1G4-OS-19a-3

Linked Open Data による絶滅危惧種情報共有の試み

Towards Knowledge Sharing of Endangered Species in a Linked Open Data Architecture

亀田 尭宙 *1 KAMEDA Akihiro　加藤 文彦 *1 KATO Fumihiro　神保 宇嗣 *2 JINBO Utsugi　大向 一輝 *3 OHMUKAI Ikki　武田 英明 *3 TAKEDA Hideaki

*1情報・システム研究機構 Research Organization of Information and Systems

*2国立科学博物館 National Museum of Nature and Science

*3国立情報学研究所 National Institute of Informatics

In this paper, we introduce efforts toward integrating endangered species information to our "LODAC species". LODAC (Linked Open Data for ACademia) project is for sharing academic knowledge in the style of LOD and consists of multiple domain-specialized sub-projects, and LODAC species is specialized to biodiversity domain. LODAC species is designed based on Names-based architecture and enables us to connect heterogeneous data like information about specimens in museums and Red list information easily. We illustrate the result of integrating red list of Japanese government and Kyoto prefecture and discussion about future work.

1. はじめに

ウェブ上でのデータ共有の基盤として，ここ数年間で Linked Open Data (LOD) は急激に発展してきた．Linked Data の原則は

1. あらゆる事物を URI で記述すること
2. HTTP URI を用いることでアクセス可能にすること
3. 記述に RDF を用いることで有用な情報を提供すること
4. さらなる情報を発見するための外部リンクを提供すること

の 4 つである．LOD においては，それに加えてオープンライセンスによる情報共有という実践が結合している [Heath 09][Heath 13].

一方，生物多様性保護は大きな社会的関心事であり [UNEP 92]，ウェブをデータ共有基盤として用いた生物多様性情報共有の試みも広まっている [EOL][GBIF].

しかし，依然としてドメインを超えたデータの共有が難しいという問題があり，そこを前述の LOD によるアプローチによって解決していきたいと考えている．

1.1 LODAC Species

我々はこれまで学術情報に関する LOD 基盤の構築を LODAC（Linked Open Data for ACademia）プロジェクトとして行ってきた．対象は博物館情報に始まり [嘉村 10]，生物多様性情報に拡張され [武田 12]，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の閲覧が可能になった．さらにデータの追加とデータモデルの整理を経て [南 13]，今回，絶滅危惧種情報とのデータ統合を試みた．

2. 絶滅危惧種情報のデータ化と統合

絶滅危惧種に関するデータとしてはレッドリストとレッドデータブックが有名である．それぞれ，国際機関，国，都道府県など異なるレベルで作られ，大抵の場合，名前と保全状況についての最低限の情報を速やかにレッドリストとして編纂し，後に詳細な情報を加えてレッドデータブックとして発行するという手順を踏んでいる．我々はまず，国レベル，都道府県レベルでのレッドリストのデータ化を試みた．具体的には環境省の生物多様性センターが 2012 年から 2013 年にかけて編纂した第 4 次レッドリスト *1（以下，環境省レッドリストと呼ぶ）と，2013 年に編纂された京都府レッドリスト *2（以下，京都府レッドリストと呼ぶ）をデータ化した．

表 1: レッドリスト 1 項目のデータ化例

```
<http://lod.ac/species/Oceanodroma_castro> a speciesOnto:ScientificName;
speciesOnto:hasCommonName <http://lod.ac/species/クロコシジロウミツバメ>;
    speciesOnto:hasSuperTaxon <http://lod.ac/species/鳥類>;
rdfs:label "Oceanodroma castro";
cnsvOnto:hasRedListEntry redlist:jibis-redList2012_tyorui-17.

redlist:jibis-redList2012_tyorui-17 a cnsvOnto:RedListEntry;
    rdfs:comment "クロコシジロウミツバメ"@ja;
    cnsvOnto:ofSpecies <http://lod.ac/species/Oceanodroma_castro>;
    cnsvOnto:currentStatus cnsv:CR;
    cnsvOnto:ofArea "日本"@ja.
```

環境省は学名，和名，保全状況，その種のカテゴリーといった情報を公開している．そこで学名「Oceanodroma castro」を持つ種が，和名として「クロコシジロウミツバメ」を持ち，「絶滅危惧 IA 類（CR）」の保全状況で,「鳥類」にカテゴライズされていることは Turtle 形式の RDF で表 1 のように記述できる．前半は学名を主語としたトリプルであり，species オントロジの語彙 *3 を用いて和名とカテゴリを記述している．カテゴリに関してはデータ源の記載を尊重してそのまま記載している．例えば「汽水・淡水魚類」といったカテゴリが環境省レッドリストに存在するが，そもそも魚類というカテゴリ自体が，系統分類学的ではない用語であり，体系の異なる用語をマッピングすることは困難であるし，また LOD 化の段階でするべき

連絡先: 亀田 尭宙, 情報・システム研究機構, 東京都千代田区一ツ橋 2-1-2, kameda@nii.ac.jp

*1 生物多様性情報システム 絶滅危惧種検索 http://www.biodic.go.jp/rdb/rdb_f.html

*2 京都府改訂版レッドリスト 2013 京都府ホームページ http://www.pref.kyoto.jp/kankyo_red/shiryuou5.html

*3 http://lod.ac/ns/species

ではないと考えている．それを speciesOnto:hasSuperTaxon というプロパティで単に上位の分類群として登録することで，異なる分類体系の情報を共存させることを可能にしている．実際，この種については LODAC Species 内の既存の情報と統合され，NCBI *4 や DBpedia *5 といった他のデータベースへのリンクや，ミズナギドリ目ウミツバメ科に属するといった他の上位分類群の情報，[EOL] 等から取ってきた関連する写真が閲覧できるようになっている（図 1）．また学名を主語としたトリプルの一つとして，保全情報オントロジ語彙 *6 の cnsvOnto:hasRedListEntry プロパティを用いてレッドリストのエントリを参照しており，その項目の情報を後半で記述している．このように情報を分けているのは，複数のレッドリスト情報が一つの種について存在し得るからである．そして，レッドリストのエントリを主語としたトリプルで具体的な保全状況や指定されている地域などの情報を記述している．

京都府レッドリストについても同様であるが，そちらは学名が記載されていなかったため，和名を主語として情報を記述した．

統合に際しては，LODAC Species が名前ベースのアーキテクチャを採用し，各生物種に対応する URI が `http://lod.ac/species/`種名 のような形式になっているので，種名が一致すれば自動的に統合されるようになっている．

3. 結果と考察

3.1 統合に成功した比率

環境省レッドリスト 5690 件 には学名と和名が存在するが，それぞれを用いて LODAC Species との統合を試みたところ，学名を介して 3294 件 (57.9%) が既存のデータと統合された．一方，和名を介しては 4145 件 (72.8%) の統合に成功した．その和集合は 4711 件 (82.8%) となっている．京都府レッドリスト 1871 件 については和名のみを用いたが，1598 件 (85.4%) という比較的高い割合で統合に成功した．全体的に高い割合で統合に成功したのは，生物の種名が比較的ゆれがなく使われていること，LODAC Species のデータが広い範囲をカバーしていることを示している．和名の方が効率的に統合できた一因は，和名の方が表記ゆれが少ないことではないかと考えている．

3.2 失敗例の分析

元のデータと統合できなかったデータについて，統合すべきデータがそもそも入っていなかったのか，データがあるにもかかわらず何らかの理由で統合に失敗しているのか，を正確に知る術は無い．しかし，統合に失敗した名前について調査することで，より統合率を向上させるのに役立つ知見が得られたので，以下に報告する．

3.2.1 統合すべきデータが入っていなかったと思われるもの

例えば，京都府版レッドリストにある「ヨドゼゼラ」に関しては，2010 年に新種「Biwia yodoensis」の発見の論文が出ており *7，この論文の著者である細谷が琵琶湖生物多様性画像データベースに和名として「ヨドゼゼラ」を記載している *8 ことから，比較的新しい種であるために，元のデータの中に対応する種が含まれていなかったと考えられる．また，同じく京都版レッドリストにある「ルイスムネボソヨツメハネカクシ」については，「Boreaphilus lewisianus」という学名が付けられている種の発見自体は 1874 年と古いが，和名がつけられたのが柴田らによって 2013 年に編纂された『日本産ハネカクシ科総目録』においてであるため *9，元のデータの中に対応する種が含まれていなかったと考えられる．

これらの例については，データベースや図鑑，目録といった形で発行される新しい情報を積極的に入力する他，そのフローを効率化するために各分野の分類学者と協力していく必要がある．

3.2.2 複数の和名を持つ種

京都府版レッドリストにある「イモリ」や「モモンガ」は族や科の名前である一方，種として「モモンガ」といった場合には「ニホンモモンガ」（「ホンドモモンガ」ともいう）のことを指し，種として「イモリ」といった場合には「アカハライモリ」を指す．例えば，Wikipedia の生物分類表テンプレートには「和名」という項目があり，それぞれの種について「アカハライモリ, ニホンイモリ, イモリ」「モモンガ、ニホンモモンガ、ホンドモモンガ」という形で名前が列挙されている *10．また，イモリについては京都府がウェブ上に公開している 2002 年版レッドデータブック *11 において学名「Cynopus pyrrhogaster (Boie, 1826)」が付されているため，「Cynops pyrrhogaster」という名前を持つアカハライモリであると推測され *12，それは環境省のレッドリストにおいても「準絶滅危惧 (NT)」指定されていることがわかる．

これらの例については，DBpedia や他のレッドリストデータから同じ種を指す複数の和名を抽出し，それらを関連付けることで解決できると考えられる．

3.2.3 ミススペルと表記ゆれ

前述の例の前半「Cynopus pyrrhogaster」と同様の学名を用いた論文，文書はウェブ上に他にも存在したが，タンパク質に関するデータベースにおいて，Cynopus は Cynops のミススペルであるとされている *13．また，発見者名や年号を後ろにつけるのは学名において多くみられる表記法であり，1700 万以上の学名についての情報を提供している Global Names Index *14 では，「Cynops pyrrhogaster」「Cynops pyrrhogaster (Boie, 1826) 」「Cynops pyrrhogaster Boie」の 3 つを表記グループ（Lexical groups）としてまとめている．一方，和名についてはこういった形の表記ゆれはないことが，前述のように効率的に統合できた一因だと考えられる．

また環境省レッドリストにある「Aerobryum speciosum

*4 National Center for Biotechnology Information `http://www.ncbi.nlm.nih.gov/EOL`

*5 `http://dbpedia.org/`

*6 `http://lod.ac/ns/cnsv`

*7 Kawase, S. and Hosoya, K: Biwia yodoensis, a new species from the Lake Biwa / Yodo River Basin, Japan (Teleostei: Cyprinidae). Ichthyological Exploration of Freshwaters, 21: 1-7. (2010)

*8 ヨドゼゼラ `http://www.lberi.jp/root/jp/62pick/tayosei_db/data/Biwia-yodoensis/index.htm`

*9 柴田 泰利, 丸山 宗利, 保科 英人, 岸本 年郎, 直海 俊一郎, 野村 周平, Volker Puthz, 島田 孝, 渡辺 泰明, 山本 周平: 日本産ハネカクシ科総目録（昆虫綱：甲虫目）Bulletin of the Kyushu University Museum No. 11, pp. 69-218 (2013) `http://www.museum.kyushu-u.ac.jp/bulletin/011/11-69.pdf`

*10 モモンガ - Wikipedia `https://ja.wikipedia.org/wiki/`モモンガ, アカハライモリ - Wikipedia `https://ja.wikipedia.org/wiki/`アカハライモリ

*11 京都府レッドデータブック RED データベース　両生類 `http://www.pref.kyoto.jp/kankyo/rdb/bio/amphibian.html`

*12 詳しくは次節を参照

*13 MyHits Entry taxid:8329 `http://myhits.isb-sib.ch/cgi-bin/view_cla_entry?name=taxid:8329`

*14 `http://gni.globalnames.org/`

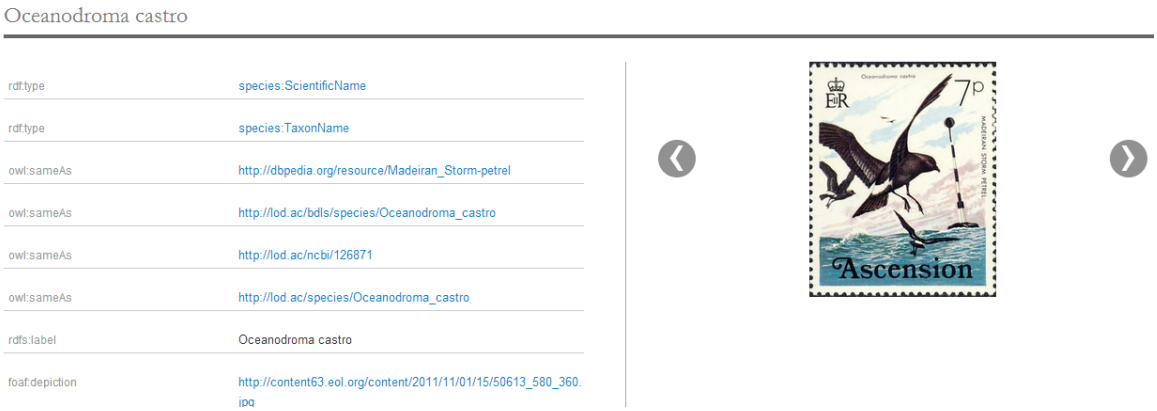

図 1: 生物種一項目の HTML 表示例

(Dozy & Molk.) Dozy & Molk. var. nipponicum Nog.」については，非常に近い表記の「Aerobryum speciosum (Dozy et Molk.) Dozy et Molk. var. nipponicum Nog.」が米倉らの作成した YList *15 に記載されており，それが LODAC Species にも含まれているが，記号表記「&」とラテン語表記「et」の差異やスペースの数の差異といった細かな差異によって統合に失敗している．

これらの例については前節と同様，ミススペルや表記ゆれについて扱っているデータベースの情報を追加する他に，すでにあるエントリとの表記上での類似度を計算し，統合先として推薦するという手法が考えられる．我々はすでにオープンソースの検索エンジンである Apache Solr*16 を用いて類似の文字列を検索するシステムを実装しており *17，今後異なる情報源からのデータの統合の際にこのエンジンを活用したいと考えている．

3.2.4 同名異種

統合できたものについては統合成功としているが，その統合が適切だったかどうかについては十分に検討できていない．異なる種に同じ学名が付けられることはほぼ無いと思われるが *18，和名については昆虫類のカマキリと淡水魚類のアユカケの別名であるカマキリが同名になってしまっているといった例が実際に存在し *19，それらを区別する仕組みが必要と考えられる．

4. おわりに

本研究では，生物情報を共有する LOD 基盤として構築を進めてきた LODAC Species へ，生物多様性に関する重要なデータである絶滅危惧種情報の統合を行った．学名や和名を手がかりとして多くの絶滅危惧種情報を既存の情報と統合できた一方，失敗例の分析を通して，より効率的な統合のためにデータの追加や統合手法の改善の必要性が示された．

*15 米倉浩司, 梶田忠:BG Plants 和名－学名インデックス (YList), (2003) http://bean.bio.chiba-u.jp/bgplants/ylist_main.html

*16 https://lucene.apache.org/solr/

*17 http://lod.ac/apps/solr-search/

*18 同じ界においては異なる種に同じ学名がつくことは命名規約上許されていないが，植物のアセビ属の属名は，動物のモンシロチョウ属と同じ「Pieris」であるなど，界をまたぐと制約は無いため，学名における同名異種も原理的にはありうる．

*19 http://lod.ac/species/カマキリ

参考文献

[Heath 09] Bizer, Christian; Heath, Tom; Berners-Lee, Tim: Linked Data—The Story So Far, International Journal on Semantic Web and Information Systems 5 (3), pp. 122, Solving Semantic Interoperability Conflicts in CrossBorder EGovernment Services (2009).

[Heath 13] Tom Heath, Christian Bizer: Linked Data: Evolving the Web into a Global Data Space, (邦訳 : Linked Data: Web をグローバルなデータ空間にする仕組み, (2011) 武田 英明 監訳, 大向 一輝, 加藤 文彦, 嘉村 哲郎, 亀田 尭宙, 小出 誠二, 深見 嘉明, 松村 冬子, 南 佳孝 訳 (2013)).

[UNEP 92] UNEP CBD, Convention on Biological Diversity, (1992).

[EOL] Encyclopedia of Life. http://www.eol.org

[GBIF] Global Biodiversity Information Facility. http://www.gbif.org/

[嘉村 10] 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).

[武田 12] 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, 人工知能学会全国大会 (第 26 回) 論文集, No. 3C2-OS-13b-3, 山口 (2012).

[南 13] Y. Minami, H. Takeda, F. Kato, I. Ohmukai, N. Arai, U. Jinbo, M. Ito, S. Kobayashi and S. Kawamoto: Towards a Data Hub for Biodiversity with LOD, in H. Takeda, Y. Qu, R. Mizoguchi and Y. Kitamura eds., Semantic Technology - Second Joint International Conference, JIST 2012, Nara, Japan, December 2-4, 2012. Proceedings, Vol. 7774 of LNCS, pp. 356361, Springer (2013).