Summer Internships at AUC Libraries: Transforming Bibliographic Data and Digital Access

Mark Muehlhaeusler, AUC Libraries



Program Overview

- Collaboration with Department of Computer Engineering
- Three rounds of internships with growing participation
- 52 applications in 2024 (increased from previous year)
- Team-based approach following Industrial Training Guidelines
- 8-week duration with minimal supervision

Internal supervision team:

Irina Schmid, head of Digitization Center

Mohamed Khalil, head of Library IT

Usama Salama, head of Metadata

Mutual Benefits
For the Library

- Technical expertise for specialized projects
- Innovation in bibliographic data management
- Enhanced digital collection accessibility
- Cost-effective solutions



Mutual Benefits
For students

Exposure to specialized library data standards:

- MARC format and MARC-8 encoding
- ContentDM metadata structures
- ISBN, LCCN, and OCLC number systems

Domain-specific technical challenges:

- Arabic OCR and text processing
- Bibliographic data normalization
- Metadata extraction from historical materials
- Entity recognition in multilingual contexts

Focus Project 1: Electronic Availability Analysis

Problem: Identifying print materials with electronic surrogates
Approach:

- Processing bibliographic data in MARC format
- Analyzing unique identifiers (ISBN, LCCN, OCLC numbers)
- Overlap analysis with external databases like HathiTrust

Outcomes: Data-driven collection management decisions

Focus Project 2: Facial Recognition for Historical Images

Problem: Identifying and tagging faces in large historical image collections
Approach:

- Dual-track workflow: manual tagging interface and automated batch processing
- Python libraries for facial recognition

Outcomes: Enhanced searchability of historical photograph collections
GitHub: https://github.com/omarelwaliely/LibraryFacialRecognition

Focus Project 3: Enhanced Metadata for Historical Newspapers

Problem: Limited searchability of historical newspaper collections
Approach:

- Integration with ContentDM metadata structure
- Florence AI for entity recognition
- Arabic language processing

Outcomes: Detailed bilingual metadata records
Project Link: https://colab.research.google.com/drive/16QzqJ2LrTieiec9mqTG-mfudbl3YiYZT

Uncovering AI Biases

1. Model limitations with Arabic script and regional content
2. Hallucinations associated with Arabic/Islamic content
3. Entity recognition challenges with Arabic names and cultural references
4. Need for culturally-adapted and region-specific models

CFO

Program Evolution and Future Directions

- Increased selectivity in candidate selection
- Refined project scoping and supervision
- Focus on deliverables
- Pivot to AI applications

Conclusion

- Successful technical projects benefiting the library
- Valuable learning experiences for students
- Insights on AI model biases
- Model for academic library-computer science collaborations

Thank you. Questions?

mark.muehlhaeusler@aucegypt.edu

