**Meeting Minutes**

**Professional Computing Project – Team 03**

**Date:** September 17, 2025  
**Duration:** 60 minutes  
**Purpose:** Review current development of the automated census document search system, discuss metadata integration, filtering strategies, and plan next steps with emphasis on library-based searches.

**Attendees**

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

**Key Discussion Points**

**1. System Development Update**

* The prototype now integrates **AI-generated queries → Google Search → PDF retrieval**.
* Current process: AI generates queries, searches return ~15 results, PDFs are extracted if available. Metadata is then compared and ranked for relevance
* Limitations: Many false positives from academic papers; discussed filtering out repositories like ResearchGate and Academia.edu

**2. Metadata & Testing**

* Metadata has been expanded and refined; AI can correct minor errors (e.g., misspelled "census").
* Chris emphasized need for testing in the next 2–3 weeks before Week 11, to avoid delays.
* Plan: Chris will personally test batches of documents to provide focused feedback

**3. Library Search Strategy**

* Prototype extended to search within national library portals (tested for Texas and France).
* Workflow: After general Google/PDF search, system can switch into library mode for deeper retrieval.
* Challenge: Each library uses a different HTML structure; requires manual configuration for accurate extraction
* Discussion on creating a priority list of ~20 libraries (e.g., British Library, French National Library, German National Library, US Census Bureau Library) for consistent searches
* Colonizer relationship noted: e.g., Indonesian census also found in Dutch libraries. Chris to provide colonizer mapping as metadata field.

**4. Technical Considerations**

* Speed: Each search cycle takes ~2–3 minutes. Parallelization possible if sufficient RAM is available.
* Hardware: Chris has high-performance setup (96GB RAM), suitable for running multiple libraries in parallel. For student testing, scaling may be limited
* Internet/Latency: Current bottleneck lies in AI (Gemini) query processing, not local hardware. Paid tier may help with stability but not guaranteed speed boost.

**Decisions Made**

1. **Filtering** – Exclude low-value academic repositories (ResearchGate, Academia.edu).
2. **Priority Libraries** – Chris will provide a ranked list of ~20 key libraries for hard-coded implementation.
3. **Metadata Enhancement** – Add “Colonizer” field to metadata for query optimization.
4. **Testing Schedule** – Chris to dedicate time for hands-on weekly testing; aim for session next Wednesday (Sep 23).

**Action Items**

|  |  |  |
| --- | --- | --- |
| **Task** | **Responsible** | **Deadline** |
| Provide top 20 priority libraries list | Client | 17 Sep. 25 |
| Add colonizer field into metadata schema | Team | Next week |
| Continue refining Google + Library integration (HTML extraction) | Team | Next week |
| Parallelization testing with available RAM | Team | Next week |

**Next Steps**

* Implement library prioritization and colonizer-based search logic.
* Conduct structured test run with Chris next week (Week 9).

**Meeting Minutes Prepared by:** Hazel Wang  
**Date Prepared:** September 17, 2025  
**Next Meeting:** Sprint 3 Discussion – Thursday | Client Meeting - Wednesday