# INFORMATION COLLECTION STRATEGIES TO SUPPORT STRENGTHENED SAFEGUARDS

L. COSTANTINI
IAEA Division of Safeguards Information Technology,
Vienna, Austria

J. HILL[1]
Nuclear non-proliferation and safeguards contractor,
Canberra, Australia.

**Abstract**

Part 1 of the Strengthened Safeguards System, approved by the IAEA Board of Governors in 1995, envisages the collection and analysis of a wide range of information on States' nuclear activities, beyond that used in classical safeguards. That information includes open source reports in the media and journals. The IAEA Division of Safeguards Information Technology (SGIT) has established a number of collections of open-source reports held in electronic form. Some of these are very large and comprise general news rather than nuclear-specific material. Special-purpose search mechanisms for use with Verity's TOPIC/Search 97 software (TOPIC Trees) have been designed to search for reports relevant to the subjects covered by the various sections of the IAEA's standard State File format. Where State File sections deal with nuclear fuel-cycle processes, the relevant search trees draw heavily on the IAEA's Physical Model. The trees and the associated collections of reports are accessible from throughout the Department of Safeguards as a tool to help Country Officers and State Evaluation Groups to review, search, and select information useful for State Files and Evaluations. Users do not need a detailed understanding of either the Search 97 software or the Physical Model (although this is needed for report selection). Additional software tools from the same source have been put into service to facilitate the handling and organization of the selected reports. It is now possible to construct an electronic State File that incorporates links to the original documents upon which that material in the File is based.

## 1. Introduction

The IAEA Board of Governors approved the implementation of Part 1 of Strengthened Safeguards in June 1995. Since then, the collection and analysis of information beyond that provided by States parties and acquired by inspectors under NPT Safeguards Agreements has been an integral part of IAEA safeguards. The Agency has formally established internal structures and procedures to facilitate the effective use of open-source and other information not previously used in safeguards.

Over this period the IAEA Division of Safeguards Information Technology (SGIT) has been building its collections of electronically held open source information. Some of these collections are quite nuclear-specific, such as material from the Monterey Institute in California, and nuclear news collections provided voluntarily by a number of Member States. Others are completely general news sources. Several of these collections contain many more reports than could possibly be reviewed, or even skimmed through by a human analyst.

So a need arose for computerised search facilities to identify nuclear-relevant items from those collections. More specifically, the need was for search mechanisms to identify reports that would be useful to inspectors responsible for preparing State Files and State Evaluations, and for making the comparisons with declarations needed to identify questions and apparent inconsistencies.

[1] The Australian Support Program funded part of Mr Hill's contribution to this work.

This paper describes how the IAEA Division of Safeguards Information Technology addressed that need.

## 2. Choice of software

Of the software available to the Agency to help with the searching and analysis of substantial collections of reports, Search 97 from Verity was chosen for this particular application. Search 97 uses a function called "Concept Retrieval", a technology which enables users to search for subjects or concepts in documents, rather than individual words or phrases. Search 97 treats specific words and phrases as evidence of the presence of a concept in a report.

For use in Search 97, search terms are encapsulated in a component called a "TOPIC Tree". In the design of TOPIC Trees, each concept is subdivided into different sub-concepts in a manner similar to the subdivision of the branches of a tree. The designer attributes importance weights to sub-concepts to reflect the fact that some words, phrases or other concepts are more important than others in expressing the overall concept.

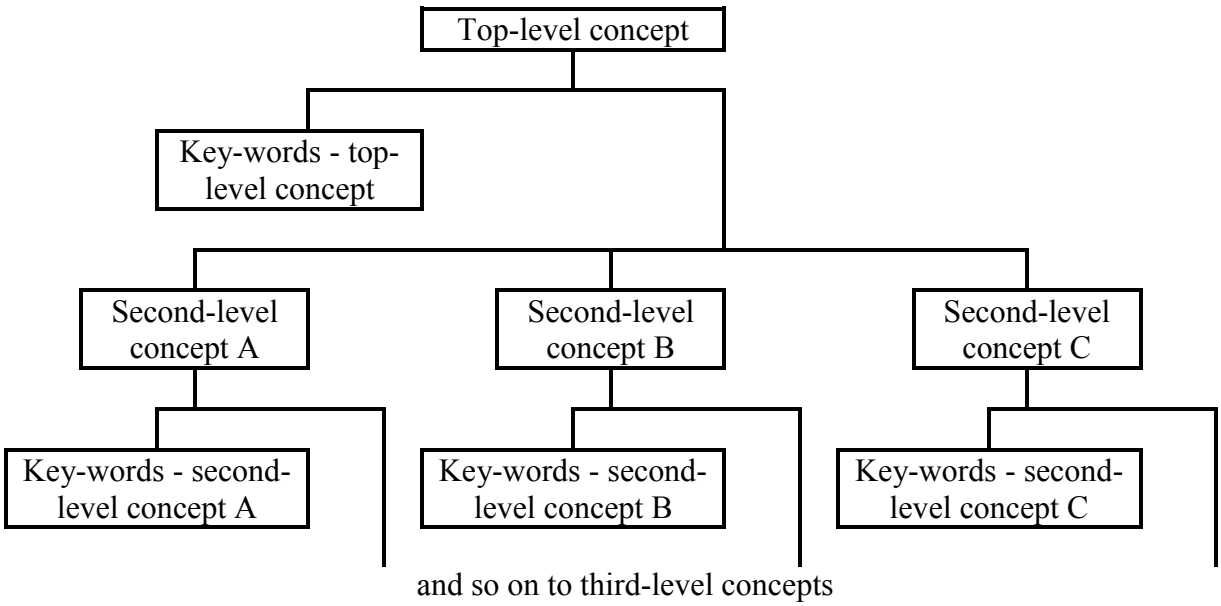
Once appropriate TOPIC Trees have been designed, Search 97 is very straightforward to use. That was an important consideration in its selection, since it was expected that personnel from all over the Department of Safeguards would carry out these searches on a routine basis. It became possible to adopt an approach whereby special-purpose search mechanisms (the TOPIC Trees) could be designed for use by a large number of users who then did not need to be familiar with the details of the search software. Such an approach seems to be unusual if not unique to the Agency. So there was no body of professional experience with a similar approach upon which to fall back while carrying out this task. User feedback has been, and continues to be, a very important guide in improving and refining the product.

## 3. Search 97 and the Physical Model

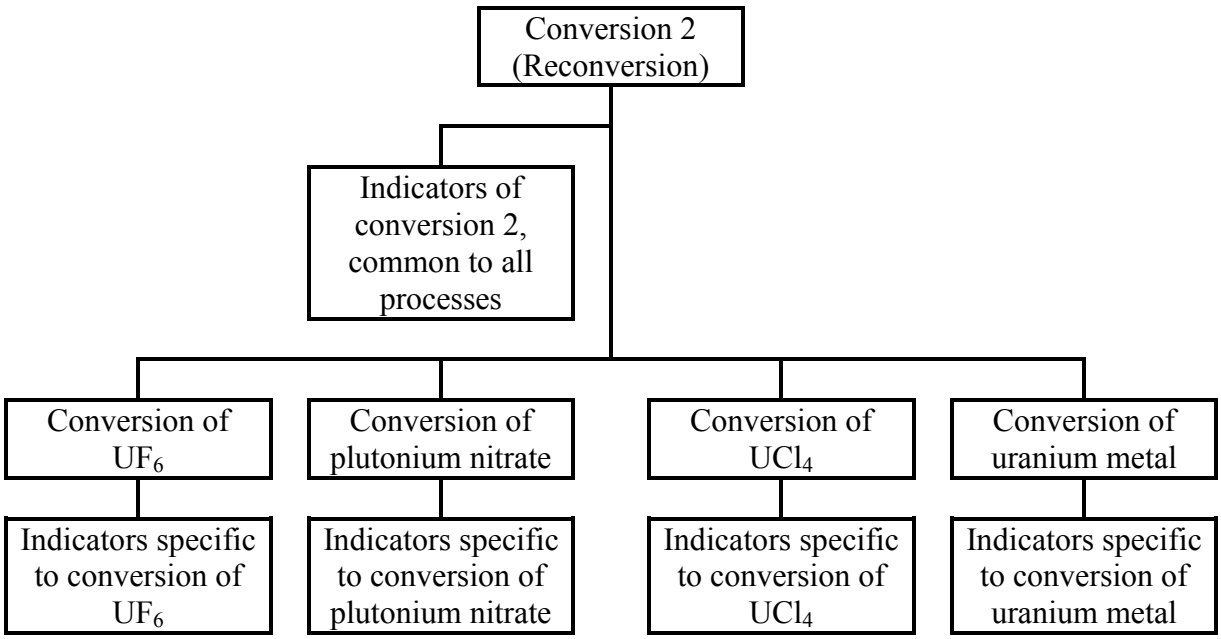
The task began with the use of Search 97 to find reports dealing with each of the proliferation-relevant fuel-cycle processes. The Agency's Physical Model describes each of the nuclear fuel-cycle processes that a state with a nuclear weapons program might need to use, from the mining and milling of uranium onwards. At each stage the Model identifies "indicators", which are potential observables that would suggest that a state was undertaking (or envisaged undertaking) the fuel cycle step in question.

It is particularly appealing to use Search 97 with the Physical Model. Both require the presentation of information in a tree-like form. This form proceeds from the most general information (for example indicators, or key words, relevant to all uranium enrichment processes) at the top level of the tree (on the trunk). At the second level of generality (or on the biggest branches of the tree, coming from the trunk) we would find indicators or key words relevant to major groups of processes (for example enrichment using $UF_6$, enrichment using $UCL_4$ and enrichment using uranium metal). And at the lowest level of the tree we find indicators or key words specific to individual processes such as centrifuge enrichment.

The similarities in structure between a TOPIC search tree and the Physical Model is evident from the following. In TOPIC, key words for a search are embedded in a tree-like structure that describes the logic of the search:




The Physical Model identifies indicators, or possible observables, associated with each fuel cycle step, include (for example) those associated with equipment acquisition (or manufacture), environmental emissions, and training and technology. The indicators of the Physical Model are also laid out with a tree-like structure, eg:




So it appeared a promising concept to use the Physical Model, and the structure of its description of the various fuel-cycle processes, as the basis for TOPIC search trees to find open-source reports dealing with those processes.

From that point on, the detailed design of TOPIC Trees diverged from the structure of the Physical Model, because a key word is very different animal from an indicator. The following considerations were relevant at this point in tree design, and they came more fully into play at the testing and commissioning stage:

- the search tree must be equipped with essentially all possible synonyms for terms used. The media from which Agency open-source material are taken will typically use more colloquial terms than readily come to the mind of a nuclear professional. Fortunately TOPIC systems handle synonyms easily. Once lists of synonyms have been defined, a defined list can be called upon by any search tree;

- some terms that appear at first sight to be nuclear-specific technical terms cannot easily be used in the most obvious searches. For example, "enrich" and its derivatives cannot easily be used to search for reports about the process of enriching uranium, because so many reactors are described as using fuel enriched to a particular value, and, more generally, enriched uranium continues to be referred to as enriched throughout its life – long after the process of enriching it is over;

- key words are not usually sufficiently specific to the subject of the search to be used by themselves without returning many too many irrelevant reports. (A certain number of irrelevant reports are unavoidable if useful material is not to be missed.) To ensure that key words are being used in a relevant context, most must have a "proximity condition" attached to them. Such a condition means that if the key word is found in a report, the report is only returned if another identified key word is found within a prescribed distance of it (eg five words or less away from it, within the same paragraph, etc). In practice, prescribing maximum numbers of words distance between key words was found to be more useful than prescribing that the key words be found within the same sentence or paragraph.

- there is a need to address the importance weighting to be given to keywords at this stage. The Physical Model weights indicators as "strong", "medium", "weak" or "less than weak" according to how unambiguously the existence of the indicator suggests the existence of the process in question, i.e. according to how *specific* the indicator is to the process in question. When dealing with words in text, it is necessary also to give consideration to how specific the word is to the indicator. For this reason, a five-point scale replaced the four-point scale of the Physical Model for this work. A score of 1 indicates that the presence of a key word (perhaps including a proximity condition) shows unambiguously that the report addresses the process in question. A score of 5 indicates that the presence of the key word only suggests rather tenuously that the report addresses the process. The thinking behind this was that "1". "2" and "3" were needed to correspond to "strong", "medium" and "weak", but that there was also a need to extend the scale downwards to allow for key words that were not very specific to the relevant indicator.

After entry into TOPIC, each new search tree was subjected to a testing process, which often resulted in substantial changes to the tree. The procedure comprised an examination of the reports retrieved by the tree to ensure that:

- straightforward errors, such as typographical mistakes, were eliminated. It did, however, prove necessary to include some common spelling mistakes (eg "hexaflouride") in synonyms to ensure that relevant reports were captured;

- the tree had made as much use as possible of the relevant information contained in the report when selecting the report for retrieval. For this purpose, retrieved reports were examined for
  - words that might serve as new key-words; and
  - excessively restrictive proximity conditions that did not allow combinations of key-words to contribute to the retrieval of the report in the manner expected;

- the report was relevant to the subject. If not, then consideration was given to whether some or all of the key-words found were too general in meaning for the purpose, and, if so, whether they should be eliminated or used with new or more restrictive proximity conditions.

- the importance weights attributed to sub-concepts resulted in the returned reports being ranked broadly in accordance with expectations. Surprising rankings also sometimes from errors in structuring the tree.

Initial testing was carried out using a report collection of manageable size (so that a new search with a modified tree could be carried out quickly), and containing exclusively nuclear reports (so that no surprises arose from totally irrelevant reports).

Later testing was carried out using much larger and more general collections.

During testing, the initial "top-down" approach becomes more "bottom-up" in nature. Nonetheless, the Physical Model provides a very useful starting point.

It should be noted that none of the Physical Model search trees has so far been tested against collections of reports expected to have a high technical content. When they are tested against, for example, INIS, it is expected that many more modifications will be required. Indeed, it may turn out that trees are needed for searching technical literature that are quite different from those appropriate for searching general news. To a degree it can be argued that the search mechanism that is optimal for searching material from a given source is only optimal for that source.

## 4. Current status and use

A series of TOPIC Trees has been designed and introduced into routine service in the Department of Safeguards. As well as the processes covered by the Physical Model, these trees cover certain subject headings in the standard State File structure, and other subjects of interest to the Department. For the trees unrelated to the Physical Model, where there was no pre-existing structure of concepts and sub-concepts, it was necessary to create such a structure. In general that proved a straightforward exercise. Trees currently in service include those for:

- mining and milling of uranium and thorium;
- conversion (including re-conversion);
- uranium enrichment;
- fuel fabrication;
- heavy water manufacture;
- manufacture of reactor grade graphite;

- research reactors (and critical facilities);
- power reactors;
- plutonium production reactors;
- reprocessing;
- waste handling and disposal;
- research centres and laboratories;
- illicit trafficking;
- nuclear policies and programmes;
- nuclear law and regulation;
- nuclear import and export; and
- nuclear co-operation agreements.

In addition, a search tree exists for each relevant state, which will normally be used in conjunction with one of the above, so as to only return reports relevant to a particular country.

Quality Assurance procedures are being introduced whereby the performance of the trees is checked against a "representative" collection of reports, amongst which the reports dealing with the concepts covered by the trees have been identified in advance. It is also possible to assess their performance by comparison with search procedures used by other organizations.

SGIT's mandate is to collect, store, process and disseminate open-source and non-safeguards information for use in the overall country-by-country evaluation process and for use in other products, such as ad-hoc reports written at the request of various high level Agency officials. SGIT uses the above mention TOPIC Trees to organize, cull, and finally select relevant open-source information for State Evaluation Reports.

## 5. Managing the derived information for accessibility

As the amount of retrieved information continues to grow, attention has shifted from the problem of finding potentially relevant reports to that of managing retrieved material so that it can be readily accessed. To achieve that, reports are stored, by subject, in directories that are available for browsing. Directories and sub-directories are typically arranged in a hierarchical structure that goes from a general topic to more specific detail.

But the manual creation of directories is expensive, labour-intensive and time-consuming. And maintaining directories requires an ongoing commitment of resources. It is also difficult to incorporate external data. For that reason SGIT has acquired one of Verity's newest tools, "Knowledge Organiser". The Knowledge Organiser uses TOPIC Trees to categorize and store vast bodies of information in directories together documents dealing with the same subject. The hierarchy into which the directories are organised is called a knowledge tree. In general a knowledge tree divides into a number of different categories representing the subjects of interest. These categories are linked to relevant documents available in the open source collections. The most basic category reflects the country with which the report deals. If, however, the country has a developed nuclear programme, then this main category subdivides into further sub-categories, each of which corresponds to a chapter of the Physical Model (Enrichment, Fuel Fabrication, etc.).

Using this tool, SGIT is able to provide information to the Division of Operations in a way that makes it easily accessible for searching and navigating. This approach has not only

increases the efficiency and effectiveness of the country evaluations, but will also cut down the time needed for carrying out information processing procedures.

So use of the Knowledge Organiser makes the process of collecting and disseminating information both more timely and more efficient. It also allows SGIT to extract reports dealing with a particular TOPIC Tree subject (for example, a particular fuel-cycle process as described in the relevant chapter of the Physical Model), for a particular country, originating within a particular timeframe. Country Officers typically ask for the most recent reports dealing with countries for which they are responsible. The Knowledge Organiser allows such a service to be provided with minimal effort.

The Knowledge Organiser also enables SGIT to implement a long sought-after product, the electronic State File. As a major step forward toward a totally electronic State File, we have adapted Knowledge Organiser to provide SGIT's contribution to the State File in electronic form. Working closely with the Divisions of Operations, we have provided a secure method for storing selected documents in electronic format, highlighting relevant portions, incorporating the relevant sections into a summary document, and storing the complete collection of summary and source documents on the secure ring. This selected, highly relevant information has been manually reviewed and extracted by SGIT staff and incorporated into the State File structure. The software maintains this structure, providing not only an easy mechanism for browsing but, in addition, all the advanced search capabilities of a sophisticated text retrieval system. Access is restricted by country, for users approved by the Director of the responsible Division of Operations. The electronic State File interface allows the State Evaluation Group to review, search, and select information that will be useful for constructing the State Evaluation report.

## 6. Observations and conclusions

It is possible to devise standardized search mechanisms drawing on all the information in the Physical Model, that can be used Department-wide without the need for users to have a detailed familiarity with the Physical Model or the Search 97 program. Certainly users will draw on other open sources, such as individual Web sites that they have found report well on their areas of interest, and hard-copy publications and journals. The more creatively they address the collection of information the better. Nonetheless, we would expect the main body of open-source information used in most State Files to come from the SGIT system.

Notwithstanding the brief discussion of quality assurance procedures at 4 above, it is not possible to be absolutely confident that all relevant reports will be returned by a search using one of the trees. So it is important for the tree designers to remain in touch with users, and for failures to find pertinent items to be reported, so that the trees can be continuously improved.

The Agency has not yet established any procedure for checking whether the *collections* of reports maintained by it:

- are comprehensive (or, conversely, whether there are important groups of reports never picked up in any of the collections);
- exhibit a great deal of duplication between themselves (or even internally).

If a representative collection of reports is established for testing the search trees, as discussed at 4 above, and it is established that all the reports will be retrieved by the appropriate tree, then it will be possible to use the trees to investigate:

- which of the reports contained in the special-purpose representative collection are not present in the regular collections;
- what *information* contained in the representative collection is not contained in the regular collections (i.e., the regular collections may not have exactly the same report as the representative collection, but they may have parallel reports from other services that give much the same information); and
- where the regular collections cover the same ground. It may be that some of the very large collections entirely embrace the material in some smaller collections.