# Scope of work

Team 3-CITS3200 Professional Computing-Documentation Identification Software
By Rania Khan, Hazel Wang, James Felstead, Peter Fang,Chunyu Zheng

### Objectives and Project Overview

The objective of this initiative is to develop a scholarly literature discovery and retrieval system that is based on artificial intelligence (AI). It will automatically locate historical government census publications from national statistical offices, national libraries, university libraries, and digital archives internationally, extract metadata, and prioritise recommendations based on availability (e.g., convenient downloads (Portable Document Format, PDF) → remote collection requests → offline physical retrieval). The system is designed to substantially reduce search time and enhance hit rates, enabling clients to concentrate on academic research rather than manual search.

## In Scope

.

### Basic Search Functionality

- Provide assistance in the development of queries by country, year, document type, publisher, language, and province/state.
- Integrate GPT-5 Deep Research as the principal search engine.
- Conduct a parallel search of multiple sources, including statistical bureaux, libraries, and archives.
- Translate non-English search queries automatically;
- Support both exploratory searches by country/date range and specific named-document searches.
- Constructed to manage documents in the "low thousands" range.

### Source Coverage and Document Discovery

- Websites of national statistical bureaux (more than 200 countries and territories);
- Catalogues of national libraries (e.g., the British Library, the Bibliothèque nationale de France, and other significant European and American libraries);
- Digital archive platforms and university library systems;
- Items that are available for direct download in Portable Document Format (PDF) are flagged.
- Record entries in the library catalogue for objects that do not have downloadable copies.

### Metadata Extraction

- Title with multi-level subtitles; publisher/statistical agency; year; country; province/state; volume/edition;
- File type, file size, source Uniform Resource Locator (URL), and most recent update date;
- Standardised geographic names and multilingual fields.

## Prioritisation and Classification of Accessibility

- Priority 1: direct PDF download; Priority 2: remote request (with contact person/method); Priority 3: tangible retrieval (in conjunction with the library's location);
- Determine the prospective costs and difficulty of access; indicate items that necessitate special permissions.

## Nonfunctional Requirements

- Support bulk processing and 5–10 concurrent users; target a per-search response time of less than 60 seconds.
- Read-only integration with the existing document corpus; strict Non-Disclosure Agreement (NDA) compliance; local deployment and encrypted transmission/storage;
- Scalable to accommodate a burgeoning results database and additional document categories.

## User Interface (UI) and User Experience (UX)

- A principal search box that corresponds to Google's simplicity, as well as advanced fields.
- Results table with sortable columns (Title/Country/Year/Access Method/Source) and color-coded access difficulty (green represents download, yellow represents remote, and red represents physical).
- Action buttons: "Download PDF," "Contact Library," and "View Details"; in-row expansion to display metadata and library contact information;
- The search history is organised according to the difficulty of access.
- Batch search: submit a list, process it sequentially, display progress and a summary report, and save the results in a format that is organised by country and year.