

Компьютеризированная индексация баз данных INIS

Александр Невижель
Международное агентство
по атомной энергии, Вена

Международная информационная система в области ядерной энергии (INIS) — крупнейшая мировая информационная система в сфере «мирного атома». Эксплуатируется Международным агентством по атомной энергии при ООН в сотрудничестве со странами-членами.

В настоящее время при растущем объеме информации, поступающей в INIS, и недостатке персонала остро встает вопрос о повышении эффективности её функционирования. Важным аспектом функционирования базы данных INIS и одним из наиболее затратных этапов её создания и поддержки является предметизация. Компьютеризированная индексация (КИ) — один из наиболее перспективных вариантов повышения эффективности работы INIS. КИ позволяет улучшить качество работы, повысить информативность, снизить стоимость создания баз данных.

Концепция КИ

Для разработки концепции была организована рабочая группа из представителей государств-членов (Австралии, Аргентины, Беларуси, Китая, Франции, Германии, Японии, Нидерландов, Российской Федерации, Швейцарии, Турции и Соединенных Штатов). Группа занимается анализом возможностей создания системы КИ, подбором и тестированием программного обеспечения (ПО) для её реализации, выработкой предложений по внедрению.

Основные задачи КИ:

- поддержка уровня информативности БД;
- облегчение работы персонала, занятого предметизацией;
- улучшение качества индексации.

В настоящее время качество базы данных поддерживается за счёт принятых в INIS правил ввода информации. Тезаурус, как терминологический механизм контроля, используется в INIS для обработки информации и описания её на языке системы, более простом и удобном для предметизации.

Computer-assisted Indexing for the INIS database

Alexander Nevyjel
International Atomic
Energy Agency, Vienna

INIS is the world's leading information system on the peaceful uses of nuclear science and technology. The acronym INIS stands for International Nuclear Information System. INIS is operated by the International Atomic Energy Agency (IAEA), an autonomous organization within the United Nations System, in collaboration with its Member States.

INIS has identified Computer-assisted Indexing as an area where information technology could assist best in maintaining database quality and indexing consistency, while containing production costs. Subject analysis is a very important but also very expensive process in the production of the INIS database. Given the current necessity to process an increased number of records, including subject analysis, without additional staff, INIS as well as the Member States need improvements in their processing efficiency. Computer-assisted Indexing is a promising way to achieve this.

Concept and Design

For the concept development and design, a working group of Member States (Australia, Argentina, Belarus, China, France, Germany, Japan, Netherlands, Russian Federation, Switzerland, Turkey and the United States) was set up to analyse how Computer-assisted Indexing could best assist both INIS and the Member States in the database production, identify and test potential technologies and develop an implementation proposal, if appropriate.

The main objectives of Computer-assisted Indexing (CAI) have been defined as

- maintaining database quality
- saving of subject analysis manpower
- improving indexing consistency

The quality of the INIS database at present is defined by the inputting rules as described in the INIS Reference Series. The Thesaurus as a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained system language is the essential tool for subject analysis in INIS. The planned CAI

Библиографические записи (включая название и краткое описание, возможную классификацию и ссылку на полный текст) анализируются системой КИ и формируется перечень предлагаемых КИ дескрипторов (ключевых слов). При выводе пользователю (через графический интерфейс)



предложенные дескрипторы сортируются по релевантности содержанию документа. Система позволяет пользователю удалять дескрипторы из предложенного списка и задавать дополнительные, когда это требуется.

Функционирование системы КИ

Работа КИ основывается на электронных библиографических записях, получаемых напрямую от издателей или в результате каталогизации изданных работ. Записи хранятся в специальном формате INIS (основанном на XML); предусмотрено ПО для преобразования в данный формат.

Подлежащие анализу записи должны содержать название и краткое описание. Дополнительные поля, такие как название журнала / конференции, можно использовать для того, чтобы

system should meet the present rules for indexing quality and even improve the consistency of the subject analysis.

The bibliographic records (including title and abstract, potentially also classification and/or link to full text) should be analyzed by the CAI system, resulting in a list of suggested descriptors. Within the working platform (graphical user interface) of the CAI system the suggested descriptors should be sorted by their relevance for the content of the document. The CAI system should allow the subject specialist to accept or reject descriptors from the suggested list and to assign additional descriptors when necessary.

Functionality Concept

The CAI system will work on electronic bibliographic records, which are collected from publishers or produced by the descriptive cataloguing process. The records are in the INIS format (XML); data format conversion facilities have to be provided.

The records to be analyzed must carry at least a title and abstract. Additional fields such as journal title, conference title, etc. can be used to find additional suggested descriptors and/or for automatic assignment of a subject category. Full text analysis can be envisaged for enhanced subject analysis, but the ranked weight of descriptors from the full text must be lower than those from title and abstract.

In the analysis process semantic/linguistic methods will be applied for word stemming and the character set will be simplified. Special attention will be necessary for mathematical and chemical formulas and the identification of isotopes (Picture 1). Context sensitive rules should be applied, i.e. that a certain word/phrase gives a suggested descriptor only if there are other certain words/phrases in the adjacent context or if there is a match with a certain subject category.

Within the working platform (graphical user interface) of the CAI system the suggested descriptors should be sorted by importance (by their relevance for the content of the document). The assignment of weights to the suggested descriptors will be necessary, based on the field where selected, usage count in the Thesaurus, interrelation with

подобрать дополнительные ключевые слова и автоматического назначения предметной категории. Для улучшения предметизации может быть проведен анализ всего текста документа, но вес ключевых слов из основного текста, должен быть ниже веса ключевых слов из названия и краткого описания.

Для проведения морфологического поиска в процессе анализа применяются семантические и лингвистические методы, упрощается набор используемых символов. Особое внимание уделяется математическим и химическим формулам и идентификации изотопов. (Рис.1)

other descriptors found in the document, frequency of occurrence in the document, consistency with classification and other factors. These weights will allow the system to rank (sort) the suggested descriptors by their relevance to the subject content of the document, so that the subject specialist sees the most important suggested descriptors first. The subject specialist clearly sees the highlighted context from which the terms were selected (Picture 2). The system allows the subject specialist to accept or reject descriptors from the suggested list and to assign additional descriptors from the Thesaurus when necessary (Picture 3).

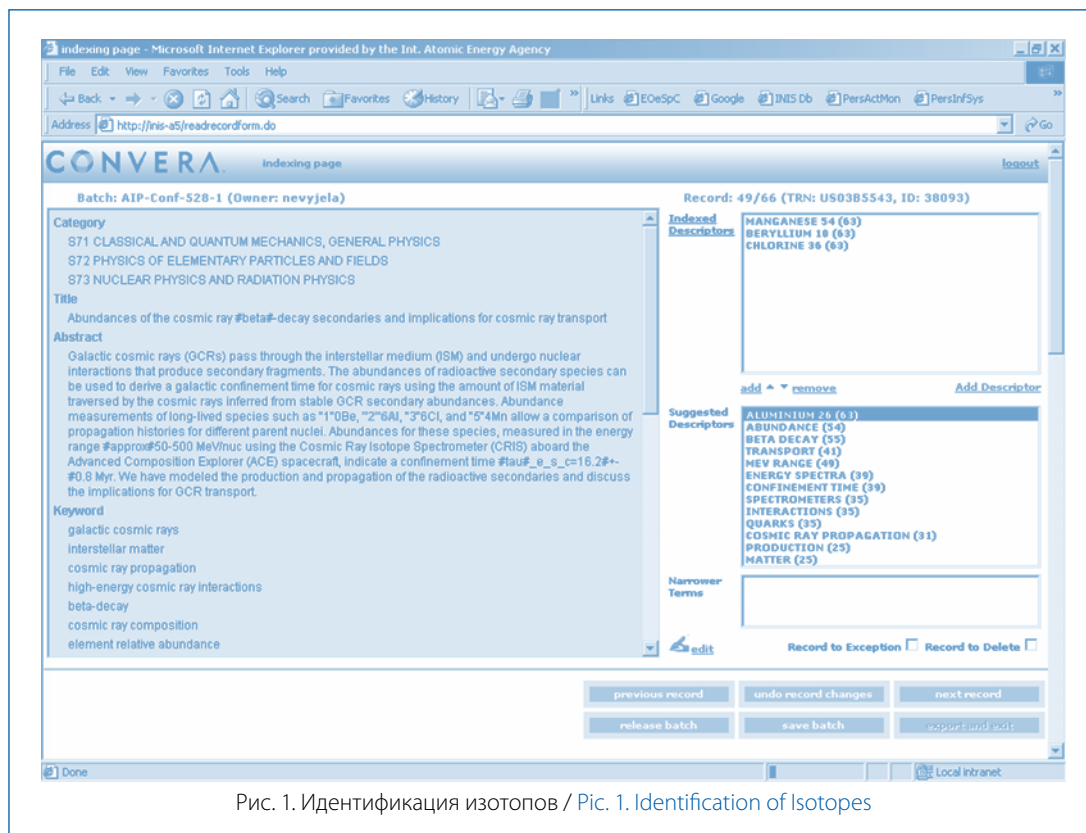


Рис. 1. Идентификация изотопов / Pic. 1. Identification of Isotopes

Используются правила чувствительности к содержанию, т. е. слово / фраза рассматриваются в качестве предлагаемого дескриптора только, если в тексте встречаются синонимы или если оно совпадает с наименованием предметной категории.

При выводе пользователю (через графический интерфейс) дескрипторы сортируются по важности (релевантности содержанию документа). Рассчитываются веса предложенных де-

Software Evaluation

Based on this functionality concept, specifications for software providers considered to be candidates were worked out and a standard test package has been prepared for a coordinated evaluation of software products. Eight software packages from different software providers have been analysed and evaluated in terms of their suitability and performance in Computer-assisted Indexing and thesaurus maintenance.

скрипторов с учетом выбранных полей записи, вхождения в тезаурус, взаимосвязи с другими найденными в документе дескрипторами, частотой вхождения в документ, последовательности классифицирования.

The evaluation of the software packages showed that semantic/linguistic analysis (elimination/transcription of document- and character-formats, sub- and superscripts, special characters, etc. resulting in pure ASCII terms and phrases, and elimination

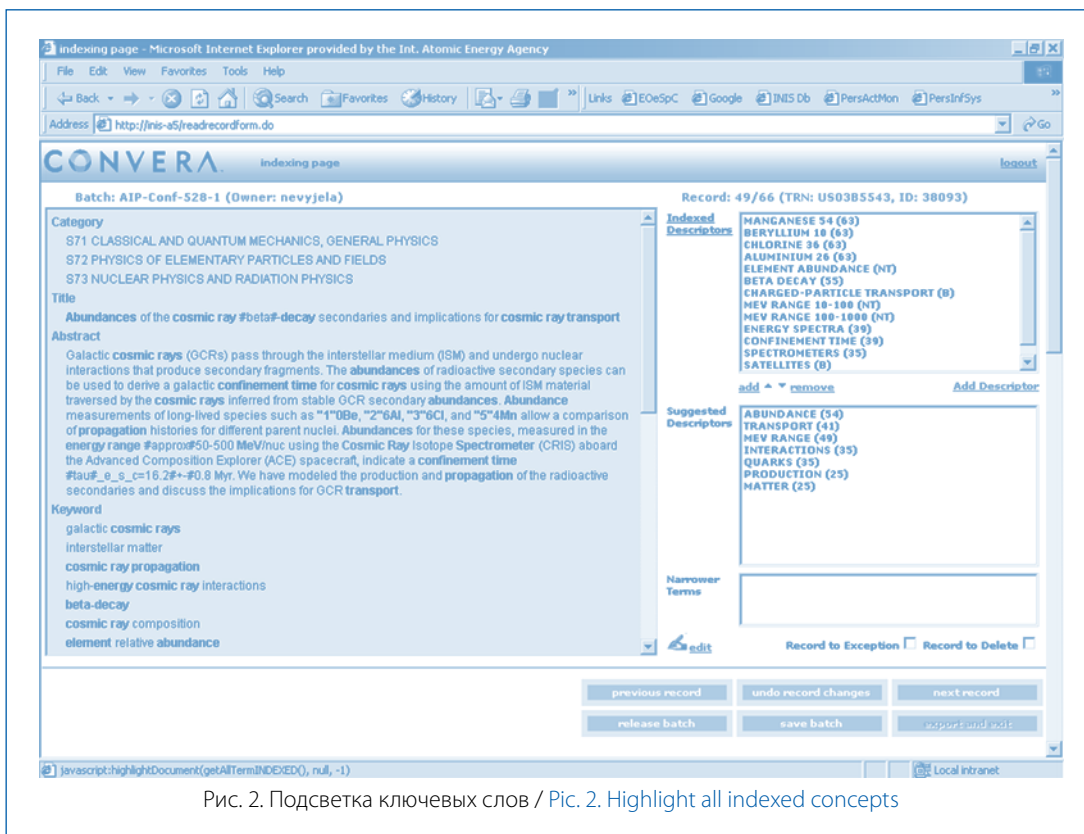


Рис. 2. Подсветка ключевых слов / Pic. 2. Highlight all indexed concepts

Наличие веса дескрипторов позволяет системе ранжировать их по релевантности содержанию документа так, что самые важные дескрипторы будут в начале списка. Выбранные дескрипторы подсвечиваются в тексте документа. (Рис. 2). Система позволяет добавлять или удалять дескрипторы из списка и задавать дополнительные дескрипторы из тезауруса. (Рис. 3).

Сравнение программного обеспечения для системы КИ

Создана тестовая база данных для объективной оценки предлагаемого ПО, проанализированы и оценены 8 программных пакетов от различных разработчиков. В первую очередь учитывались их работоспособность, возможность автоматизированной индексации и работы с тезаурусом.

of grammatical variations of terms and phrases) is state-of-the-art and provided by all investigated software packages.

Context sensitive rules are also provided more or less by all software packages, based on statistical text analysis methods and fuzzy logic, in some advanced cases supported by a logical rule base.

Pattern recognition methods (names/entities finder, supported by Boolean logic combined with context operators) however are only implemented in a few of these software packages. These methods are necessary to identify isotopes, elementary particles, chemical compounds, etc., which is essential for the subject analysis of scientific literature in the subject scope of INIS.

Thesaurus maintenance is included in most of the investigated software packages; however the

Оценка представленного ПО показала, что семантический и лингвистический анализ реализован на очень высоком уровне всеми разработчиками. Правила чувствительности к содержанию в большей или меньшей степени также реализованы во всех представленных пакетах ПО.

user-friendliness shows significant differences in the software packages investigated.

The functionality concept together with the specifications, process and workflow descriptions worked out by the INIS Secretariat formed the basis for a request for quotations sent by the Agency's

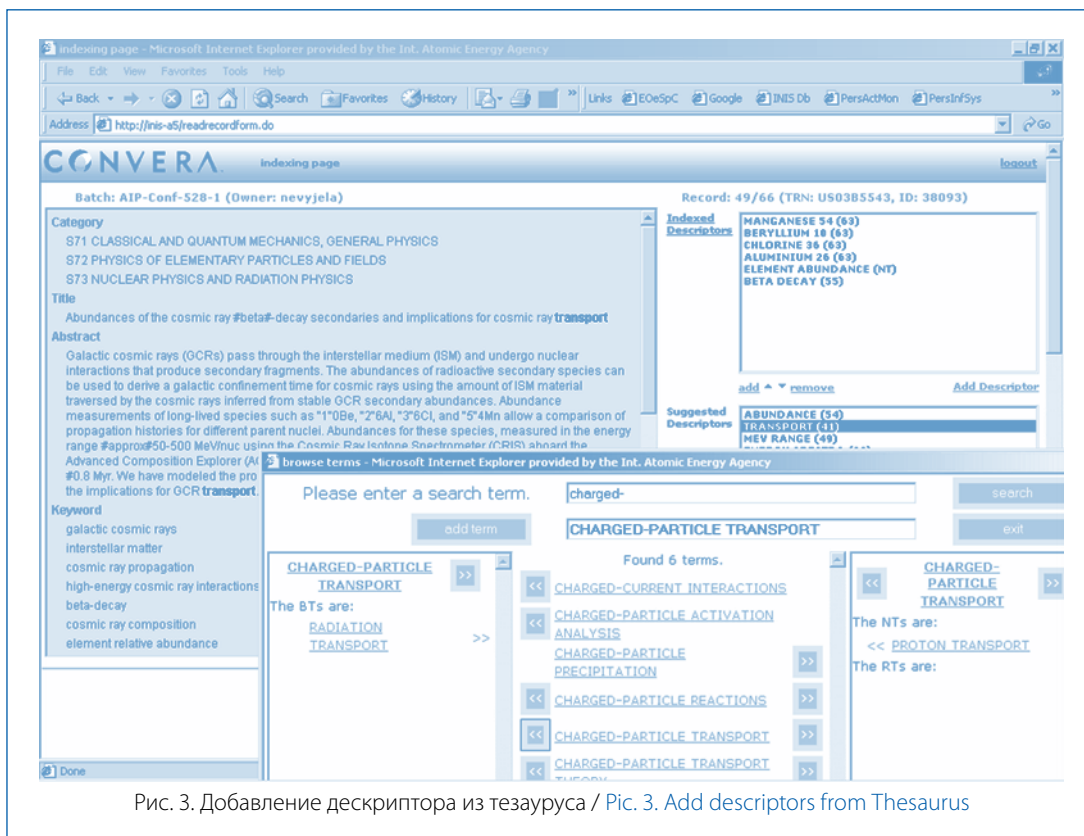


Рис. 3. Добавление дескриптора из тезауруса / Pic. 3. Add descriptors from Thesaurus

Методы распознавания по образцу (поиск по названию или заголовку, с использованием Булевой логики и совмещённый с контекстными операторами) реализованы только в нескольких пакетах ПО. Данные методы поиска необходимы для работы с изотопами, элементарными частицами, химическими соединениями, которые важны для предметизации документов в базе INIS. Работа с тезаурусом доступна в большинстве из представленных пакетов программ; однако степень дружелюбности пользовательского интерфейса существенно различается.

Были отобраны три компании (Data Harmony, Convera и RecomMind), из которых на основе оценки ряда критериев была выбрана компания Convera AG (Швейцария).

Procurement and Supply Section to three short-listed software suppliers: Data Harmony, Convera and RecomMind.

Based on the quotations from these three suppliers the software from Convera (Convera AG, Wil, Switzerland) has been chosen.

Implementation

Draft concepts for system interfaces (Pictures 1...3), protocols and data formats, the user interfaces, thesaurus handling and the project plan with time table and milestones, were developed in close cooperation with Convera and approved by the INIS Secretariat.

The first test version of the CAI system was installed on the servers of the INIS Secretariat in May

Реализация

Фирмой Convera разработаны схемы интерфейсов системы (Рис. 1...3), протоколы и форматы данных, варианты пользовательского интерфейса, настройки тезауруса; согласованы план, расписание, основные этапы работ над проектом. Вся документация утверждена секретариатом INIS.

Пилотный вариант системы установлен на серверах секретариата INIS в мае 2004 года. С мая по июнь проводились тренинги персонала по работе с системой индексации, тезаурусом и общим администрированием системы. Затем работа продолжилась с новой бета-версией системы. Система получила положительную оценку в секретариате INIS, и в августе 2004 года была окончательно установлена версия 1.0 системы.

Для идентификации дескрипторов в тексте (название, краткое содержание, ключевые слова) выявилась необходимость во внедрении «скрытых терминов», которые определяют фразы или части текста и указывают дескриптор, который может быть предложен. Скрытая информация используется только для работы системы индексирования и не выводится пользователю.

Разработаны ряд правил, позволяющих упростить индексацию и повысить производительность системы.

Расширение тезауруса INIS до терминологической базы знаний привело к увеличению словаря более чем на 60000 терминов (21147 дескрипторов и 38903 «скрытых» терминов).

Разработана улучшенная версия системы КИ 1.10 со следующими технологическими улучшениями: улучшение индексирования и тезауруса, повышение эргономики, дополнительные возможности контроля (статистика, управление технологическим потоком). Система одобрена и установлена в апреле 2005 года.

С июня 2004 года по сентябрь 2006 обработано 135923 записи (администрированием системы занимаются лишь 5 сотрудников INIS). По предварительным оценкам, производительность системы предметизации может повыситься более чем на 100 %.

2004. Training on CAI indexing, thesaurus and taxonomy handling and system administration was provided to INIS staff during May and June. After extensive tests the beta-version was installed in June 2004 and was used for production indexing from then on. The results of the tests was analysed, verified and approved by the subject specialists at the INIS Secretariat. The final acceptance of the CAI system Version 1.0 was in August 2004.

To support the identification of descriptors in the free text (title, abstract, free keywords) the introduction of "hidden terms" was necessary, which identify phrases or character strings of free text and point to the valid descriptor, which should be suggested. Hidden terms are character patterns representing the different appearances of a concept in the free text, which is indexed by one or more descriptors. These "hidden terms" are CAI internal only, they will not be printed in any appearance of the thesaurus but handled in the background of the CAI process and of the planned new search engine to facilitate the identification of concepts in free text for indexing and retrieval.

INIS specialists worked out some rules enabling to simplify indexing and to raise productivity of the system.

Expansion of the INIS Thesaurus to a terminological knowledge base has resulted in a vocabulary of more than 60 000 terms (21 147 valid descriptors and 38 903 forbidden and "hidden terms").

For the second project phase CAI 1.10 further developments and upgrades of this system have been implemented, i.e. handling improvements (indexing and thesaurus browsing, ergonomics), monitoring improvements (statistics, workflow management) and improvements in indexing capabilities. The CAI software Version 1.10 was approved by the INIS Secretariat in April 2005.

During the period June 2004 — September 2006 a total of 135 923 records have been processed (indexed) with the Computer-assisted Indexing System by the five subject specialists of the INIS Secretariat. First experiences show, that a performance enhancement of more than 100 % can be achieved in the subject analysis process.