CHILLI – Collaborative Human-centered Machine Learning for Archival Practice: Enhancing Accessibility, Transmission, and Curation

Direction
This Ph.D. project will take place at ISIR (Institute for Intelligent Systems and Robotics), Sorbonne Université. It will be directed by Baptiste Caramiaux, CNRS researcher at Sorbonne Université, a member of ISIR and HCI Sorbonne group. The project will be co-directed by Sarah Fdili Alaoui, associate professor at Université Paris-Saclay, member of the Ex-Situ Inria team, and the LISN lab.

Context
Archiving involves collecting, preserving, and managing historical records and documents. Archiving is a collaborative practice of data collection, curation, and transmission (Ernst 2012). Archives embody collective knowledge about one or several communities of practice. With the increasing volume of digital records and the need for efficient management, there is a growing interest in machine learning and artificial intelligence techniques for archiving, even if their application remains marginal and difficult (European Commission, 2022).

Machine learning has several advantages. First, it can assist the archival practice by automating tasks such as document classification, metadata extraction, and data analysis. Second, machine learning can help retrieve implicitly biased narratives, which often take the share of dominant positions, and may help mitigate them (Jo et al., 2020). Third, it can provide practitioners with a tool to collect and manage their own archival collections (Birhane et al., 2022).

In this context, we would like to investigate the potential of the collaborative use of machine learning to enhance practitioners’ collection, curation and access to these records. We mean by collaborative use, the integration of diverse human expertise and machine learning technologies to record, curate, navigate, and transmit archival material. We are particularly interested in applying collaborative human-centered machine learning in the context of archival of intangible cultural heritage such as dance and music, particularly popular forms where dance and music cannot be separated. These dance practices are often transmitted without formalisation, within a community of practice, and thus less represented by archives, sometimes considered official.

Our main research questions are: What is the potential of machine learning in dance music archival practice? How can it be used in collaboration with human expertise, especially dance practitioners, to facilitate the collection, curation, access and transmission of cultural content?

Objectives
This Ph.D. contributes to HCI and has three objectives:

1. To review the literature on the application of human-centered machine learning in archival practice;
2. To design collaborative human-centered machine learning systems focused on the archival of dance-music material, in particular considering generation strategy;
3. To critically assess the collaborative human-centered machine learning systems concerning what is kept and what is discarded.

Positioning and preliminary work
This research project is positioned in a research on archiving and therefore the problems underlying classification and which execute this classification (Bower, Star, 1999). The use of machine learning in this context brings forward the challenge of data curation and model authoring, which has been explored in interactive machine learning and teaching (Sanchez et al., 2021). However, work on stakeholder collaboration in this context remains scarce. Some work has been done on crowdsourced data curation (for instance in the PI@ntNet project (Joly et al., 2016)), but
model creation is generally not allowed. Furthermore, this technology needs to be deployed and made accessible to non-experts, which poses challenges in terms of human-machine interface and explainable machine learning. The Ph.D. proposal relies on preliminary work conducted through two internships and the collaboration with a Paris-based dance collective. In this project, we have recorded data sets made of motion capture data of Dance Hall and Voguing dance styles, together with the music accompanying each performance. We have a sustained collaboration with dancers from this community whose dance styles have never been archived before. We aim to collaborate with these practitioners to train our models and assess them critically with them.

**Methodology**

The research will adopt a mixed-method approach, combining qualitative and quantitative data collection and analysis methods. The study will begin with a comprehensive literature review of relevant articles on the application of human-centered machine learning in archival practice. In addition, interviews will be conducted with practitioners investigating how they archive their work in order to gain insights into what they decide to document and what they discard.

The study will then proceed to the development of a collaborative machine learning system for the archival of dance pieces. The system will use machine learning algorithms for data compression and representation. The system will leverage web technology and the Marcelle toolkit (François et al., 2021). As archiving is a collaborative process, training machine learning algorithms on archived data will be performed collaboratively through data curation, data structuration, machine teaching, and collaborative model quality assessment.

The prototyped system will be assessed with professional artists comparing how the system views and represents cultural data and how the practitioners do so themselves. We will conduct qualitative interview studies to assess the relevance of the system-based representation, the inherent narratives in the data, and the generated narratives through the machine learning pipeline.

**Expected outcomes**

This study will contribute to the understanding of the potential of collaborative user-centered machine learning in archival practice and its application to intangible cultural heritage. The study will also provide a critical reflection on the biases of machine learning systems regarding archival practice, examining what the system values and disregards.

**Partnership**

The project will take place at ISIR. It will be co-directed 50-50 between Baptiste Caramiaux (expert in Interactive Machine Learning) and Sarah Fdili Alaoui (expert in dance and technology research). Both directors are HCI researchers.

**References**

- European Commission (2022), Opportunities and challenges of artificial intelligence technologies for the cultural and creative sectors. *Publications Office of the European Union*
- Joly, A., et al. (2016). A look inside the Pi@ntNet experience: The good, the bias and the hope. *Multimedia Systems, 22*