

Creating a Cross-Lab Curation Portal Featuring ML/AI Metadata Extraction

Juliane Schneider, David ML Brown, Amanda Casella, Eric Stephan
Pacific Northwest National Laboratory

The Challenge

There are tens of thousands of unclassified DOE historical documents regarding plutonium processing inaccessible to the nonproliferation community. To make these accessible they need to be:

- Digitized
- Annotated with rich metadata
- Stored in a system with appropriate access

Given the limited time available for nuclear science SMEs for curation, metadata extraction needed to be automated

The Workflow

Digitized documents loaded to SharePoint = *Accessible to all four laboratories in the project*

Documents converted to JSON-LD and stored in CosmosDB = *Structured documents allow for better AI/ML performance*

Metadata is extracted using AI/ML and made available to curators in SharePoint = *reduced curator workload*

Curators review and edit metadata using SharePoint = *reviewing extracted metadata saves time and effort*

Documents and metadata are pushed to a repository = *search and discovery by the nonproliferation community*

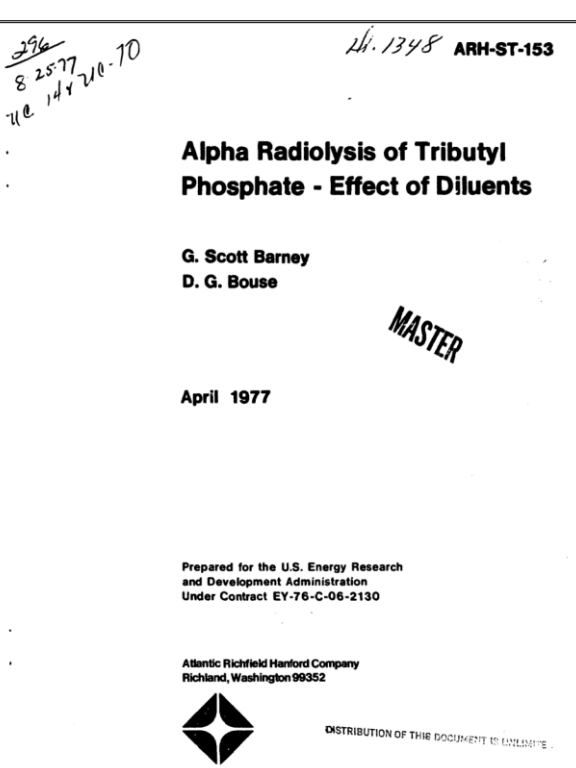
Status/Next Steps

Completed:

- 100 documents digitized
- Nuclear Science taxonomy created
- Basic metadata automatically extracted

Next steps

- Curate documents
- Make documents available
- Train generative AI to extract taxonomy-based metadata
- Use generative AI to ask questions on the corpus of extraction documents for education and discovery



Title: Alpha Radiolysis of Tributyl Phosphate – Effect of Diluents
Authors: G. Scott Barney, D.G. Bouse
Publication Date: April 1977
Report No: ARH-ST-153
Publisher: Atlantic Richfield Hanford Company

Chemicals: tri-n-butyl phosphate (TBP)
Hazards: Uncontrolled Reaction
Equipment: Scrub Column
Unit Processing: Solvent cleanup, Solvent Extraction
Facility: 222-S (Hanford analytical facility)

LLM extracts
Taxonomy concepts
from text

Nuclear Scientist-created taxonomy

