Management of Free Text for NDIC: An Overview of the FTM Project

Nancy J. Pruett and Thomas Kinsella
Office of Research and Development,
National Drug Intelligence Center, &
Intelligence Systems Support Office (OASD C3I)

SUMMARY: The FTM Project is a joint project to build a prototype which retrieves and extracts against free (unstructured) text and to evaluate the prototype's usefulness for the National Drug Intelligence Center (NDIC). NDIC has strong needs for retrieval and extraction from large multi-source textual databases.

SPONSORS: FTM is jointly sponsored by the Intelligence Systems Support Office (ISSO) of OASD/C3I, the National Drug Intelligence Center (NDIC) and the Office of Research and Development (ORD). The prototype is being built and evaluated at the Federal Intelligent Document Understanding Laboratory (FIDUL) in Tyson's Corners, VA, in an unclassified environment which replicates the hardware and software environment at NDIC.

PARTICIPANTS: As of April 15, contractors include BETAC, PRC, IDI, HNC and Lockheed Martin. BETAC and IDI are providing the technical leadership, the integration and the GUI. PRC is providing TIPSTER expertise (from the TIPSTER SE/CM), and the testing and evaluation leadership (through FIDUL). HNC and InQuery (via Sovereign Software) are the first detection tools being integrated. Lockheed Martin is supplying the TIPSTER compliant document manager, document viewer and extraction system used on other TIPSTER demonstration projects. Information on lessons learned is shared among all the participants and sponsors.

OBJECTIVES: The project objectives (from the Memorandum of Agreement) are to:

- Prototype and demonstrate new technology which might be incorporated into production systems at NDIC. The NDIC OASIS/IDEF model and the current NDIC free text management architecture development will guide project activity. Results will be reported in monthly updates as well as in a final report.
- Assemble a prototype which includes at least two TIPSTER detection tools working on source material which represents NDIC data (unclassified).
- Expand the prototype to include at least two TIPSTER extraction tools, and use the extraction tools to load a database. The database will serve as the basis for analytical tools.
- Expand the prototype to include conversion of hard copy documents (or the simulation of them), including degraded OCR data.
- Develop (or acquire) a minimal user interface and other interfaces which are necessary to demonstrate integrated use of detection and extraction tools, and any other analysis tools to be demonstrated.
- Develop test measures appropriate for testing these tools in an end-to-end system; test and evaluate the tools.
- Use the TIPSTER Architecture in the development, and work closely with the TIPSTER Architecture Working Group (AWG) to improve the Architecture.
- Explore the implications of the prototype for an end-to-end architecture, i.e., expanding the TIPSTER architecture to include end-to-end text processing (e.g., OCR, machine translation and database loading).

CONCEPT OF OPERATIONS: The prototype will allow users to choose among detection tools, do sophisticated searches on particular topics, maintain chosen references to documents in project files, view those documents either within the detection tool or through a common viewer, "copy and paste" from full text information into word processing documents while retaining the source information; and extract particular fields and relationships from project files or any other TIPSTER collection.

THE HARDWARE AND SOFTWARE ENVIRONMENT: The hardware environment is a DEC ALPHA 2100 server and two Celeris 5100DP (dual Pentium) workstations. The software environment includes the Digital OSF Operating Environment on the server, Windows NT on the workstations, eXceed (an x-windows emulator for Windows NT), Oracle, and MS Word. The GUI is being developed in Visual Basic.

SCHEDULE: The prototype will be completed by September 30. Usability testing, some modifications, and testing against degraded OCR data will occur by January, 1997.

POSSIBLE FUTURE ADDITIONS: Once the prototype is complete, other detection and extraction tools, GUI's, or alternative TIPSTER-compliant document managers could easily be added to the FTM prototype, providing an excellent unclassified environment for continuing to evaluate and demonstrate TIPSTER technology. Other possible modifications include fine-tuning the GUI to the NDIC users, and integrating the system into the NDIC environment.