3C1-OS-13a-3

DashSearch LD: 探索的検索の Linked Data への適用

DashSearch LD: Exploratory Search for Linked Data

後藤 孝行[*1] 濱崎 雅弘[*2] 武田 英明[*1]
Takayuki Goto Masahiro Hamasaki Hideaki Takeda

[*1]国立情報学研究所
National Institute of Informatics (NII)

[*2]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

Although the bigger number of datasets gathered as LOD makes the better chance for data sharing and re-using, it also makes more difficulty to understand the datasets themselves. Since each dataset has its own data structure, we should understand them one by one. Since entities in datasets are interconnected, we furthermore understand interconnection between datasets. So understanding data is crucial to exploit LOD. In this paper, we show a novel system called DashSearch LD to understand and use LOD with exploratory search approach. The user interactively explores datasets by viewing and selecting entities in them. She just operates widgets in the screen with the mouse, e.g., moving and overlapping them, in order to check entities, draw detail data on them, and obtain other entities linked by them.

1. はじめに

近年，Linked Open Data (LOD) は，その規模を急速に拡大し，また非常に多様性を帯びてきた．LOD は，多様なデータを共有し再利用することで，データの創発的な価値を生み出す大きな可能性を秘めている．ただし，LOD として集積されたデータ集合の膨大なサイズは，多様なデータの組み合わせを生み出す機会を作る一方，データ集合自体の理解を困難にしており，その多様な活用を難しくしている．各データ集合は独自のデータ構造を持っているため，データを利用するためにはそれらを一つずつ理解する必要がある．また，データ集合内のエンティティは相互接続されているので，さらにデータ集合感の相互接続の理解も必要である．つまり，LOD を活用するためにはデータの理解が必要不可欠であるにも関わらず，その規模が拡大することによって利用者はその把握が困難になり，活用しきれないという状況になりつつある．

そこで本稿では，探索的検索アプローチによって，LOD を理解し利用する DashSearch LD と呼ぶ新しいシステムを提案する．DashSearch LD は，対話的にエンティティを選択や閲覧を行いながらデータ集合を探索する．ユーザは，ウィジェットを動かしたり，重ね合わせたりするだけで，例えば，エンティティの確認や，詳細表示，他のエンティティとの繋がりを得ることができる．

2. 関連研究

総じて LOD は，巨大かつ複雑なデータベースから構成されており，これを SPARQL クエリ言語によって検索を行う．しかし，適切な SPARQL クエリを作成するには， データ構造をよく理解する必要がある．データの理解には，二つの側面がある．一つは，スキーマの理解である．例えば，どの様な種類のクラスが存在し表現されているのか，また，どのように相互接続されているのかといった構造の理解である．もう一つは，データの分布の理解である．データ集合のクラス定義の中には非常に使われている，すなわち多くのエンティティが存在するものがあれば，その一方で，ほとんど使われていない，ほとんどエンティティが存在しないものもある．また，同じプロパティでも，常に値をもっているものもあれば，値が欠落しているものもある．このような，偏った分布は，実際のデータ集合ではよく見かけるものである．LOD を活用するためには両方の理解が重要であるが，既存の研究では主にそれらのいずれかに焦点を当てている．

Kiefer らの iSPARQL[Kiefer 07] や Russell らの NITELIGHT[Russell 08] はグラフインタフェースによって SPARQL クエリの作成を支援することができる．グラフインタフェースは，ユーザにとって直感的なだけでなく，RDF の性質に適している．しかし，これらは，語彙の事前登録が必要で，横断的な LOD グラフの場合，新たな語彙が現れる場合があるので，そのまま LOD に適用するのは難しい．そして，スキーマの理解が必要である．

Tummarello らの Deri Pipes[Phuoc 09] は，フロー図形式の GUI インタフェースを用意することで RSS でのマッシュアップアプリケーションの迅速な開発を可能にする Yahoo! Pipes[*1]のセマンティックバージョンである．DERI Pipes は，RSS の代わりに RDF トリプルと SPARQL クエリを操作することができる．アプリケーション作成には使いやすくて便利ではあるが，スキーマの理解には何度も操作が要求される．Jarrar らの MashQL[Jarrar 11] も同様のシステムであるが，SPARQL クエリの構築によりスマートなアプローチをとっている．スキーマがシステムによって検出されることで，ユーザは視覚的に指定することができる．これは，データからスキーマを検出する点で有利であるが，まだその解釈は，ユーザ任せである．

Deligiannidis らは，RDF トリプルの可視化とグラフの横断によって探索を行える，Paged Graph Visualization (PGV) というシステムを提案している．これは，データ構造とデータの理解の両方を支援することができるが，個々の関係をすべて表すことは，目障りになる場合もあり，むしろ包括的な理解を妨げる．

LOD を活用し新たなマッシュアップサービスの構築を支援するためには，スキーマの理解と全体的なデータの傾向の両方を手軽に確認できるシステムが必要になる．

連絡先: 後藤孝行，国立情報学研究所，東京都千代田区一ツ橋 2-1-2，tygoto@nii.ac.jp

*1 http://pipes.yahoo.com/pipes/

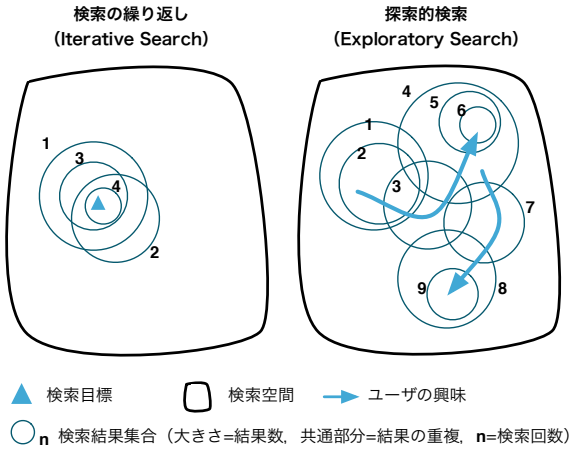

図 1: 検索の繰り返しと探索的検索 ([White 09] を基に作成)

3. 探索的検索

探索的検索 (Exploratory Search)[Marchionini 06] とは，事実検索 (Fact retrieval) や質問応答 (Question answering) のように 1 回の質問で答えを得るような参照型 (Lookup) の情報検索ではなく，探索目的を少しずつ明確化しながら新しい知識を獲得していく学習 (Learn) や調査 (Investigate) のような情報検索のことである．探索的検索では，従来の情報検索が想定する検索の繰り返し (Iterative Search) のように，検索目標は一定でそこにむかって検索結果をしぼりこんでいくような検索プロセスモデルとは異なり，探索的閲覧 (Exploratory Browsing) と絞り込み検索 (Focused Searching) の異なるモデルが出現する [White 09]．探索的閲覧とは検索空間を広げる方向をもつモデルで，絞り込み検索とは検索空間を絞り込む方向をもつモデルである．図 1 の右の図が示すように，探索的閲覧によって検索空間を遷移しつつ，途中絞り込み検索によって検索結果を絞り込み，また探索的閲覧によって検索空間を遷移するといった行為が確認できる．これによって，ユーザは情報要求の具体化だけでなく，検索空間を理解し検索に最適なクエリの入力が可能になる．

我々は，メタデータ検索においてこの探索的検索行為を支援するため，多様な検索視点で試行錯誤しながら情報を探し出す「探索的メタデータ検索 (Explorative Metadata Search)」を提案している [後藤 11]．探索的メタデータ検索は，ウィジェットと呼ぶ視覚的オブジェクトを用いるこ とで直接操作による検索式やファセット検索を実現し，検索過程の中で検索空間の理解と情報要求の具体化をおこなっていく．そして，このコンセプトを実装した検索インタフェース DashSearch を開発している．この DashSearch を LOD に適用することで，探索的検索行為が LOD でも行いやすくなり，データ集合の理解につながる．

4. DashSearch LD の機能

DashSearch LD は，ウェブブラウザ上で動作する (図 2)．ツールバーに登録してあるウィジェットをフィールドに配置することで検索が利用できるようになる．ウィジェットには，SPARQL Endpoint の機能を持つ，エンドポイントウィジェットと，プロパティとその値を表示しリソースの絞り込みに利用できるメタデータウィジェットがある．

エンドポイントウィジェットにおいて，キーワード検索を行った後，メタデータウィジェットをエンドポイントウィジェットに重ねると，検索結果のリソースが持つプロパティとその値をメタデータウィジェットに表示することができる．またプロパティの値の件数も表示されるため，どのようなエンティティが多く存在するのかを知ることができる．

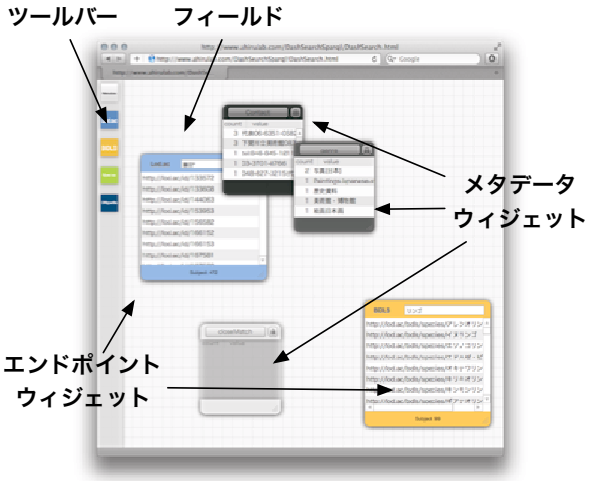

図 2: DashSearch LD

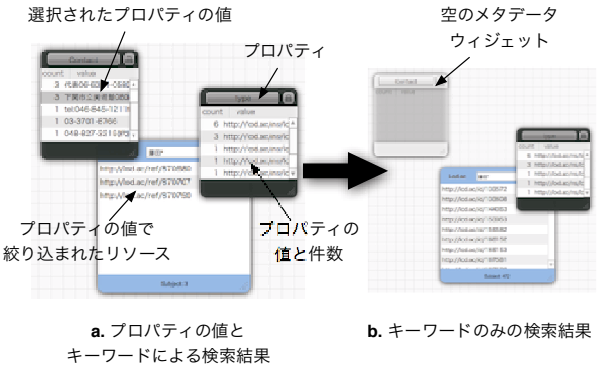

図 3: プロパティの値による絞り込み

表示されているプロパティの値を選択すると，その値によって検索結果を絞り込むことができる (図 3-a)．絞り込みに利用したメタデータウィジェットをエンドポイントウィジェットから離すとプロパティによる検索条件が外れる (図 3-b)．このように，ウィジェットをマウス操作によって移動させるだけで，検索条件を追加したり外したりすることができ，また，メタデータウィジェットによる検索条件は複数用いることができる*2ので，多様な条件の組み合わせを気軽に試すことができる．

メタデータウィジェットの上に，メタデータウィジェットを重ねることができる．そして，下に位置するメタデータウィジェットのプロパティの値を選択すると，リソースの絞り込みだけでなく，上に重ねたメタデータウィジェットのプロパティとその値を絞り込む (図 4)．このように複数のメタデータウィジェットを重ねることで，ファセット検索を実現することができ，これによって，選択したプロパティの値に関係する他のプロパティの値を知ることができる．

検索結果から得られたプロパティの値を他の Endpoint に用いることで，メタデータを媒介とした関連検索を実現する (図

*2 条件の組み合わせは論理積を表す

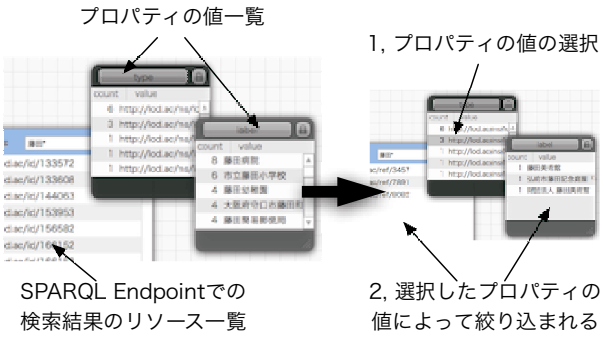

図 4: ファセット検索

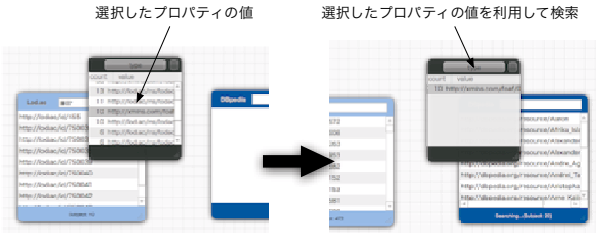

図 5: 関連検索

5). これによって，他の SPARQL Endpoint において，選択したプロパティの値を持つリソースが存在するかを調べることもできる．

複数のエンドポイントウィジェットをまたぐようにメタデータウィジェットを重ねると，異なる検索結果に共通するプロパティとその値を知ることができる (図 6)．

このように，DashSearch LD を利用することで，データ集合のスキーマやデータの分布の理解，また異なるデータ集合間の関係性を把握することができるようになる．

4.1 実装

DashSearch を LOD に応用するにあたって，膨大なプロパティにどう対応するかが大きな課題になる．データ集合ごとに使われているプロパティが異なっているため（共通するものもある），事前に利用するプロパティを決めることはできない．そこで，DashSearch LD では，エンドポイントウィジェットにおけるキーワード検索*3に該当するリソースが持つプロパティとその値を最大 10,000 件取得して，それを集計することでデータ集合が持つプロパティ情報を取得している．そして，DashSearch LD は，一度取得したプロパティ情報を任意で保持することができる．通常，エンドポイントウィジェットからメタデータウィジェットを離すと，プロパティ情報はメタデータウィジェットから消えてしまうが，ホールドボタンを押すと，プロパティ情報を保持し続ける (図 7)．これによって，一旦キーワード検索を行わなくても，プロパティ情報によるメタデータ検索を行うことができる．

次に，膨大なプロパティを利用するには，プロパティとその値を素早く確認する必要がある．そこで，DashSearch LD ではプロパティボタンを押すと，リスト表示されたプロパティ一覧が表示され，プロパティを選択するとその背景に選択したプロパティの値一覧が表示される (図 8)．これによって，プロパティを選択しながらその値を確認することができるようになり，膨大なプロパティ情報を効率よく閲覧することができる．

*3 label を対象にした全文検索

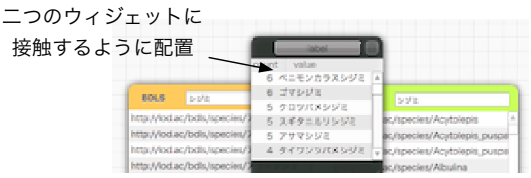

図 6: 異なる検索結果に共通するプロパティの値の表示

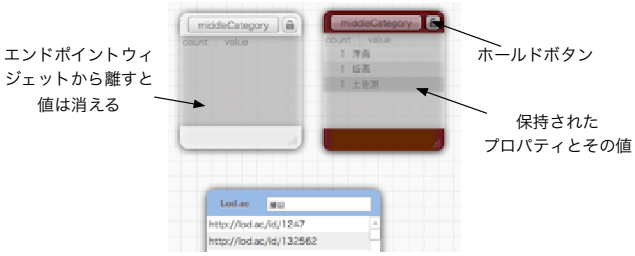

図 7: プロパティの保持

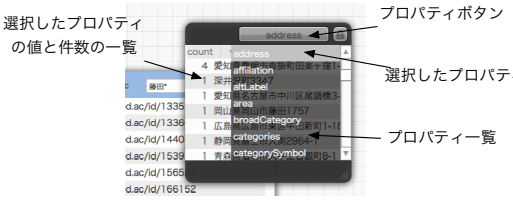

図 8: プロパティの選択

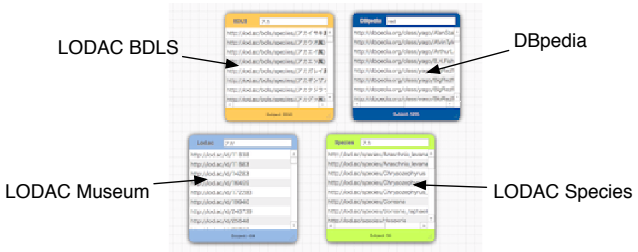

図 9: 多様な SPARQL Endpoint

DashSearch LD は，Endpoint の URL を指定するだけで，新たなエンドポイントウィジェットを追加することができる．これによって，多様な SPARQL Endpoint を検索に取り込むことができ，様々なデータの組み合わせを確認することができる (図 9)．

5. 議論

多くの多種多様なデータ集合が LOD として公開されるようになったことで，複数のデータを組み合わせて新たな情報発見の視点を提供するマッシュアップが容易に作成できるようになった．そして，一つのデータでは得られなかった価値を生み出している [Berners-Lee 10]．

その一方で，時間や場所にデータをマッピングするような決まり切ったマッシュアップが氾濫し，大規模な LOD 集合にもかかわらず，価値のある組み合わせは多種多様であるとはいいがたい状況でもある．この原因として我々は，LOD があまりにも多様で膨大なデータ集合のため全体像がつかみにくく，どのような組み合わせが可能なのかを検討できないため，意外な組み合わせが生まれにくくなっているのではと考えた．

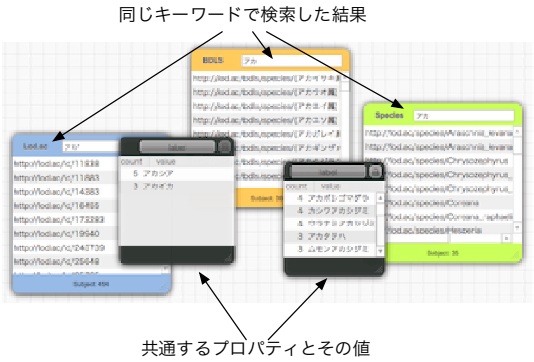

図 10: 異なるデータ集合の関連性の調査

DashSearch LD は，異なるデータ集合において同じ検索キーワードで共通プロパティと値を見つけることができるなど，大まかにデータ集合を理解することができる．これによって例えば,「アカ」,「アオ」などの色を表す言葉が，異なるデータベースにおいて共通して存在しているのかをすぐに確認することができる (図 10)．美術の所蔵データ，生物学辞書，標本データを色にまつわる言葉にマッピングすることで，意外な発見を導きだせるかもしれない．このような，ちょっとした思いつきをすぐ試せることが有益なデータの組み合わせの発見につながると考える．もちろん，LOD は公開データであるので，任意のデータベース間で共通するアイテムや語彙を列挙することは不可能ではない．だがそのようなアプローチでは，やはりまた膨大な情報に埋もれてしまうだろう．提案手法は探索的にデータベース間の共通性を発見できるところに優位性があると考える．

LOD はデータベースを横断してマッシュアップされることで，データの価値がさらに高まるとされている．組み合わせて価値があがると簡単にわかるようなものであれば，データの持ち主がすでにやっているであろう．つまりマッシュアップは本質的にデータの持ち主では到底気付けなかったであろう意外な組み合わせを見つけることが求められる．そのためには数多くの試行錯誤が必要となる．提案手法による探索的メタデータ検索は，まさにそのような思考錯誤プロセスを支援するものであり，本システムを利用することで意外なデータを組み合わせ，新しいマッシュアップサービスが生まれてくることを期待する．

6. まとめ

我々は，LOD のデータ理解のため，探索的検索を LOD で可能にする DashSearch LD を開発した．DashSearch LD によって，LOD を試行錯誤しながら情報探索する過程の中で，プロパティ間の関係把握や，異なる Endpoint に共通するプロパティ，また，共通するプロパティの値などを簡単に確認することができるようになり，データ集合の理解を支援することができた．

今後は，データの性質に応じて表示を変更するなど，よりマッシュアップを支援するようなツールを目指して開発していきたいと考えている．

参考文献

[Berners-Lee 10] Berners-Lee, T.: オープンデータとマッシュアップで変わる世界, http://www.aoky.net/articles/tim_berners_lee/the_year_open_data_went_worldwide.htm (2010)

[Jarrar 11] Jarrar, M. and Dikaiakos, M. D.: A Query Formulation Language for the Data Web, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 99, No. PrePrints (2011)

[Kiefer 07] Kiefer, C., Bernstein, A., and Stocker, M.: The Fundamentals of iSPARQL: A Virtual Triple Approach For Similarity-Based Semantic Web Tasks, in *the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference* (2007)

[Marchionini 06] Marchionini, G.: Exploratory search: from finding to understanding, *Communications of the ACM*, Vol. 49, pp. 41–46 (2006)

[Phuoc 09] Phuoc, D. L., Polleres, A., Morbidoni, C., Hauswirth, M., and Tummarello, G.: Rapid semantic web mashup development through semantic web pipes, in *the 18th World Wide Web Conference (WWW2009), Madrid, Spain, April* (2009)

[Russell 08] Russell, A., Smart, P. R., Braines, D., and Shadbolt, N.: NITELIGHT: A Graphical Tool for Semantic Query Construction, in *Semantic Web User Interaction Workshop (SWUI 2008)* (2008)

[White 09] White, R. and Roth, R.: *Exploratory Search: Beyond the Query-Response Paradigm*, Morgan & Claypool Publishers (2009)

[後藤 11] 後藤 孝行, 濱崎 雅弘, 武田 英明, 塚田 浩二, 安村 通晃, 視覚的オブジェクトを用いた探索的メタデータ検索, 情報処理学会論文誌, Vol. 52, No. 4, pp. 1504–1514 (2011)