**TẠP CHÍ THÔNG TIN VÀ TƯ LIỆU**
**Journal of Information and Documentation**
ISSN 1859-2929

**Số 1/2005**
**No. 1/2005**

# Tìm tin theo từ khoá trong các cơ sở dữ liệu thư mục của Trung tâm Thông tin KHCN Quốc gia

**Ths. Nguyễn Thị Đào**
*Trung tâm Thông tin KHCN Quốc gia*

***Tóm tắt:*** *Nhận dạng một số tồn tại trong việc sử dụng các từ khoá để tìm tin tại các CSDL của Trung tâm Thông tin KHCN Quốc gia. Giới thiệu kết quả tìm tin và nêu sự phụ thuộc của độ đầy đủ tìm tin vào từ khoá được sử dụng. Đưa ra các kiến nghị để tiếp tục hoàn thiện các từ khoá để phục vụ việc tìm tin.*

Khi vào tìm tin trong các CSDLTM, từ khoá là một trong những điểm truy cập phổ biến và hữu hiệu nhất. Người dùng tin có thể tra tìm theo nhiều điểm truy cập khác nhau như: Tác giả, nhan đề, ký hiệu phân loại, từ khoá, ký hiệu kho,... . Chất lượng từ khoá quyết định rất nhiều đến hiệu quả tìm tin. Nếu từ khoá định không chính xác hoặc không thống nhất sẽ gây mất tin và nhiễu tin, vì thế, tới nay, các CSDL của Trung tâm đã sử dụng từ khoá kiểm soát (TKKS) theo Bộ từ khoá KHKT đa ngành. Bài viết này sẽ trình bày việc tìm tin theo từ khoá tại Trung tâm Thông tin KHCN Quốc gia.

## I. Một số tồn tại trong việc sử dụng từ khoá

TTTTKH&CNQG đã tiến hành xây dựng CSDLTM vào cuối những năm 80. Các CSDLTM này phản ánh vốn tài liệu có trong các kho của Trung tâm và cũng là nguồn tra cứu quan trọng của người dùng tin khi muốn tiếp cận tài liệu của Trung tâm. Nhìn chung, chất lượng từ khoá trong các CSDLTM đã được Trung tâm rất chú trọng và càng ngày càng có nhiều biện pháp để chuẩn hoá nhằm nâng cao hiệu quả tìm tin. Tuy nhiên do việc sử dụng từ khoá trải qua nhiều giai đoạn khác nhau nên hiện tại trong các CSDLTM của Trung tâm đang tồn tại một số vấn đề như:

- Không thống nhất về mặt thuật ngữ: trong một CSDL, cùng một khái niệm có thể sử dụng các thuật ngữ khác nhau do trước đây sử dụng từ khoá tự do (TKTD), từ 1996 sử dụng TKKS, Ví dụ: Vật liệu composit (TKTD) - Vật liệu tổ hợp (TKKS); Đạo hàng (TKTD) - Dẫn đường (TKKS); Vật liệu kết dính (TKTD) - Chất kết dính (TKKS); Các nước đang phát triển (TKTD) - Nước đang phát triển (TKKS).

- Các cụm từ hoặc thuật ngữ lúc thì được tách thành những từ đơn/khái niệm đơn giản, lúc thì để nguyên. Ví dụ: Thép cacbon bền nhiệt (TKTD) - Thép cacbon%Thép bền nhiệt (TKKS); Ung thư phổi (TKTD) - Ung thư%Phổi (TKKS); Chế biến thực phẩm (TKTD)- Thực phẩm%Chế biến (TKKS); Sản xuất đường mía (TKTD) - Đường mía%Sản xuất (TKKS); Cấp thoát nước (TKTD) - Cấp nước%Thoát nước.

- Không thống nhất cách viết về chữ số, từ gốc nước ngoài được Việt hoá,… Ví dụ: Thế kỷ XX - Thế kỷ 20; Bêtông – Bê tông; Ôtô- Ô tô; Oxi hoá – Oxy hoá.Như vậy, về mặt nội dung, từ khoá trong các CSDLTM của Trung tâm đang có nhiều thuật ngữ được viết theo nhiều cách khác nhau, từ đó gây nên sự mất tin, thiếu chính xác trong tìm tin. Sau khi có Bộ từ khoá, việc thống nhất cách viết các từ khoá có tốt hơn. Với sự phát triển của các ngành khoa học, năm 2004 Trung tâm đã cho bổ sung, cập nhật thêm từ khoá và in thành Từ điển Từ khoá khoa học và công nghệ. Như vậy, ở góc độ xử lý trong các CSDL của Trung tâm sẽ tồn tại ba giai đoạn sử dụng từ khoá khác nhau:

+ Giai đoạn từ đầu cho đến 1996: sử dụng TKTD;

+ Giai đoạn từ 1996 đến 2004: sử dụng từ TKKS theo Bộ Từ khoá đa ngành khoa học tự nhiên và công nghệ, xuất bản năm 1996 (TKKS 1996);

+ Giai đoạn từ 2004 trở đi: sử dụng TKKS theo Từ điển Từ khoá khoa học và công nghệ, xuất bản năm 2004 (TKKS 2004).

Nhưng điều đáng nói ở đây là trong bộ TKKS 2004 có rất nhiều thuật ngữ đã được chỉnh sửa theo hướng không tách nhỏ làm cho từ khoá trở nên thông dụng, chính xác hơn, thân thiện với người dùng tin hơn. Nhưng khác với bộ TKKS 1996, hầu hết những từ khoá này lại trùng hợp với cách đánh TKTD trước đây. Ví dụ: Hệ thống điều khiển (TKTD) - Hệ điều khiển (TKKS 1996) - Hệ thống điều khiển (TKKS 2004); Xuất nhập khẩu (TKTD) - Xuất khẩu%Nhập khẩu (TKKS 1996) - Xuất nhập khẩu (TKKS 2004).

Ngoài ra, có một số thuật ngữ trong TKKS 2004 được sử dụng khác với TKKS 1996 do thêm hoặc lại bỏ đi chữ "học". Ví dụ: Sinh lý người (TKKS 1996)- Sinh lý học người (TKKS 2004); Hoá học hữu cơ (TKKS 1996) - Hoá hữu cơ (TKKS 2004).

Về mặt hình thức từ khoá:

Trong các CSDLTM đều mắc phải các lỗi chính tả tiếng Việt như:

- Đánh máy sai lỗi chính tả. Ví dụ: Mage (Magie), An Độ (ấn Độ); Châu A (Châu á).
- Không thống nhất giữa "y" và "i". Ví dụ: Kĩ thuật và Kỹ thuật; Xử lí nhiệt và Xử lý nhiệt.
- Đặt dấu không thống nhất (do thời kỳ đầu dùng phông chữ "vnload" sau này đổi sang ".VnTime". Ví dụ: Hoá học và Hóa học; Thủy nhiệt và Thuỷ nhiệt; Tiêu hóa và Tiêu hoá,.

**II. Hiệu quả tìm tin theo từ khoá**

Để đánh giá hiệu quả tìm tin theo từ khoá, chúng tôi đã lấy 4 ví dụ tìm, trong đó, mỗi ví dụ đều được tìm theo thuật ngữ chính xác (phương pháp tìm đơn giản nhất), cụ thể là bằng từng từ khoá, chưa có toán tử và đánh giá theo tính đầy đủ của cuộc tìm. Phương pháp đánh giá độ đầy đủ của tài liệu được dựa vào công thức: R (recall) =A/A+C (trong đó A: là tài liệu phù hợp được tìm ra; C: Tài liệu phù hợp không được tìm ra).

Kết quả 4 cuộc tìm trong CSDL BOOK được trình bày trên bảng dướI đây:

| Cuộc tìm | Khoá tìm (từ khoá) | A | A + C | R |
|---|---|---|---|---|
| 1 | Hệ điều khiển | 147 | 578 | 25,4% |
| | Hệ thống điều khiển | 431 | 578 | 74,6% |
| 2 | Hóa hữu cơ | 396 | 607 | 65,2% |
| | Hoá hữu cơ | 25 | 607 | 4,2% |
| | Hoá học hữu cơ | 27 | 607 | 4,4% |
| | Hóa học hữu cơ | 159 | 607 | 26,2% |
| 3 | Hóa lập thể | 21 | 40 | 52,5% |
| | Hoá lập thể | 2 | 40 | 5% |
| | Hóa học lập thể | 16 | 40 | 40% |
| | Hoá học lập thể | 1 | 40 | 2,5% |
| 4 | Sinh lý người | 24 | 30 | 80% |
| | Sinh lý học người | 6 | 30 | 20% |

Rõ ràng, nếu không thống nhất về mặt thuật ngữ sẽ bị mất tin rất nhiều, thậm chí có những cuộc tìm độ đầy đủ chỉ còn 2,5%, có nghĩa là 97,5% số tài liệu phù hợp với yêu cầu tin đã không được tìm thấy.

**III. Kết luận và kiến nghị**

Tìm tin theo từ khoá là khá hiệu quả và chất lượng cuộc tìm phụ thuộc vào chất lượng từ khoá. Tuy nhiên, hiện nay hầu như ở các CSDL của các trung tâm thông tin và thư viện nói chung đều đang tồn tại một số vấn đề về từ khoá như: thuật ngữ dùng không thống nhất, các từ gốc nước ngoài được Việt hoá không giống nhau, dấu tiếng việt để không đúng, lỗi chính tả,....Những điều này sẽ ảnh hưởng rất nhiều đến hiệu quả tìm tin. Bởi vậy, điều quan trọng mà chúng tôi muốn đề cập đến trong bài báo này là:

- Nên định kỳ xem xét để phát hiện và khắc phục những cái sai, cái không thống nhất về từ khoá trong các CSDL;
- Cần dựa vào công cụ kiểm soát để định từ khoá cho thống nhất trong tất cả các CSDL ít nhất trong phạm vi của một cơ quan và lý tưởng là trong cùng một hệ thống thông tin thư viện chuyên ngành, ví dụ như hệ thống thông tin KH&CN, hệ thống thông tin thuỷ sản, hệ thống thông tin thư viện đại chúng,...
- Khi xây dựng công cụ kiểm soát từ khoá, cần có sự tham gia thực sự của các chuyên gia xử lý. Có như vậy các công cụ kiểm soát mới có giá trị vì các thuật

ngữ được sử dụng trong đó luôn có độ ổn định, thông dụng và phù hợp với xu hướng phát triển của các ngành khoa học, thân thiện với người dùng tin.

- Trong quá trình xứ lý phải luôn lựa chọn từ khoá phản ánh đúng với chủ đề nội dung tài liệu. Nếu thuật ngữ đó chưa có trong công cụ kiểm soát ta cần xem xét cẩn thận vì giữa ngôn ngữ tự nhiên (được dùng trong tài liệu) và ngôn ngữ tư liệu (ở đây là ngôn ngữ từ khoá) có sự khác nhau. Nhưng nếu thực sự là thiếu thì chúng ta phải đưa vào CSDL để sau cập nhật vào công cụ kiểm soát. Có như vậy các bộ từ khoá mới không bị lạc hậu so với sự phát triển của khoa học;
- Với cấu trúc của khổ mẫu MARC21, từ khoá được đưa vào nhiều trường khác nhau trong khối 6XX: Các trường truy cập chủ đề. Trong đó có trường 650 (Thuật ngữ chủ đề có kiểm soát) dùng để điền những thuật ngữ có trong công cụ kiểm soát (như khung đề mục chủ đề/từ điển từ chuẩn), trường 651 (Chủ đề địa danh có kiểm soát) và trường 653 (Thuật ngữ chủ đề không kiểm soát) để ghi những từ khoá không được lấy ra từ một công cụ kiểm soát nào. Trong đó cấu trúc của các trường 650, 651 có rất nhiều trường con như: $aThuật ngữ chính, $v Đề mục con hình thức, $x Đề mục con chung, $y Đề mục con thời gian, $z Đề mục con địa lý. Trong khổ mẫu này, giữa các từ khoá sẽ có mối quan hệ ngữ cảnh, chính phụ rõ ràng, phản ánh đúng chủ đề nội dung tài liệu và tránh được sự gây nhiễu trong tìm tin với điều kiện là chúng ta phải nắm rất vững phương pháp định từ khoá và có sự trợ giúp của phần mềm máy tính.

Với những công cụ kiểm soát ngôn ngữ từ khoá ngày càng được chuẩn xác và với sự trợ giúp của các phần mềm mạnh cùng với những kinh nghiệm xử lý của các cán bộ thông tin thư viện, từ khoá sẽ thật sự trở thành một trong những công cụ hữu hiệu nhất giúp chúng ta tìm tin dễ dàng, linh hoạt và hiệu quả trong các CSDL, nhất là trong môi trường nối mạng Internet.

**Tài liệu tham khảo**


1. Bộ từ khoá đa ngành khoa học tự nhiên và công nghệ. - H., 1996
2. Từ điển từ khóa khoa học và công nghệ. – H., 2004
3. Đoàn Phan Tân. Thông tin học. – H., 2001



**Information retrieval by keywords in NACESTI's bibliographic databases / Nguyen Thi Dao // J. of Information and Documentation. - 2005, N.1. – pp**



**Nguyen Thi Dao**



**Abstracts:** Identifies some problems in using keywords for accessing information in NACESTI's bibliographic databases; Presents the results of information retrieval and indicates the dependence of information recall on used keywords; Puts forth some recommendations on further improving the controlled keywords for information retrieval.