3C2-OS-13b-3

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

Towards LOD of Species to Build the Biology Information Infrastructure

武田 英明[*1*2] Hideaki Takeda
南 佳孝[*1] Yoshitaka Minami
加藤 文彦[*1] Fumihiro Kato
大向 一輝[*1*2] Ikki Ohmukai
新井 紀子[*1] Noriko Arai
神保 宇嗣[*3] Utsugi Jimbo
伊藤 元己[*4] Motomi Ito
小林 悟志[*5] Satoshi Kobayashi
川本 祥子[*6] Shoko Kawamoto

[*1] 国立情報学研究所 National Institute of Informatics
[*2] 総合研究大学院大学 Graduate University for Advanced Studies
[*3] 国立科学博物館 National Museum of Nature and Science, Tokyo
[*4] 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 Department of General Systems Studies, Graduate School of Arts and Sciences, The University of Tokyo
[*5] 国立極地研究所 National Institute of Polar Research
[*6] 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 Database Center for Life Science, Research Organization of Information and Systems

Linked Data is becoming a popular method to publish data on the Web. As a part of LODAC (Linked Open Data for ACademia) project which aims to construct a framework to share and integrate academic data, we are developing Linked Data for biodiversity information. As a pilot study, the latest species name list of Japanese butterflies was converted into Linked Data format, integrated with relevant biodiversity data such as museum specimen collections and extended for collection data of bryophyta deposited at National Institute of Polar Research. This integrated dataset is available via our website and a SPARQL endpoint. This system will facilitate use of heterogeneous biodiversity and its relevant resources.

1. はじめに

Web が普及し，多くの情報がインターネットを通じて入手可能になった．しかし，その情報は人間が読むことを前提に作られており，コンピュータを通じて利用しているも関わらず，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 の発明者である Tim Berners-Lee は，Web の前提として，人間だけでなく，コンピュータもその内容を処理可能であることが必要だと考えており，その仕組みとしてセマンティック Web を提唱している[Berners-Lee01]．しかし，人間が読む情報の共有という点で Web は大変強力であり，Web が急速に普及したことに比べて，セマンティック Web は必ずしも発展・普及したとはいえない．

ところが，Web が社会にあまねく普及し，膨大な情報が Web にのるようになって，再びセマンティック Web の考え方，すなわちコンピュータが処理可能な形式による情報の公開の重要性が認識されるようになった．特に大量のデータにおいてはこの必要性が強く認識されるようになった．そこで，概念的な定義ではなくて個別の情報をコンピュータが処理できる仕組みとして Linked Data という方法が提唱された．Linked Data はセマンティック Web の分野で開発されてきた言語(RDF, RDFS, OWL)を利用するが，主に個別の情報，データを記述する手段としてそれを用いる．

Linked Data におけるデータは RDF を用いて記述される．RDF はシンプルで柔軟性があり，多様なデータの記述が可能である．このような理由から，近年，Linked Data が情報流通の仕組みとして普及しつつある．ヨーロッパや米国では，すでに新しい情報公開・共有の仕組みとして認知されつつあり，我が国でも様々な研究や活動が行われている[武田 11]．

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[UNEP92][環境省 10]．これらの問題を解決するには，対象生物のみではなく，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，Linked Data 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

連絡先：武田英明，国立情報学研究所，千代田区一ツ橋 2-1-2，takeda@nii.ac.jp

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）」プロジェクト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 Linked Data で構築することを目標に，2010 年 4 月に開始し，2010 年 12 月には博物館情報を対象とした Web サイトを公開した[深見 10][嘉村 10a][嘉村 10b]．その後，関連する美術館等のデータを拡充し，現在も活動を継続している[深見 11][Kamura11]．

本研究では，LODAC プロジェクトでこれまで構築してきた情報基盤について，分野を超えて拡張し，生物多様性情報を Linked Data 化して提供することにより，将来，生物情報基盤として活用することを目指す．

3. 生物多様性情報基盤整備の現状

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．

本研究は，中でもその中核をなす種の多様性に着目する．このレイヤーでは，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）とよばれる[Bisby00]．

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，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）が挙げられる．

生物多様性情報は，様々なデータベースを通じて，データの種別や目的に特化した形式で公開されている．そのため，ある課題の解決のためには，複数の異なる Web サイトにわたる検索やデータの統合が必要であり，その統合には多かれ少なかれ摺り合わせが必要である．実際，同一種の情報であっても，異なる Web サイトに掲載されている場合，同一の検索キーでは，それぞれから適切な検索結果を得られないことがある．このため，相互運用性の確保は生物多様性情報学分野で最重要な課題の 1 つにされている[Edwards00]．Linked Data は，そのような多様な情報の相互運用性の担保に有効だが，生物多様性情報学分野で実際に用いられているリソースは，国内では存在しない．海外では，Linked Data を提供しているプロジェクトとして GeoSpecies があるものの，まだ発展途上の研究プロジェクトであり，実用化には至っていない．

4. Linked Data 化のプロセス

生物多様性情報の Linked Data 化では，1）基本となるデータの選定，2）データ公開のための構成の決定，3）関連データとのリンクと公開，4）対象データの拡張の順で作業を行った．

4.1 データの選定

本研究で対象とする生物多様性のデータは，様々な組織から複数の Web サイトで公開されており，分類群，データの種類が多岐にわたっている．

本研究では，まず，対象分類群として蝶類を選定した．その理由として，パイロット研究に適当な種数であること，一般によく知られており科学的データをはじめとした様々な情報がリッチであり，Linked Data の利用が様々な場で期待できること，基本情報がデータベースの形で公開されていること，分類学者とも連携可能なことがあげられる．

様々な多様性情報をリンクする最も重要な要素は生物の名前，すなわち種名である．そこで，本研究ではまず種名情報の整備を目標とし，ソースとして日本産蝶類和名学名便覧を選択した[猪俣 11]．日本産蝶類全種にあたる 327 種・亜種について，所属分類群・学名・和名などが記されたもので，専門家が最新の知見に基づいて編纂しているため，基本となるデータとして適当であると判断した．

4.2 データの構成

蝶類の種名データは，分類体系を構成する要素，すなわち分類群に関する項目（界名・門名・綱名・目名・科名・亜科名・族名・亜族名・属名・亜属名・種小名）と種を表現する要素，すなわち種名に関する項目（学名・著者・出版年・和名・和名の別名）に大別できる．そこで，データを公開するために，分類体系を表現することと種名に関する情報を表現することを考えた．

まず，分類体系を表現するために，各分類群名に対して URI を定義し，さらに分類体系の階層性を表現するため，木構造の根にあたる界名以外の分類群名に項目については，上位階層の分類群を指し示す URI にリンクした．分類に関するデータについて，図 1 に Lepidoptera 鱗翅目（チョウ目）の例を示す．

次に，種に関する情報を表現するために，学名に URI を定義し，和名，著者，出版年に加え，所属する分類群を指し示す URI にリンクした．種に関するデータについて，図 2 に Papilio xuthus アゲハ（ナミアゲハ）の例を示す．

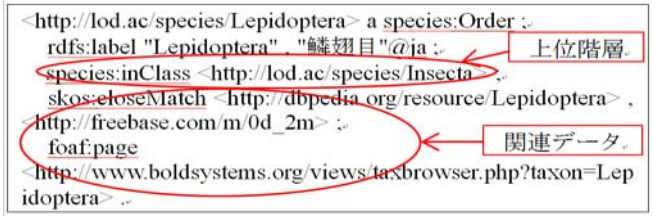

図 1 分類に関するデータ（Lepidoptera の例）

<http://lod.ac/species/Papilio_xuthus> rdfs:label "Papilio xuthus"@en , "アゲハ"@ja ;
species:inKingdom <http://lod.ac/species/Animalia> ;
species:inPhylum <http://lod.ac/species/Arthropoda> ;
species:inClass <http://lod.ac/species/Insecta> ;
species:inOrder <http://lod.ac/species/Lepidoptera> ;
species:inFamily <http://lod.ac/species/Papilionidae> ;
species:inSubFamily <http://lod.ac/species/Papilioninae> ;
species:inTribe <http://lod.ac/species/Papilionini> ;
species:inGenus <http://lod.ac/species/Papilio> ;
species:inSpecificEpithet <http://lod.ac/species/xuthus> ;
species:author "Linnaeus" ;
species:namedYear "1767" ;
skos:closeMatch <http://freebase.com/m/03c7d4n> ,
<http://dbpedia.org/resource/Papilio_xuthus> ;
foaf:page
<http://www.boldsystems.org/views/taxbrowser.php?taxon=Papilio+xuthus> ;
species:commonName "ナミアゲハ"@ja , "アゲハチョウ"@ja ;
species:scientificName "Papilio xuthus" .

和名。
分類群。
著者，出版年。
関連データ。

図 2 種に関するデータ（Papilio xuthus の例）

4.3 関連するデータとのリンク

蝶類に関する多様性情報リソースは数多く，前述した GBIF，EoL，CoL，BOLD，S-Net，GeoSpecies などに保存されているほか，DBpedia や Freebase など既存の Linked Data リソースにも蝶に関する情報が存在する．そこで，これらの Web サイトに対して，分類名と学名をキーに検索し，該当した Web ページへのリンクを作成した．図 1，図 2 で付与されているのは，GoogleRefine を用いて変換した RDF（Resource Description Framework）であり，BOLD，DBpedia，Freebase のデータに対しては GoogleRefine 上でリンクを生成できた．その他の Web サイトに対しては，Web サイト毎にスクリプトを記述してスクレイピングを行い，必要な情報を取得した．このうち，S-Net のデータは，各博物館の所蔵標本情報である．そのため，S-Net のデータには，標本を所蔵する博物館の情報が含まれていたため，その情報を活用して LODAC の博物館情報にリンクした．また，標本は 1 種に対して複数存在するので，各標本データに URI を定義し，種の情報へリンクした．Linked Data 化した S-Net の標本情報について，図 3 に Papilio xuthus の例を示す．

specimen:k1118470 speciesOnto:species <http://lod.ac/species/Papilio_xuthus> .
specimen:k1118470 foaf:page <http://www.science-net.kahaku.go.jp/specimen/collection/collection_details.do?division=collect&Search_Mode=1&Conf_Name=integration&Said_Number=10&View=0&Data_Id=1118470&Class_Name=OMPIM> .
specimen:k1118470 speciesOnto:collectionGround "中国 浙江省 四明山 " .
specimen:k1118470 speciesOnto:collectionDate "1979 年 06 月**日" .
specimen:k1118470 crm:P55_has_current_location <http://lod.ac/id/458869> .
specimen:k1118470 speciesOnto:museumName "橿原市昆虫館" .

種情報へリンク。
博物館情報へリンク。

図 3 S-Net のデータ（Papilio xuthus の例）

4.4 データの拡張

データを拡張する対象として，国立極地研究所に収蔵されている蘚苔類の標本データについて，同様に Linked Data 化することを考えた．標本データには，蝶類データの項目以外に，亜綱名，亜種名，ID，採集日，採集者，緯度・経度，地域番号，地域名，収蔵場所，分布していた状況，標高，採集者，（掲載標本集：CBM，HIRO，HIRU，HYO，KOCH，NICH，NIPR，TNS）といった情報が記されている．そこで，前節までと同様のプロセスで分類に関するデータ，種に関するデータ，標本に関するデータについて，Linked Data 化を行った．

5. Linked Data 化の結果

本研究で，生物多様性の情報について Linked Data 化を行った結果，図 4 に示すように，関連する情報を一覧で提供できるようになった．また，リンクした他の生物多様性情報の Web ページをインラインフレーム（iframe）によって表示することで，関連する情報の俯瞰が可能になった．さらに，種情報の Web ページに DBpedia の画像を表示することで，掲載情報が利用者の探している蝶の種類かどうかを特定できるようになった．

また，本研究では，Linked Data 自体の外部インタフェースとして SPARQL Endpoint を公開した．これにより，複合的，横断的な検索が可能になった．

図 4 結果表示例

6. BDLS の Linked Data 化

生物種に関わる異なる情報源として BDLS の Linked Data 化も行った．BDLS (Building Dictionary for Life Science)とは，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て一つの辞書として構築したものである[1]．主に生物種を中心とするタクソン情報（学名，和名）と用語（日本語と英語）が含まれている．

6.1 Linked Data 化の方針

BDLS は個別の情報に必ず出典が付されている．このため，この出典ごとに Named Graph として表現した．語彙については，タクソン間の関係などは geospecies[2]で定義されている語彙を参考にして構築した（図 5 参照）．

[1] http://lifesciencedb.jp/bdls/
[2] http://lod.geospecies.org/

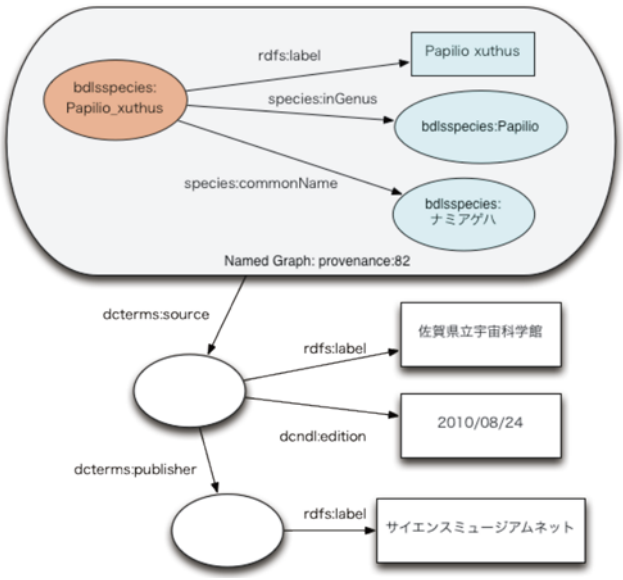

図 5 BDLS のデータモデル

6.2 結果と応用

上記の方法で RDF Store に登録を行った. 現在, 6,366,545 トリプルがある. HTML によるインタフェースと SPARQL Endpoint を用意している.

このデータを利用した応用例として, 文献検索システムとの連携を作った. ここでは, 種名で検索するとその和名や同属の別種も含めた検索結果を得られる. (図 6 参照).

図 6 BDLS データを検索に利用した例

7. おわりに

本研究では, 生物情報基盤構築に向けて, 生物関連データの Linked Data 化を行った. これにより, 分類体系・種名・種の特徴・標本に関する情報を柔軟に組み合わせて閲覧できるようになった.

本システムが生物多様性分野に与えるインパクトとして, 以下の 2 点があげられる. 1 点目は, 既存のシステムで困難だった複雑な検索が可能となったこと, 2 点目は, 分野横断型の情報システム作成を可能にする基盤が構築できたことである. これらは, 生物多様性情報学の課題である相互運用性の向上ともつながっている.

今後は, 更なるデータの拡充を考え, データを容易に追加できる仕組みや, 目的に応じて利用しやすいインタフェースを備えたアプリケーションの開発を目指す. 情報とシステムを整備することで, 本研究の成果が, 社会問題として生物多様性保全問題を扱う際に不可欠な情報基盤になると期待される.

謝辞

本研究は, LODAC プロジェクトでの議論を経て遂行した. プロジェクトチームの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

[Berners-Lee01] T. Berners-Lee, J. Hendler, James and O. Lassila: The Semantic Web, Scientific American, May 2001, p. 29-37.

[Bisby00] Bisby, F.A.: The quiet revolution: biodiversity informatics and the Internet, Science, Vol. 289, No. 5488, pp. 2309-2312, (2000)

[Edwards00] Edwards, J.L., Meredith A.L., and Nielsen E.S.: Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science, Vol. 289, No. 5488, pp. 2312-2314, (2000)

[Kamura11] 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

[UNEP92] UNEP CBD, Convention on Biological Diversity, (1992)

[猪俣 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)

[嘉村 10a]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.

[嘉村 10b]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.

[環境省 10] 環境省, 生物多様性国家戦略 2010, (2010)

[武田 11] 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.

[深見 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 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.

[深見 11] 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.