Unlocking Value: AI for Enhanced Metadata in STI
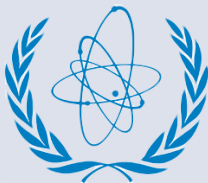
Brian Bales
INIS, International Atomic Energy Agency (IAEA)

אוניברסיטת חיפה
UNIVERSITY OF HAIFA
جامعة حيفا
Amit Niv
Haifa University Library

UC3
UC Curation Center
Dione Mentis
John Chodaki
COMET Project
California Digital Library

OU
The
UNIVERSITY
of
OKLAHOMA
Dr. Manika Lamba
University of Oklahoma
School of Library &
Information Studies

"The Metadata Uncertainty Principle: Extracting Schrödinger's Cat with AI"

Over the past several years, the IAEA's International Nuclear Information System (INIS) has increasingly relied on artificial intelligence to extract, summarize, and index scientific literature. AI now supports the harvesting of bibliographic metadata from source documents, as well as quality assurance and subject indexing, improving efficiency, speed, and consistency. Recently, INIS was provided with 115 historical works of the Austrian Physicist Erwin Schrödinger, for inclusion in its repository. This offered an opportunity to test the limits of AI metadata extraction. Using a state-of-the-art model, a very large, restrictive prompt, and human-in-the-loop review, the approach sought to eliminate uncertainty. However, the experiment revealed a paradox. Instead of eliminating uncertainty, the process exposed its structural persistence. This presentation proposes a "Metadata Uncertainty Principle" to help understand the limits of AI extracted metadata and for designing systems to govern it.

"Voice to Text Metadata Extraction for Periodicals"

In this presentation, Amit Niv will outline a process to extract metadata for cataloging periodicals using Gemini's voice to text capabilities.
This process was devised as a student project and later refined by library staff.

"Transforming Metadata Quality Through AI and Community Collaboration: COMET's arXiv Parsing Pilot"

COMET is a collaborative initiative advancing PID metadata quality through community-driven enrichment. Launched in 2024, COMET is now in a pilot phase, demonstrating how the COMET Model works through targeted enrichment projects. This session focuses on the project to 'improve affiliations parsing of preprints'. Using the arXiv corpus as the test case, COMET developed a fine-tuned large language model (LLM) that significantly outperforms existing tools in extracting author-affiliation metadata from preprints. The project illustrates the potential of the COMET Model as a whole by addressing gaps in institutional attribution and has the potential to increase research discoverability across open infrastructures such as DataCite and OpenAlex.

"Metadata Enrichment of Low Resource Data using LLM Agent"

Missing data fields and lack of metadata standards for low-resource digital collections limit both discoverability and accessibility in large-scale repositories. In this talk, I will present a case study using dissertation data from the HathiTrust Digital Library to demonstrate how incomplete metadata records can be semantically enriched using an LLM-based agent. This case study will demonstrate how metadata enrichment using LLM-based agents is particularly valuable for large-scale digital repositories, as it creates additional metadata access points that may not have been originally anticipated, especially for low-resource content types.