XÂY DỰNG VÀ QUẢN LÝ DỮ LIỆU THUẬT NGỮ ANH-VIỆT CHUYÊN NGÀNH ĐỊA CHẤT BẰNG PHẦN MỀM MÃ NGUỒN MỞ

BÙI BẢO TRUNG[1], PHẠM NGUYỄN HÀ VŨ[2]

[1]*Trung tâm Nghiên cứu Đô thị, Đại học Quốc gia Hà Nội, 144 Xuân Thủy, Quận Cầu Giấy, Hà Nội*
[2]*Khoa Địa chất, Trường Đại học Khoa học Tự nhiên, 334 Nguyễn Trãi, Quận Thanh Xuân, Hà Nội*

Tóm tắt: *Dữ liệu thuật ngữ Anh-Việt chuyên ngành địa chất được xây dựng để hỗ trợ cán bộ nghiên cứu, sinh viên thuộc các ngành địa chất, tài nguyên môi trường sử dụng trong học tập và nghiên cứu khoa học. Dữ liệu được quản lý nhờ phần mềm mã nguồn mở Geoterms chạy trên hệ điều hành Window của Microsoft có giao diện đơn giản, dễ sử dụng, phù hợp với người không có trình độ tin học cao. Phần mềm Geoterms không chỉ quản lý dữ liệu thuật ngữ ngành địa chất, tài nguyên và môi trường mà còn có thể mở rộng, kết nối với từ điển thuộc nhiều lĩnh vực chuyên môn khác. Người dùng có thể cài đặt sử dụng phần mềm thông qua hướng dẫn sử dụng miễn phí.*

I. MỞ ĐẦU

Hiện nay, Việt Nam đã chính thức trở thành thành viên của tổ chức WTO và đang trong quá trình nhanh chóng phát triển hội nhập quốc tế. Vì vậy đào tạo nguồn nhân lực chất lượng cao đối với Việt Nam chính là chìa khóa để phát triển kinh tế, hội nhập quốc tế. Đây là thách thức lớn đặt ra cho ngành giáo dục Việt Nam. Ngành tài nguyên và môi trường là ngành đa lĩnh vực thực hiện chức năng, nhiệm vụ quản lý và điều tra cơ bản trên bảy lĩnh vực gồm tài nguyên nước, tài nguyên khoáng sản địa chất, môi trường, khí tượng thủy văn và biến đổi khí hậu, đo đạc và bản đồ, quản lý tổng hợp và thống nhất về biển đảo. Việc tiếp cận tri thức chung của thế giới trong công tác nghiên cứu và học tập là yêu cầu cần thiết để có thể cập nhật những tri thức mới, sử dụng những nguồn tài liệu phong phú từ kho tàng tri thức của nhân loại, chủ yếu là bằng tiếng Anh. Do đó, cán bộ nghiên cứu và học viên phải làm chủ ngoại ngữ để có thể khai thác hiệu quả nguồn tài liệu nước ngoài và đáp ứng được yêu cầu hội nhập quốc tế. Hiện nay sự tiếp cận tiếng Anh cơ bản của sinh viên nói riêng và cán bộ nghiên cứu khoa học nói chung khá thuận lợi với nguồn tài liệu phong phú, đa dạng nhưng việc sử dụng tiếng Anh chuyên ngành thì còn nhiều hạn chế. Điều này là do sự tiếp cận với các từ điển chuyên ngành còn gặp nhiều khó khăn, nguồn cung cấp tài liệu thường ít và mỗi từ điển chuyên ngành được xây dựng thường chỉ tập trung trong phạm vi chuyên ngành hẹp theo chuyên môn chính của tác giả [1, 2, 4-10]. Trong nghiên cứu và học tập lại đòi hỏi phải tiếp cận nguồn tài liệu chuyên ngành bằng tiếng nước ngoài rất đa dạng và phong phú, đòi hỏi người sử dụng ngoại ngữ cần phải dùng một lúc vài từ điển chuyên ngành khác nhau trong lĩnh vực địa chất, quản lý tài nguyên và môi trường. Đồng thời, trong quá trình học tập và nghiên cứu, cùng với sự phát triển của ngành, người sử dụng sẽ phải cập nhật thêm các thuật ngữ mới mà các từ điển giấy không thể kịp thời cập nhật theo nhu cầu sử dụng của mỗi cá nhân. Vì vậy, xây dựng dữ liệu thuật ngữ Anh-Việt chuyên ngành được lưu trữ dưới dạng phần mềm từ điển nhưng đồng thời có khả năng cập nhật và bổ sung thêm thuật ngữ mới đơn giản và thuận tiện là một yêu cầu thực tế trong nghiên cứu khoa học và học tập trong các ngành địa chất, tài nguyên và môi trường.

Phần mềm quản lý từ điển chuyên ngành Anh-Việt Geoterms là công cụ hỗ trợ hiệu quả cho cán bộ và sinh viên nghiên cứu khoa học để làm chủ ngoại ngữ trong lĩnh vực chuyên ngành của mình.

Trong bài báo này, các tác giả sẽ giới thiệu cách xây dựng dữ liệu thuật ngữ và phần mềm quản lý dữ liệu thuật ngữ chuyên ngành Geoterms. Các thông tin về hướng dẫn sử dụng phần mềm và cách cài đặt các độc giả có thể thông qua địa chỉ buibaotrung89@gmail.com.

II. XÂY DỰNG DỮ LIỆU CHUYÊN NGÀNH TRÊN CƠ SỞ CÁC PHẦN MỀM MÃ NGUỒN MỞ

Dữ liệu thuật ngữ chuyên ngành Anh-Việt được xây dựng với mục đích đơn giản, thuận tiện, được khai thác và quản lý trên cơ sở phối hợp các phần mềm mã nguồn mở nhằm đảm bảo các yêu cầu về vấn đề bản quyền.

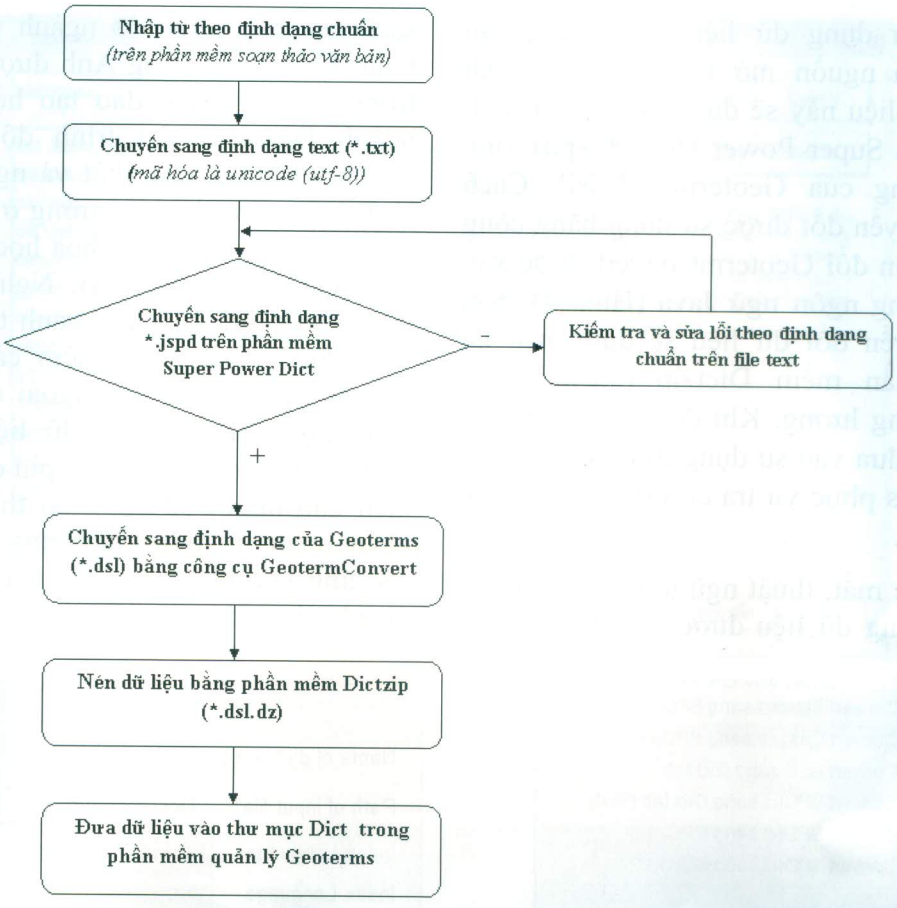
Phần mềm mã nguồn mở là những phần mềm được cung cấp dưới cả dạng mã và nguồn, không chỉ là miễn phí về giá mua mà còn miễn phí về bản quyền: người dùng có quyền sửa đổi, cải tiến, phát triển, nâng cấp theo một số nguyên tắc chung quy định trong giấy phép phần mềm nguồn mở (ví dụ General Public Licence – GPL) mà không cần xin phép, điều mà họ không được phép làm đối với các phần mềm nguồn đóng (tức là phần mềm thương mại). Tiện ích mà mã nguồn mở mang lại chính là quyền tự do sử dụng chương trình cho mọi mục đích, quyền tự do để nghiên cứu cấu trúc của chương trình, chỉnh sửa phù hợp với nhu cầu, truy cập vào mã nguồn, quyền tự do phân phối lại các phiên bản cho nhiều người, quyền tự do cải tiến chương trình và phát hành những bản cải tiến vì mục đích công cộng.

Xây dựng dữ liệu thuật ngữ phải nhằm đảm bảo để người không có trình độ cao về tin học có thể tự xây dựng được bộ dữ liệu thuật ngữ chuyên ngành của mình và bổ sung thêm thuật ngữ bên cạnh nguồn thuật ngữ đã sẵn có nhưng vẫn đảm bảo quy định của pháp luật về sở hữu trí tuệ. Cách tiếp cận này sẽ đáp ứng nhu cầu cần bổ sung thêm thuật ngữ chuyên ngành bằng tiếng nước ngoài trong quá trình học tập và nghiên cứu.

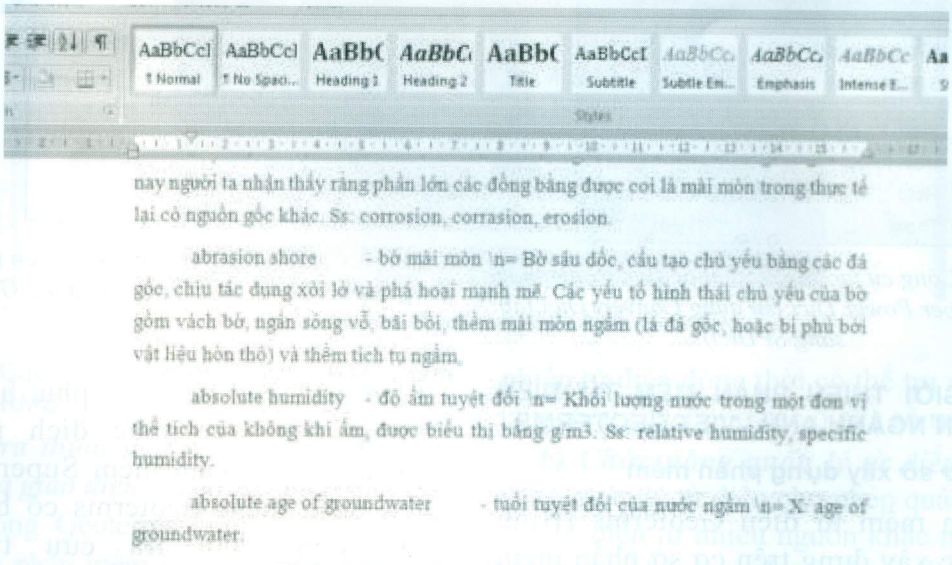
Mỗi thuật ngữ chuyên ngành được xây dựng bao gồm phần thuật ngữ tiếng Anh, phần thuật ngữ tiếng Việt tương ứng và phần giải thích thuật ngữ bằng tiếng Việt. Trong trường hợp có hình ảnh minh họa thì có thể đưa ra hình ảnh minh họa trong phần mềm quản lý dữ liệu thuật ngữ. Dữ liệu thuật ngữ được xây dựng dựa trên phần mềm từ điển mã nguồn mở Super Power Dict của Bùi Đức Tiến [2]. Xây dựng dữ liệu theo phần mềm này đơn giản, nhanh hơn và dễ kiểm soát, khắc phục được lỗi xảy ra trong quá trình nhập liệu (Hình 1). Phần mềm này cho phép xây dựng dữ liệu trên các phần mềm soạn thảo văn bản dạng word hoặc text theo cú pháp định dạng có sẵn như sau:

[từ tiếng anh]<dấu tab>[- (nghĩa tiếng việt1)][\n= (giải thích mức 1)][\n+(giải thích mức 2)][\n-(nghĩa tiếng việt 2)][\n=(giải thích mức 2)]...

Trong đó, các ký hiệu dấu (-), (=), (+), (@) là dấu hiệu để định dạng chữ và ký hiệu (\n) là ký hiệu báo xuống dòng trong phần mềm từ điển. Để hiện thị hình ảnh (chỉ sử dụng ảnh có định dạng đuôi *.jpg) thì sử dụng cú pháp: ~tenhinh.jpg xen trong cú pháp trên (Hình 2). Cách thức nhập liệu như trên trong phần mềm soạn thảo văn bản word hoặc text cho phép phân chia dữ liệu thành nhiều phần để nhiều người cùng xây và sau đó lắp ghép vào một dữ liệu thống nhất. Cách thức xây dựng như vậy đáp ứng được yêu cầu đơn giản, thuận tiện và nhanh chóng. Sau khi soạn thảo xong ở định dạng văn bản word thì dữ liệu sẽ được lưu dưới định dạng file văn bản text (*.txt) với encoding là unicode (utf-8). File văn bản text này sẽ được chuyển đổi thành file dữ liệu từ điển (tạo từ điển) của phần mềm Super Power Dict để có thể sử dụng để tra cứu trong phần mềm từ điển này bằng công cụ chuyển đổi có trong phần mềm (Hình 3, 4, 5).



Hình 1. *Sơ đồ khối xây dựng cơ sở dữ liệu thuật ngữ.*

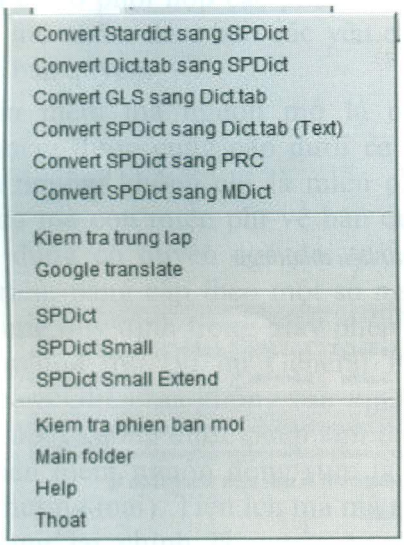


Hình 2. *Soạn thảo dữ liệu trên word.*

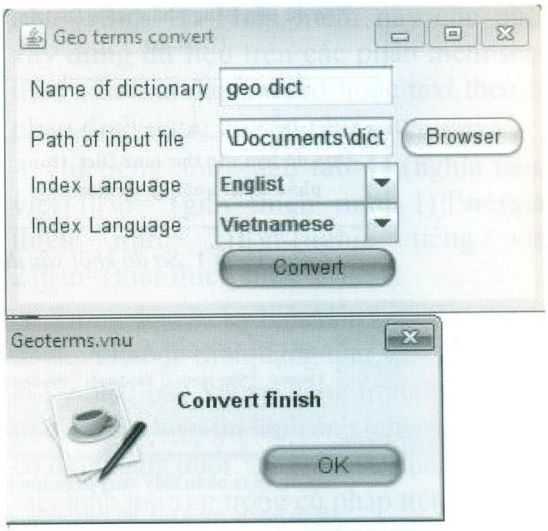
Người sử dụng có thể bổ sung dữ liệu theo hai cách. Một là bổ sung vào dữ liệu dạng word hoặc text ban đầu trong trường hợp số lượng thuật ngữ phải bổ sung lớn để tiết kiệm thời gian. Hai là có thể sử dụng phần mềm Super Power Dict để bổ sung cập nhật thêm từ trong trường hợp chỉ bổ sung ít thuật ngữ.

Để sử dụng dữ liệu này trên phầm mềm mã nguồn mở Geoterms thì định dạng dữ liệu này sẽ được chuyển từ định dạng của Super Power Dict (*.jspd) sang định dạng của Geoterms (*.dsl). Cách thức chuyển đổi được sử dụng bằng công cụ chuyển đổi GeotermConvert được xây dựng bằng ngôn ngữ Java (Hình 4). Sau khi chuyển đổi dữ liệu sẽ được nén lại bằng phần mềm Dictzip (*.dsl.dz) để giảm dung lượng. Khi đó, dữ liệu đã sẵn sàng để đưa vào sử dụng trong phần mềm Geoterms phục vụ tra cứu ở dạng từ điển (Hình 1).

Trước mắt, thuật ngữ được đưa vào sử dụng trong dữ liệu được lựa chọn từ các sách giáo trình chuyên ngành và tài liệu tham khảo bằng tiếng Anh được sử dụng trong chương trình đào tạo hệ cử nhân ngành Địa chất đạt trình độ quốc tế, ngành Kỹ thuật Địa chất và ngành Quản lý Tài nguyên và Môi trường ở Khoa Địa chất, Trường Đại học Khoa học Tự nhiên (ĐH Quốc Gia Hà Nội). Nghĩa và giải thích thuật ngữ chuyên ngành trong tiếng Việt được tham khảo thêm các từ điển chuyên ngành hiện có. Ngoài ra, dữ liệu còn được tích hợp với dữ liệu từ điển Anh-Việt thông dụng miễn phí để tạo điều kiện cho người sử dụng có thể sử dụng phần mềm từ điển Geoterms vào trong các lĩnh vực khác ngoài lĩnh vực chuyên ngành nêu trên.



Hình 3. *Công cụ tạo file dữ liệu từ điển trong phần mềm Super Power Dict (sử dụng Convert Dict.tab sang SPDict).*

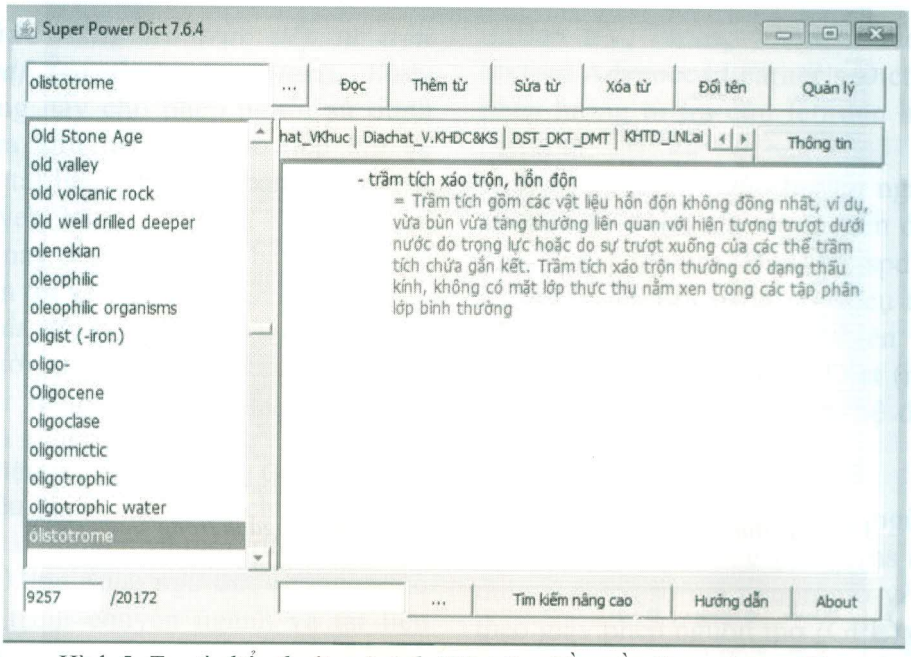


Hình 4. *Công cụ chuyển đổi file từ điển từ định dạng Super Power Dict sang định dạng Geoterms.*

III. GIỚI THIỆU PHẦN MỀM TỪ ĐIỂN CHUYÊN NGÀNH ANH - VIỆT GEOTERMS

1. Cơ sở xây dựng phần mềm

Phần mềm từ điển Geoterms (Hình 6) được xây dựng trên cơ sở phần mềm mã nguồn mở Goldendict (giấp phép GPL) có sửa lỗi và thêm các chức năng sửa lỗi không hiển thị ảnh trên Wiki, giao diện chính của chương trình, tìm hiểu và xây dựng công cụ chuyển đổi cơ sở dữ liệu từ điển phù hợp với chương trình, có mục đích phi lợi nhuận. So với phần mềm Super Power Dict, phần mềm Geoterms có bổ sung thêm tính năng tra cứu từ trên wikipedia cũng như các nguồn tra cứu từ điển trực tuyến tương tự. Phần mềm Geoterms còn có thể sử dụng được các dữ liệu của các từ điển khác như StarDict, Babylon, Lingvo và Dictd.



Hình 5. *Tra từ điển thuật ngữ Anh-Việt trên phần mềm Super Power Dict.*

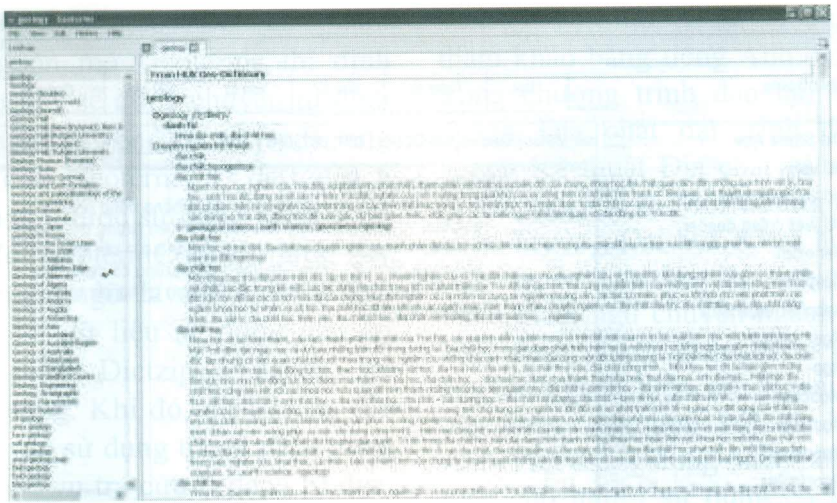
Hình 6. *Giao diện chính của phần mềm Geoterms.*

2. Các tính năng cơ bản của GEOTERMS

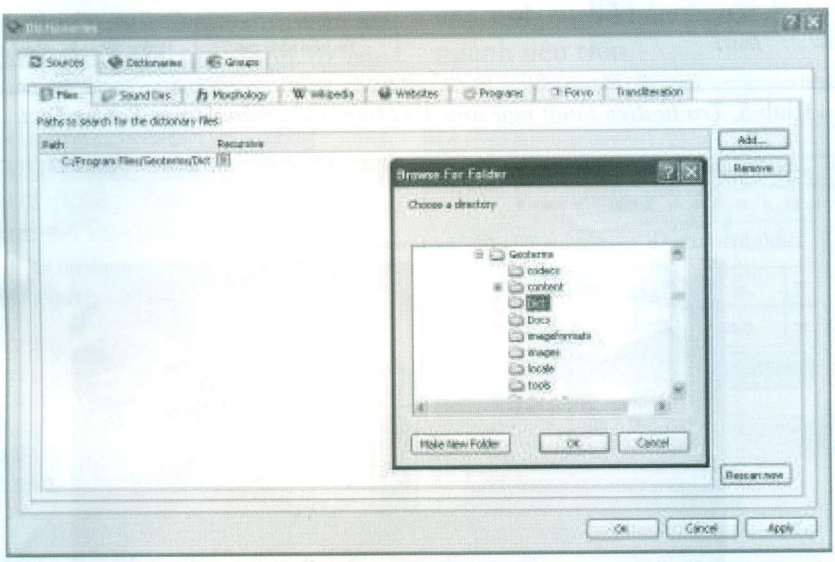
a) Tra thuật ngữ chuyên ngành Anh-Việt qua giao diện: Thao tác tra từ qua giao diện trong Geoterms hoàn toàn tương tự như các phần mềm khác, người dùng nhập từ cần tra, ấn Enter hoặc chọn từ chính xác, lập tức nghĩa của từ sẽ xuất hiện và có trích dẫn nguồn của từ đó (Hình 7).

Trong Geoterms người sử dụng có thể tra từ trong từng từ điển hoặc tra từ trong nhiều từ điển đồng thời có thể tra nhiều từ cùng bằng cách mở các ô cửa sổ mới.

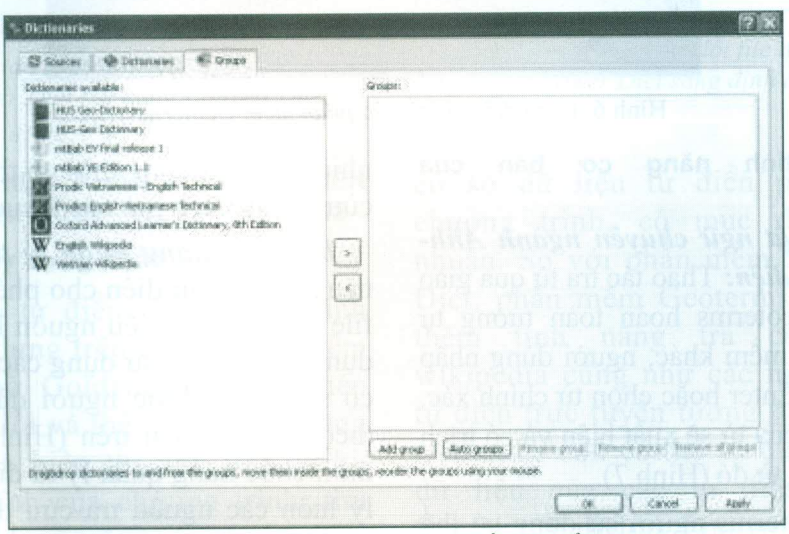
b) Chức năng quản lý từ điển: Chức năng quản lý từ điển cho phép quản lý các file từ điển từ nhiều nguồn khác nhau; sử dụng hay không sử dụng các từ điển số đã có sẵn hoặc được người dùng xây dựng theo cách đã nêu trên (Hình 8, 9). Đồng thời, chức năng quản lý từ điển cũng quản lý luôn các nguồn tra cứu thuật ngữ trực tuyến như Wikipedia.



Hình 7. *Giao diện tra cứu thuật ngữ Anh-Việt chuyên ngành trong Geoterms.*



Hình 8. *Giao diện duyệt các nguồn từ điển trong Geoterms.*



Hình 9. *Giao diện quản lý sử dụng các nguồn từ điển trong Geoterms.*

c) Chức năng tra trực tiếp từ trên Word, pdf (có định dạng text), Web: Chức năng này cho phép người sử dụng có thể tra cứu từ điển trực tiếp khi sử dụng các tài liệu dạng văn bản trên word, pdf và web bằng công cụ scan popup. Công cụ này được sử dụng bằng cách bôi đen từ cần tra và bấm Ctrl+C+C.

d) Xuất dữ liệu: Chức năng này cho phép người sử dụng có thể xuất dữ liệu thuật ngữ đang tra sang các định dạng HTML và xuất dữ liệu để in.

3. Dữ liệu thuật ngữ trong Geoterms

Dữ liệu thuật ngữ các chuyên ngành địa chất, tài nguyên và môi trường hiện đã có trên 40.000 thuật ngữ được thu thập từ các giáo trình chuyên ngành và tài liệu tham khảo bằng tiếng Anh đang sử dụng trong chương trình đào tạo hệ cử nhân ngành Địa chất đạt trình độ quốc tế, ngành Kỹ thuật Địa chất và ngành Quản lý Tài nguyên và Môi trường ở Khoa Địa chất, Trường Đại học Khoa học Tự nhiên (ĐH Quốc Gia Hà Nội). Nghĩa và giải thích thuật ngữ chuyên ngành bằng tiếng Việt được tham khảo từ các Từ điển địa chất Anh - Việt của Viện Khoa học Địa chất và Khoáng sản [8], từ điển Dầu khí Anh-Nga-Việt của Tổng hội Địa chất Việt Nam [7], từ điển Địa chất Anh - Việt của Vũ Khúc [9], từ điển Thuật ngữ Các Khoa học-Trái đất của Lê Như Lai [5], từ diển giải thích Khoa học Địa chất Anh-Việt và Việt-Anh của Phan Cự Tiến [6], từ điển giải thích Địa sinh thái - Địa môi trường - Địa kỹ thuật của Bùi Học [1], từ điển thuật ngữ Môi trường của Ủy ban Bảo vệ môi trường của Mỹ (EPA) [8], từ điển thuật ngữ Phát triển của Chương trình Phát triển Liên Hiệp Quốc (UNDP) [4]. Dữ liệu thuật ngữ chuyên ngành được kết hợp với thuật ngữ Anh-Việt phổ thông của Hồ Ngọc Đức trong dữ liệu HUS-Geo Dictionary. Ngoài ra, dữ liệu trong Geoterms còn sử dụng các thuật ngữ từ dữ liệu từ điển được cung cấp miễn phí như mtBab EV, Prodict English-Vietnamese, Oxford Advanced Learner's Dictionar cho phép lượng từ tra cứu lên đến hàng trăm nghìn từ.

Dữ liệu thuật ngữ chuyên ngành -Geo Dictionary được xây dựng trên định dạng của Super Power Dict (*.jspd) có thể được chuyển đổi bằng công cụ thích hợp để sử dụng được trên phiên bản cho android của Super Power Dict (mSPDict) dùng cho smartphone chạy hệ điều hành Android [3].

IV. KẾT LUẬN

Phần mềm từ điển thuật ngữ chuyên ngành Anh-Việt trong Geoterms được xây dựng trên cơ sở mã nguồn mở và tuân thủ theo giấy phép nguồn mở (GPL) chạy trên hệ điều hành Window nhằm phục vụ cho việc tra thuật ngữ trong lĩnh vực khoa học trái đất, tài nguyên và môi trường. Phần mềm Geoterms là công cụ hỗ trợ tốt cho cán bộ nghiên cứu khoa học, sinh viên học tập thuộc khoa học trái đất để hướng tới hợp tác quốc tế trong nghiên cứu khoa học. Việc sử dụng các phần mềm mã nguồn mở cũng đáp ứng được yêu cầu tuân thủ về các yêu cầu pháp luật về bản quyền và sở hữu trí tuệ trong sử dụng dữ liệu cũng như sử dụng phần mềm Geoterms.

Việc xây dựng dữ liệu thuật ngữ Anh-Việt chuyên ngành địa chất, tài nguyên và môi trường bằng các phần mềm mã nguồn mở được thực hiện theo cách thức đơn giản, tiện lợi, phù hợp với người không có trình độ tin học cao. Dữ liệu được sử dụng tra cứu thông qua phần mềm sử dụng mã nguồn mở Geoterms cho phép kết hợp sử dụng với nhiều nguồn từ điển khác nhau cũng như tra cứu trực tuyến với các nguồn từ điển trực tuyến. Dữ liệu thuật ngữ chuyên ngành Anh-Việt HUS-Geo Dictionary được quản lý trên phần mềm mã nguồn mở Geoterms đã đáp ứng nhu cầu và hỗ trợ tích cực cho các nhà khoa học và sinh viên trong học tập và nghiên cứu khoa học tại Khoa Địa chất, Trường

Đại học Khoa học Tự nhiên (Đại học Quốc gia Hà Nội), Trung tâm Nghiên cứu Đô thị (Đại học Quốc gia Hà Nội)...

Xây dựng dữ liệu theo cách tiếp cận trên cho phép không chỉ xây dựng dữ liệu thuật ngữ ngành địa chất, tài nguyên và môi trường mà còn có thể mở rộng ra cho nhiều lĩnh vực chuyên môn khác.

Xây dựng dữ liệu thông qua phần mềm Super Power Dict của Bùi Đức Tiến tạo ra cơ hội để sử dụng dữ liệu trên điện thoại smartphone và máy tính bảng sử dụng hệ điều hành android thông qua phiên bản mSPDict.

Lời cảm ơn: Bài báo được hoàn thành dưới sự hỗ trợ của đề tài TN-13-25. Các tác giả bài báo xin được cảm ơn Đề tài QG.TĐ 11.07 đã hỗ trợ cung cấp dữ liệu thêm về thuật ngữ chuyên ngành cho phần mềm.

Quá trình xây dựng dữ liệu thuật ngữ chuyên ngành có sự tham gia rất nhiều của các bạn sinh viên và sự động viên hỗ trợ của các đồng nghiệp tại Khoa Địa chất, Trường Đại học Khoa học Tự nhiên - ĐHQGHN. Các tác giả xin chân thành cảm ơn đồng nghiệp và các bạn sinh viên.

VĂN LIỆU

1. Bùi Học *(Chủ biên)*, 2008. Từ điển giải thích Địa sinh thái - Địa môi trường - Địa kỹ thuật. *Nxb Xây dựng. Hà Nội.*

2. Bùi Đức Tiến, 2000. Phần mềm từ điển SUPER POWER DICT.

3. Bùi Đức Tiến, 2015. Hướng dẫn sử dụng mSPDict.

4. Chương trình Phát triển Liên Hiệp Quốc (UNDP), 2011. A glossary of common development terms.

5. Lê Như Lai, 2003. Từ điển Thuật ngữ Các Khoa học - Trái đất. *Nxb Xây dựng. Hà Nội.*

6. Phan Cự Tiến, 2006. Từ điển giải thích Khoa học Địa chất Anh - Việt và Việt - Anh. *Nxb Văn hóa - Thông tin. Hà Nội.*

7. Tổng hội Địa chất Việt Nam, 2004. Từ điển Dầu khí Anh-Nga-Việt. *Nxb Lao động Xã hội. Hà Nội.*

8. Viện Khoa học Địa chất và Khoáng sản, 2001. Từ điển Địa chất Anh - Việt (có bổ sung, sửa chữa). *Nxb Bách khoa. Hà Nội*

9. Vũ Khúc, 2005. Từ điển Địa chất Anh-Việt. *Nxb Khoa học và Kỹ thuật. Hà Nội.*

10. Ủy ban Bảo vệ môi trường của Mỹ (EPA), 2009. Từ điển thuật ngữ môi trường.

SUMMARY

Data development and management for specialized English-Vietnamese terms using open source software

Bùi Bảo Trung, Phạm Nguyễn Hà Vũ

Term data English- Vietnam of geology is built to support researchers and students in the fields of geology, environmental resource use in scientific research. The data is managed using open source software Geoterms run on Microsoft Windows operating system has a simple interface, easy to use, suitable for people without high-level information. Software Geoterms allow not only manager database of geological terms, resources and environment, but also can expand, connect with dictionaries in many other professional fields. Users can install the software over via free manuals.

Người biên tập: TS. Đào Thái Bắc.