisme
c. Sexualité, Amour, Manuel historique
d. Sexualité, Pin-up, Années 50-60

1. Introduction

L'indexation dans le domaine des sexualités n'est pas une tâche facile (G3 2022). Les termes disponibles dans les thésaurus, comme celui de la BNF, manquent parfois de précision, ou ne sont pas présents. Le Centre Maurice Chalumeau en Sciences des Sexualités (CMCSS) nous a mandaté afin de les aider à réfléchir à des solutions leur permettant d'indexer au mieux les documents du fond Michel Froidevaux.

2. Méthodologie

Pour mieux comprendre les méthodes d'indexation, nous avons :
- Récolté le nombre de documents indexés avec des termes d'indexation sur un échantillon de 90 (cf. nuage de mots ci-dessus).
- Comparé les notices d'un échantillon aléatoire de 60 documents indexés au CMCSS avec celles d'autres institutions francophones.

3. Résultats

Le graphique ci-dessous nous montre que le CMCSS possède une collection unique dans le monde francophone et surtout au sein de l'UNIGE.

Nombre de monographies en commun avec le CMCSS
Sur un échantillon aléatoire de 60 monographies

Lors de nos comparaisons, cet ouvrage nous a interpelé, car seul le CMCSS a ajouté des termes d'indexations à leur notice. Le champ des termes d'indexation est vide chez les autres institutions, ce qui pourrait traduire un manque de connaissance ou un manque d'intérêt.

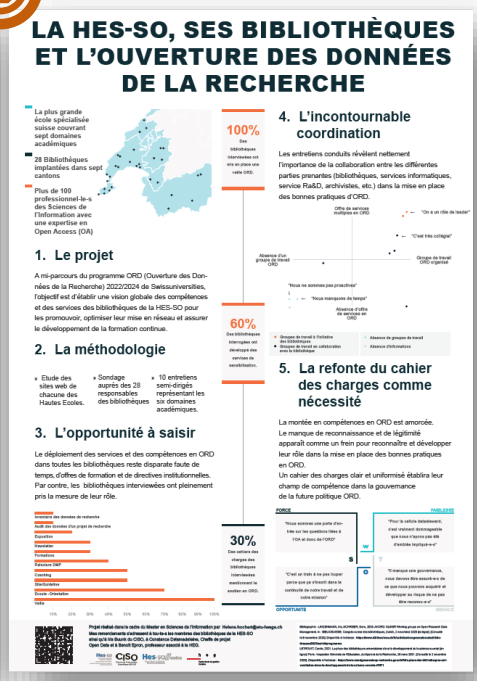
Ici, malgré le fait que le terme «asexuelle» est utilisé dans le titre et qu'il parle notamment d'asexualité, aucun de ces deux termes n'apparaissent dans les champs d'indexation des institutions*. Au contraire, on y retrouve les termes «Abstinence sexuelle» et «Chasteté».

4. Conclusion

Indexer le sexuel n'est pas simple. Les termes ne sont pas assez précis, sont manquants ou peuvent même être maladroits ou problématiques. Les représentations des sexualités évoluent rapidement depuis plusieurs décennies, il faudrait donc que les thésaurus officiels comme celui de la BNF puissent suivre ces évolutions pour pouvoir correctement indexer ce type de collection.

Remerciements
Nous remercions le CMCSS pour tout le travail de réflexion fourni et pour leur chaleureuse collaboration.
Nous remercions également Benoît Epron pour son suivi et ses encouragements.

Références
G3, 2022. Penser la place des sciences des sexualités en bibliothèque universitaire. Le G3 ou le Représentant [en ligne]. Disponible à l'adresse : https://g3lerepresentant.wordpress.com/2022/06/29/penser-la-place-des-sciences-des-sexualites-en-bibliotheque-universitaire/ [Consulté le 26.11.2023].
LORULOT, André. 1967. Tricheries et mensonges de l'amour. Verdalle : Éditions de l'Idée Libre.
PRADA, Juan Manuel de. 2000. Corfou. Madrid : Valdemar.
TONNAC, Jean-Philippe de. 2006. La Révolution asexuelle. Paris : Albin Michel.



LA HES-SO, SES BIBLIOTHÈQUES ET L'OUVERTURE DES DONNÉES DE LA RECHERCHE

La plus grande école spécialisée suisse couvrant sept domaines académiques

28 Bibliothèques implantées dans sept cantons

Plus de 100 professionnel-le-s des Sciences de l'Information avec une expertise en Open Access (OA)

1. Le projet

A mi-parcours du programme ORD (Ouverture des Données de la Recherche) 2022/2024 de Swissuniversities, l'objectif est d'établir une vision globale des compétences et des services des bibliothèques de la HES-SO pour les promouvoir, optimiser leur mise en réseau et assurer le développement de la formation continue.

2. La méthodologie

- Étude des sites web de chacune des Hautes Ecoles.
- Sondage auprès des 28 responsables des bibliothèques
- 10 entretiens semi-dirigés représentant les six domaines académiques.

3. L'opportunité à saisir

Le déploiement des services et des compétences en ORD dans toutes les bibliothèques reste disparate faute de temps, d'offres de formation et de directives institutionnelles. Par contre, les bibliothèques interviewées ont pleinement pris la mesure de leur rôle.

4. L'incontournable coordination

Les entretiens conduits révèlent nettement l'importance de la collaboration entre les différentes parties prenantes (bibliothèques, services informatiques, service R&D, archivistes, etc.) dans la mise en place des bonnes pratiques d'ORD.

5. La refonte du cahier des charges comme nécessité

La montée en compétences en ORD est amorcée. Le manque de reconnaissance et de légitimité apparaît comme un frein pour reconnaître et développer leur rôle dans la mise en place des bonnes pratiques en ORD.
Un cahier des charges clair et uniformisé établira leur champ de compétence dans la gouvernance de la future politique ORD.

Bibliography : LAUDERDALE, Lily; ALLAMAN, Sara. 2019. AHORD Working group on Open Research Data. Bibliography. 4. BIBLIOGRAPHY. Congrès suisse des bibliothèques. [Zurich]. 2 novembre 2023 [en ligne]. [Consulté le 4 novembre 2023]. Disponible à l'adresse : https://www.bibliotheksverband.ch/fileadmin/Bibliothekskongress/Bibliothekskongress2019/Bibliography.pdf
Bibliography : Bibliothèques suisses. Bibliothèques suisses. [s.d.]. [Consulté le 3 novembre 2023]. Disponible à l'adresse : https://www.bibliotheques.ch/fr/bibliotheques-suisses-liste-de-bibliotheques-de-toutes-les-bibliotheques-suisses-classees-par-canton.html
Groupe de travail ORD. Groupe de travail ORD. [s.d.]. [Consulté le 3 novembre 2023]. Disponible à l'adresse : https://www.bibliotheques.ch/fr/groupe-de-travail-ord.html
Open Data et Research Epron, professeur associé à la HES. [s.d.]. [Consulté le 3 novembre 2023]. Disponible à l'adresse : https://www.hes-so.ch/fr/personnes/epron-benoit-2.html
SWISSUNIVERSITIES. 2022. Swissuniversities. Swissuniversities. Swissuniversities. [s.d.]. [Consulté le 3 novembre 2023]. Disponible à l'adresse : https://www.swissuniversities.ch/fr/organisation/organisation-suisse-des-hautes-ecoles/organisation-de-la-conference-des-recteurs-des-universites-suisses-crus
TOSI, Claudio Vittorio. Wikipedia. Wikipedia. [s.d.]. [Consulté le 3 novembre 2023]. Disponible à l'adresse : https://fr.wikipedia.org/wiki/Claudio_Vittorio_Tosi

Projet réalisé dans le cadre du Master en Sciences de l'information par Helena Bochsler@etu.hesge.ch
Mes remerciements s'adressent à tou-te-s les membres des bibliothèques de la HES-SO
ainsi qu'à Ms Bauer du CISO, à Constance Zabaleguicha, Chefe de projet
Open Data et à Benoît Epron, professeur associé à la HES.

...tionale sur
...ériques