MỘT SỐ Ý KIẾN VỀ VIỆC XÂY DỰNG BỘ ĐỀ MỤC CHỦ ĐỀ TIẾNG VIỆT

TS. Nguyễn Thu Thảo
Trung tâm Thông tin KHCNQG

1. MỞ ĐẦU

Trong khoảng 10 năm vừa qua, ngành thư viện nước ta đã có nhiều tiến bộ đáng kể, trên phương diện nhận thức cũng như phương diện hoạt động thực tế.

Nói riêng về việc xây dựng các phương tiện ngôn ngữ tư liệu, thì cách đây 5-10 năm trở về trước, cùng với việc áp dụng tin học và ào ạt xây dựng các CSDL ở hầu hết các cơ quan TTTV, chúng ta chỉ dám nghĩ đến việc biên soạn các phương tiện kiểm soát từ khóa để tạo điều kiện tìm tin hiệu quả hơn so với các hệ thống mục lục tra cứu thủ công trước đó. Nhưng gần đây, việc giao lưu và hợp tác với các nước tiên tiến đã thúc đẩy nhu cầu về trao đổi thông tin, tương hợp với các hệ thống thông tin thư mục của các nước, và nâng cao hơn nữa chất lượng tìm tin. Và từ đó, mong muốn tạo lập những phương tiện ngôn ngữ tư liệu mạnh hơn đã trở nên rõ ràng và khả thi. Việc giao lưu quốc tế ngày càng phát triển cũng góp phần hé mở những khả năng với tới những nguồn tài trợ kinh phí cho công việc này.

Những điều kiện nói trên vừa là tạo cơ hội, vừa là tạo thách thức cho chúng ta, khiến chúng ta phải nghĩ đến bộ ĐMCĐ, một phương tiện ngôn ngữ tư liệu mạnh hơn, với hiệu quả cao hơn so với trước.

2. SỰ TƯƠNG QUAN VỀ CHỨC NĂNG CỦA CÁC NGÔN NGỮ TƯ LIỆU: ĐMCĐ, TỪ KHÓA VÀ PHÂN LOẠI

Bộ đề mục chủ đề (ĐMCĐ) là một phương tiện ngôn ngữ tư liệu mạnh, cho phép tìm tin với hiệu quả cao. Nguyên nhân chủ yếu của tính ưu việt này là do trong mỗi bộ ĐMCĐ có sẵn sự kết hợp giữa từng đối tượng nghiên cứu của các ngành khoa học với các phương diện nghiên cứu thường gặp trong các tài liệu khoa học, tạo thành các ĐMCĐ, dùng làm các chỉ mục để mô tả. Khi thực hành

việc mô tả tài liệu, nếu trong bộ ĐMCĐ có chứa ĐMCĐ trùng hợp với chủ đề của tài liệu, thì ta có thể áp dụng ngay ĐMCĐ đó làm chỉ mục (index) cho tài liệu đó. Nếu không, ta có thể tạo một ĐMCĐ mới, với sự kết hợp theo quy tắc đã định. Những ĐMCĐ mới này sẽ được tập hợp lại theo định kỳ để xem xét và cập nhật vào bộ ĐMCĐ.

Theo nguyên tắc nói trên, nếu so sánh ĐMCĐ với từ khóa, thì có thể nói rằng với chức năng mô tả nội dung tài liệu và tìm tin trong một hệ thống tìm tin thì từ khóa là một phương tiện kém hiệu quả hơn so với ĐMCĐ. Khi mô tả tài liệu, so với ĐMCĐ, thì từ khóa không thể hiện được sự kết hợp nói trên (giữa đối tượng nghiên cứu và phương diện nghiên cứu). Mỗi từ khóa (mô tả đối tượng nghiên cứu hoặc phương diện nghiên cứu) là một chỉ mục độc lập. Trong quá trình tìm tin, việc sử dụng các toán tử logic để kết hợp các chỉ mục nói trên có thể đưa đến sự kết hợp khác với nội dung trong tài liệu gốc, gây ra độ nhiễu tin lớn hơn.

Tuy nhiên, so với một khung phân loại thì tình hình lại không phải như vậy. Việc phân loại tạo khả năng đáp ứng những tình huống tìm tin khác so với ĐMCD hoặc từ khóa. Việc tìm các tài liệu về một đối tượng nghiên cứu nào đó có thể dễ dàng thực hiện bằng ĐMCĐ hoặc từ khóa, nhưng lại không dễ thực hiện bằng ký hiệu phân loại. Bởi vì, nguyên tắc thông thường của khung phân loại là hệ thống hóa nội dung các môn ngành khoa học phân cấp hình cây, trong đó một đối tượng nghiên cứu có thể được phân chia vào nhiều ngành khác nhau, với những ký hiệu phân loại rất khác nhau. Qua đó cũng có thể thấy rằng trong một tình huống tìm tin khác: chẳng hạn là một yêu cầu về hệ thống hóa tài liệu theo môn ngành khoa học, thì có thể thực hiện dễ dàng bằng ký hiệu phân loại, nhưng lại không dễ thực hiện bằng ĐMCĐ hoặc từ khóa.

Như vậy, có thể nói rằng phân loại và ĐMCĐ là 2 ngôn ngữ tư liệu có thế mạnh và nhược điểm ngược nhau, nhưng lại có thể bổ sung cho nhau một cách hữu hiệu, cho phép thỏa mãn những tình huống tìm tin khác nhau, thường gặp trong thực tế. Còn từ khóa thì có chức năng đáp ứng các tình huống tìm tin tương tự như chức năng của ĐMCĐ, nhưng yếu kém hơn về độ chính xác tìm tin. Cho nên, nếu ta có một bộ ĐMCĐ tốt, thì không cần dùng từ khóa nữa.

Trong những năm qua, chúng ta sử dụng từ khóa như một giai đoạn trung gian, khi chưa đủ trình độ, thời gian cũng như tiềm lực tài chính để biên soạn một bộ ĐMCĐ. Tới nay, đã đến lúc có thể tiến lên một giai đoạn mới, thì những bộ từ khóa đã được biên soạn trước đây sẽ được coi như đã hoàn thành nhiệm vụ. Để tiến vào giai đoạn mới, những bộ từ khóa cũ có thể được sử dụng như những nguồn cung cấp từ vựng phong phú và thiết thực cho các bộ ĐMCĐ.

Trong cấu trúc MARC21 có nhóm trường 6XX là nhóm trường dành cho các điểm truy cập nội dung tài liệu theo ĐMCĐ. Trong đó, sự bố trí các trường con và tính lặp của chúng đã thể hiện rất rõ sự kết hợp giữa các đối tượng nghiên cứu với các phương diện nghiên cứu của các đối tượng nghiên cứu đó. Nếu chúng ta áp dụng tốt cấu trúc này, chúng ta sẽ tận dụng tối đa sức mạnh của nó cho chất lượng tìm tin. Ngược lại, nếu áp dụng không tốt, ví dụ, dùng từ khóa để điền vào nhóm trường này (một số nơi đang làm như vậy), thì hiệu quả sẽ bị hạn chế nhiều.

3. YÊU CẦU ĐỐI VỚI MỘT BỘ ĐMCĐ

Một bộ ĐMCĐ có chất lượng cao sẽ phải đảm bảo các yêu cầu tối thiểu sau đây:

a/ Bao quát được tương đối đầy đủ các đối tượng nghiên cứu thuộc nhiệm vụ của một thư viện, hoặc một hệ thống thư viện. Các đối tượng đó phải được diễn đạt sao cho đáp ứng các yêu cầu của hoạt động tra cứu;

b/ Tạo được sự kết hợp giữa các đối tượng nghiên cứu nói trên với các phương diện nghiên cứu thường gặp, lập ra các ĐMCĐ mẫu;

c/ Thiết lập được 3 loại quan hệ ngữ nghĩa chủ yếu giữa các ĐMCĐ (quan hệ tương đương quy ước, quan hệ phân cấp và quan hệ liên đới);

d/ Được tổ chức và trình bày với hình thức thân thiện với người sử dụng.

4. VỀ VIỆC BIÊN SOẠN MỘT BỘ ĐMCĐ TIẾNG VIỆT

Để biên soạn một bộ ĐMCĐ tiếng Việt, trước hết, cần tham khảo kỹ lưỡng ***Tiêu chuẩn quốc tế ISO 2788-1986 "Hướng dẫn biên soạn và phát triển thesauri đơn ngữ"***, trong đó bao gồm các phương pháp tiến hành các giai đoạn:

thu thập và lựa chọn vốn từ, xử lý từ vựng, kiểm soát từ, xây dựng các quan hệ ngữ nghĩa và trình bày sao cho thân thiện với người dùng.

Trong số các giai đoạn được đề cập trong ISO này, đáng lưu ý nhất là 2 giai đoạn: Thu thập, lựa chọn vốn từ, và xử lý từ vựng. Dưới đây xin giới thiệu một số nét chính về 2 giai đoạn này.

a/ Về việc thu thập và lựa chọn vốn từ, ISO 2788 có hướng dẫn 2 phương thức thực hiện công đoạn này, gồm: Quy nạp (inductive) và Suy diễn (deductive).

Theo phương thức suy diễn, vốn từ được thu thập từ hoạt động định chỉ mục trên thực tế mà không qua kiểm soát từ. Khi vốn từ đủ lớn, các nhóm chuyên gia sẽ xem xét, chọn lọc vốn từ này rồi mới tiến hành các công việc tiếp theo (sắp xếp, lập quan hệ ngữ nghĩa, trình bày, v.v.)

Theo phương thức quy nạp, từ vựng được kiểm soát ngay từ đầu, đưa vào danh mục ngay khi nó xuất hiện trong tài liệu và được sắp xếp theo các nhóm quan hệ ngữ nghĩa ngay.

Trên thực tế, người ta thường áp dụng kết hợp cả 2 phương thức trên, lần lượt theo từng giai đoạn. Tuy nhiên, với điều kiện thực tế hiện nay của nước ta, theo những hình dung về tiết kiệm thời gian, công sức và kinh phí, thì phương thức suy diễn nên được áp dụng ở giai đoạn đầu tiên. Trong giai đoạn này, có thể sử dụng các bộ ĐMCĐ và các bộ từ khóa đã được xây dựng ở những hệ thống thông tin khác nhau (kể cả của nước ngoài) để làm nguồn cung cấp từ vựng ban đầu.

Trong số các nguồn cung cấp từ vựng tốt nhất, có thể kể đến LCSH, RAMEAU, hoặc SEARS, là những bộ từ vựng có diện bao quát tổng hợp và các quan hệ ngữ nghĩa phong phú. Nhưng cho dù chúng ta sử dụng nguồn nào đi nữa, thì sau đó ta vẫn phải xem xét lại, và chỉ chọn lọc những ĐMCĐ nào thuộc phạm vi nhiệm vụ của ta mà thôi. Chính vì vậy, nếu ta chọn được những bộ ĐMCĐ của những hệ thống càng tương đồng về chức năng với hệ thống của ta thì càng có lợi cho ta. Nếu có nhiều bộ ĐMCĐ phù hợp với ta về nội dung, thì việc chọn bộ nào còn có thể phụ thuộc vào khả năng tài trợ của tổ chức hoặc quốc gia có bản quyền.

Đó cũng là một điều hợp lý, vì việc xây dựng một bộ ĐMCĐ đòi hỏi rất nhiều công sức, thời gian và kinh phí.

b/ Về việc xử lý từ vựng

Việc sử dụng một bộ ĐMCĐ của nước ngoài làm nguồn cung cấp từ vựng cho bộ ĐMCĐ tiếng Việt được coi là một giải pháp khả thi và tiết kiệm. Tuy nhiên có một khó khăn đáng kể ở đây là việc dịch các ĐMCĐ từ tiếng nước ngoài sang tiếng Việt.

Việc dịch này hoàn toàn không giống như dịch một văn bản bằng ngôn ngữ tự nhiên thông thường. Đây là việc dịch các đơn vị từ vựng của ngôn ngữ nhân tạo, với những quy định chặt chẽ, nhằm đạt tới một mục tiêu hết sức riêng biệt, đó là tìm tin với hiệu quả cao. Để đạt được mục tiêu này, cần tham khảo, nắm vững và áp dụng các hướng dẫn trong ***ISO 5964-1986 "Hướng dẫn biên soạn và phát triển thesauri đa ngữ"***. ISO này đề cập đến các trường hợp tương quan ngữ nghĩa phức tạp giữa các ngôn ngữ tự nhiên và phương thức xử lý các trường hợp đó để đáp ứng các yêu cầu của một ngôn ngữ nhân tạo.

Tuy những quy định mà ISO 5964 hướng dẫn là những quy định quan trọng nhất cho tất cả mọi ngôn ngữ trên thế giới, nhưng việc áp dụng cụ thể cho từng ngôn ngữ thường có một số chi tiết riêng biệt. Vì vậy, sau khi nắm vững các quy định đó, mỗi quốc gia còn phải nghiên cứu để phát triển thêm một số điểm để thích ứng với điều kiện cụ thể theo ngôn ngữ của quốc gia đó. Việt Nam cũng không phải là ngoại lệ. Và chúng ta cũng phải giải quyết những chi tiết riêng biệt đối với tiếng Việt.

Trong việc xây dựng các phương tiện ngôn ngữ tư liệu, nếu chúng ta tuân thủ các tiêu chuẩn quốc tế ngay từ bây giờ, thì trong tương lai, chúng ta mới có cơ hội làm cho những phương tiện của chúng ta tương hợp với quốc tế, và qua đó, góp phần hữu hiệu vào việc khai thông dòng chảy thông tin xuyên quốc gia của chúng ta với thế giới.