# ARCHITECTURE OVERVIEW

*TIPSTER SE/CM*

*tipster@tipster.org*

## THE TIPSTER ARCHITECTURE

The TIPSTER Architecture is a software architecture for providing Document Detection (i.e. Document Retrieval and Message Routing) and Information Extraction functions to text handling applications. The high level architecture is described in the Architecture Design Document.

## PURPOSE OF THE ARCHITECTURE

The TIPSTER Architecture is intended to facilitate the deployment into the workplace of advanced Document Detection and Information Extraction software. It provides a component and module design which has been jointly developed by a significant number of providers of advanced software of this type. In addition, this design meets the requirements of a number of US Government agencies.

The Architecture was developed to meet the need for US Government agencies with similar text handling requirements to share some software modules and knowledge sources that meet these requirements. Use of the Architecture for Government procurements will also shorten the development process for new text handling applications, because a basis for design would already exist and be understood by vendor and customer alike. Finally, the Architecture will allow systems to be upgraded in a modular fashion as new text handling technology becomes available. Similarly, the research community can take advantage of the Architecture to facilitate the testing of new ideas in advanced text handling.

## SCOPE OF THE ARCHITECTURE

The Architecture has been designed to meet a large number of text handling requirements for CIA, DIA, and NSA. It meets, however, only those requirements having to do with Document Detection and Information Extraction functions. Most requirements for other functions, such as Machine Translation or Optical Character Recognition must be met outside the TIPSTER Architecture. Selected requirements in these areas may be part of TIPSTER

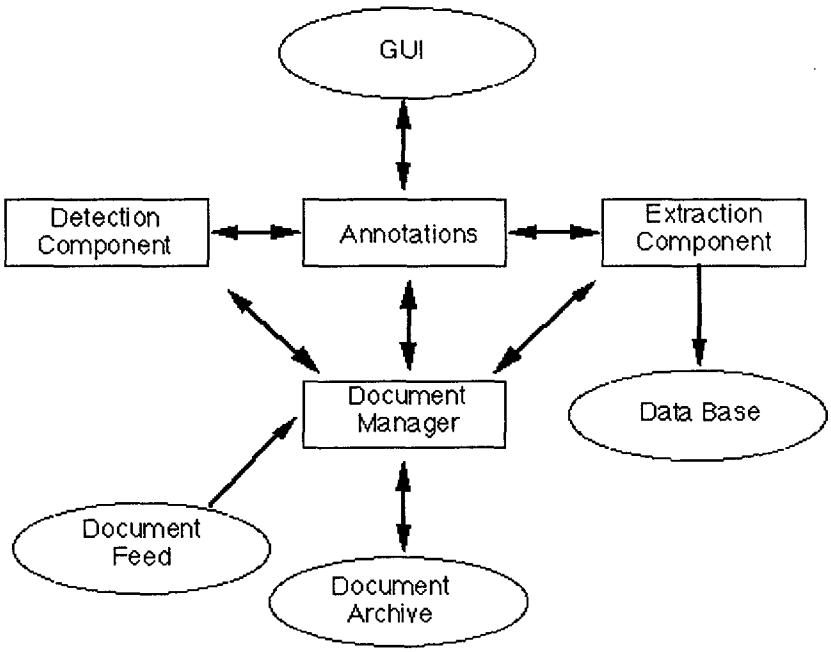


**Figure 1** Architecture Overview

Phase III as the Architecture is expanded. In addition, User Interface (GUI) requirements are not covered by the Architecture, but, are unique to the specific application. Analytical tools, such as link analysis tools, timelines, or other displays showing document clustering are considered part of the User Interface or the application. These tools lie outside the Architecture, but use information about document relevancy, relationships between documents, phrase lists, name lists, and relational or object data base records which has been exported by the functionality residing within the TIPSTER Architecture.

## ARCHITECTURE COMPONENTS

There are four components: Detection, Extraction, Annotation, and Document Management.

- Detection encompasses the technology which does document retrieval and document or message routing.
- Extraction encompasses the technology which identifies specific entities and the relationships between entities in free text so they can be use to build a database.
- Annotation allows these two components to share information at a component level. Primarily, at present, it is the method for recording and passing forward the information developed by the Extraction component. Items of specific types, such as personal names, places, or organization names, for example, can be located in the text by appropriate annotators, and the text locations and data types can be passed to any other component or part of the application, through Annotations, for further processing or viewing.
- The Document Management component handles the document storage and archive. This function can be performed by existing document managers or Commercial off the Shelf (COTS) products, such as a standard Data Base Management System (DBMS), with the addition of a wrapper to be compatible with the TIPSTER Architecture.

The TIPSTER Architecture is explained in more detail in the "TIPSTER Text Phase II Architecture Concept" in this volume.