



Government data management and analysis – taming the rising flood of information

Quantum leaps in computing, data storage, and high bandwidth communications technology has created a flood of data that threatens to drown governmental agencies. According to a study conducted by the University of California, Berkeley, approximately 5 exabytes of information is generated annually worldwide – about enough information to fill 37,000 libraries the size of the Library of Congress, or about thirty feet of books for every man, woman, and child on earth. The United States alone produces about 40% of this new information. Today, the need for processing, analyzing, and managing huge amounts of data is critical for governments the world over.

In response, governments are urgently looking for informatics solutions – technologies and processes for acquiring, managing, and analyzing huge amounts of information. The value and the risks are high – buried in mountains of data are the keys to discovering and preventing acts of terror, diplomatic events, and security crises. The difficulty of this task is made even greater by the global reach of information – much of what might be valuable to the viewer goes unrealized because it's in a foreign language. And, increasingly, that language is one of the many dialects of Arabic, spoken by over 300 million people in over 20 countries.

Finding the right information.

A key component to any governmental intelligence operation is the exploitation of a wide variety of structured and unstructured data, such as public and private documents, broadcast transcripts, e-mail, trade and scientific papers, and much more. This information comes from a variety of sources: human (HUMINT), communications monitoring (COMINT) and even “spy” satellites and other technology assets (IMINT). Searching this disparate set of formats and files for specific information in Arabic using Western-based search tools is fraught with drawbacks,

inefficiencies, and errors. Most problematic is the technical differences between Arabic and Western languages; Arabic has no short vowels and utilizes rarely-used diacritics – signs that act as keys to pronunciation. In addition, our research indicates that over 60% of Arabic words have more than one meaning. The sheer richness and variety of Arabic makes it one of the world's most difficult languages for comprehensive and effective searching and analysis.

More than any other organization in the world, the US intelligence community is awash in data, much of it still unanalyzed and even unprocessed. Part of the challenge is the quandary that comes from “not knowing what you don't know”. In other words, intelligence professionals might not know what they're looking for and consequently won't ever find it. Often this involves relational context – the unanticipated alignment of people, places, and things. For intelligence agencies, this can mean the difference between missing the connection between suspects' names, a foreign city, and a mode of transportation and uncovering and stopping a terrorist incident.

This problem is compounded by the frequent multiple spellings of proper nouns – names in Arabic can have dozens of correct spellings and if an analyst is not finding them all, critical information can be missed. Conventional search applications are not enough – it requires an application with built-in “intelligence” that “knows” key linguistic features so that it can uncover the right information, even if it wasn't asked for. Not knowing the variations of someone's name – or that variations even exist – could result in someone on a terrorist watch list getting into the country. For example, a Google search for the name “Ussama” yields 75,300 results. A search for the same name, but spelled “Osama” yields 16,400,000 results.



Extracting value from Arabic information

Data analysis and information extraction are two of the most powerful economic, diplomatic, and security tools available to governmental departments and agencies. Increasingly, the worldwide war on terror is becoming a war of information. COLTEC is a leading provider of Arabic language tools specifically designed to help government agencies process and manage documents, files, and transcripts to find relevant information, analyze context, and create valuable, indexed databases from massive quantities of structured and unstructured data.

With a growing shortage of trained, experienced Arabic translators and analysts available worldwide, our software fills a critical gap that enables governments to effectively monitor and act upon information in many of the world's most active and important regions.

What makes COLTEC different – and better.

We are the only Arabic language processing software developer in the world with a truly mature foundation in NLP, a foundation that has been under continuous development for over ten years. Staffed with a core of experienced, highly trained multi-language scientists, COLTEC's fundamental advantage is our patent-pending Natural Language Processing (NLP) model. Developed by our founder, Dr. Taghride Anbar, this revolutionary model is the basis for the processing, analysis, and management of Arabic as a linguistic system suitable for computer manipulation. This powerful, fundamental methodology is the very core of our business – it turns complex linguistic rules into flexible, logical communication tools designed specifically for computer use.

Our strengths include:

- Strong linguistic foundation and infrastructure
- Native Arabic-speaking developers
- Strong balance of linguistic and statistical expertise
- Ability to “identify, analyze, and even generate” Arabic words
- Processing speeds multiple times faster than our competitors, for example our Word Identification tool is able to process over 600,000 words/second

- A fully linguistically processed balanced corpus of Modern Standard Arabic (MSA) that includes over 150 million words
- Lexical Database uses proprietary compression techniques that result in a total size smaller than 3 MB

As a result, we are uniquely positioned to help governmental agencies worldwide process, analyze and manage Arabic data more efficiently and effectively than any other company or software.

We offer a range of products ideally suited to provide Arabic data management and analysis capabilities to governmental agencies. These include:

Arabic Search Plug-In (ASPI) specifically designed to help Arabic & non-Arabic speakers search Arabic data. It takes into account the unique writing features of Arabic, as well as issues related to Ambiguity & affixations. No other Arabic search tool is based on a language model as subtle and sophisticated as ASPI.

Arabic Entity Extraction Tool (ANEE) uses advanced discovery search methodologies to extract critical information from large amounts of structured and unstructured data – unlike search tools, it analyzes concept and context and delivers results based on meaning.

Word Conversion (WORDCON) accurately and reliably converts Arabic characters into Latin ones and vice versa, enabling crucial information, once found, to be converted for more thorough examination and analysis. This also enables users to search Arabic names with all their different Latin spellings for example (ie. Mohamed, Muhamad, Mohammed, etc.). This is often vital for intelligence operations facing a shortage of Arabic speaking employees.



Tel.: + (202) 3336-0326 / 3336-0736 Fax: + (202) 3336-0782