LỰA CHỌN PHẦN MỀM VÀ KHỔ MẪU DỮ LIỆU SỐ PHỤC VỤ XÂY DỰNG SƯU TẬP SỐ PHỤC VỤ NÔNG THÔN MIỀN NÚI

Ths. Cao Minh Kiểm, Ths. Đào Mạnh Thắng
Trung tâm Thông tin KH&CN Quốc gia

Trình bày kết quả đánh giá lựa chọn phần mềm, khổ mẫu dữ liệu số và siêu dữ liệu để xây dựng sưu tập số về công nghệ kỹ thuật nông nghiệp thích hợp trong khuôn khổ một dự án ứng dụng công nghệ thông tin và truyền thông phục vụ phát triển kinh tế - xã hội nông thôn miền núi. Phần mềm thư viện số Greenstone Digital Library đã được chọn sử dụng để tạo lập và phổ biến sưu tập số về công nghệ thích hợp nông thôn.

I. Giới thiệu

Đảng và nhà nước ta đang rất quan tâm đến việc cải thiện và nâng cao chất lượng đời sống của người dân các vùng nông thôn, vùng sâu, vùng xa. Tạo điều kiện cho người dân truy cập thông tin là một trong những phương pháp giúp người dân nâng cao hiểu biết, tiếp cận với các phương pháp sản xuất kinh doanh tiên tiến, góp phần phát triển kinh tế - xã hội nông thôn, miền núi. Tuy nhiên, có một thực tế là người dân những vùng này hầu như không có điều kiện tiếp cận với các loại thông tin, đặc biệt là thông tin khoa học và công nghệ. Do đó, ứng dụng công nghệ thông tin và truyền thông (CNTT-TT) được coi là giải pháp hữu hiệu để giúp người dân có khả năng truy cập vào các nguồn tin trên. Vì người dân ở các vùng nông thôn, vùng sâu, vùng xa thường có hạn chế về năng lực sử dụng CNTT-TT nên cần phải tạo ra những công cụ dễ sử dụng, dễ tiếp cận. Trong khuôn khổ dự án "Xây dựng công trình mẫu về ứng dụng công nghệ thông tin và truyền thông nông thôn", nhóm thực hiện dự án đã xây dựng một sưu tập số về công nghệ thích hợp nông thôn để giúp người dân các vùng nông thôn tiếp cận với các nguồn tin KH&CN, phục vụ sản xuất kinh doanh của các địa phương.

Trong bài viết này, chúng tôi trình bày quá trình nghiên cứu lựa chọn và ứng dụng phần mềm thư viện số Greenstone Digital Library (GSDL) để xây dựng sưu tập số cho nông thôn.

II. Lựa chọn phần mềm

1. CDS/ISIS

CDS/ISIS là phần mềm lưu trữ và tìm kiếm thông tin văn bản do UNESCO phát triển. CSDL xây dựng bằng CDS/ISIS có thể được đưa lên mạng Internet bằng những công cụ được cung cấp bởi một số tổ chức sử dụng ISIS. Thí dụ, Thư viện Khoa học Điện tử Trực tuyến (SciELO) đã xây dựng thư viện số toàn văn các tạp chí khoa học dựa trên các công cụ ISIS được phát triển bởi Trung tâm Thông tin Y học Mỹ Latin và Ca-ri-bê (BIREME) [4]. Tổ chức Nông lương Thế giới (FAO) cũng sử dụng CDS/ISIS để tổ chức các CSDL thư mục của mình trên Web [5].

Tại Việt Nam, CDS/ISIS được sử dụng rộng rãi tại các trung tâm thông tin và thư

viện trong thập kỷ 90 của thế kỷ trước. Thí dụ, Trung tâm Thông tin KH&CN Quốc gia (NACESTI) đã xây dựng CSDL toàn văn báo cáo kết quả nghiên cứu (KQNC) bằng phần mềm CDS/ISIS dành cho Windows (phiên bản 1.52).

2. Greenstone Digital Library

Greenstone Digital Library (GSDL) là phần mềm dùng để xây dựng và phổ biến các bộ sưu tập số. GSDL được phát triển bởi Dự án Thư viện số New Zealand tại trường Đại học Waikato với sự hợp tác của UNESCO và Human Info NGO. GSDL là phần mềm nguồn mở, đa ngôn ngữ, tuân theo giấy phép sử dụng công cộng chung GNU (General Public License) [6]. Mục đích chính của phần mềm GSDL là nhằm cung cấp cho người dùng một công cụ mạnh và miễn phí để xây dựng và phổ biến các sưu tập số.

GSDL rất dễ cài đặt. Nó có thể chạy trên hầu hết các phiên bản của nhiều hệ điều hành Windows, Unix/Linux và Mac OS-X. Nhiều tổ chức trên thế giới đã sử dụng GSDL để tạo và xuất bản các bộ sưu tập trên web. Nó cũng có thể được sử dụng cho công tác lưu trữ điện tử của các đơn vị [7]. Tại Việt Nam, một số nơi đã sử dụng GSDL để tạo các bộ sưu tập số của mình. Thí dụ, Đại học Quốc gia Tp. Hồ Chí Minh đã sử dụng GSDL để tạo một số bộ sưu tập số theo những chủ đề khác nhau [8].

3. So sánh giữa CDS/SIS và GSDL

Để lựa chọn phần phềm, chúng tôi xem xét theo một số tiêu chí sau: hỗ trợ tiếng Việt; hỗ trợ đưa CSDL lên World Wide Web; quản lý toàn văn; quản lý siêu dữ liệu; chức năng tìm kiếm; chức năng duyệt tài liệu; lưu kết quả tìm kiếm; in tài liệu từ cơ sở dữ liệu. Kết quả so sánh được tóm tắt trong Bảng 1.

Hiện tại, một số chuyên gia đã đưa ra một số phương thức để chuyển đổi CSDL CDS/ISIS sang GSDL và đưa lên INTERNET [9,10]. Tuy nhiên, trong khuôn khổ dự án điều này không thể thực hiện được vì do dữ liệu tiếng Việt trong CDS/ISIS được lưu trên bộ mã ký tự 8 bit (TCVN 5712-1999), và không thể chuyển đổi chính xác bằng công cụ có sẵn của GSDL sang mã UTF-8 của dữ liệu lưu trữ trong CSDL của GSDL.

Bảng 1. So sánh các tính năng của CDS/ISIS và GSDL

Tiêu chí	CDS/ISIS	GSDL
Hỗ trợ tiếng Việt	- Hỗ trợ tiếng Việt ASCII 8-bit (TCVN 5712-1999). - Cần phát triển bổ sung các chức năng cho phép hiển thị thanh tác vụ và các thông báo bằng tiếng Việt. - Để truy cập Internet, cần có máy chủ Webisis, một máy chủ proxy đặc biệt để thực hiện chuyển đổi từ mã ASCII sang Unicode (TCVN 6909-2001) và ngược lại.	- Hỗ trợ Unicode (TCVN 6909-2001); - Có sẵn giao diện tiếng Việt. - Không cần chuyển đổi bộ mã ký tự, GSDL mặc định sử dụng Unicode. - Dữ liệu bên trong GSDL lưu trữ theo UTF-8

Tiêu chí	CDS/ISIS	GSDL
Hỗ trợ xuất bản CSDL trên web	- CDS/ISIS tự thân không có tính năng cài đặt chạy trên Internet mà cần phải có công cụ đặc biệt (Webisis, WWW-ISIS, Genisis,...)	- Có thể cài đặt chạy trên Internet
Quản lý toàn văn	- Một biểu ghi CDS/ISIS có tối đa 200 trường, và có thể chứa tới 32.000 ký tự; - Không có sẵn khả năng quản lý toàn văn. - Sử dụng liên kết siêu văn bản để kết nối tới tệp tin toàn văn (Thí dụ tệp tin PDF hay MS. Word);	- Không giới hạn số ký tự trong một biểu ghi. - Có khả năng quản lý toàn văn. Tạo chỉ mục toàn văn, do đó có thể tìm kiếm toàn văn. - Có thể sử dụng liên kết siêu văn bản để kết nối đến tệp tin toàn văn.
Quản lý siêu dữ liệu	- Người dùng phải tạo Bảng Định nghĩa các trường (FDT) để định nghĩa cấu trúc (các yếu tố mô tả) của biểu ghi thư mục.	- Có sẵn một số bộ siêu dữ liệu để lựa chọn (Thí dụ: Dublin Core). - Người dùng có thể tạo ra bộ siêu dữ liệu riêng của mình
Tìm kiếm	- CDS/ISIS có khả năng tìm kiếm mạnh mẽ sử dụng các loại toán tử khác nhau (toán tử Bool, tìm lân cận, chặt cụt, ...) - Hai hình thức tìm kiếm: tìm nâng cao và tìm đơn giản có hướng dẫn; - Đối với Webisis, cần tạo mẫu tìm kiếm, một việc không đơn giản đối với những người không phải là lập trình viên.	- Khả năng tìm kiếm mạnh mẽ sử dụng toán tử Bool và ký hiệu chặt cụt. - Hai hình thức tìm kiếm: tìm đơn giản và tìm theo mẫu. - Dễ dàng tạo mẫu tìm kiếm sử dụng các công cụ sẵn có.
Duyệt tài liệu	- CDS/ISIS có giao diện duyệt tài liệu riêng của mình. Người dùng có thể duyệt lần lượt từ đầu hoặc từ bất kỳ một biểu ghi nào. - Không có chức năng duyệt theo hệ thống thứ bậc hoặc duyệt theo chủ đề.	- GSDL sử dụng trình duyệt web (Thí dụ: Internet Explorer, Netscape) để duyệt bộ sưu tập. - Cho phép duyệt bộ sưu tập theo các cách khác nhau (sắp sếp theo danh sách, theo hệ thống thứ bậc, theo chủ đề, theo thời gian,...)
Lưu kết quả tìm kiếm và in CSDL	- Sử dụng các công cụ lưu kết quả và in tài liệu riêng của mình. - Người dùng cần được đào tạo để sử dụng các chức năng này của ISIS	- Sử dụng tính năng lưu kết quả và in ấn mặc định của các trình duyệt Internet.

Dựa trên so sánh này, nhóm dự án đã chọn GSDL để tiến hành xây dựng sưu tập số dành cho nông thôn bởi tính đơn giản, dễ sử dụng, đồng thời có nhiều tính năng mạnh.

III. Lựa chọn khổ mẫu dữ liệu số và phương pháp nhập siêu dữ liệu

1. Lựa chọn khổ mẫu dữ liệu số

Việc lựa chọn khổ mẫu tệp tin của bộ sưu tập số cần phải tính đến các chuẩn dữ liệu hiện có. Mục đích của việc lựa chọn này là nhằm đảm bảo tính bền vững của dữ liệu theo thời gian. Tại Việt Nam, đã có một số cuộc thảo luận về chuẩn dữ liệu của các thư viện điện tử [11,12], tuy nhiên cũng chưa có tài liệu nào đề cập đến định dạng phù hợp cho GSDL. Chúng tôi đã xem xét lựa chọn khổ mẫu tài liệu số dựa trên bảy tiêu chí do Thư viện Quốc hội Hoa Kỳ đưa ra, bao gồm: Tính công khai hoặc tính mở (Disclosure); Sự chấp nhận hoặc Tính phổ biến (Adoption); Tính trong suốt (Transperancy); Dữ liệu đặc tả nhúng hoặc mô tả nhúng (Self-documentation); Sự phụ thuộc bên ngoài (External dependencies); Ảnh hưởng của sáng chế độc quyền (Impact of patents) và các cơ chế bảo vệ kỹ thuật (Technical protection mechanisms) [3].

Tính công khai hoặc Tính mở liên quan đến mức độ mà nhà sản xuất cung cấp những thông tin về đặc tính kỹ thuật và công cụ liên quan đến khổ mẫu để đánh giá và xem xét tính toàn vẹn và bền vững của nội dung số. Sự bảo quản lâu dài nội dung số sẽ không thể thực hiện được nếu không có sự hiểu sâu sắc về cách thức thông tin số được trình bày, mã hoá tới từng bit và byte trong tệp tin. Thông thường, những chuẩn không độc quyền, chuẩn mở có nhiều thông tin được bộc lộ công khai để chúng ta nghiên cứu và xem xét hơn các chuẩn độc quyền. Vì thế việc được một tổ chức tiêu chuẩn hoá công nhận có thể chưa quan trọng bằng việc có đầy đủ tài liệu kỹ thuật về khổ mẫu.

Sự chấp nhận hoặc tính phổ biến đề cập đến mức độ mà khổ mẫu này đã được sử dụng bởi những người tạo ra nó, sử dụng bởi người phổ biến khổ mẫu và bởi những người dùng khác. Nếu khổ mẫu được chấp nhận rộng rãi, sẽ ít có nguy cơ bị lỗi thời hơn khổ mẫu chưa được chấp nhận rộng rãi và những công cụ cho việc chuyển đổi, di trú dữ liệu cũng dễ kiếm hơn và có thể không đòi hỏi sự đầu tư đặc biệt để chuyển đổi dữ liệu.

Tính trong suốt đề cập đến mức độ mà việc trình bày thông tin số có thể được nhận biết, phân tích trực tiếp bằng các công cụ cơ bản, kể cả việc đọc bằng những chương trình soạn thảo văn bản đơn giản không mã hoá (Text-only editor). Những khổ mẫu dữ liệu trong đó thông tin được trình bày đơn giản và trực tiếp sẽ dễ dàng được di trú sang loại khổ mẫu khác. Tính trong suốt còn đề cập đến nội dung văn bản (kể cả siêu dữ liệu nhúng trong tệp tin). Dữ liệu nhập bằng mã ký tự chuẩn UNICODE (thí dụ UTF-8) được coi là trong suốt. Nhiều khổ mẫu dữ liệu số sử dụng kỹ thuật mã hoá (encryption) hoặc nén dữ liệu (compression). Hai kỹ thuật này mâu thuẫn với yêu cầu về tính trong suốt. Mặc dù vậy vì những lý do thực dụng, một số khổ mẫu dữ liệu video và âm thanh vẫn áp dụng kỹ thuật nén hoặc mã hoá để lưu trữ dữ liệu ngay từ lúc tạo ra dữ liệu số.

Mô tả nhúng hoặc Dữ liệu đặc tả nhúng liên quan đến việc đối tượng số chứa siêu dữ liệu (siêu dữ liệu mô tả, siêu dữ liệu kỹ thuật, siêu dữ liệu hành chính) nhúng bên trong nó (tương tự như một cuốn sách có trang tên) thay vì siêu dữ liệu mô tả độc lập với đối tượng (như phiếu thư mục hoặc biểu ghi thư mục của mục lục trực tuyến, cơ sở

dữ liệu thư mục). Những siêu dữ liệu này cho phép người sử dụng có thể biết những thông tin về đối tượng mà không cần phải sử dụng công cụ tra cứu bên ngoài và có thể theo dõi được quá trình tạo lập và thay đổi của đối tượng.

Sự phụ thuộc bên ngoài liên quan đến mức độ mà một khổ mẫu phụ thuộc vào phần cứng, phần mềm, hệ điều hành cụ thể nào đó để có thể sử dụng được. Điều này rất quan trọng để dự báo mức độ phức tạp liên quan đến phát triển của thiết bị, công nghệ trong tương lai. Thí dụ, một số khổ mẫu được tạo ra chỉ có thể hoạt động trên một số môi trường kỹ thuật nhất định sẽ có thể có khó khăn trong tương lai khi có nhu cầu thay đổi thiết bị hoặc khi tốc độ phát triển công nghệ của thiết bị liên quan xảy ra quá nhanh. Một số dữ liệu khoa học sẽ có thể không đọc được nếu không có những phần mềm chuyên dụng để phân tích và trình bày.

Vấn đề tác động của sáng chế độc quyền có thể hạn chế khả năng của tổ chức trong việc duy trì nội dung số trên khổ mẫu nào đó. Việc phải mua bản quyền, mua giấy phép để chuyển mã dữ liệu từ khổ mẫu lỗi thời sang khổ mẫu mới có thể đòi hỏi chi phí lớn và có thể gây khó khăn. Khi phí bản quyền được tính dựa trên mức độ sử dụng thì chi phí có thể sẽ không lường trước được.

Cơ chế bảo vệ kỹ thuật liên quan đến bảo vệ nội dung số và cung cấp dịch vụ cho người sử dụng trong nhiều năm sau. Người lưu giữ dữ liệu phải có khả năng nhân bản hoặc sao dữ liệu số sang môi trường mới, di trú dữ liệu sang môi trường hoặc khổ mẫu mới đồng thời đối mặt với vấn đề phát triển và thay đổi công nghệ. Vì thế dữ liệu lưu giữ phải không được bảo vệ bằng một cơ chế kỹ thuật đặc biệt nào (như mã hoá) mà nó hạn chế khả năng của người lưu giữ tài liệu thực hiện những bước cần thiết, chuyển đổi khổ mẫu để làm cho nó có thể tiếp cập được với những thế hệ tương lai.

Nhóm nghiên cứu đã xem xét, đánh giá kỹ càng những khía cạnh trên để lựa chọn được khổ mẫu dữ liệu phù hợp, có thể lưu trữ lâu dài, có khả năng thay đổi, đảm bảo sự truy cập của thế hệ tương lai. Nhóm dự án đã lựa chọn so sánh ba loại khổ mẫu tài liệu điện tử là: MS Word, HTML và PDF. Kết quả so sánh được thể hiện trong Bảng 2.

Bảng 2. So sánh giữa MS Word, PDF và HTML

Tiêu chí	MS Word	PDF	HTML
Tính công khai (hoặc tính mở)	không	một phần	có
Sự chấp nhận hoặc Tính phổ biến	Được sử dụng rộng rãi trong lĩnh vực văn bản hành chính	Được sử dụng rộng rãi để phổ biến văn bản, tài liệu	Được sử dụng rộng rãi trên web
Tính trong suốt	Cần có phần mềm MS Word để tạo và đọc tài liệu	Có các công cụ đơn giản để đọc. Các công cụ tạo tài liệu phức tạp hơn nhiều	có

Tiêu chí	MS Word	PDF	HTML
Dữ liệu đặc tả nhúng hoặc Mô tả nhúng	Một phần	Một phần	Có
Sự phụ thuộc bên ngoài	Có	Có	Không hoặc một phần
Ảnh hưởng của sáng chế độc quyền	Có	Có	Không
Các cơ chế bảo vệ kỹ thuật	Có	Có	Không

Hiện nay, hầu hết các tài liệu điện tử của Việt Nam sử dụng khổ mẫu MS Word. Tuy nhiên, MS Word thuộc sở hữu của Công ty Microsoft, do đó nhóm dự án không chọn chuẩn này cho sưu tập số của mình.

PDF là từ viết tắt của Portable Document Format, do công ty Adobe Systems Incorporated phát triển. Nó là ngôn ngữ thể hiện văn bản được định dạng. Tài liệu PDF có thể được cấu trúc hoặc có thể dưới dạng đơn giản. Tệp tin PDF có thể chứa văn bản, hình ảnh, đồ họa, và các nội dung đa phương tiện khác như video và âm thanh. PDF cũng hỗ trợ chú thích, siêu dữ liệu, liên kết siêu văn bản và đánh dấu văn bản. PDF/A là khổ mẫu dùng để trình bày và bảo quản dữ liệu, được phê chuẩn trong tiêu chuẩn ISO 19005-1 về Quản lý tài liệu - Khổ mẫu tệp tin tài liệu điện tử dành cho bảo quản lâu dài - Phần 1: Sử dụng PDF (PDF/A) [14]. Mặc dù PDF là một khổ mẫu tốt để trình bày thông tin, song nhóm dự án không sử dụng chuẩn này bởi các tài liệu trong bộ sưu tập của sưu tập số sử dụng tiếng Việt, mà hiện tại phần mềm GSDL chưa cho phép chuyển đổi chính xác tài liệu PDF sang khổ mẫu HTML.

Do đó, nhóm dự án đã lựa chọn khổ mẫu HTML cho các tài liệu trong sưu tập số. Ưu điểm chính của HTML là nó có thể dễ dàng được nhập vào GSDL và tạo chỉ mục bởi các công cụ phần mềm nhúng (plug-in) sẵn có của GSDL. Các plug-in này có khả năng nhận dạng và chuyển đổi chính xác tiếng Việt.

2. Phương pháp nhập siêu dữ liệu

Siêu dữ liệu hoặc Dữ liệu đặc tả (Meta-data) là những thông tin mô tả các đặc tính của dữ liệu như nội dung, định dạng, chất lượng, điều kiện và các đặc tính khác nhằm tạo thuận lợi cho quá trình tìm kiếm, truy nhập, quản lý và lưu trữ dữ liệu [15]. Chúng tôi quyết định chọn siêu dữ liệu Dublin Core để làm khổ mẫu siêu dữ liệu cho dữ liệu số.

Dublin Core là chuẩn siêu dữ liệu mặc định của phần mềm GSDL. Siêu dữ liệu Dublin Core được giới thiệu bởi Dublin Core Metadata Element Initiative năm 1995 [16]. Tháng 6/2000, Dublin Core được Ủy ban Tiêu chuẩn hóa/Hệ Thống Tiêu chuẩn Xã hội Thông tin Châu Âu (CEN/ISSS) chấp nhận là một tiêu chuẩn. Tháng 9/2001, Dublin Core được công bố là một tiêu chuẩn quốc gia của Hoa Kỳ ANSI/NISO Z39.85-2001 [17].

Có hai cách nhập siêu dữ liệu cho các tài liệu trong thư viện số GSDL:

- Nhúng siêu dữ liệu vào trong tài liệu bằng cách sử dụng thẻ siêu dữ liệu HTML (meta tags);

- Nhập siêu dữ liệu thông qua Giao diện

thủ thư (GLI).

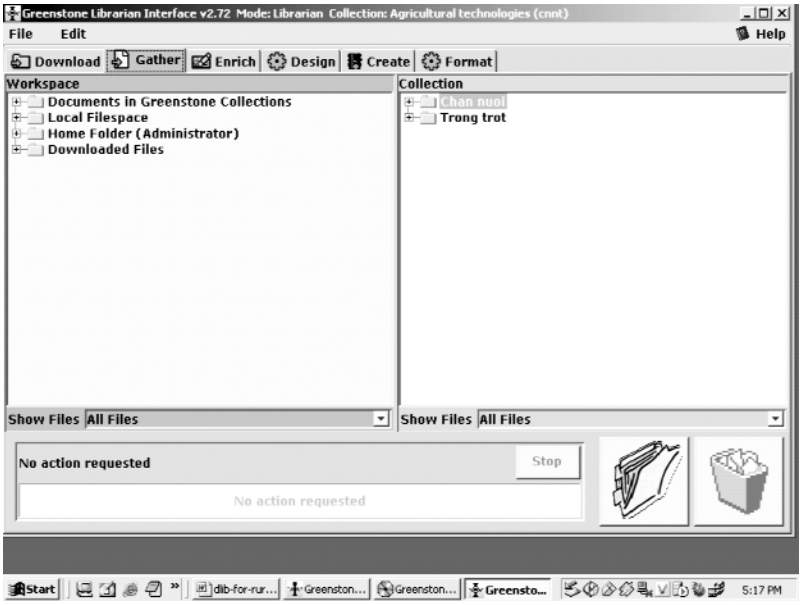
Theo cách thứ nhất, các thẻ siêu dữ liệu được tạo ra trong phần tiêu đề (HEAD) của tệp tin HTML. Nội dung của siêu dữ liệu được thể hiện trong dấu "< >". Cách nhập được thể hiện như dưới đây:

```
<html>
<head>
<meta name="GENERATOR" content="Microsoft FrontPage 5.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="Title" content="Kỹ thuật thâm canh giống đậu tương 3 vụ: DT 84">
<meta name="Creator" content="Nguyễn Văn A.">
<meta name="Subject" content="Trồng trọt">
<meta name="Subject" content="Đậu tương">
<meta name="Subject" content="Giống đậu tương DT 84">
<meta name="Publisher" content="Trung tâm Giống cây trồng Vụ Bản tỉnh Nam Định">
<meta name="Date" content="2006">
<title>Kỹ thuật thâm canh giống đậu tương 3 vụ: DT 84</title></head>
<body>
Kỹ thuật thâm canh giống đậu tương 3 vụ: DT 84
(Techniques of intensive cropping for soybean variety DT-84)
.....
</body>
</html>
```

Nếu nhập siêu dữ liệu theo cách này thì không cần phải nhập trực tiếp siêu dữ liệu trong Giao diện thủ thư (GLI). GSDL có khả năng tự động triết rút và đưa các siêu dữ liệu có sẵn trong các thẻ HTML vào trong chỉ mục của bộ sưu tập GSDL. Tuy nhiên, nhược điểm của phương pháp này là chưa có công cụ nào hỗ trợ việc tạo ra các thẻ siêu dữ liệu trong tệp tin HTML. Toàn bộ công việc này phải làm thủ công, do đó có thể dẫn đến sai sót và không nhất quán. Bên cạnh đó, người nhập không thể tham khảo các siêu dữ liệu đã được nhập trước đây.

Phương pháp thứ hai là sử dụng Giao diện thủ thư (GLI) để nhập siêu dữ liệu, thông qua mục Enrich. GLI là giao diện đồ họa cho phép tạo lập, chỉnh sửa các bộ sưu tập số, và xuất bộ sưu tập ra đĩa CD-ROMs. Nó cho phép nhập hoặc gán các siêu dữ liệu. GLI cho phép người dùng sử dụng các tính năng của mình thông qua một giao diện thân thiện và dễ sử dụng. Giao diện GLI được thể hiện trong Hình 1.

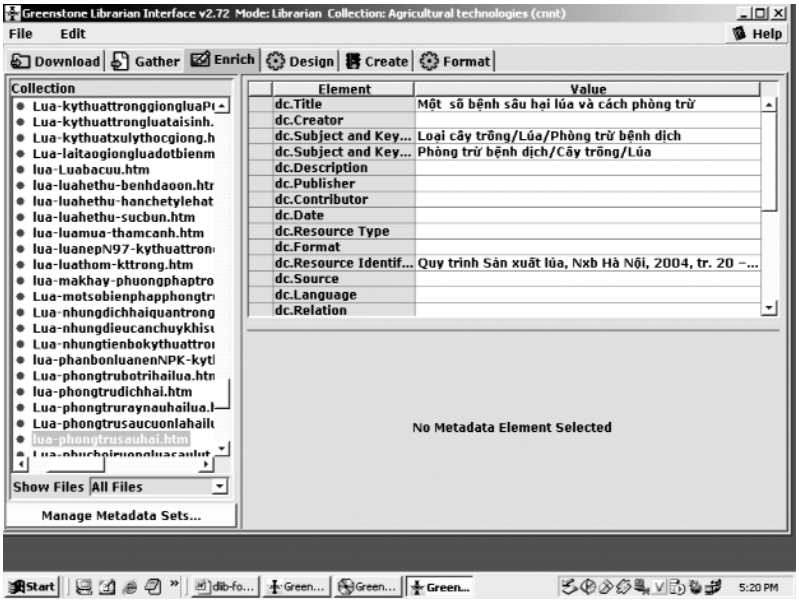
Trong giao diện GLI, người dùng có thể lựa chọn một số chuẩn siêu dữ liệu cho bộ sưu tập của mình. Trong khuôn khổ dự án này, chúng tôi sử dụng chuẩn siêu dữ liệu Dublin Core. Cách nhập siêu dữ liệu trong GLI rất trực quan và đơn giản. Bên cạnh đó, giao diện GLI cũng hỗ trợ tiếng Việt nên rất thuận tiện cho người dùng. Cách nhập siêu dữ liệu trong GLI được thể hiện tại Hình 2.



Hình 1. Giao diện GLI

3. Duyệt tài liệu trong sưu tập số

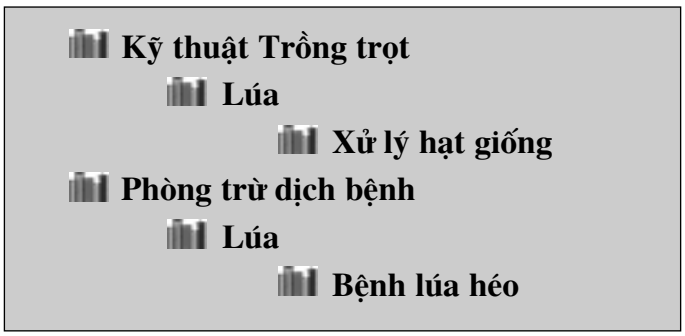
Duyệt bộ sưu tập số trong GSDL rất thuận tiện. Người dùng có thể duyệt tài liệu theo chủ đề. Thí dụ: duyệt theo chủ đề giống cây trồng, vật nuôi, v.v.. Điều này rất quan trọng đối với sưu tập số dành cho nông thôn, bởi người dùng chủ yếu là nông dân có trình độ CNTT hạn chế nên việc tìm kiếm toàn văn đối với họ là việc khó khăn. Để tạo khả năng duyệt tài liệu theo hệ thống chủ đề, chúng tôi sử dụng ký hiệu "/" hoặc "|" để tạo hệ thống thứ bậc. Thí dụ, nếu chúng ta nhập vào trường dc.Subject and keywords với nội dung như sau:



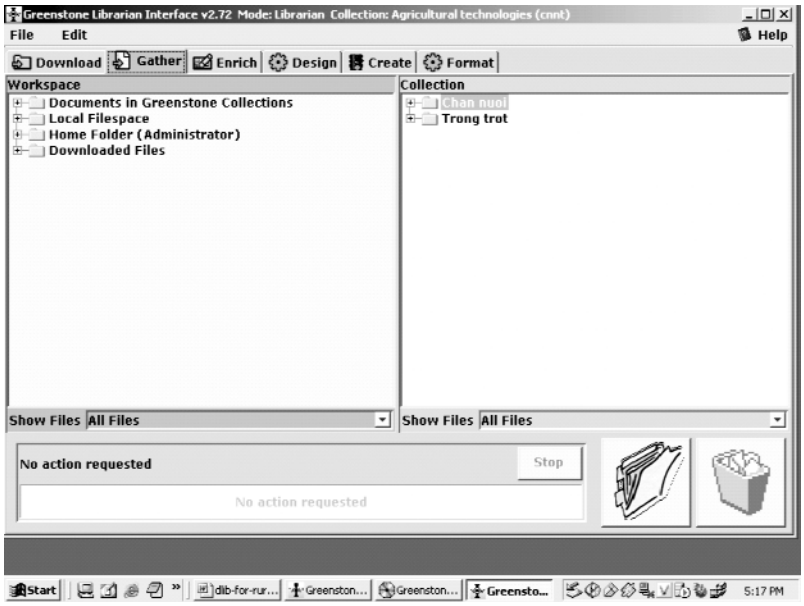
Hình 2. Màn hình nhập siêu dữ liệu

Siêu dữ liệu: Kỹ thuật Trồng trọt/Lúa/Xử lý hạt giống
Phòng trừ dịch bệnh/Lúa/bệnh lúa héo

Thì màn hình sẽ hiển thị tài liệu theo hệ thống thứ bậc chủ đề như sau:



Thí dụ về hiển thị để duyệt tài liệu theo cách trên được thể hiện trong Hình 3.



Hình 3. Màn hình duyệt tài liệu trong sưu tập số

IV. Kết luận

Sưu tập số về công nghệ thích hợp nông thôn là một nguồn tin tham khảo quan trọng đối với người dân các vùng nông thôn. Để xây dựng sưu tập này, phần mềm GSDL đã được lựa chọn sử dụng nhờ có một số ưu điểm so với phần mềm CDS/ISIS về khả năng quản lý toàn văn và quản lý siêu dữ liệu trong sưu tập số.

Trong ba loại khổ mẫu tài liệu được xem xét, chúng tôi chọn khổ mẫu HTML cho sưu tập số bởi đây là một chuẩn mở và nó đảm bảo tính tương thích và tính bền vững của sưu tập số. Giao diện GLI được sử dụng để tạo lập, chỉnh sửa bộ sưu tập số và nhập siêu dữ liệu. Quá trình tạo lập tương đối đơn giản nhờ sự trực quan và các công cụ thuận tiện mà phần mềm GSDL cung cấp.

Trên cơ sở phần mềm GSDL, với khổ mẫu dữ liệu số là HTML và sơ đồ siêu dữ liệu Dublin Core, chúng tôi đã xây dựng được một sưu tập số về công nghệ thích hợp cho nông thôn. Hệ thống rất dễ khai thác, phù hợp với trình độ cũng như kỹ năng sử dụng các công cụ tin học của khu vực nông thôn.

Tài liệu tham khảo

1. CDS/ISIS database software: CI. http://www.unesco.org/webworld/isis/. Viewed June 11, 2007.

2. Cao Minh Kiem and Middleton M.R. (1998) An evaluation of textual storage and retrieval software: CDS/ISIS and InMagic. Program: electronic library & information systems, Vol. 32, no. 3, pp. 283-302.

3. Mahmood, K. (1997) The best library software for developing countries: more than 30 plus points of Micro CDS/ISIS. Library software review, Vol. 16, no. 1, pp. 12-16

4. Carlos Henrique Marcondes, Luís Fernando Sayão (2003). Nuclear Information Center, Nuclear Energy National Commission, Brazil The SciELO Brazilian Scientific Journal Gateway and Open Archives: A Report on the Development of the SciELO-Open Archives Data Provider Server. D-Lib Magazine, March 2003, Vol. 9, no. 3. (ISSN 1082-9873). http://www.dlib.org/dlib/march03/marcondes/03marcondes.html

5. Introduction to Webagris. http://www.fao.org/agris/tools/WebAGRIS/WebAGRIS_En.htm

6. Trang chủ phần mềm Greenstone. http://www.greenstone.org/. Truy cập ngày 11/6/2007.

7. Bainbridge D., Osborn W., Witten I.H. and Nikols D.M. (2006). Extending greenstone for institutional repositories. Digital Libraries : Achievements, challenges and opportunities. Proceedings of the 9th International Conference on Asian Digital Libraries, ICADL 2007, Kyoto, Japan, November 27-30, 2006.

8. Đại học Quốc gia TP. Hồ Chí Minh. http://gralib.hcmuns.edu.vn

9. Morete P. and Rose J.. Creating digital libraries based on CDS/ISIS databases. http://greenstonesupport.iimk.ac.in/Documents/CDS-ISIS_to_DL.pdf

10. Jayakanth F., Shivaram B.S., Venkatlakshmi K. and Singh S. (2004) Integrating CDS/ISIS databases with Greenstone Digital Libraries Software (GSDL). Available at Digital Library of Information science and Technology. http://dlist.sir.arizona.edu/1689/

11. Cao Minh Kiểm và Lê Xuân Định (2002). Một số suy nghĩ về chuẩn dữ liệu in liên kết. Tạp chí Thông tin & Tư liệu, 2002, Số 3, trang7-13.

12. Cao Minh Kiểm (2006). Tìm hiểu một số vấn đề về chuẩn khổ mẫu dữ liệu cho thư viện điện tử. Tài liệu trình bày tại Hội thảo Tiêu chuẩn hóa hoạt động Khoa học & Công nghệ, tổ chức tại Trung tâm Thông tin KH&CN Quốc gia, Hà Nội, 16-17/11/2006.

13. Sustainability of Digital formats. Planning for Library of Congress collections. http://www.digitalpreservation.gov/formats/intro/intro.shtml

14. International Standard Organization. ISO 19005-1:2005. Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1) http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=38920&scopelist=PROGRAMME

15. Nghị định số 64/2007/NĐ-CP về ứng dụng CNTT trong hoạt động của cơ quan nhà nước

16. Dublin Core Metadata Initiative. http://dublincore.org/

17. The National Information Standard Organization. Dublin Core Metadata Element set. Bethesda, Maryland: NISO Press, 2001. ISSN 1041-5653