The story of a research collaboration

Use of LLMs for metadata extraction from NATO archive documents

Jan Švec

Faculty of Applied Sciences
University of West Bohemia

honzas@fav.zcu.cz

FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

LINDAT
CLARIAH-CZ

CLARIN, DARIAH & Libraries 2026 Workshop
10 November 2026

October 2025: an email starts the project

Lauren Sukin to Michal Smetana, 2 October 2025

"There is no finding aid."

"I would love for them to all be OCR'd, as a first step."

- PDFs organised by year, with uneven existing OCR.
- **First research task**: find mentions of public opinion and read the documents.
- **Cost and data protection** were concerns from the outset.

3 October: Michal contacts Jiří Kocian, who forwards the request to our team.

The NATO collection behind the project

30,819

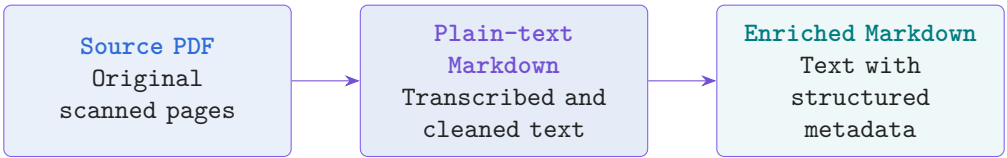
PDF files

290,045

scanned pages

- **Data collection** from https://archives.nato.int/ and NATO Library in Brussels.
- **Material**
  Released English and French archival documents; declassified and publicly available documents.
- **NATO bodies**
  North Atlantic Council, Defence Planning Committee, Nuclear Planning Group and Secretary General's Private Office.

The document path: PDF and two Markdown outputs



- **Source PDF**: scanned pages retained for verification.
- **Plain-text Markdown**: searchable transcription with cleaned layout.
- **Enriched Markdown**: text with fields, summaries, entities, topics, source links and interlinks.

Initially: Qwen3-VL and Gemma4 in Ollama.
Now: qwen3.8-27b for both OCR and metadata extraction.

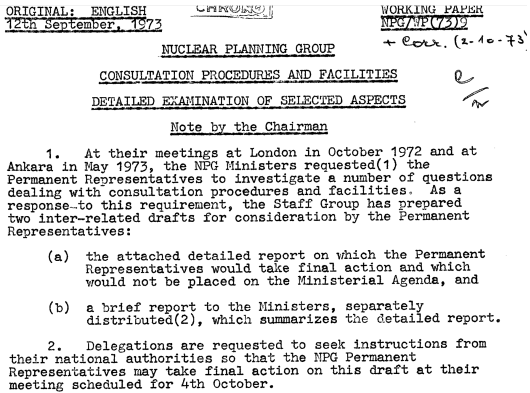
Two weeks hands-on in Oxford: developing metadata extraction

Working with Lauren Sukin and Michal Smetana.
Intensive day-to-day collaboration.

- **Proof-of-concept implementation**
  Fields, summaries, abstracts, topics and source links.
- **Validation**
  Checking extracted text and metadata together.
- **User interface**
  Obsidian provides the wiki-like presentation and navigation experience.

NPG/WP(73)9: the scan and its transcription

Source PDF



ORIGINAL: ENGLISH
12th September, 1973

WORKING PAPER
NPG/WP(73)9
+ Corr. (2-10-73)

NUCLEAR PLANNING GROUP

CONSULTATION PROCEDURES AND FACILITIES

DETAILED EXAMINATION OF SELECTED ASPECTS

Note by the Chairman

1. At their meetings at London in October 1972 and at Ankara in May 1973, the NPG Ministers requested(1) the Permanent Representatives to investigate a number of questions dealing with consultation procedures and facilities. As a response to this requirement, the Staff Group has prepared two inter-related drafts for consideration by the Permanent Representatives:

(a) the attached detailed report on which the Permanent Representatives would take final action and which would not be placed on the Ministerial Agenda, and

(b) a brief report to the Ministers, separately distributed(2), which summarizes the detailed report.

2. Delegations are requested to seek instructions from their national authorities so that the NPG Permanent Representatives may take final action on this draft at their meeting scheduled for 4th October.

Plain-text Markdown

WORKING PAPER
NPG/WP(73)9

- Corr. (2-10-73)

NUCLEAR PLANNING GROUP

CONSULTATION PROCEDURES AND FACILITIES

DETAILED EXAMINATION OF SELECTED ASPECTS

Note by the Chairman

1. At their meetings at London in October 1972 and at Ankara in May 1973, the NPG Ministers requested(1) the Permanent Representatives to investigate a number of questions dealing with consultation procedures and facilities. As a response to this requirement, the Staff Group has prepared two inter-related drafts for consideration by the Permanent Representatives:

(a) the attached detailed report on which the Permanent Representatives would take final action and which would not be placed on the Ministerial Agenda, and

(b) a brief report to the Ministers, separately distributed(2), which summarizes the detailed report.

2. Delegations are requested to seek instructions from their national authorities so that the NPG Permanent Representatives may take final action on this draft at their meeting scheduled for 4th October.

The same document with metadata and an abstract

NPG-WP(73)9_ENG

Metadata

item_type:	**working_paper**
title:	**Consultation Procedures and Facilities: Detailed Examination of Selected Aspects**
summary:	This Nuclear Planning Group working paper examines selected aspects of consultation procedures and facilities for nuclear weapons use, making recommendations to strengthen the consultation process and proposing future work in this area.
Nº pages:	**10**
source:	pdf/NPG-WP(73)9_ENG.pdf
sender:	Nuclear Planning Group
recipient:	NPG Permanent Representatives
intent:	To present a detailed report on consultation procedures and facilities for nuclear weapons use for final action by NPG Permanent Representatives at their meeting on 4th October 1973
author_name:	Joseph M.A.H. Luns
author_institution:	Nuclear Planning Group
date:	**1973-09-12**
archive_code:	NPG/WP(73)9
citekey:	NPG-WP-73-9_ENG

Abstract

Consultation Procedures and Facilities: Detailed Examination of Selected Aspects

This Nuclear Planning Group working paper, prepared in response to requests from NPG Ministers at their meetings in London (October 1972) and Ankara (May 1973), examines selected aspects of procedures and facilities used by the Alliance for political consultation in time of tension or hostilities on questions of using nuclear weapons. The document makes recommendations for specific actions to strengthen the consultation process and develops proposals for future NPG work. It addresses three main areas: the content of selective release messages, the identification and handling of messages related to nuclear consultation, and the classification and precedence of selective release messages. Regarding message content, the paper discusses both the substance (kinds of information to be included) and the format (use of plain language versus technical abbreviations), noting difficulties experienced during WINTEX 73 in understanding release request messages. The paper recommends that Major NATO Commanders use plain language wherever possible in drafting selective release messages addressed to political authorities. On identification of nuclear consultation traffic, the paper notes that while selective release messages carry distinctive labels (WHITE DOT for exercises, BLUE DOT for actual crises), other essential consultation messages lack easily recognizable distinguishing marks, and recommends inserting the words 'Nuclear Consultation Message' at the beginning of such messages. The document also addresses the handling of nuclear consultation messages to ensure rapid processing by responsible officials. Delegations were requested to seek instructions from their national authorities so that NPG Permanent Representatives could take final action at their meeting scheduled for 4th October 1973.

Topics, entities and tags create reading paths

Topics with supporting passages

Categories/Topics

- #Topics/Alliance-Governance-and-Institutions/Political-Consultation-Mechanisms — The entire document is dedicated to examining and improving consultation procedures: *"This report examines selected aspects of procedures and facilities to be used by the Alliance for political consultation in time of tension or hostilities on questions of using nuclear weapons, makes recommendations for some specific actions to strengthen this consultation process."* It proposes specific mechanisms including message identification, precedence, AIGs, and special operating signals.
- #Topics/Nuclear-Planning-Deployment-and-Control/Nuclear-Release-Authority — The document substantively examines the release message process integral to nuclear release authority: *"the content of selective release messages and identification and handling of messages related to nuclear consultation"* and discusses *"the kinds and substance of the information to be included in such messages"* for *"requests for release of nuclear weapons,"* including format, terminology, and handling procedures for release requests.

Selected entities

- Cities
  - #Entities/City/Belgium/Brussels
  - #Entities/City/Turkey/Ankara
  - #Entities/City/United-Kingdom/London
- Persons
  - #Entities/Person/Luns/Joseph-M-A-H

Conceptual tags

#Tags/ConsultationProcedures #Tags/SelectiveRelease #Tags/CommunicationProtocols #Tags/PoliticalConsultation #Tags/WINTEX/73 #Tags/ReleaseProcedures #Tags/TimeFactors #Tags/NuclearWeapons

#ConsultationProcedures
#SelectiveRelease
#WINTEX/73

The documents online: a research reading path

https://nato-obsidian.kky.zcu.cz

1. **Search** for a subject, person or phrase.
2. **Assess** the metadata and abstract before reading.
3. **Explore** topics, entities and document references.
4. **Verify** extracted information against the original PDF.
5. **Collect** documents in the shopping cart.
   Download the Markdown files as an Obsidian vault.
6. **PDFs remain hosted online.**

Developing Archivatorium with agentic AI

No code was written manually.
Specifications were carefully written in human language.

1. **Specify** the researcher's needs and how to check the result.
2. **Plan** the work for the AI agents.
3. **Implement and verify**: agents write code and tests, check results and refine the software.

Example specification

As a researcher, I want a link to the original PDF so I can verify the extracted text. The link must open the correct source document.

https://github.com/honzas83/archivatorium

Researcher feedback: focus groups

Evidence from recorded and transcribed discussions.

- **Search and summaries**
  Find a starting point through keyword search and document links.
- **Early research**
  Identify actors and themes; assemble a reading list.
- **Researcher requests**
  Explain collection limits; save sets; add notes and corrections.
- **Conditional trust**
  Use generated descriptions as leads; check originals.

Document validation: the research assistants' reviews

Three completed XLSX review sheets

Scores and written comments on generated text and metadata.

- **Summaries**: generally rated highly; some missed important content.
- **Abstracts**: sometimes omitted attachments or overstated what was agreed.
- **Metadata and OCR**: most errors originated in OCR: wrong dates, missing/duplicate references, misread or repeated text.
- **Topics**: ratings differed across sheets; some assignments lacked substantive support.

Review of Qwen3-VL outputs motivated the switch to qwen3.8-27b, now undergoing validation.

A project that began with a request for OCR

Conclusion & What's next

300 GB

New PDF batch from NATO Library, September 2026.

Better OCR

Better Markdown and handwriting transcription.
Far fewer OCR variants in metadata.

Agentic AI

Build tools.
Search archives.