統計数理(2013)
第 61 巻 第 2 号 217–231

特集「環境リスクと統計解析—データ基盤構築と解析—」
[総合報告]

ビッグデータ時代の環境科学
—生物多様性分野におけるデータベース統合, 横断利用の現状と課題—

大澤 剛士[1,3] ・神保 宇嗣[2,3]

(受付 2012 年 12 月 28 日; 改訂 2013 年 3 月 22 日; 採択 3 月 27 日)

要　　旨

IT の発展やデータベース公開等, 情報公開の機運が高まったことに伴い, 環境科学分野における巨大データを利用した研究は急増している. 巨大データを利用した研究を行う際には, 膨大なデータセットと, その巨大データをハンドリングする技術, そして適切な統計解析技術の全てが必要とされる. しかし, このうちデータ管理技術について, 情報が著しく不足しており, 国内においてはデータの統合, 横断利用の促進といった取り組みが遅れている. 本稿はデータ管理技術, すなわち散在するデータをどのように整理し, 利用可能な形に整備するかについて, 特に生物多様性情報関係のデータに注目し, データ記述フォーマット, 通信プロトコル, メタデータフォーマットの標準化といった技術的な概説ならびに国内における現状の紹介を行い, 国内における巨大データを利用した環境科学研究の推進に向けた課題について論じる.

キーワード: 横断利用, 生物多様性情報学, メタデータ, 標準化, データフォーマット, Darwin Core.

1. はじめに

近年の IT 技術, 特にストレージ容量の増加と観測技術の発展に伴い, かつては想像することすらできなかった巨大データ, いわゆる「ビッグデータ」の利用に注目が集まるようになってきた.「ビッグデータ」とは, 典型的なデータベースソフトウェアが把握し, 蓄積し, 運用し, 分析できる能力を超えたサイズのデータと定義される(総務省, 2012; 他にも様々な定義があるが, 本稿ではこの定義に従う). これら「ビッグデータ」は既に様々な分野において活用段階にあり, 例えばビジネスの世界では経済予測, マーケティング等, 既にそれを活用して利益を得るような取り組みが多数出てきている(Cukier, 2010).

環境科学の分野においても, 巨大データを活用した研究を行うことは, 一つの潮流となりつつある. 例えば生態学分野では, 全国〜全世界という空間スケールを研究対象としたマクロエコロジー(天野 他, 2010), 気候変動分野では, 地球温暖化の将来シナリオを利用し, 国レベルにおける食料生産(Iizumi et al., 2011)や生物分布の変化を予測した研究(Ficetola et al., 2007),

[1] 独立行政法人農業環境技術研究所 農業環境インベントリーセンター: 〒305–8604 茨城県つくば市観音台 3–1–3; arosawa@gmail.com
[2] 独立行政法人国立科学博物館 動物研究部: 〒305–0005 茨城県つくば市天久保 4–1–1
[3] 地球規模生物多様性情報機構 (GBIF) 日本ノード (JBIF)

分類学の分野では，巨大ライブラリに登録された塩基配列を参照し，対象サンプルの種名を塩基配列から同定する DNA バーコーディング（伊藤 他, 2007; Jinbo et al., 2011; 神保, 2012a, 2012b; Kato et al., 2012）等，圧倒的に巨大なデータを活用した研究が続々公表されている．このような研究を行う際には，巨大データと，巨大データをハンドリングする技術，そして適切な解析技術の全てが必要とされる．巨大データ活用の潮流は，データ収集，データ管理，解析技術の全てが発展したことにより生まれたものであると言えよう．

データ収集技術，データ管理技術，解析技術のうち，データ収集技術については，環境科学の分野においても日本語の解説等が出版され，かなり身近になってきている（例えば次世代シーケンサー：林 他, 2009; バイオロギング：日本バイオロギング研究会, 2009; リモートセンシング：(独) 農業環境技術研究所, 2011）．また，解析技術についても基礎から学ぶことができる文献等が出版され，多くの研究者にとって身近になりつつある（例えば統計モデリング：久保, 2012）．それに比してデータ管理技術については，こと環境科学の分野に限ると，専門家が極めて少なく，技術に関する情報が十分とはいえない．その影響もあり，既に多くの環境情報データベースが構築されているにも関わらず，分野を超えたデータの利活用はあまり普及していない（絹谷 他, 2008）．特に気候，水循環，農業，生態系などの分野では，各々が類似したデータを保持しているにもかかわらず，横断利用が特に普及していないと指摘されている（絹谷他, 2008; 大澤 他, 2012）．横断利用に向けた方法論も近年になって少しずつ提案されはじめているが（三橋, 2010; 大澤 他, 2011, 2012），少なくとも国内において，環境科学関連のデータの統合，横断利用の促進のような『データ管理技術』についての検討は始まったばかりである（大澤 他, 2012）．ますます巨大化するデータを効率的に活用していくための仕組みを確立し，活用していくことは，国内における今後の環境科学の発展において必須と言っても過言ではない．

本稿は，散在するデータをどのように整理し，利用可能な形に整備するか，すなわち「データのデータベース化」について，環境科学，特に生物多様性情報関係のデータに注目して技術的な概説および現状の紹介を行い，国内における情報科学としての生物多様性分野，「生物多様性情報学」の推進に向けた課題について論じたい．筆者らの専門が生態学，分類学を軸とした生物多様性情報学であるため，内容はかなり同分野に寄ったものになるが，巨大なデータを格納したデータベースをどのように構築するか，横断利用をどのように実現するかという考え方は分野を問わず共通なものであり，他分野の方々にとっても有用な情報になることを確信している．なお，本稿における「データベース」とは，統一した記述フォーマットで記述されたデータを一元化し，データ管理システム内に格納したものと定義する．

2. 国内における公開データベース

巨大データを利用する研究が広がった要因の一つとして，データベースを公開する動きが活発化したことが挙げられる．生物多様性分野においても，近年データベースの構築ならびに公開が進み，それに対する注目度はますます高まっている（Flemons et al., 2007; Claire, 2009; Huang and Qiao, 2010; Whitlock, 2011）．例えば日本では，環境省生物多様性センター J-IBIS から GIS データ化された現存植生図，レッドリスト指定種一覧等の生物多様性情報を取得することができる（http://www.biodic.go.jp/J-IBIS.html, 2012 年 12 月 27 日確認）．日本 Long Term Ecological Research（JaLTER）では，生態学メタデータ（メタデータについては後述）データベースである JaLTER MetaCat が運営され（鎌内・小川, 2008），2012 年 12 月現在で 113 件のメタデータが登録・公開されている（http://www.jalter.org/, 2012 年 12 月 27 日確認）．生物多様性情報機構（Global Biodiversity Information Facility, GBIF：菅原, 2007; 松浦, 2009, 2012）の日本ノード JBIF が運営するポータルページからは，世界中の生物多様性情報の検索が可能になっており

(http://www.gbif.jp/, 2012 年 12 月 27 日確認)，2012 年 12 月現在で約 3 億 5 千万件のデータが検索できる．JBIF 活動の一つでもあるサイエンスミュージアムネット (S-NET：松浦, 2009) が運営するポータルページからは，全国の博物館に収蔵された標本情報が日本語で公開され，約 265 万件のデータを日本語で検索できる (http://science-net.kahaku.go.jp/, 2012 年 12 月 27 日確認)．ここで挙げたのはあくまで一例であり，これ以外にも規模の大小はあるが，現在日本では，非常に多くの生物多様性情報データベースが公開されている．データベースの公開は，国からも補助がなされており，例えば日本学術振興会における科学研究費助成事業において，研究成果公開促進費というカテゴリの中に「データベース」という項目がある．生物多様性情報データベースのうち「昆虫学データベース KONCHU」(http://konchudb.agr.agr.kyushu-u.ac.jp/index-j.html, 2012 年 12 月 27 日確認)「証拠標本データベース VSPECIMENS 」(http://www.vspecimens.net/, 2012 年 12 月 27 日確認) などは，一部この助成金を受けて構築・公開されたものである．もちろん知的財産の問題や希少生物の情報等，公開に際して事前に考慮しなければならない項目は少なからずあるが，全体として生物多様性分野におけるデータベース公開は積極化していると言える．

3. 巨大データの構築

本稿で言及する巨大データには，少なくとも 2 種類，統一ルールに従って莫大な量の観測を行ったデータセットと，複数のプロジェクト等で収集されたデータセットを統合して巨大化させたデータセットがありうる．例えば前者の例としては，センサネットワークを利用し，野外に取り付けた多数のセンサから物理環境データをリアルタイムに取得したもの (Barseghian et al., 2010)，行政の事業として県全域の植生 GIS データが整備された例 (Osawa et al., 2010, 2013) 等が挙げられる．後者の例として，行政資料や博物館の収蔵標本，自身で収集した観察データ等様々なデータを一元化し，県全域における複数種の絶滅危惧植物の分布データを構築した例 (Osawa et al., 2011a, 2011b) 等がある．とはいえ，生物多様性に関する研究において，研究者や研究プロジェクトで収集されるデータセットは一般に比較的小規模であり，巨大データを作成する際には，異なるプロジェクト等で個別に収集されたデータセットを一元化する場合が多い (Neela et al., 2012) しかし先述したように，環境科学関連のデータの統合，横断利用を実現する技術は体系化されておらず (Kwon et al., 2009; 重元 他, 2009)，これらの分野に携わっている研究者は，データ整備の度に多大なる苦労を強いられている．そこで次節からは，筆者がこれまで行ってきた研究や経験等に基づき，独立したデータセットを一元化，横断利用するために何が必要なのか，その実現方法について技術的な面から概説する．

4. データベースの横断利用

4.1 記述フォーマットの標準化

独立したプロジェクト等で構築されたデータベースを統合，横断利用する際には，原則として，いずれか，あるいは双方のデータを加工する必要がある (Conover et al., 2010)．もっともシンプルな方法は，データを記述するフォーマットの標準形式を定め，そのフォーマットでデータの記述形式を統一することである (図 1)．同じデータ記述フォーマットで記述されたデータ同士は，容易に結合 (横断利用) することが可能である (図 1)．しかし，これは単純で簡単に見えて，研究等の現場で実行するには困難を伴う．理由は，「どのフォーマットを標準形式にするか」という課題による．一般に研究機関や研究室，研究者ごとで，記述内容は同じであっても，記述フォーマットは異なる．そして当然のことながら，誰もが自分が使っている形式が一番使いやすく，別の形式に変換するような手間がかかる作業は避けたいと考える．その結果として，

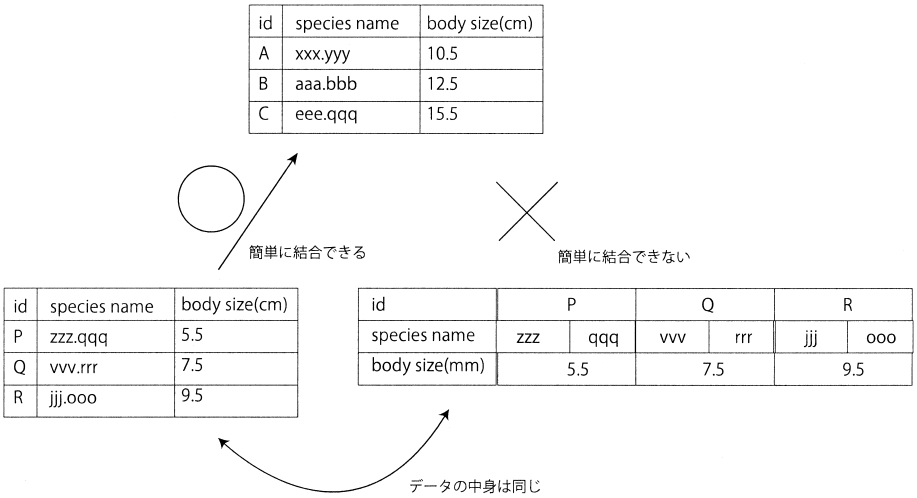

図 1. 記述フォーマットの違いが起こす問題．データの中身が同じであっても，記述フォーマットが同じであれば簡単に結合することができるし，フォーマットが異なると結合に手間がかかるようになる．

研究室ごと，研究者ごとといった具合に，非常に多種多様なデータ記述フォーマットが乱立しているというのが現状である（鎌内・小川, 2008）．実際，先述の昆虫学データベース KONCHU，証拠標本データベース VSPECIMENS それぞれのデータベースにおいて標本情報を閲覧してみると，VSPECIMENS ではタブ区切り等によって情報を区切った形式を採用しており，対して KONCHU では基本的に全ての情報を一画面に入れ込んでいる（図 2）．誰かが既に作成した巨大データに，別の研究者なりがデータを追加して別の研究を行いたいと考えた際にも，統一記述フォーマットをどちらにするかを考えなければならない（図 3）．この問題がデータベースの統合，横断利用を妨げ，さらに，既に研究で利用されたデータセットを使って別の研究を行うといった，データの二次利用（鎌内・小川, 2008）を妨げる一因にもなっている．

データフォーマットの標準化には，「誰もが納得する標準形式」と「どんなフォーマットでも標準形式に書き換えられる手順」の両方が求められる（Thessen and Patterson, 2011）そこで筆者らは，「誰もが納得する標準形式」として，国際プロジェクトである GBIF が採用しているデータ記述フォーマット Darwin Core（DwC, http://rs.tdwg.org/dwc/, 2012 年 12 月 27 日確認）を，生物多様性分野における標準として利用することを勧めている（大澤 他, 2011）．このデータ記述フォーマットは，もともと分類学における標本情報記述のために開発されたものが拡張され，生物多様性情報分野の標準形式策定を担うコミュニティである Biodiversity Information Standards（BIS/TDWG）によって標準形式として規定されたものである．現在では微生物や植生情報等，生物多様性情報に関するかなりの分野で適用可能になっており（三橋, 2010; Wieczorek et al., 2012），最近はゲノム分野との連携も始まっている（Tuama et al., 2012）．

さらに筆者らは，DwC を基盤に，原則としてどんなデータ記述フォーマットでも DwC に互換させることができる方法を確立した（大澤 他, 2011）．概説すると，国際規格である DwC の必須項目（国際ネットワーク上に公開する上で記述しなければならない項目）と研究の性質上必要な項目は DwC の項目名に従って英語で記述し，研究機関や研究者，自身の研究の目的に特異的な項目を拡張項目として付与し，そこは日本語で記述するという極めてシンプルなものである（図 4）．この方式はシンプルな分，高い柔軟性と拡張性を持っており，例えば博物館にお

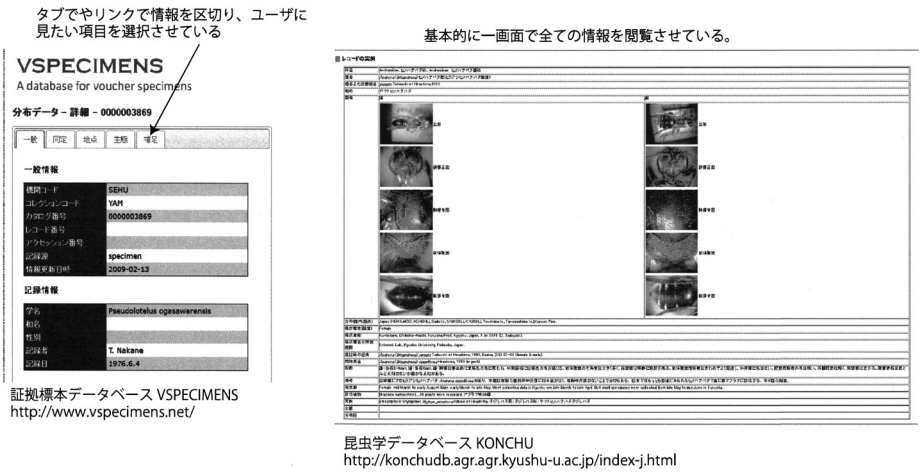

図 2. 本文で例示した 2 つの Web データベースにおける閲覧画面．いずれも昆虫標本情報の閲覧画面だが，データの形式が大きく異なることがわかる．

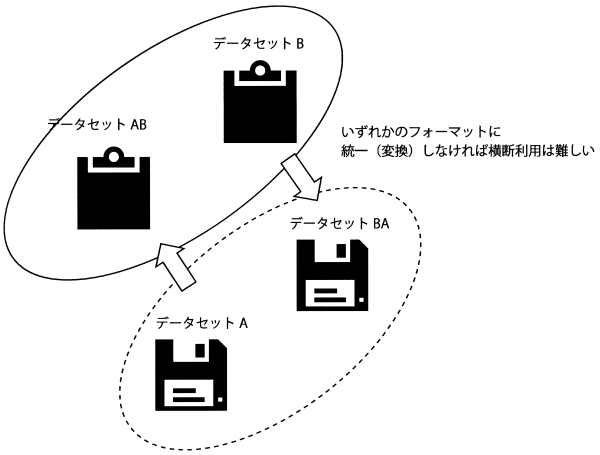

図 3. 例えばデータセット A とデータセット B があったとき，それらを統合するためには，A を B のフォーマットにそろえるか（データセット AB），B を A のフォーマットにそろえるか（データセット BA）いずれかの処理が必要になる．

ける標本ラベル，環境アセスメント等における報告書，植生調査票，市民参加イベント等による生物の観察情報等，非常に多くのデータ形式に適用可能である．この方式は 2012 年 12 月現在，生態学分野における標準フォーマットとして受け入れられはじめており，例えば東京大学が主体となっている文部科学省グリーン・ネットワーク・オブ・エクセレンス」（GRENE）事業（http://www.nara-wu.ac.jp/rigaku/grene/about_grene.html, 2012 年 12 月 27 日確認）の生物多様性分野におけるデータ作成，国立環境研，東京農工大学と共同で進めている紙媒体の生物観察情報の電子化，首都大学東京における標本情報の公開フォーマット等，様々な主体において利用されはじめている．

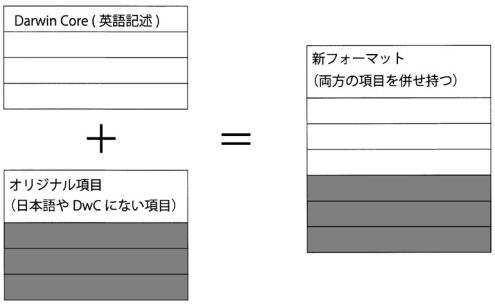

図 4. 大澤ら (2011) によって提案している，Darwin Core を柔軟に既存記述フォーマットに互換させる方法．この方法で作成された新フォーマットは，少なくとも Darwin Core 部分が共通するため，データの統合が容易に行える．

4.2 通信プロトコルの標準化

標準フォーマットで整理されたデータベースであっても，数が膨大になると，データの統合作業は煩雑になる．例えば各データベースにアクセスし，コピー＆ペーストによってデータを取得するという作業は，ミスが混入する危険も多く，効率も非常に悪い (Kwon et al., 2009; 重元他, 2009)．よって次に求められるのは，データ通信プロトコルを統一し，人間が介在せずに複数データを利用する仕組みを構築することである (Conover et al., 2010) プロトコルとは，手順や規定と訳される言葉だが，ここでは「電子データにアクセスするためのルール」を考える．同じルールに従っているデータは，コンピューターによって自動的に統合することができる．この点について筆者らは，Hyper Text Transfer Protocol (HTTP) プロトコル上で Representational State Transfer (REST) とよばれるアプローチを利用することを提案している (Richardson and Ruby, 2007; 大澤 他, 2011, 2012)．REST は，HTTP のメソッドと引数付きの URL でリソースにアクセスする手法である．筆者らは，URL に引数を与えて HTTP GET メソッドでリクエストすると，データがテキストや JSON 形式などで得られるような Web API (Application Program Interface) を開発することを提案している (大澤 他, 2012)．このプロトコルを利用してデータアクセス過程を単純化することで，プログラミングによって自動的に処理を行わせることが容易になる．つまり，いちいちデータベースにアクセスしてデータをコピー＆ペーストしたり，ウェブページのフォームを操作したり，HTML を解析してデータを抽出 (＝スクレイピング) したりする煩雑なプログラムを作ってアクセスする必要がなくなる (図 5 左)．筆者らが中心になって開発したオサムシ科標本情報閲覧システム (http://habucollection.dc.affrc.go.jp/, 2012 年 12 月 27 日確認)，農業環境情報データセンター gamsDB (http://agrienv.dc.affrc.go.jp/, 2012 年 12 月 27 日確認) は，昆虫標本情報や各種の農業環境情報へアクセスする Web API を，REST を用いて提供している (大澤 他, 2012)．また，環境省生物多様性センターによるアジアの生物多様性情報および分類学の基盤整備を行うプロジェクト東・東南アジア生物多様性情報イニシアティブ (East and Southeast Asia Biodiversity Information Initiative, ESABII, http://www.esabii.org/, 2012 年 12 月 27 日確認) の情報ポータルでは，「マッシュアップ」によって絶滅危惧種に関するデータを様々なリソースから取得し，それを政策決定者が利用可能な形にまとめて公開している (Kurashima et al., 2012)．

Web API を利用する利点として，別のサービスとの組み合わせが容易であることも挙げられる (大澤 他, 2011, 2012)．例えば先に紹介したオサムシ標本情報閲覧システムでは標本採集地点の地図が閲覧できるが，地図リソースには Google 社の Google Map を利用している (大澤

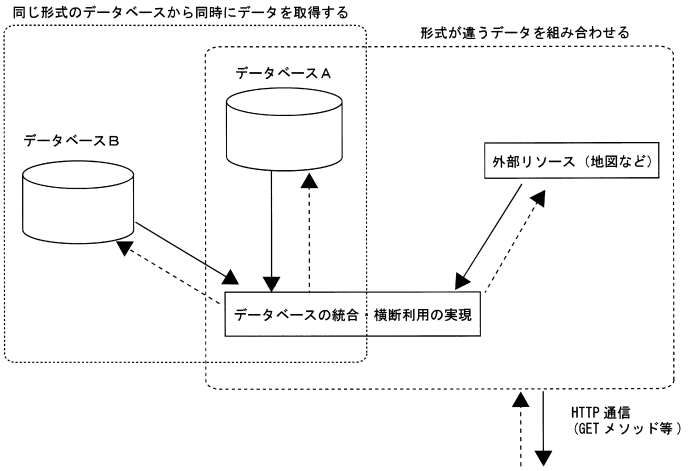

図 5. Web マッシュアップによるデータ横断利用，別リソースとの組み合わせを示した概念図．例えばデータベース A とデータベース B を組み合わせて巨大データベースを構築することや，地図のような性質が異なるリソースを組み合わせて別のサービスを構築する等，自由度の高い横断利用が実現できる．

他, 2011)．つまり，公開しているシステム本体は地図リソースを持っておらず，外部の地図を組み合わせている (図 5 右)．このように Web サービス (本稿においてはデータ配信) を組み合わせて，短期間でサービスを開発する手法は「Web マッシュアップ」と呼ばれており，現在我々が目にしている様々な Web サイトで採用されている (Benslimane et al., 2008; 長嶺 他, 2009; 田中・平藤, 2009; 下條 他, 2010; 大澤 他, 2011, 2012)．Web マッシュアップは組み合わせ次第で新しい価値を生み出す可能性がある手法であり，データの統合，横断利用の実現に加えて，製作者が意図しなかったような新しいデータの利用形態も生まれてくることが期待できる．

4.3 メタデータレベルでの共有

環境科学に関係するデータは種類が多岐にわたり，それぞれのデータを標準化された統一フォーマットで記述することは困難な場合が多い．例えば同じ生態学データであっても，生物の分布データと，二酸化炭素のフラックスデータを記述する共通フォーマットを定めることは不可能であろう．こういった状況においてデータの相互利用性や欲しい情報の検索効率を高めるには，データそのものではなく，メタデータ (二次情報) の記述フォーマットを標準化して共有することが有効である (絹谷 他, 2008; 大澤 他, 2012; Jinbo and Ito, 2012)．メタデータとは，「データのデータ」と表現されるデータのことで，例えば「いつ，どこで，誰が，どうやって取ったデータで，どんな形式 (フォーマット) で記述されたものなのか」等が記述され，データそのものとは別に作成される．具体例を出すと，書籍等の出版物のメタデータには，表題や著者名，発行年月日や出版社名等を記述する (吉野, 2011)．メタデータの概念図を図に示した (図 6)．メタデータの記述項目はデータそのものの性質に左右されないので，全く違う性質のデータであっても，メタデータは同じフォーマットに記述することができる．例えば先に例示した生物分布データと二酸化炭素フラックスデータも，作成者，タイトル，公開年月日などは共通の項目を持つ．したがって，メタデータレベルであれば，同じ「生態学」に関するデータベースに格納することは容易である．このような，メタデータを収集したデータベースは，メタデータデータベースと呼ばれる．

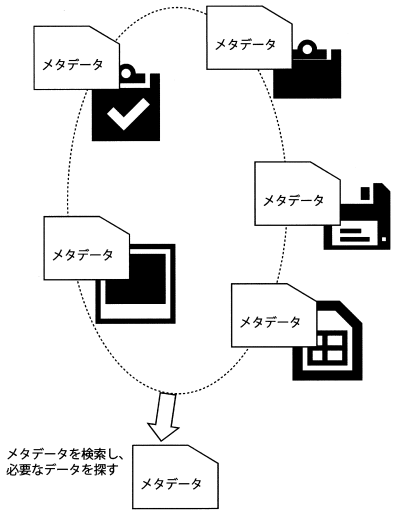

図 6. メタデータの概念図．メタデータはデータの性質に左右されないので，データの記述フォーマットやデータそのものの性質に関わらず，同じフォーマットで作成することができる．それらメタデータを検索することで，ユーザは目的のデータを見つけることができる．

メタデータデータベースの一次的な役割は，ユーザが求めるデータがどこにあるのかを探す「検索」にある．データそのものの取得ではないことに注意していただきたい．現在国内で公開されている「クリアリングハウス」の多くは，目的のデータがどこにあるかを検索するメタデータデータベースである（例えば国土情報クリアリングハウス http://nlftp.mlit.go.jp/chm/index.html，生物多様性クリアリングハウス http://www.biodic.go.jp/chm/index.html ともに 2012 年 12 月 27 日確認）．ユーザはメタデータデータベースによって，性質が異なるデータベース集合の中からであっても，求めるデータの在り処を見つけることができる（図 6）．

メタデータのデータ記述フォーマットは，既に ISO 等によって，いくつかの標準形式が定められている（例えばウェブのメタデータ形式である Dublin Core は，ISO 15836 によって国際標準になっている）．生物多様性の分野に限ると，Ecological Metadata Language（EML）というメタデータフォーマットが確立されている（Fegraus et al., 2005）．EML は Knowledge Network for Biocomplexity（KNB http://knb.ecoinformatics.org, 2012 年 12 月 27 日確認）という生物情報学の国際プロジェクトにおいて検討が進んでいるメタデータフォーマットで，主に International Long Term Ecological Research（ILTER, JaLTER はこれに含まれる）で活用されている生態学用メタデータ形式だが，先述の GBIF や，世界の生物多様性観測情報の集積と活用を目的としたプログラム Group on Earth Observations，Biodiversity Observation Network（GEO BON）においても採用されており，事実上の業界標準と言える．

4.4 メタデータとデータ利用の融合

しかし，メタデータデータベースによって求めるデータベースを見つけることができたとしても，そのデータを実際に利用するためには，原則としてそれぞれのデータベースにアクセスし，データを取得し，手元でマージするという作業が求められる．そこで筆者らは現在，データの横断利用とメタデータレベルの共有を両立する仕組み作りに取り組んでいる．具体的には，メタデータ内に実際にデータにアクセスできる WebAPI の仕様を記述する仕組みを開発

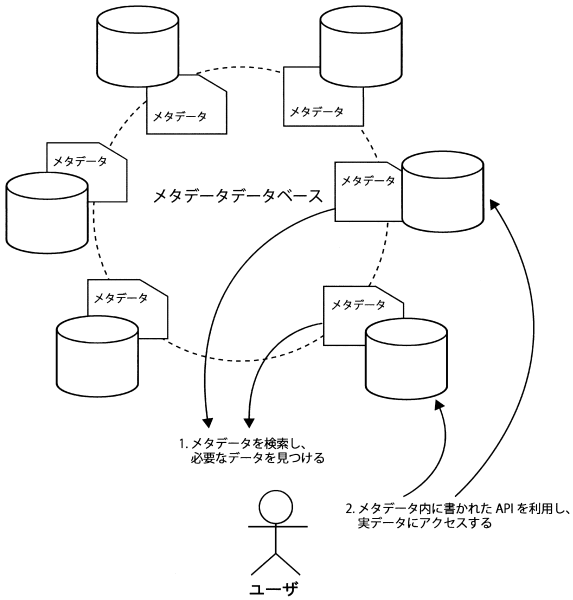

図 7. メタデータデータベースと実際のデータ横断利用を両立させる仕組みの概念図．ユーザはまずメタデータを検索し，自身が求めるデータベースを見つけることができる．メタデータにはデータ本体にアクセスするための API 仕様が記述されているので，ユーザはそれを利用し，簡単なプログラミングでデータにアクセスできる．

している (図 7)．記述された仕様に従って API を利用すれば，ユーザは先述の REST でデータベースのデータにアクセスすることができる．このメタデータデータベースを利用すれば，メタデータデータベースによって目的のデータベースを見つけた後，そのまま半自動的に複数データベースから横断的にデータを取得することができる (ある程度の技術は求められるが)．なお，農業環境技術研究所では，この方式を採用したメタデータデータベースを公開し，各種の農業環境情報を順次公開していく予定である．

4.5 次世代のセマンティック技術

これまで述べてきたデータの横断利用技術は原則として「ユーザ」が「どんなデータが必要で，どう処理するか」を明示しなければならない．横断利用するデータベースやデータ項目の数が少なければ，同一の標準形式に従っていなくても，各データベースの項目の対応関係をつけてケースバイケースで処理することは可能である．しかし，その数が多くなり分野も増えてくると，手作業でこれを行うのは困難なので，別のアプローチを考える必要がある．

この問題を解決することを目的としたアプローチの一つがセマンティックウェブと呼ばれる技術である (Berners-Lee et al., 2001)．この技術は，様々な情報をコンピューターが利用しやすい「リンク構造」で記述することで，自動処理やデータの再利用を促進させるものである．中でも Linked Data は，具体的な生データを項目 (URI) のリンクの集合で記述するもので，セマンティックウェブの実践的アプローチである (Bizer et al., 2008; ヒース・バイツァー, 2013)．どのようなデータもリンクの形で書けるので，異分野の情報が自動的に統合され単一のネットワーク上に記述されることが特徴である．ここで成立する「データ間のネットワーク」を用いることで，人間が介することなくコンピューターのみで異分野の多様なデータ間を結びつけることが

可能になり，検索効率の向上や新知見の獲得を実現する（Minami et al., 2012; 武田 他, 2012; ヒース・バイツァー, 2013）．Linked Data は，将来的に巨大データを扱う上での標準技術になる可能性があると期待されている（Thessen and Patterson, 2011）．

最近，Linked Data 形式で様々な情報を誰でも利用可能な形で公開する Linked Open Data（LOD）と呼ばれるアプローチが活発になってきている．その例としては，Wikipedia の LOD 版である DBPedia（http://dbpedia.org/, 2012 年 12 月 27 日確認）等が挙げられるほか，生物多様性の分野においては，生態学データを対象にした研究例が示されている（Reichman et al., 2011; Mai et al., 2011）．日本国内でも，情報・システム研究機構による研究プロジェクト Linked Open Data for Academia（LODAC）において，博物館の収蔵標本や生物の学名等を対象に，LOD を適用した研究が実施されている（南 他, 2012; Minami et al., 2012; 武田 他, 2012）．その成果として，生物多様性に関する情報も LOD 形式で整備され，実際に公開されている（http://lod.ac/, 2012 年 12 月 27 日確認）．

5. 国内の環境科学分野における現状と課題

以上概説してきたように，複数のデータベースを整理，統合し，横断利用するための技術的な問題は解決，あるいは解決に向けた方法が提案されており，巨大データを扱う標準的な技術は確立されつつある．しかし，冒頭に述べたように，環境科学の分野ではその普及が遅れている．では，なぜ環境科学の分野においてそれらの技術が普及しないのだろうか．本節では，その問題について筆者らの考えと，その対策案について述べる．なお，この説の内容は主に筆者らの経験に基づくため，より生物多様性分野に限った内容になることを明記しておく．

まず何より大きな問題は，担い手不足である．日本の生物多様性の分野において，情報科学の専門家が自身の研究テーマとして生物多様性分野に関わってくる例は非常に少ない．国際的な情勢に目を向けてみると，『biodiversity informatics（生物多様性情報学）』（Guralnick et al., 2007），『ecological informatics（生態情報学）』（小川・藤原, 2007）と呼ばれる分野が既に確立し，専門の学術誌も存在している（Biodiversity Informatics ISSN: 1546-9735; Ecological Informatics ISSN: 1574-9541）．それに対して日本では，これらの分野は知名度も低く，専門家はほとんどいない．Google scholar を利用して“生物多様性情報学”を検索したところ，2012 年 12 月 27 日時点で完全一致はわずか 4 件であった．“生物多様性情報”にしても，87 件であった（http://scholar.google.co.jp/, 2012 年 12 月 27 日確認）（“生物多様性”は 6,410 件，‘データベース’にいたっては 3,130,000 件がヒットした）．

では，生物多様性研究の現場でどのようにデータベースの管理が行われているのかというと，少し情報技術に長けた大学院生やポスドク，若手研究者がやむを得ず研究室や機関の情報管理を担っている場合が圧倒的に多い．残念なことに，現在の生物多様性分野では，データベースの構築やデータ管理そのものは全くといっていいほど研究者としての評価につながらず，彼らの尽力は基本的に無償労働となっている．そのため，若手研究者が情報技術を身に着けると，かえって自身の研究時間が奪われ，本人のキャリアに悪影響を及ぼすという最悪のスパイラルが形成されている．このような状況において，当該分野の若手は積極的に情報技術を学んだり，身に着けたいと考えるだろうか？これは極めて由々しき問題であるにも関わらず，対策らしい対策も行われていない．データベースそのものを査読つき論文として公開する「データペーパー」という制度が，その状況に対する対策として期待されているが（例えば Ecological Research 誌 http://link.springer.com/journal/11284, 2012 年 12 月 27 日確認），これはあくまで「データベースの作成，公開」のみを対象にしており，恒常的に存在するデータベースの維持管理等に対しては，いまだその取り組みを評価できる仕組みが存在していない．

この問題を根本的に解決するには，各研究機関や研究プロジェクトに情報管理担当者：Information Manager を専門職として置くことが必要であろう．これはいわゆる‘充て職’として誰かに併任させるのではなく，情報科学の専門的な技術と知識を有し，さらに最新技術の動向を常に追い続けるような専門家を配置しなければならない．例えばアメリカ合衆国では，演習林や臨海実験所等，生態学に関連する観測が継続的に行われている大規模調査サイトには，情報管理担当者が 1 名以上いるのが普通である．その体制を維持するための意識も高く，大型の研究プロジェクトを立ち上げる際，予算全体の 40% が情報管理に費やされ，参画メンバーに情報管理者がいることは必須条件である（Poter 私信）．台湾では，競争的研究資金のプロポーザルを評価する際，情報管理に対する項目が 25% を占める（Lin 私信）．日本における生物多様性関連のプロジェクトでは，情報管理者はおろか，予算に情報管理という項目すら存在しないのが普通である．日本でアメリカや台湾と同じ体制を構築するのは困難にしても，見習うべき点は多々ある．

次の問題は，環境科学分野における研究者，研究機関ネットワークの不備である．つまり国内では，データベース管理や，データの所在に関する情報共有が十分になされていない．例えば新規で生物多様性データベースを作るとき，標準形式についての問い合わせ先のようなものが存在していない．このため，複数の研究機関やプロジェクトにおいて類似した課題を抱え，同じ問題を起こしてしまうケースが極めて多い．この問題を解決するには，「生物多様性情報ポータル Web サイト」のような，誰もが自由にアクセスできる基本情報の集積サイトを作成するのが有効と考えられる（教科書等は頻繁な更新が難しいので，Web サイトの方が有効であろう）．そこに関連する研究者のコンセンサスのもと作成された「生物多様性データベース構築マニュアル」のようなものを置いておけば，新規作成者はそれを参照の上，安心してデータベースを作成することができるようになるだろう．もちろんポータルサイトは恒常的に運営され，マニュアルは常に更新されていなければ意味がない．国際プロジェクトである GBIF のポータルサイトでは，既にそういったマニュアルや指針をドキュメントにまとめたものを無償で公開し，頻繁に更新されている（http://www.gbif.org/communications/resources/print-and-online-resources/, 2012 年 12 月 27 日確認）．同様に，ILTER では LTER Information Management というサイトにおいて，LTER コミュニティ内における標準形式と技術的な解説を行っている（http://im.lternet.edu/, 2012 年 12 月 27 日確認）．

生態学分野においては，2006 年の時点で既に，日本では個別のデータ蓄積体制が存在するにも関わらず，それらのネットワーク化が進んでいないという指摘がなされていた（榎木 他, 2007）．この状況は，筆者らから見て，2012 年 12 月現在において解決されたとは思えない．日本国内における関連コミュニティは，何よりも先にこの問題を解決しなければならない．筆者らが関わる GBIF 日本ノード JBIF では，この役割を日本でも担えるよう検討をはじめ，GBIF において公開されているドキュメント類を和訳したものを徐々に公開しはじめた（http://www.gbif.jp/, 2012 年 12 月 27 日確認）．同じく国際プロジェクトである GEO BON の日本活動主体である J-BON では，国内における関連コミュニティの役割を整理し，国内関係者における共通認識を確立するための議論を始めた（大澤 私信）．日本は，この問題の解決に向けて少しずつ動きはじめたといえよう．

6. おわりに

近年の環境科学分野，特に生物多様性分野では，解析技術と収集技術にばかり注目が集まっており，データ管理技術がおざなりになっている印象を筆者らは受けている．しかし，高度な統計解析技術の利点を十分に活かすためにも，それに資する巨大データを適切に扱い，解析へ

の利用を容易にすることが必要である．筆者らは，統計解析，データ収集とデータ整備は補完関係にあり，どちらが主で，どちらが従という関係は存在しないと考えている．本特集のタイトルである「環境リスクと統計解析」をより高度化していくためにも，今後は国内における情報基盤が拡充していくような体制の確立が推進されることを期待する．

謝　辞

本稿を作成するにあたり，多くの方と議論させていただいた．全員の名を表することはできないが，特に USLTER Porter 氏，O'Brien 氏，台湾林業研究所 Lin 氏，(独) 国立科学博物館松浦氏，細矢氏，東京大学伊藤氏，倉島氏，(独) 国立環境研究所角谷氏，石濱氏，小川氏，東京農工大学赤坂氏，北海道大学鎌内氏，匿名の査読者に心からの謝意を表する．

参 考 文 献

天野達也 他 (2010)． 日本の保全生物学が必要とするマクロスケールからの視点，日本生態学会誌, **60**, 385–392.

Barseghian, D., et al. (2010)． Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis, *Ecological Informatics*, **5**, 42–50.

Benslimane, D., Dustdar, S. and Sheth, A. (2008)． Services mashups: The new generation of web applications, *IEEE Internet Computing*, **12**, 13–15.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001)． The semantic web, *Scientific American*, **284** (5), 34–43.

Bizer, C., Heath, T., Idehen, K. and Berners-Lee, T. (2008)． Linked data on the web, *Proceedings WWW 2008, Beijing, China*, 1265–1266.

Claire, T. (2009)． Biodiversity databases spread, prompting unification call, *Science*, **324** (5937), 1632–1633.

Conover, H. et al. (2009)． Using sensor web protocols for environmental data acquisition and management, *Ecological Informatics*, **5**, 32–41.

Cukier, K. (2010)． Data, data everywhere: A special report on managing information, *The Economist*, **394** (8671), 3–18.

(独) 農業環境技術研究所 (2011)．『農業と環境の空間情報技術利用ガイド』，創文印刷工業株式会社，東京.

榎木　勉，柴田英昭，日浦　勉，中静　透 (2007)． 日本における LTER の稼働：森林科学からのアプローチ，日本森林学会誌，**89** (5)，311–313.

Fegraus, E., Andelman, S. J., Jones, M. B. and Schildhauer, M. (2005)． Maximizing the value of ecological data with structured Metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation, *Bulletin of the Ecological Society of America*, **86**, 158–168.

Flemons, P., Guralnick, R., Krieger, J., Ranipeta, A. and Neufeld, D. (2007)． A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA), *Ecological Informatics*, **2** (1), 49–60.

Ficetola, G. F. et al. (2007)． Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog, *Diversity and Distributions*, **13** (4), 476–485.

Guralnick, R. P., Hill, A. W. and Lane, M. (2007)． Towards a collaborative global infrastructure for biodiversity assessment, *Ecology Letters*, **10**, 663–672.

林　良英，八尾　徹，五條堀孝 (2009)． 次世代シーケンサーは生命科学に新たな "革命" をもたらす，

科学，**79**(2), 231–244.
ヒース，トム，バイツァー，クリスチャン(2013)．『Linked Data Web をグローバルなデータ空間にする仕組み』(武田英明 監訳)，近代科学社，東京.
Helen, C. et al.(2010)． Using sensor web protocols for environmental data acquisition and management, *Ecological Informatics*, **5**, 32–41.
Huang, X. and Qiao, G.(2010)． Biodiversity databases should gain support from journals, *Trends in Ecology and Evolution*, **26**(8), 377–378.
Iizumi, T., Yokozawa, M. and Nishimori, M.(2011)． Probabilistic evaluation of climate change impacts on paddy rice productivity in Japan, *Climatic Change*, **107**, 391–415.
伊藤元己，神保宇嗣，吉武　啓 (2007)． DNA バーコーディング―新たな生物多様性研究手法，遺伝，**61**(4), 42–47.
神保宇嗣(2012a)．『DNA バーコーディング』，進化学辞典(日本進化学会 編)，共立出版，東京.
神保宇嗣(2012b)．『生物多様性情報プロジェクト』，進化学辞典(日本進化学会 編)，共立出版，東京.
Jinbo, U. and Ito, M.(2012)． Data discovery mechanisms for biodiversity resources in the Asia-Pacific region, *Biodiversity Observation Network in Asia-Pacific Region*: *Towards Further Development of Monitoring Activities* (eds. S. Nakano, T. Yahara and T. Nakashizuka), 195–204, Springer, Tokyo.
Jinbo, U., Kato, T. and Ito, M.(2011)． Current progress in DNA barcoding and future implications for entomology, *Entomological Science*, **14**, 107–124.
鎌内宏光，小川安紀子 (2008)． ワークショップ「生態学関連データベースにおける最近の動向と今後の展望」の報告，日本生態学会誌，**58**, 131–136.
Kato, T., Jinbo, U. and Ito, M.(2012)． DNA barcoding: A novel tool for observation of biodiversity, *Biodiversity Observation Network in Asia-Pacific Region*: *Towards Further Development of Monitoring Activities* (eds. S. Nakano, T. Yahara and T. Nakashizuka), 259–266, Springer, Tokyo.
絹谷弘子，生駒栄司，高橋　慧，吉川正俊，喜連川優 (2008)． 地球観測データに対するメタデータ処理システムの設計，電子情報通信学会 第 19 回データ工学ワークショップ論文集，**C9–6**.
久保拓弥 (2012)．『データ解析のための統計モデリング入門：一般化線形モデル・階層ベイズモデル・MCMC』，岩波書店，東京.
Kurashima, O., Jinbo, U. and Ito, M.(2012)． Development of a portal for threatened species in the Asia-Pacific region, *Biodiversity Observation Network in Asia-Pacific Region*: *Towards Further Development of Monitoring Activities* (eds. S. Nakano, T. Yahara and T. Nakashizuka), 195–204, Springer, Tokyo.
Kwon, Y., Shigemoto, Y., Kuwana, Y. and Sugawara, H.(2009)． Web API for biology with a workflow navigation system, *Nucleic Acids Research*, **37**, Web Server issue: 11–16.
Mai, G., Wang, Y., Hsia, Y., Lu, L. and Lin, C.(2011)． Linked Open Data of Ecology(LODE): A new approach for ecological data sharing, *Taiwan Journal of Forest Science*, **26**(4), 371–378.
松浦啓一 (2009)． 自然史系博物館の GBIF への貢献，日本プランクトン学会報，**56**(2), 169–173.
松浦啓一 (2012)． GBIF(地球規模生物多様性情報機構)の到達点と展望，日本動物分類学会誌，**32**, 31–37.
三橋弘宗 (2010)． 生物多様性情報の整備法，『保全生態学の技法』(鷲谷いづみ，宮下直，西広淳，角谷拓 編)，103–128，東京大学出版会，東京.
南　佳孝 他 (2012)． 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，人工知能学会研究会資料，SIG-SWO-A1103–02.
Minami, T. et al.(2012)． Towards a data hub for biodiversity with LOD, *The 2nd Joint International Semantic Technology Conference, Special Track on Database Integration*.
長嶺貴一，池田宗平，鎌田十三郎，草野直樹 (2009)． マッシュアップデータの選択的閲覧における効

率的な部分更新，*DBSJ Journal*, **7**, 1–6.

Neela, E. et al.（2012）. The user's view on biodiversity data sharing— Investigating facts of acceptance and requirements to realize a sustainable use of research data, *Ecological Informatics*, **11**, 25–33.

日本バイオロギング研究会（2009）.『バイオロギング—最新科学で解明する動物生態学—』，京都通信社，京都.

小川安紀子，藤原章雄（2007）. USLTER のエコロジカル・インフォマティクス技術の動向，日本森林学会誌，**89**(5)，360–364.

Osawa, T., Mitsuhashi, H., Niwa, H. and Ushimaru, A.（2010）. High diversity at network nodes: River confluences enhance vegetation diversity, *The Open Ecology Journal*, **3**, 48–58.

大澤剛士，栗原 隆，中谷至伸，吉松慎一（2011）. 生物多様性情報の整備と活用方法—Web 技術を用いた昆虫標本情報閲覧システムの開発を例に—，保全生態学研究，**16**，231–241.

Osawa, T., Mitsuhashi, H., Niwa, H. and Ushimaru, A.(2011a). The role of river confluences and meanderings in preserving local hot spots for threatened plant species in riparian ecosystems, *Aquatic Conservation: Marine and Freshwater Ecosystems*, **21**, 358–363.

Osawa, T., Mitsuhashi, H., Uematsu, Y. and Ushimaru, A.(2011b). Bagging GLM: Improved generalized linear model for the analysis of zero-inflated data, *Ecological Informatics*, **6**, 270–275.

大澤剛士，神山和則，桑形恒男，須藤重人（2012）. Web API を活用した個別データベースシステムの横断利用，農業情報研究，**21**(1)，1–10.

Osawa, T., Mitsuhashi, H. and Niwa, H.（2013）. Many alien invasive plants disperse against the direction of stream flow in riparian areas, *Ecological Complexity*, **15**, 26–32, DOI: 10.1016/j.ecocom.2013.01.009.

Reichman, O. J., Matthew, B. J. and Schildhauer, M. P.（2011）. Challenges and opportunities of open data in ecology, *Science*, **331**, 703–705.

Richardson, L. and Ruby, S.（2007）.『RESTful Web サービス』(山本陽平 監訳)，オライリー・ジャパン，東京.

柴田英昭（2008）. 日本長期生態学研究ネットワーク(JaLTER)と森林立地研究の関連性と可能性，森林立地，**50**(2)，111–116.

重元康昌，桑名良和，權 娟大，菅原秀明（2009）. 生物分野における Web API の適用，情報知識学会誌，**19**(2)，86–91.

下條 彰，福田将之，井垣 宏，中村匡秀（2010）. 異なるライフログをマッシュアップするためのデータ変換・集約アクセス API の実装，電子情報通信学会技術研究報告，**109**(450)，85–90.

総務省（2012）. 情報通信白書平成 24 年度版.

菅原秀明（2007）. 地球規模生物多様性情報機構(GBIF)とその活動，遺伝，**61**(4)，48–54.

武田英明 他（2012）. 生物情報基盤構築のための生物種データの Linked Data 化の試み，*The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, 3C2-OS-13b-3.

田中 慶，平藤雅之(2009). 農業モデルにおける Web 地図サービスを利用した地図インタフェース，農業情報研究，**18**, 98–109.

Thessen, A. E. and Patterson, D. J.（2011）. Data issues in the life sciences, *Zookeys*, **150**, 15–51.

Tuama, E. O. et al.（2012）. Meeting report: Hackathon-workshop on Darwin core and MIxS standards alignment(February 2012), *Standards in Genomic Sciences*, **7**, 166–170.

Whitlock, M. C.（2011）. Data archiving in ecology and evolution: best practices, *Trends in Ecology and Evolution*, **26**(2), 61–65.

Wieczorek, J. et al.（2012）. Darwin Core: An evolving community-developed biodiversity data standard, *PLoS ONE*, **7**(1), e29715, doi:10.1371/journal.pone.0029715.

吉野知義（2011）. ONIX：書籍流通における出版社のメタデータ標準化，カレントアウェアネス，No. 308, 11–14.

Proceedings of the Institute of Statistical Mathematics Vol. 61, No. 2, 217–231 (2013) 231

Environmental Sciences and BIG Data —Current Status and Future Perspective on Biodiversity Informatics in Japan—

Takeshi Osawa[1,3] and Utsugi Jinbo[2,3]

[1]Natural Resources Inventory Center, National Institute for Agro-Environmental Sciences
[2]Department of Zoology, National Museum of Nature and Science
[3]Japan Node of Global Biodiversity Information Facility

Recently many large databases have been developed for research in environmental science fields. To make meaningful use of such large databases, we need three components: 1) large databases, 2) data management techniques and 3) analyzing techniques. However, 2) data management techniques are often viewed as unimportant especially in biodiversity science fields. In this review we explain the technical aspects of data management as well as show the current status of information technology in biodiversity fields in Japan. Accordingly, we discussed future perspectives in developing an infrastructure on biodiversity informatics in Japan.

Key words: Biodiversity informatics, cross-use, Darwin Core, data format, metadata, standardization.