CÁC PHƯƠNG PHÁP ÁNH XẠ THUỘC TÍNH HỘP THÔNG TIN WIKIPEDIA ĐẾN WIKIDATA

Tạ Hoàng Thắng[a*]

[a]*Khoa Công nghệ Thông tin, Trường Đại học Đà Lạt, Lâm Đồng, Việt Nam*

Nhận ngày 04 tháng 01 năm 2016
Chỉnh sửa ngày 31 tháng 03 năm 2016 | Chấp nhận đăng ngày 10 tháng 04 năm 2016

Tóm tắt

Wikidata là một cơ sở dữ liệu trực tuyến mở lưu trữ các tài nguyên chung của các dự án liên quan do tổ chức Wikimedia quản lý. Việc đồng nhất hóa các hộp thông tin (infobox) của Wikipedia được nêu trong kế hoạch giai đoạn 2 của Wikidata. Theo đó, các hộp thông tin sẽ được đồng nhất hóa để tránh tình trạng đa dạng dữ liệu giữa các dự án ngôn ngữ. Đồng thời, nhóm phát triển Wikidata cũng lên kế hoạch phát triển hệ thống tự động diễn dịch các thuộc tính của các hộp thông tin Wikipedia. Bài báo này nhắm đến việc đưa ra một vài phương pháp tiếp cận để ánh xạ các thuộc tính của hộp thông tin đến Wikidata, từ đó nâng cao khả năng phát triển làm giàu nội dung cho các bài viết Wikipedia. Chúng tôi tập trung chủ yếu vào việc ánh xạ các thuộc tính ở Wikipedia Tiếng Việt và Wikipedia Tiếng Anh.

Từ khóa: DBPedia; Infobox Property; Mapping; Wikidata; Wikipedia.

1. GIỚI THIỆU

Được biết đến như là bách khoa toàn thư mở trực tuyến lớn nhất thế giới, Wikipedia không ngừng phát triển nội dung bài viết để phục vụ cho mục đích giáo dục, nâng cao trình độ hiểu biết và chia sẻ kiến thức của mọi người trên thế giới. Wikipedia hiện có 291 dự án ngôn ngữ khác nhau với hàng triệu bài viết thuộc các chủ đề đa dạng. Để quản lý một lượng kiến thức nhân loại khổng lồ, Wikipedia hoàn toàn phụ thuộc vào đội ngũ thành viên đông đảo với hơn 56 triệu người dùng. Các bài viết ở các ngôn ngữ khác nhau đều có liên kết ngôn ngữ (interwiki link) để độc giả có thể tham khảo nội dung kiến thức giữa các phiên bản ngôn ngữ khác nhau. Việc duy trì các liên kết ngôn ngữ cùng các nguồn tài nguyên chung (hình ảnh, nội dung media, các tập luật ngữ

* Tác giả liên hệ: Email: thangth@dlu.edu.vn

nghĩa) này cực kỳ phức tạp, do đó Wikipedia đã tổ chức thành lập Wikidata năm 2012 [1], một máy chủ để lưu trữ các loại dữ liệu chung. Dữ liệu ở Wikidata lưu trữ theo nhiều tập dữ liệu phong phú, được kết nối với nhau được xem là mỏ tài nguyên để khai thác và rút trích các tập luật ngữ nghĩa [2]. Các bài viết ở Wikipedia thông thường đều chứa các hộp thông tin, mô tả vắn tắt nội dung của một bài viết. Một hộp thông tin bao gồm nhiều thuộc tính (hay còn gọi là các tham số), mỗi thuộc tính chứa giá trị dữ liệu đi kèm. Các hộp thông tin ở các dự án ngôn ngữ đều được đặt tên là Bản mẫu (Template) và cũng liên kết với nhau thông qua các liên kết ngôn ngữ được lưu trữ tại Wikidata. Trong Hình 1 là hộp thông tin "*Thông tin đơn vị hành chính Việt Nam*" với các thuộc tính mô tả về tỉnh Lâm Đồng như thuộc tính *tên* có giá trị "*Lâm Đồng*", thuộc tính *diện tích* có giá trị *9.773,5 km*^{2}, ...

```
{{Thông tin đơn vị hành chính Việt Nam
| tên = Lâm Đồng
| logo =
| hình = Da Lat, view to Xuan Huong lake 2.jpg
| diện tích = 9.773,5 km²
| dân số = 1.246.200 người
| thời điểm dân số = 2013
| mã hành chính = [[ISO 3166-2:VN|VN-35]]
| mã bưu chính = [[Mã bưu chính Việt Nam|67xxxx]]
| mã điện thoại = [[Mã điện thoại Việt Nam|63]]
| biển số xe = [[Biển xe cơ giới Việt Nam|49]]
| web = [http://www.lamdong.gov.vn/ Tỉnh Lâm Đồng]}}
```

Hình 1. Mã nguồn hộp thông tin và giao diện hiển thị của bài viết về tỉnh Lâm Đồng tại Wikipedia

Trong nghiên cứu trước đó [3], chúng tôi đề xuất ra mô hình làm giàu nội dung Wikipedia dựa trên các tập luật ngữ nghĩa được rút trích chủ yếu từ các hộp thông tin. Vì vậy, trong báo báo này, chúng tôi sẽ tập trung vào việc ánh xạ các hộp thông tin ở các phiên bản ngôn ngữ tới Wikidata, từ đó có thể hiểu được sự tương quan giữa các hộp thông tin ở các ngôn ngữ khác nhau để nhằm mục đích làm giàu dữ liệu và đồng bộ hóa nội dung bài viết ở các bài viết Wikipedia. Chúng tôi đề xuất các phương pháp ánh xạ với trường hợp cụ thể là dựa trên 2 ngôn ngữ tiếng Anh và tiếng Việt. Độ chính xác của các thông tin bài viết và giá trị dữ liệu các thuộc tính sẽ không đề cập đến trong bài báo này.

2. CÁC NGHIÊN CỨU LIÊN QUAN

DBPedia là tổ chức đã ánh xạ thuộc tính hộp thông tin Wikipedia đến các bản ánh xạ (mapping) và bản thể học (ontology) do DBPedia tự định nghĩa [4, 5, 6]. DBPedia rút trích nội dung ngữ nghĩa từ Wikipedia ở nhiều dự án ngôn ngữ và phân loại nội dung thành các tập dữ liệu khác nhau, lưu trữ ở dạng bộ ba RDF (RDF triples). DBPedia cũng thực hiện việc ánh xạ từ dữ liệu họ thu thập được đến Wikidata. Vì vậy, có thể xem DBPedia là cầu nối quan trọng để chúng tôi có thể kế thừa nhằm nâng cao việc thực thi ánh xạ các thuộc tính thông tin đến Wikidata. Tuy nhiên, DBPedia chưa có phép người dùng khắp nơi trên thế giới tham gia dự án một cách dễ dàng, vì vậy còn nhiều hộp thông tin ở nhiều ngôn ngữ, đặc biệt ngôn ngữ hiếm và có ít người sử dụng.

Tác giả Thanh Nguyên và cộng sự phát triển WikiMatch, một hệ thống để ánh xạ hộp thông tin ở tiếng Việt, tiếng Bồ Đào Nha và Tiếng Anh [7]. Nghiên cứu này cũng dùng phương pháp dịch thuật, từ điển và một mô hình kịch bản ánh xạ cố định để nâng cao tính hiệu quả. Nghiên cứu của Eytan và đồng nghiệp có cùng hướng nghiên cứu, tập trung ở tiếng Anh, tiếng Tây Ban Nha, tiếng Pháp và tiếng Đức [8]. Navigli cũng phát triển hệ thống BabelNet để thực thi việc ánh xạ đa ngôn ngữ [9]. Tương tự, Bouma cũng phát triển hệ thống ánh xạ thuộc tính giữa tiếng Hà Lan và tiếng Anh. [10] Các nghiên cứu này đều mang tính độc lập với DBPedia với các kết quả riêng. Do đó, chúng tôi muốn đề xuất một phương pháp lai, tận dụng các kết quả có được của DBPedia cùng một số phương pháp ánh xạ để đề xuất cho báo cáo này.

3. MỘT SỐ PHƯƠNG PHÁP ÁNH XẠ THUỘC TÍNH HỘP THÔNG TIN WIKIPEDIA ĐẾN WIKIDATA

3.1. Phương pháp kế thừa nội dung ánh xạ ở DBPedia

Chúng tôi kế thừa việc ánh xạ hộp thông tin Wikipedia đến các bản ánh xạ và bản thể học (ontology) của DBPedia, điều đó tạo điều kiện thuận lợi để tập trung nâng cao hiệu quả và mở rộng dữ liệu việc ánh xạ dựa trên dữ liệu nghiên cứu đã có. DBPedia ánh xạ một hộp thông tin cùng các thuộc tính của thành một bản thể học. Ví dụ, hộp thông tin Template: Infobox settlement ở Wikipedia tiếng Anh được ánh xạ thành Mapping en:Infobox settlement và bản thể học Settlement ở Wikidata. Mỗi thuộc tính của hộp thông tin cũng được ánh xạ tương ứng với một thuộc tính của bản thể học.

Dựa vào liên kết ngôn ngữ[1] ở Wikidata, Bản mẫu: Thông tin khu dân cư ở Wikipedia Tiếng Việt có liên kết ngôn ngữ với Template:Infobox settlement ở Wikipedia tiếng Anh. Theo Bảng 1, các bản ánh xạ và bản thể học tiếng Việt ở DBPedia là chưa có, tuy nhiên chúng ta có thể suy luận dựa vào tiếng Anh. Giữa DBPedia và Wikidata đã có sự ánh xạ các thuộc tính với nhau do kế hoạch phát triển của DBPedia.

Bảng 1. Thông tin ánh xạ hộp thông tin Settlement với DBPedia trên 2 ngôn ngữ tiếng Việt và tiếng Anh

Wikipedia tiếng Anh[2]	Template:Infobox settlement
Bản ánh xạ tiếng Anh ở DBPedia[3]	Mapping en:Infobox settlement
Bản thể học tiếng Anh ở DBPedia[4]	Settlement
Wikipedia tiếng Việt	Bản mẫu:Thông tin khu dân cư
Bản ánh xạ tiếng Việt ở DBPedia	Không có
Bản thể học tiếng Việt ở DBPedia	Không có

[1] https://www.wikidata.org/wiki/Q5683132#sitelinks-wikipedia

[2] https://en.wikipedia.org/wiki/Template:Infobox_settlement

[3] http://mappings.dbpedia.org/index.php/Mapping_en:Infobox_settlement

[4] http://mappings.dbpedia.org/server/ontology/classes/Settlement

Dựa theo tính bắc cầu, chúng tôi có thể suy luận sự tương ứng giữa các thuộc tính ở dự án tiếng Việt và dự án tiếng Anh với Wikidata. Chẳng hạn, Bản mẫu:Thông tin khu dân cư ở Wikipedia tiếng Việt dùng song thuộc tính *"ngày thành lập"* và *"established date"* và Wikipedia Tiếng Anh chỉ dùng *"established date"*. Thông qua bản ánh xạ Settelement của DBPedia, thuộc tính *"established date"* được ánh xạ thành thuộc tính ontology *"foundingDate"* với Wikidata tương ứng là *"Wikidata:P571"*. Do đó có thể suy luận, thuộc tính *"ngày thành lập"* tương ứng với thuộc tính ở Wikidata là "*P571"*.

3.2. Phương pháp ánh xạ trực tiếp thuộc tính hộp thông tin Wikipedia với Wikidata

Đối với các thuộc tính mà DBPedia chưa thể ánh xạ được tới Wikidata, chúng tôi phải dùng phương pháp trực tiếp để ánh xạ các thuộc tính này. DBPedia tạo thống kê chi tiết về việc ánh xạ mỗi hộp thông tin từ Wikipedia (đa phần với hộp thông tin tiếng Anh), bao gồm tổng số thuộc tính, số lượng thuộc tính được ánh xạ, chưa ánh xạ, tỉ lệ phần trăm các thuộc tính đã được ánh xạ và nhiều thông số khác[5]. Từ đó, có thể tận dụng bảng thống kê này để nắm bắt các thông tin thuộc tính chưa được ánh xạ. Hình 2 chỉ rõ nội dung vừa nêu trên.

Bước 1. Xếp hạng tần suất sử dụng thuộc tính

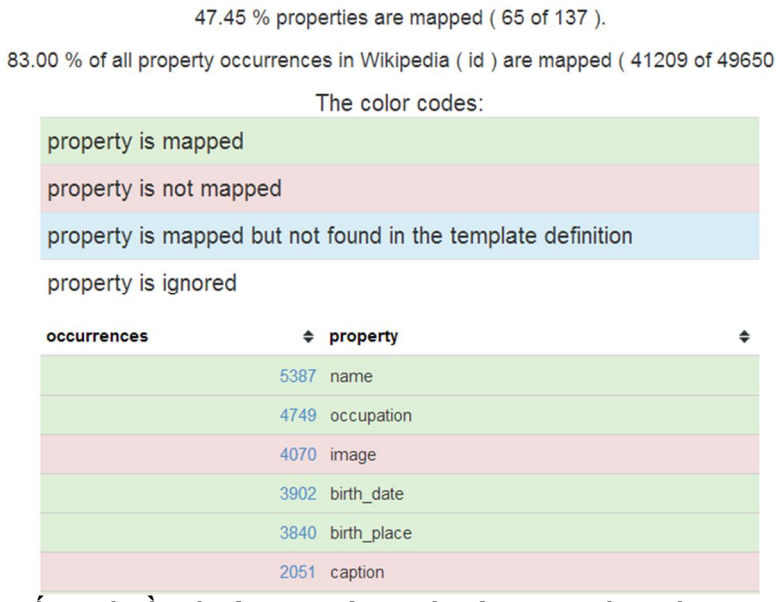
Đầu tiên, việc giảm bớt các thuộc tính ít sử dụng để giảm nhiễu là cần thiết [11]. Để làm được điều này, chúng ta có thể chọn ngẫu nhiên một lượng bài viết vừa đủ (100 bài), từ đó xếp hạng tần suất các thuộc tính được sử dụng, đối những thuộc tính còn lại chưa được sử dụng thì nên xếp hạng chúng sau cùng.

Bước 2. Tạo tập thuộc tính đồng nghĩa ứng với mỗi thuộc tính Wikipedia

Tiếp đến nên thực thi việc phân tách các thuộc tính thành các tên có ý nghĩa hơn đến mức có thểv[11]. Thuộc tính có thể là tên viết tắt hay tên viết liền, vì vậy chúng tôi dựa vào tính năng *Tự động kiểm tra chính tả và sửa lỗi* của Google (Spell-check and

[5] http://mappings.dbpedia.org/server/templatestatistics/en/?template=Infobox_settlement

automatic corrections) để tìm kiếm từ gốc. Ngoài ra, có thể áp dụng phương pháp xây dựng hệ thống nhận dạng chứa các cơ sở dữ liệu từ viết tắt riêng để tăng khả năng tìm kiếm từ gốc. Chẳng hạn, thuộc tính *official_name*, *officialname*, *officialName* được chức năng sửa lỗi Google hiển thị thành *official name*. Hay như từ viết tắt *latd* trong tiếng Anh phải nhận dạng là *latitude* (vĩ độ).


47.45 % properties are mapped (65 of 137).

83.00 % of all property occurrences in Wikipedia (id) are mapped (41209 of 49650

The color codes:

property is mapped

property is not mapped

property is mapped but not found in the template definition

property is ignored

occurrences	property
5387	name
4749	occupation
4070	image
3902	birth_date
3840	birth_place
2051	caption

Hình 2. Thống kê về việc ánh xạ các thuộc tính của hộp thông tin Person ở Wikipedia tiếng Anh tới DBPedia

Dựa vào từ điển đồng nghĩa để mở rộng tập thuộc tính đồng nghĩa, trước hết chúng ta phải dò thuộc tính thuộc ngôn ngữ [12,13,14] nào để áp dụng từ điển đồng nghĩa trong ngôn ngữ đó tương ứng. Để dò thuộc tính thuộc ngôn ngữ, chúng tôi đề xuất nên dùng một số module như Language Detection API [6] hay Google Language Detection.

Bước 3. Dò tập thuộc tính với Wikidata

Ở Wikidata, mỗi thuộc tính đều được gán 1 chỉ số chỉ mục (index) kèm theo nội dung thuộc tính ở các ngôn ngữ khác nhau, chẳng hạn *P18* là thuộc tính *Hình ảnh* trong tiếng Việt và *Image* trong tiếng Anh theo Hình 3.

[6] https://detectlanguage.com/

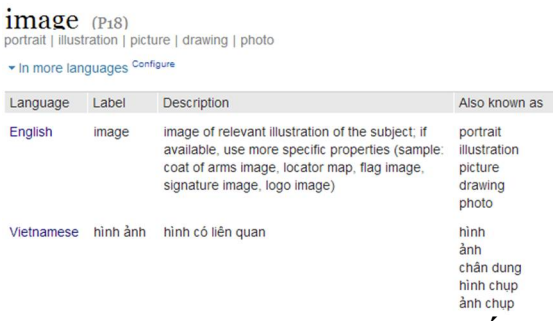


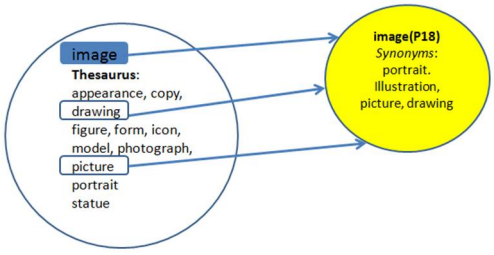
image (P18)
portrait | illustration | picture | drawing | photo
In more languages Configure

Language	Label	Description	Also known as
English	image	image of relevant illustration of the subject; if available, use more specific properties (sample: coat of arms image, locator map, flag image, signature image, logo image)	portrait illustration picture drawing photo
Vietnamese	hình ảnh	hình có liên quan	hình ảnh chân dung hình chụp ảnh chụp

Hình 3. Thuộc tính P18 ở Wikipedia ở 2 ngôn ngữ tiếng Việt và tiếng Anh

Trong Hình 3, cũng cho thấy tập các đồng nghĩa với thuộc tính *Image* đó là *portrait, illustration, picture, drawing* và *photo*. Từ đó hình thành nên tập thuộc tính Wikidata.

Công việc cuối cùng là so khớp giữa 2 tập thuộc tính, tập thuộc tính đồng nghĩa Wikipedia và tập thuộc tính Wikidata. Hình 4 cho thấy cách so khớp tuyệt đối đúng theo từ, từ đó có thể suy luận ra thuộc tính *Image* ở hộp thông tin Wikipedia tương ứng với thuộc tính *Image (P18)* ở Wikidata. Một số trường hợp, một thuộc tính Wikipedia có thể dò ra được nhiều thuộc tính Wikidata, lúc đó chúng ta có thể dùng phương pháp xếp hạng tần suất so khớp hay dò tìm dữ liệu và kiểu dữ liệu kèm theo thuộc tính trên các hộp thông tin ở các bài viết để xác định kết quả cuối cùng. Tuy nhiên, để tăng độ chính xác chúng ta cũng nên áp dụng thêm các phương pháp giám sát bán tự động khác.



Hình 4. Sự ánh xạ tập thuộc tính đồng nghĩa với Wikidata

3.3. Sự phụ thuộc nội dung từ dự án Tiếng Anh ở dự án ngôn ngữ tiếng Việt

Wikipedia Tiếng Anh là dự án có nhiều lượng bài nhất với hơn 4.9 triệu bài[7]. Vì vậy, ở các dự án khác, chẳng hạn như Wikipedia tiếng Việt, thay vì viết mới nội dung

[7] https://meta.wikimedia.org/wiki/List_of_Wikipedias

thì có một cách thông thường dễ hơn đó là dịch nội dung từ tiếng Anh sang. Do đó, nội dung các hộp thông tin cũng trong tình trạng tương tự. Ngoài ra ở Wikipedia tiếng Việt, các biên tập viên còn thêm các thuộc tính của các hộp thông tin bằng tiếng Việt bên cạnh các thuộc tính bằng tiếng Anh trước đó. Như vậy, hộp thông tin tiếng Việt luôn chưa 2 thuộc tính song ngữ tương ứng với 1 thuộc tính của hộp thông tin ở Wikipedia tiếng Anh. Bảng 2 cho thấy một vài thông tin thuộc tính của hộp thông tin Khu dân cư so với nội dung tiếng Anh.

Bảng 2. So sánh các thuộc tính của hộp thông tin Khu dân cư ở tiếng Việt và Tiếng Anh với nội dung DBPedia tiếng Anh và Wikidata

Wikipedia Tiếng Việt	**Wikipedia Tiếng Anh**	**Bản thể học DBPedia bằng tiếng Anh**	**Wikidata**
founder người sáng lập	founder	OntologyProperty:founder Wikidata:P112	P112 (dựa vào DBPedia)
official name tên chính thức	official name	OntologyProperty:foaf:name	P1448 (theo cách so ở mục 2)
hình ảnh	image	Không tìm thấy	P18 (theo cách so ở mục 2)

Tuy nhiên, đôi khi cũng có trường hợp ở dự án tiếng Việt chỉ có 1 thuộc tính tiếng Việt khớp với 1 thuộc tính tiếng Anh ở dự án tiếng Anh, hoặc ở nội dung tiếng Việt có thuộc tính không khớp với bất cứ dự án nào tiếng Anh hay ngược lại. Ngoài ra, còn có thể là thuộc tính tiếng Việt khớp với Wikidata theo cách so khớp ở mục 2, từ đó ở Wikidata lại có thể suy luận ngược về để tìm thuộc tính tương ứng ở tiếng Anh.

Như vậy, trong phần này chúng tôi đề xuất ở mỗi hộp thông tin tiếng Việt nếu có chứa thuộc tính song ngữ nên ưu tiên thuộc tính có nội dung tiếng Anh để so khớp với DBPedia hay Wikidata từ đó suy luận nhằm tiết kiệm thời gian so khớp đối với nội dung tiếng Việt.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã chỉ ra các phương pháp hữu dụng để ánh xạ các thuộc tính hộp thông tin Wikipedia đến Wikidata. Chúng tôi đề xuất các phương pháp ánh xạ lai dựa trên 2 ngôn ngữ chính là Wikipedia Tiếng Anh và tiếng Việt để từ đó thấy được tiềm năng phát triển hệ thống thực thi. Chúng tôi tin rằng báo cáo này sẽ là đòn bẩy để thực thi

việc ánh xạ cho tất cả ngôn ngữ ở Wikipedia theo một phương pháp tổng quát hơn trong tương lai. Đồng thời kết quả ánh xạ mang lại sẽ được sử dụng trong quá trình rút trích quan hệ ngữ nghĩa từ đó có thể phát triển nội dung bài viết Wikipedia ở các ngôn ngữ theo phương thức tự động hoặc bán tự động. Ngoài ra, việc ánh xạ còn giúp kiểm soát nội dung dữ liệu hộp thông tin giữa các dự án ngôn ngữ để chống phá hoại nội dung.

TÀI LIỆU THAM KHẢO

[1] Vrandečić, D., & Krötzsch, M. *Wikidata: a free collaborative knowledgebase*. Communications of the ACM, 57(10), 78-85. (2014).

[2] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. . Introducing Wikidata to the linked data web. In *The Semantic Web–ISWC 2014* (pp. 50-65). Springer International Publishing. (2014).

[3] Ta, T. H., & Anutariya, C. *A Model for Enriching Multilingual Wikipedias Using Infobox and Wikidata Property Alignment*. In Semantic Technology (pp. 335-350). Springer International Publishing. (2014).

[4] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C.. *DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia*. Semantic Web. (2014).

[5] Aprosio, A. P., Giuliano, C., & Lavelli, A.. *Automatic Mapping of Wikipedia Templates for Fast Deployment of Localised DBpedia Datasets*. In Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (p. 1). ACM. (2013)

[6] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. *DBpedia-A crystallization point for the Web of Data*. Web Semantics: science, services and agents on the world wide web, 7(3), 154-165. (2009).

[7] Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen and Juliana Freire. Multilingual schema matching for Wikipedia infoboxes. Proceedings of the VLDB En-dowment, Volume 5 Issue 2, October 2011, Pages 133-144, (2011).

[8] Eytan Adar, Michael Skinner and Daniel S. Weld. *Information Arbitrage Across Multi-lingual Wikipedia*. WSDM '09 Proceedings of the Second ACM International Con-ference on Web Search and Data Mining. Pages 94-103, (2009).

[9] Navigli, R., & Ponzetto, S. P.. *BabelNet: Building a very large multilingual semantic network*. In Proceedings of the 48th annual meeting of the association for compu-tational linguistics (pp. 216-225). Association for Computational Linguistics. (2010, July)

[10] Bouma, G., Duarte, S., & Islam, Z. Cross-lingual alignment and completion of Wikipedia templates. In Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (pp. 21-29). Association for Computational Linguistics. (2009, June).

[11] Wu, F., & Weld, D. S.. *Automatically refining the wikipedia infobox ontology*. In Proceedings of the 17th international conference on World Wide Web (pp. 635-644). ACM. (2008, April)

[12] Schulze, B. M. U.S. Patent No. 6,167,369. Washington, DC: U.S. Patent and Trademark Office. (2000).

[13] Schmitt, J. C. U.S. Patent No. 5,062,143. Washington, DC: U.S. Patent and Trademark Office. (1991).

[14] Vietnamese Wordnet. (n.d.). Retrieved April 04, (2016), from http://viet.word-net.vn/wnms

SOME APPROACHES FOR MAPPING WIKIPEDIA INFOBOX PROPERTIES TO WIKIDATA

Ta Hoang Thang[a*]

[a]*The Faculty of Information Technology, Dalat University, Lamdong, Vietnam*
[*]*Corresponding author: thangth@dlu.edu.vn*

Article history
Received: January 04th, 2016
Received in revised form: March 31st, 2016
Accepted: April 10th, 2016

Abstract

Wikidata is an open, online database which stores the common resources of other Wikimedia projects. Unifying Wikipedia infoboxes was described in Phase II of Wikidata plan which aims to augment auto-translation to Wikipedia infobox templates and deals with the diversity of Infobox data in all languages. In this paper, we offer some approaches to map Infobox properties to Wikidata for improving our enrichment model. Our results can be a valuable resource for Wikidata to alignInfobox properties. We mainly focus on how to map Vietnamese and English properties to Wikidata.

Keywords: DBPedia; Infobox Property; Mapping; Wikidata; Wikipedia.