**MEETING MINUTES**

**Professional Computing Project Team 03**

**MEETING OVERVIEW**

**Date:** August 27, 2025  
**Duration:** Approximately 60 minutes  
**Purpose:** Review progress on automated document search and retrieval system for census data and statistical publications, address technical challenges, and establish next phase implementation strategy

**Attendees:**

* Dr. Chris Parsons
* Rania Khan
* Hazel Wang
* James Felstead
* Chunyu Zheng
* Peter Fang

**KEY DISCUSSION POINTS**

**Technical System Development Progress**

The team has made significant progress on developing an automated system to find and retrieve census documents and statistical publications from various national websites. **Major breakthrough achieved** by switching from OpenAI's deep research model to **using Gemini with Google search integration**, which proved substantially more effective for web crawling and document discovery.

**System achievements:**

* Successful document retrieval from high-quality sources (Canadian and Australian census websites)
* Discovery of large compressed documents matching project requirements precisely
* Improved search accuracy through AI model optimization

**Document Search Methodology and Challenges**

**Multi-language search strategy developed(potential):**

1. **English searches** (documents often referenced internationally in English)
2. **French searches** (extensive French library collections)
3. **Native language searches** using translation tools

**Technical obstacles identified:**

* **Problematic websites** flooding search results with irrelevant data (World Bank datasets, Statistica)
* **Anti-robot mechanisms** on news sites and some statistical sites
* **Metadata extraction issues** for incorrectly uploaded historical documents
* **Special character handling problems** in non-English searches

**Data Organization and Quality Assurance**

**Comprehensive data organization framework:**

* **Standardized naming conventions** using underscores between metadata elements
* **Simple folder structures** organized by country and year
* **SharePoint integration** with full access rights for shared storage
* **Quality assurance measures** emphasizing robust logical sequences and filtering mechanisms

**Search Optimization and Filtering Strategies**

**Technical improvements implemented:**

* **Exact search parameters** using Google's functionality
* **File size filtering** to distinguish target documents (50MB+) from irrelevant small files (40KB)
* **Website exclusions** for problematic sites (Wikipedia, World Bank, UN organizations)
* **Focus shift** from broad internet searches to 200-500 specific library websites

**Strategic approach:** Targeting national libraries and institutional repositories for improved accuracy and reliability.

**Implementation Testing Strategy**

**Poland selected as primary test case:**

* **Rationale:** Challenging language characteristics with simpler document structure compared to India
* **Strategic advantage:** More focused testing while developing robust filtering mechanisms
* **Planned outcome:** Comprehensive metadata examples to refine automated system accuracy

**DECISIONS MADE**

1. **Website Filtering Implementation:** Hard-coded filters will exclude problematic websites like Statistica and World Bank datasets to improve search result quality
2. **Testing Focus Shift:** Moving from India documentation to Poland for initial testing due to simpler table structure and language processing advantages
3. **Metadata Collection Strategy:** Implementing "greedy" approach to metadata collection for comprehensive data capture
4. **Optional Province Field:** Implement flexible province field handling for documents with varying geographic scope
5. **Search Scope Limitation:** Focus on 200-500 specific library websites rather than general internet searches
6. **AI Integration:** Continue using Gemini AI (Pro version trial) for intelligent decision-making beyond basic programmed logic

**ACTION ITEMS**

|  |  |  |
| --- | --- | --- |
| **Task** | **Responsible Party** | **Deadline** |
| Provide comprehensive list of websites to exclude from searches | Dr. Chris Parsons | EOW |
| Send document coding/naming convention reference sheet | Dr. Chris Parsons | EOW |
| Email specification of required metadata fields | Dr. Chris Parsons | EOW |
| Provide minimum 10 Polish documents for testing | Dr. Chris Parsons | EOW |
| Continue testing and refining search algorithms with Gemini | Peter/James | Ongoing |
| Implement file size and page count filtering capabilities | James | Ongoing |
| Develop UI connection for result filtering and selection | Rania | Ongoing |

**NEXT STEPS**

* **Immediate** combines the weekly testing setup with technical deliverables
* **Short-term** focuses on algorithm refinement and system improvements
* **Medium-term** emphasizes full deployment and scaling objective

**PROJECT SCOPE SUMMARY**

**Technical Framework:**

* **AI Integration:** Gemini with Google search functionality
* **Document Processing:** Automated retrieval from 200-500 targeted library websites
* **Quality Control:** Multi-layer filtering and metadata validation
* **Storage Integration:** SharePoint with standardized organization structure

**Testing Strategy:**

* **Primary Test Case:** Poland (language complexity, structural simplicity)
* **Document Volume:** Weekly provision of 5-10 test documents
* **Success Metrics:** Accurate metadata extraction and document matching
* **Scalability Planning:** Extension to complex multi-document countries

**CONCLUSION**

The development team has achieved significant technical breakthroughs with the Gemini AI integration and established a comprehensive testing framework using Poland as the primary case study. The strategic shift from broad internet searches to targeted library focus positions the project for improved accuracy and efficiency in the upcoming implementation phase.

**Meeting Minutes Prepared by:** Rania Khan   
**Date Prepared:** August 27, 2025  
**Next Phase:** Wednesday (Weekly recurring)