Enhancing Research Data Usability through TurboCurator: An AI-Driven Tool for Creating FAIR Metadata

www.icpsr.umich.edu

Part of the Institute for Social Research at the University of Michigan.

By Jeannette Jackson, Aalap Doshi, Maggie Levenstein, Jenny Li

Background & Context

ICPSR, one of the world's largest social science data archives, hosts over 20,000 social science datasets from global researchers and agencies. In 2022, ICPSR received the U.S. National Science Foundation grant #1946932 ("The Research Data Ecosystem") to modernize its platform. TurboCurator is one of the resulting tools.

How does TurboCurator improve data?

By combining ICPSR's precision and expertise, AI's creativity and power, worldwide metadata standards, and human choice.

Objectives

- Simplify metadata creation for depositors and curators
- Advance FAIR principles with AI-suggested Titles, Descriptions, and Keywords
- Provide transparency and training in metadata standards
- Keep human oversight central to final decisions

Insights

- Interviews identified Title, Description, and Keywords as most valuable metadata components
- Current workflows are slowed by iterative curator-depositor exchanges
- Users prefer transparent, supportive tools over "black-box" automation

Solution: TurboCurator

TurboCurator, co-developed with Harvard's Dataverse, is an AI-assisted metadata tool that:

- Uses Azure OpenAI models
- Ingests depositor research materials
- Applies ICPSR's thesaurus and rule-based checks
- Suggests Titles, Descriptions, and Keywords for depositor review
- Explains metadata rules behind each suggestion

System Architecture and Workflow

Input: Depositor materials (abstracts, methods sections, press releases, etc.)

Processing: Text extraction --> AI: LLM suggestions --> Rule-based alignment with metadata standards

Output & Human-in-the-Loop: Interactive interface shows suggestions with explanations; depositors accept, edit, or reject before final submission to repository

Key components include a domain thesaurus for precision, adjustable AI parameters for balancing creativity and precision, transparent rule-based training that serve as micro-training for depositors, and a feedback loop for refinement.

Anticipated Impact

- Reduced curation time and costs
- Easier depositor experience
- Higher metadata quality and discoverability
- Scalable within ICPSR's Research Data Ecosystem and beyond
- Continuous improvement through depositor feedback

Next steps & Future Work

- Full integration into ICPSR's deposit system
- Expansion to more metadata fields
- Continual refinement through user feedback and error tracking

Limitations and Risks

Limitations and risks include occasional generic or missing outputs, depositor over-reliance on AI, the need for continuous tuning and curator oversight, and potential disruptions from changes in AI models.

Conclusion

TurboCurator shows how AI can support metadata creation while preserving transparency and oversight, improving workflows and metadata quality for a sustainable, FAIR-compliant research data ecosystem.

Lessons Learned

- Depositors value guidance and explanation of suggestions
- AI plus domain expertise produces the best results
- Human-in-the-loop design builds trust and sustainability

The Research Data Ecosystem: a National Resource for Reproducible, Robust, and Transparent Social Science Research in the 21st Century, is funded by the U.S. National Science Foundation as a Mid-scale RI-2 Infrastructure Award (#1946932).

U.S. National Science Foundation

INSTITUTE FOR SOCIAL RESEARCH UNIVERSITY OF MICHIGAN