# Developing Processing Algorithms for Automated Error Identification and Repair of Multilingual Records Incorporating Compound Transliteration and Diacritic Restoration

**JENNIFER PROCTOR**

## Abstract

In this digital age, processing demands for digital archival content have grown exponentially with the rate of creation enabled by the societal digital revolution. 30 years ago, the Bush Presidential Library ingested only 20 GB of digital records at the end of his administration. Today, Obama's contains more than 10,000 times that, enough, if printed, to reach further than the moon. This is further compounded by the challenges placed on archivists needing to not only initially process, but also maintain accessibility to digital records despite the serious challenges of digital fragility. Archives are facing a need for speed to handle scale without compromising accuracy, increased diversity as collections from more people and places achieve greater visibility, and a recognition of the value of connectivity for findability. In aid of these values, OCLC, is exploring the potential of Machine Learning and Artificial Intelligence.

OCLC's Cyrillic Language data is a massive number of MARC records mixing Cyrillic Russian, Ukrainian and Romanian with English, however the records contain significant inconsistencies and gaps that make it difficult to search, connect, and assess accurately. At the scale of work where 40 million records formed 'a modest subset', manual identification of errors is impossible and tasking enough record managers fluent in all the languages of the world would be wholly impractical.

We present a case study where several methods were employed in order to address these needs, including complex computational linguistics, hashing, nominalization, and other preprocessing, as well as the machine learning modeling involved in creating a big data model for automated triage of bulk ingests in multilingual library catalog records involving compound transliteration with the additional need for automated diacritic restoration.

## About the author:

***Jennifer Proctor*** is a Faculty Research Specialist at the Applied Research Lab for Intelligence and Security, a University Affiliated Research Center for the Department of Defense, and a PhD Candidate in Information Science at University of Maryland College Park specializing in digital curation. Her research focuses on computational approaches to managing historical collections, with projects relating to semi-automated conversion of metadata to linked data, machine learning and natural language processing for curation of historical photographs and full text records, and R&D of tools for improving the declassification process of the US government.