There is a growing need for multilingual information as the economy becomes more and more
global. Strategic, commercial, scientific and technical intelligence are indispensable for
companies wishing to play a major role in the new net-economy. It is also evident that e-
commerce cannot be efficient if the customers are not able to express their need in their own
language.

It is now clear that information written in English is not sufficient to give a global view of a
subject or for making the right decision. It is also not sufficient to monitor an international
domain, as is the case for the IAEA safeguards department.

I shall first examine the policy in the United States and in the European Union with respect to
multilingual information retrieval and cross-language interrogation. I will then define
multilingual and cross-language systems in the context of information retrieval and
information filtering.

Multilingual systems are systems that can process documents in several languages, but
interrogation must be done in the document’s language.

Crosslingual systems are systems that enable the user to interrogate in one language,
documents written in another language or in several languages.

The implementation of a multilingual information retrieval systems raise several problems. A
multilingual information retrieval system must not only be a system that have its
documentation, screens and error messages localized in several languages but also a system
which can index various languages. Some of these languages, like English, need little
linguistic processing to be used by unsophisticated fulltext systems. Other languages, like
French, German or Russian, require more powerful linguistic tools. Some languages cannot be
processed by simple fulltext retrieval systems even using palliatives like truncation or
proximity operators because the significant part of the words are squeezed between
uninformative parts, as in Arabic or agglutinative languages. We will give examples to
illustrate the difficulties encountered when supporting certain languages.

Several approaches can be used for crosslingual interrogation, including latent semantic
indexing(LSI), MT systems, or multilingual reformulation.
Latent semantic indexing is an evolution of the classical vector space model of Gerard Salton. The term-document matrix is submitted to a singular value decomposition. The singular values in the diagonal matrix are sorted in a decreasing order. Only the largest values are kept. This process decreases the number of vector space dimensions to a few hundred, which produces implicit semantic equivalences to retrieve documents. If some documents in the database are in several languages (translated documents), the matrix reduction will produce implicit equivalences between words in these different languages.

The second approach is based on a translation of queries or documents using a machine translation system (MT). Thus, the problem of crosslingual interrogation is transformed into an automatic translation followed by a monolingual interrogation. MT systems are configured for human use and they generally give only one translation for each sentence. Hence, the mistranslation of a single word can cause the interrogation to fail.

The last approach is based on bilingual dictionaries, and all possible translations are tried. Some of the ambiguities are removed by using morphosyntactic analysis, the others are resolved using the implicit semantic of the database during query-document comparison.

The effectiveness and limitations of these approaches will be discussed using examples.

Examples of multilingual information management will be presented including access to mixed-language databases, access to mixed-language documents, multilingual dynamic hypertext, and comparison of documents in several languages treating the same theme.