博物館が所蔵する生物標本情報の Linked Open Data 化の試み

南 佳孝[1]　武田 英明[1]　加藤 文彦[1]　大向 一輝[1]
新井 紀子[1]　神保 宇嗣[2]　伊藤 元己[3]

[1]国立情報学研究所　[2]国立科学博物館
[3]東京大学大学院総合文化研究科広域科学専攻広域システム科学系

近年，Web の普及により，多くの情報を入手できるようになった．生物多様性の分野に着目すると，生物の種名や分布，各種の特徴や保全状況といった生物的な情報が公開されている．しかし，現状では，それらの情報の形式や公開場所は分散しているため，関連が弱い．そこで本研究では，Linked Open Data(LOD)の技術を用いて，生物多様性の情報を統合的に利用するための基盤を構築し，デジタル・アーカイブとして活用することを考えた．我々はすでに和名を含む種名情報に関する LOD を構築してきたが，今回は，ここに博物館が所蔵する生物標本情報を LOD 化して統合する．この結果，標本情報に欠落していた分類群情報の補完や多様な種名での検索などが可能になった．

Study of Museum Specimen Information with Linked Open Data

Yoshitaka Minami[1]　Hideaki Takeda[1]　Fumihiro Kato[1]　Ikki Ohmukai[1]
Noriko Arai[1]　Utsugi Jinbo[2]　Motomi Ito[3]

[1]National Institute of Informatics
[2]National Museum of Nature and Science, Tokyo
[3]Department of General Systems Studies, Graduate School of Arts and Sciences, The University of Tokyo

Availability of information on biodiversity has dramatically improved thanks to Web. But there exist a lot of different information sources that are not well associated since the field of biodiversity contains various biology domains from molecular biology to ecology. We are building the data site of species and taxon names with LOD to interlink such information sources. In this paper, we describe how specimen information in museums can be integrated into the LOD. As a result, specimen information is semantically improved, i.e., incomplete and wrong taxon information can be complimented and various names for specimen are integrated to enable flexible search for specimen.

１．はじめに

近年，Web の普及により，インターネットを介して多くの情報を入手できるようになったが，それらの情報は，人間が読むことを前提に作られており，コンピュータがその内容を処理することは容易ではない．Web を発明した Tim Berners-Lee は，コンピュータがその内容を処理できる仕組みとしてセマンティック Web を提唱した[1]が，現状では，必ずしもセマンティック Web が発展・普及したとは言えない．

ところが，社会に Web が浸透したことで，膨大な情報が Web 上で提供されるようになったため，これらの情報をコンピュータで処理するセマンティック Web の重要性が注目されるようになってきた．それを実現するための手法として，Linked Open Data（LOD）[2]が提唱された．

LOD は，RDF などの言語を用いて記述されるシンプルで柔軟性がある仕組みで，多様なデータを記述することができる．そのため，欧米や米国では，既に新しい情報公開・共有の仕組みとして認知されつつあり，情報流通の仕組みとして普及しつつある．また，我が国においてもさまざまな研究や活動が行われている[3].

本研究では，生物学の中でも生物多様性の分野に焦点をあてた．この分野は，現在，生物多様性の損失や保全など，地球環境問題の 1 つとして社会問題にもなっている[4][5]．これらの問題を解決するには，地球規模の観測から人間活動まで様々な情報を横断的に利用できる基盤が必要である．しかし，生物の種名や分布，各種の特徴や保全状況といった生物的な情報でさえ，現状では形式や公開場所が分散しており関連が弱い．

そこで，本研究では，LOD の技術を用いて，分散的に公開されている生物多様性の情報を統合的に利用できるようにすることを考えた．

２．LODAC プロジェクト

筆者らが所属する情報・システム研究機構は，国立情報学研究所，国立極地研究所，統計数理研究所，国立遺伝学研究所からなり，その 4 研究所の分野を超えた研究を活性化するために新領域融合研究センターが設立された．LODAC（Linked Open Data for ACademia）プロジェクトとは，センターのプロジェクトの 1 つである「異分野研究資源共有・協働基盤の構築（略称：サイエンス 3.0 基盤構築）」のサブプロジェクト「学術リソースのためのオープン・ソーシャル・セマンティック Web 基盤の構築」の通称である．LODAC プロジェクトは，広く学術に関する情報・データを共有する仕組みを LOD で構築することを目標に，2010 年 4 月に開始し，2010 年 12 月には，LODAC Museum として Web サイトを公開した[6][7][8]．その後，関連する美術館などのデータを拡充[9][10]するなどの活動を展開している．

生物多様性に関する情報としては，日本産蝶類和名学名便覧[11]を基に，分類体系・種名・種の特徴・標本に関する情報について LODAC Species として LOD 化し[12]，さらに，情報・システム研究機構ライフサイエンス統合データベースセンターが 100 近くの多様な生物に関わる辞書を統合した生物学辞書[1]も LOD 化した[13]．

本研究では，LODAC プロジェクトで構築してきた情報基盤に，標本情報を LOD 化してリンクすることにより，生物情報基盤を構成するデジタル・アーカイブとして活用することを目指す．

３．S-Net について

生物には，分子レベルから生態系レベルまで多層のレイヤーが存在し，生物多様性もこうした多層レイヤーから構成されている．中でもその中核をなす種の多様性は，主に個体や種の名称・特徴といった情報が扱われ，大きく分けて (1) 生物名の目録の情報（種名情報），(2) 標本や観察記録などの情報（分布情報），(3) それぞれの生物種の特徴を示す情報（種情報）からなる．このようなデータを情報技術により保存・解析・活用することを目的とした横断的分野は，生物多様性情報学（biodiversity informatics）[14]とよばれる．

このような生物多様性情報は，生物分類学の研究成果として，18 世紀より紙媒体に蓄積されてきたが，情報技術が発達した現在では，膨大な情報を扱うデータベースに重要な情報ストレージとして蓄積されている．その例としては，グローバルなものとして地球規模生物多様性情報機構（The Global Biodiversity Information Facility : GBIF，種名・分布情報），Encyclopedia of Life（EoL，種情報），Catalogue of Life（CoL，種名情報），Barcode of Life Data Systems（BOLD，DNA・標本情報）などが挙げられる．一方，国内では国立科学博物館が運営するサイエンスミュージアムネット（S-Net，標本情報，GBIF と連携）がある．

S-Net は，全国の科学系博物館の情報を横断検索できる Web サイトである．S-Net では，各博物館の Web ページ上にある情報を検索できる「Web 情報検索」と，各博物館が保有する標本情報と採集に関する情報を検索できる「自然史標本情報検索」の 2 種類の検索を行うことができる．特に，「自然史標本情報検索」では，全国の 55 館の協力博物館から収集した情報を一意の形式で提供しており，人文科学の観点からも，利用者に有益なサービスとなっている．

しかし，個別の情報を見てみると，多数の博物館から収集されているため，分類群に関するデータが記載されていないなど情報の偏りや表記揺れがある．また，その情報は，人間が読むことを前提に作られており，コンピュータが利用しやすい形式になっていない．

そこで，本研究では，S-Net で提供されている標本情報を対象に，LOD 化することによって，それらの問題点を改善することを考えた．

４．標本情報の LOD 化

S-Net で扱われている標本情報は，学名，一般名（和名），分類群に関する項目，種の命名者，採集地，最終日，採集者番号，標本の性別，グローバルユニーク番号，データ種別，タイプ標本，所蔵博物館，備考というデータで構成されている．また，S-Net が提供しているサービスは，検索サービスのみであるため，クローリングによってデータを取得する．そして，取得した標本データ 1 件につき 1 つの URI を生成し，一般名（和名），採集地，採集日，所蔵博物館のデータに対して LOD 化する．

具体的には，図 1 に示すデータ構造に基づいて，LOD 化を行う．まず，各標本情報に LODAC Museum で定義している固有の ID を割り当て，URI を生成する．このとき生成する URI は，LODAC Museum でのデータの管理方法に倣って，標本 URI と標本参照 URI を作成する．URI には，それが標本情報であることを判別できるように，rdf:type の定義を行う．そして，生成した URI から種情報へのリンク，情報が記載された S-Net の Web ページへのリンク，LODAC プロジェクトで定義した機関 URI へのリンクを生成し，一般名（和名），採集地，採集日，所蔵博物館名については，リテラル（文字列）のリンクを生成する．

[1] http://lifesciencedb.jp/bdls/

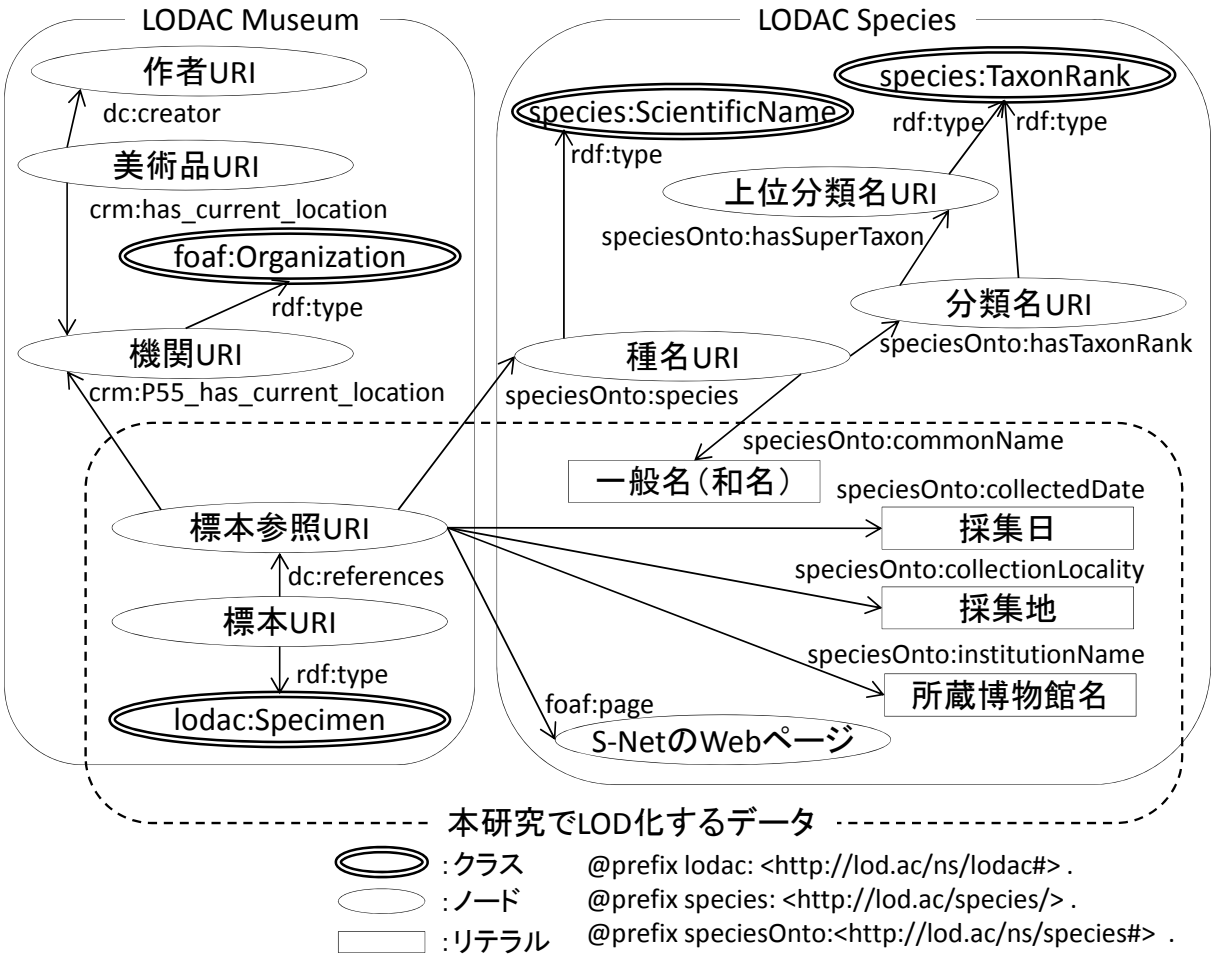

図 1　データ構造

５．LOD 化の問題と対処

S-Net のデータは，全国の協力博物館から情報を収集しているという性質のため，前述した（1）分類群に関する情報の有無の他に，（2）学名の表記揺れや（3）所蔵博物館名の記述方法の違いがあった．これらの問題について，どのような事例があるのかを確認した上で，LOD 化する場合の対処法を考えた．

（1）分類群に関する情報の有無

まず，分類群に関する情報の有無については，界名から種小名まで全て記載されているデータや属名・種小名のみといったデータなど，様々な場合があった．

一方、LODAC Species では，これまでに種名および分類群に関する情報を LOD 化してきた．省略されている分類群の情報は，適切なリファレンスがあれば一意的に補完できる．

そこで，S-Net の標本情報から LODAC Species の種名 URI へのリンクを適切に記述すれば，S-Net のデータを LOD 化したあと，標本情報から分類群に関する情報を参照できると考えられる．S-Net で記述されている分類群に関する情報については今回の LOD 化の対象から除外した．

（2）学名の表記揺れ

次に，種を示す学名の表記揺れについて対応を考えた．S-Net で提供されているデータについて，*Papilio xuthus*（和名：アゲハチョウ）の例を挙げる．動物の学名は，国際動物命名規約により属名（Papilio）と種名（xuthus）を組み合わせて表記する（二語名法）と定められている．しかし，「Papilio xuthus」と記述されている場合の他に，「Papilio xuthus Linnaeus」や「Papilio xuthus Linnaeus,1767」のように学名の後ろに命名者や年号を付加することもできる．これらは正しい学名の表記揺れと言える．また，「Papilio xuthus xuthus」のように属名・種名に亜種名を付加した 3 語での表記もある．一方で「Papilio xuthus B1292175」は，学名としては正しくないが，博物館で管理されている識別番号を学名の後ろに付加したものである．また，動物の学名では，命名者や年号が括弧で括られている場合もあるが，誤ってこれを省略する表記

もある．このように，学名だけ見ても多様な表記揺れが存在する．

この問題に対して，多様に表記されている学名全てに URI を割り当て，owl:sameAs で LODAC Species の種名 URI にリンクする方法を採用した．この方法を採ることによって，オリジナルのデータを改変する必要無く，既存のデータとリンクする事ができ，種名とリンクしている分類群などの各種データを参照することができるようになる．

（3）所蔵博物館名の記述方法の違い

S-Net の所蔵博物館名を確認すると，LODAC Museum でこれまで扱ってきた美術館名や博物館名と，標本情報が所蔵されている博物館名の記述方法が異なることがわかった．例えば，LODAC プロジェクトでの博物館名が「北九州市立自然史・歴史博物館(いのちのたび博物館)」となっているのに対し，S-Net のデータでは，「北九州市立自然史博物館」と表記されていた．また，国立科学博物館を指すと思われる名称が LODAC Museum では、「国立科学博物館」，「独立行政法人国立科学博物館附属自然教育園」，「国立科学博物館産業技術史資料情報センター」，「国立科学博物館分館」と複数存在し，S-Net のデータにも「国立科学博物館植物研究部」，「国立科学博物館（動物）」，「国立科学博物館（動物･人類）」，「国立科学博物館（植物）」と複数存在することがわかった．このような記述方法は，博物館が，組織や施設，コレクションと言った目的ごとに異なる分類を行うことから発生したと考えられる．

LOD 化において，リテラルでそのままデータを登録するには問題は無いが，これらのデータをリンクして活用することを考えると，非常に利用しにくいデータとなる．そこで， LODAC Museum で管理している美術館名や博物館名を含む機関 URI と比較し，該当する博物館の有無を確認した．該当する博物館が無い場合は，所蔵博物館について機関 URI として新規追加し，機関 URI のデータとリンクする事を考えた．

LODAC Museum[2]と LODAC Species[3]では，それぞれで SPARQL endpoint を公開している．そこで，この 2 つの SPARQL endpoint を活用してリンクを生成する事を考えた．この処理には、Silk[4]というツールを利用した．

Silk は，2 つの異なるデータソースのデータ項目間のリンクを生成するツールである．SPARQL endpoint を利用でき，リンクするプロパティはもちろん，2 つのデータソースを比較・マッチングするときに，様々な条件を設定することができる．

本研究では，S-Net のデータを登録した後，LODAC Museum と LODAC Species の SPARQL endpoint を用いて，LODAC Museum の約 20 万件と LODAC Species に登録した S-Net の約 120 万件のデータを対象に Silk でリンクを生成した．合計で約 2400 億回のマッチング処理が行われ，処理時間は約 11 時間であった．

６．LOD 化の結果

上記の方法で，S-Net のデータを LOD 化し，LODAC プロジェクトの RDF Store に登録した．トリプル数は，9,754,474 となった．

LODAC プロジェクトでは，HTML によるインタフェースを備えており，SPARQL Endpoint も公開している．そのため，個別の標本情報を閲覧することはもちろん，標本情報に関連する情報についてリンクをたどって閲覧したり，ある種に関する標本情報を一覧として取得したりすることができる．

図 2に，LODAC の HTML インタフェースでのアゲハチョウに関するデータの表示例を示す．左側が LODAC Museum に登録したアゲハチョウの標本の内容である．ページ内の左側の列がプロパティ，右側の列がオブジェクトを示している．3 行目は，図 2の右上に示す LODAC Species のアゲハチョウの種名 URI へのリンクを示している．種名 URI には，分類群や関連 Web ページへのリンク，画像などがリンクされており，それらのデータを閲覧することができる．10 行目は，図 2の右下に示す S-Net の Web ページへのリンクを示しており，データを比較すると，同一のものであることがわかる．

図 3は，Trypoxylus 属の標本を所蔵する博物館を検索する SPARQL クエリ例である．これまで，LODAC Museum では，博物館が所蔵する美術品などの作品の検索を行うことはできたが，他の所蔵品を検索することはできなかった．しかし，本研究で S-Net のデータを登録し，機関 URI とリンクする事によって，美術品と一見関係の無かった生物標本情報を検索できるようになった．さらに，これを応用して，ある地域の博物館に絞り込んだり，採集日で絞り込むといった柔軟な検索もできるようになった．

７．おわりに

本研究では，S-Net で提供されている標本情報を対象に，LOD 化を行った．標本情報はおもに種名・分類群情報，博物館情報，採集地点情報から構成される．本研究は，LODAC プロジェクトでこれまで構築してきた種名・分類群および博物館のデータとリンクすることで，情報の偏りや表記揺れを補完する働きを実現できた．

[2] http://lod.ac/sparql
[3] http://lod.ac/species/sparql
[4] http://www4.wiwiss.fu-berlin.de/bizer/silk/

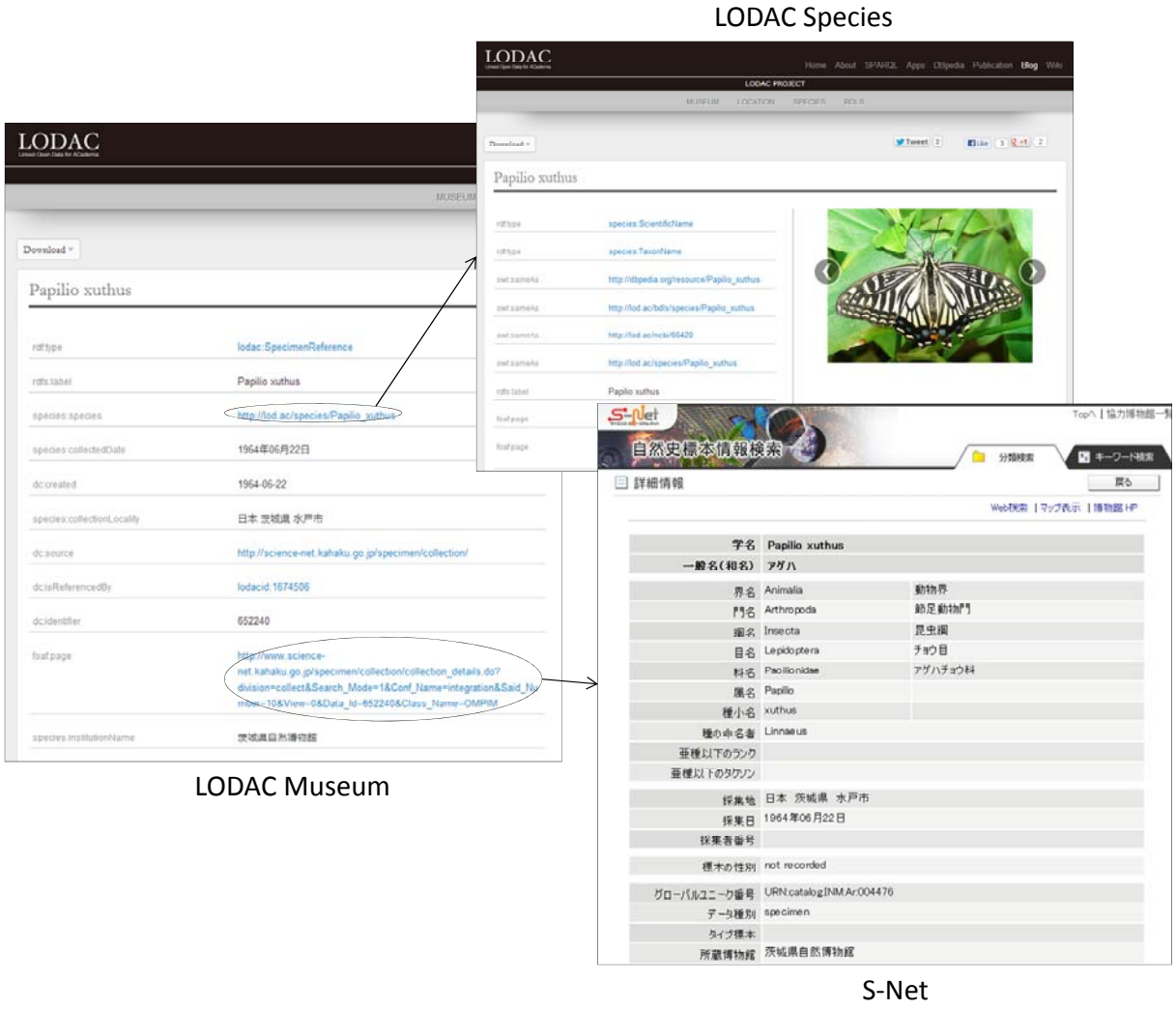

図 2　結果表示例

分類群に関する情報については，LODAC Species とリンクすることによって，S-Net のデータに欠けていた情報を補完ができ，LODAC Museum と LODAC Species の SPARQL endpoint を利用して，同属の標本を抽出するといった検索も可能になった．

学名の表記揺れについては，それぞれの学名に URI を付与することによって，関連する学名で異なる表記のものを一覧することができるようになった．

所蔵博物館名の記述方法の違いについては，機関 URI のデータとリンクする事によって，異なる目的のデータセットを繋ぐことができた．本研究が LOD の観点から注目すべき点は，「美術品」と「生物標本」といった異質なデータに関連が生まれた点にある．今回，博物館情報を利用した LODAC Museum は，博物館および美術品データを LOD 化したものであるが，美術品も生物標本も異質でありながら双方は所蔵品という概念で抱合される．すなわち，人文系と自然科学系の博物館の所蔵品をシームレスに扱うための基盤となり得ると考えられる．特に，人文科学における考古学資料と自然科学系の標本資料は，どちらもその資料を採取した場所や日時，所蔵機関などが連携できると考えられ，そこからつながる様々なデータのハブになると考えられる．

このように，生物標本情報を LOD 化し，これまでに構築してきた情報とリンクすることによって，コンピュータが利用しやすい形式になった．そのため，要求に応じた柔軟な検索が可能になり，未知なデータとつながる基盤としても今後，大きな可能性があると考える．また，本研究の成果は，デジタル・アーカイブとして標本情報の利用価値の向上，そして，相互運用性の向上にもつながると考える．

```
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX crm: <http://purl.org/NET/cidoc-crm/core#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX lodac: <http://lod.ac/ns/lodac#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX speciesOnto: <http://lod.ac/ns/species#>

SELECT *
WHERE {
 SERVICE <http://lod.ac/species/sparql> {
  ?species speciesOnto:hasSuperTaxon
    <http://lod.ac/species/Trypoxylus> .
  FILTER regex(str(?species), '^http://lod.ac/species')
 }

 SERVICE <http://lod.ac/sparql> {
  ?specimen a lodac:Specimen;  dc:references ?specimenRef .
  ?specimenRef speciesOnto:species ?species ;
   crm:P55_has_current_location/dc:references ?locationRef .

  ?locationRef lodac:address ?address ;
   rdfs:label ?locationLabel ;
   geo:lat ?lat ;
   geo:long ?long .
 }
}
LIMIT 10
```

図 3　SPARQL クエリ例

今後の課題としては，データの拡充はもちろん，本研究で得られた知見を活かして，SPARQL を扱えないユーザでも利用できるインタフェースを備えたアプリケーションの開発や LOD 化を半自動化するツールの開発などにも着手したいと考える．

謝辞

本研究は，LODAC プロジェクトにおいて議論を重ねて遂行した．プロジェクトチームのメンバー全員に感謝の意を表します．また，国立遺伝学研究所の菅原秀明先生，国立科学博物館の松浦啓一先生には，本研究を支援していただいた．そして，日本産蝶類和名学名便覧の編纂メンバーである猪又敏男氏，植村好延氏，矢後勝也氏，上田恭一郎氏には，データ利用の快諾をいただいた．東京大学の倉島治氏には，本稿に有益なコメントをいただいた．みなさまに感謝の意を表します．なお，本プロジェクトに関する GBIF 日本ナショナルノードの活動は，JST および文部科学省のナショナルバイオリソースプロジェクト（NBRP）の支援を受けている．

参考文献

[1] T. Berners-Lee, J. Hendler, James and O. Lassila: The Semantic Web, Scientific American, May 2001, p. 29-37.

[2] Bizer, C., Heath, T. and Berners-Lee, T.: Linked Data –The Story So Far, International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), pp.1—22, 2009..

[3] 武田英明, 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: 日本における Linked Data の普及にむけて, 2011 年度人工知能学会全国大会, 人工知能学会, 2011.6.

[4] UNEP CBD, Convention on Biological Diversity, 1992.

[5] 環境省，生物多様性国家戦略 2010，2010.

[6] 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: Linked Open Data による多様なミュージアム情報の統合, 人文科学とコンピュータシンポジウム じんもんこん 2010, 情報処理学会, 2010.12.

[7] 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: LOD.AC: Linked Open Data によるミュージアム情報の結合, 第 3 回知識共有コミュニティワークショップ, 情報社会学会, 2010.12.

[8] 深見嘉明, 小林巌生, 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: Linked Open Data とコミュニティが拓くオープンガバメント, 第 3 回知識共有コミュニティワークショップ, 情報社会学会, 2010.12.

[9] Kamura T., Takeda H., Ohmukai I., Kato F., Takahashi T., Ueda H.: Building Linked Data For Cultural Information Resources In Japan, Demonstration at Museum and the Web 2011, 2011.4.

[10] 深見嘉明, 小林巌生, 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: Linked Open Data によるボトムアップ型オープンガバメントの試み, 情報処理学会研究報告, DD, 2011-DD-79(1), 1-8, 2011.1.

[11] 猪又敏男, 植村好延, 矢後勝也, 神保宇嗣, 上田恭一郎: 日本産蝶類和名学名便覧, http://binran.lepimages.jp, 2010.

[12] 南佳孝, 加藤文彦, 大向一輝, 武田英明, 新井紀子, 神保宇嗣, 伊藤元己: 生物情報基盤構築に向けた生物関連データの Linked Data 化の取り組み, 第 26 回セマンティックウェブとオントロジー研究会，人工知能学会，SIG-SWO-A1103-02, 2011.12

[13] 武田英明, 南佳孝, 加藤文彦, 大向一輝, 新井紀子, 神保宇嗣, 伊藤元己, 小林悟志, 川本祥子: 生物情報基盤構築のための生物種データの Linked Open Data 化の試み，The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 人工知能学会, 3C2-OS-13b-3, 2012.

[14] Bisby, F.A.: The quiet revolution: biodiversity informatics and the Internet, Science, Vol. 289, No. 5488, pp. 2309-2312, (2000)