

Global Biodiversity Information Facility (GBIF)

2011 年 GBIF は 180 万種 10 億件の生物多様性情報を提供する

国立遺伝学研究所 生命情報・DDBJ 研究センター 教授 菅原秀明

“It is the right initiative with the right goals at the right time”

“In our view, if it did not exist, it would need to be created”

(第 3 者による GBIF 評価から(2005))

経緯

GBIF は経済協力開発機構(OECD)から始まりました。OECD はその名称から受ける印象と異なって、加盟国に対して科学技術政策の提言をする機能も持っています。世紀の変わり目に、OECD はバイオに関する科学技術政策について2つの重要な提言をしました。一つには、メガサイエンス・フォーラムに設けられた Working Group on Biological Informatics の報告(1999 年)に基づいて、GBIF と Global Neuroinformatics Capability (GNC) の設立を提言しました。*

もう一つは、Working Party on Biotechnology の Task Force on Biological Resource Centres (BRC)の報告(2001 年)に基づいて、Global BRC Network (GBRCN)の設立を提言しました。**

*) http://www.gbif.org/GBIF_org/facility/OECD_Endorsement

***) <http://www.wdcm.org/brc.pdf>

OECD の提言の中で、GBIF 構想が最も早く実現しました。それは、1992 年のリオ地球サミットで提案された「生物多様性条約(CBD)」を巡る議論の過程で、各国政府が、網羅的な生物多様性情報の必要性を認識していたからでしょう。事実、OECD の提言が公になった後も、1996 年 OECD に Working Group の設置を提案した米国のみならずオランダ、オーストラリア、メキシコなど多数の国の熱意が続き、2001 年 3 月に GBIF が発足しました。なお、GBIF は地球規模生物多様性情報機構と訳されています。国内では科学技術振興機構と国立環境研究所に加え、2003 年から国立遺伝学研究所を中核機関として、東京大学大学院総合文化研究科ならびに科学博物館と連携して、GBIF 日本ノードとして活動しています。

組織

GBIF の組織は、理事会と執行委員会のもとに、常設の科学委員会、予算委員会、ノード(各国に設けられる GBIF 対応窓口)管理者委員会と資格審査委員会などの臨時委員会が設けられています(図 1)。図 1 の右側にある事務局は国際入札の結果 2001 年に選定されたコペンハーゲン大学動物博物館で活動しています。事務局長は、OECD の Working Group 以来米国 National Science Foundation に在籍しながら一貫してリーダーシップを発揮してきた James Edwards が努めています。科学委員会には4つの小委員会 DADI、DIGIT、ECAT ならびに OCB が設けられています。

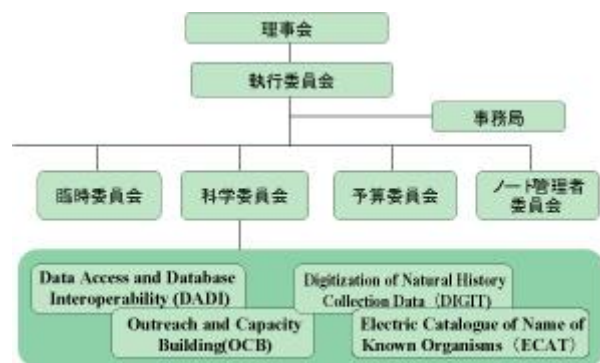


図 1: GBIF の組織

- DADI - 生物多様性の課題を、多様な情報資源をオンラインで組合わせて解決するために必要なデータやメタデータなどの標準を提案します。
- DIGIT - 標本データと観測データのデジタル化を促進します。
- ECAT - 生物に関する文献や研究成果を連携するキーである学名の網羅的電子カタログをドメインごとの努力とも協調して整備します。
- OCB - ソフトウェアの提供やトレーニングによって生物多様性情報に関する技術格差を解消します。また、データを巡る権利関係を解きほぐします。

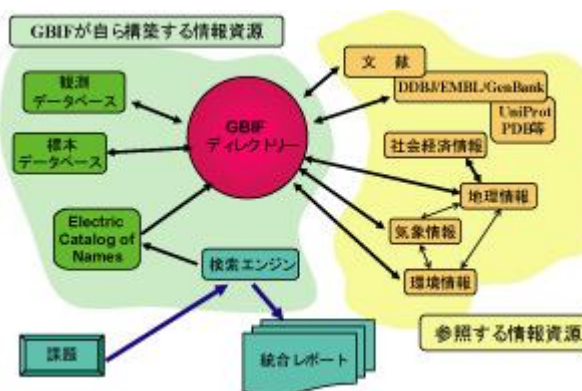
GBIF の運営方針は年 2 回開催される総会で決定されます。総会のメンバーは、Memorandum of Understanding (MOU) にサインした 47 の国ならびに国に準ずる組織体と 31 の NGO です。こうした組織と活動は、26 カ国の拠出金を中心に行われており、総会ではこれら 26 カ国が投票権を有しています。2006 年度の予算はおおよそ 300 万ユーロですが、日本は米国とともにこの予算の多くを担っています。GBIF 総会が扱うこの予算に加えて、総会メンバーがそれぞれ生物多様性情報資源と環境の充実に投資しており、実データの殆んどは各メンバーの努力によって産み出されていることを忘れてはなりません。

P

機能

2006 年 2 月には、GBIF のデータ・ポータルサイト (<http://www.gbif.net/>) から、各国に分散した 159 のサイトが提供する 98,300 種 8,660 万件余りのデータを、統合検索することが可能になっています。GBIF は、2003~2006 年の第 1 期試行期間の後、2007 年から第 2 期に入り、2011 年には既知の生物種を 100% 網羅する 10 億件のデータを提供することを目指しています (図 2 の「GBIF 自ら構築する情報資源」)。加えて、既存の文献情報、分子生物学の情報資源、地理情報、環境情報、気象情報さらには社会経済情報 (図 2 の「参照する情報資源」) との連携を実現する情報環境の構築を目指しています。この情報環境によって、誰でも、何処でも、何時でも、目的にあわせて必要な情報資源を組合せて、多様な生物の検索、分析、保全、応用を進めることができるようになります。

図 2: 生物多様性問題解決環境: Web サービスによる統合



例えば、特定地域での生物種の動向を、自然史博物館に蓄積されてきた標本データ、塩基配列データ、社会経済情報、環境情報などを組合わせて、予測を試みるとしましょう。現状では、始めに、インターネットから必要な情報資源を探し出し、次に、情報資源をそれぞれ加工して、さらに、組合わせる工夫を凝らして初めて分析にとりかかることができます。これに対して、GBIF が提供する生物多様性問題解決環境が成熟してくれば、図 2 にあるように基礎データを効率よく統合して入手することができます。図 2 の左下からにあるように、課

題を GBIF の検索エンジンに与えると、検索エンジンが学名 (Electric Catalogue of Names) を参照した上で、GBIF のディレクトリーを介して、標本データ、DDBJ/EMBL/GenBank (<http://www.insdc.org/>) の塩基配列データ、地理情報、社会経済情報および環境情報などをそれぞれ取得した上で、とりまとめて報告を戻してくれるようになります。

GBIF 内外の情報資源を活用するパイロットプロジェクトも具体に進められています。国立遺伝学研究所でも、国内植物標本情報と Google Map との連携(図3)、GBIF 全データと DDBJ との連携を試行しています。

図 3: GBIF データを Google map に表示
(国内で保存されている植物標本の一部)



仮想データベースのための情報技術

Web サービスと標準化

GBIF は実データがインターネットの多数のサイトに分散しているいわば仮想データベースを提供しています。しかも、各サイトではデータベース管理システムをはじめとしてそれぞれ多様な情報技術が使われています。それでは、どのような情報技術によって、この仮想データベースひいては図 2 の問題解決環境を構築しているのでしょうか。GBIF は、各サイトの個性を GBIF の仲間として相互理解可能なように包み込む (wrap) 技術として Web サービス (<http://www.xml.nig.ac.jp/>) を採用しました。各サイトは、GBIF が提供するツールを使って、GBIF 標準のデータ構造で GBIF 標準の通信プロトコルで、手元のデータを広く一般に提供することになります。標準データ構造と標準プロトコルとして、現在のところ、ノードのおよそ 3 分の 2 が DarwinCore と DiGIR (<http://digir.sourceforge.net/>) を、ヨーロッパを中心とするおよそ 3 分の 1 が ABCD と BioCASE (<http://www.biocase.org/>) を使っています。国立遺伝学研究所の GBIF 日本ノードは国内の生物多様性情報を MySQL に格納し、GBIF のツールを使って、DarwinCore と DiGIR に対応しています。

Web サービスは、GBIF から参照する情報資源でも採用が広がっているので、図 2 の左右の情報資源の連携のためにも有望な技術です。

Globally Unique Identifier (GUID)

GBIF は、学名を多様な情報資源を連携させるキーと設定してきましたが、異なるサイトからの情報の相互参照が可能になるデータが GBIF によって増えれば増えるほど、1 件 1 件のデータの源である標本や菌株を特定する必要があることが明らかになってきました。また、学名の背景にある分類群の概念もサイト間で必ずしも共通ではありません。そこで、物理的実体である標本や菌株から始めて、研究機関や、さらに概念まで、生物に関するあらゆる対象 (オブジェクト) に対して地球上で一意的識別子 (GUID) を付与する枠組みの検討が始まりました。GUID を定義して運用できれば、データの源である物理的実

体を厳密に特定することが可能になり、再現性がありかつ無駄な重複が無い研究の促進につながります。

GUID の技術的基盤には、Digital Object Identifier (DOI、 <http://www.doi.org/>) や Life Science Identifier (LSID、 <http://lsid.sourceforge.net/>) といった候補があるので、生物材料の変異や一部または全部の複製・増殖といった特徴に十分注意を払いかつ GUID の適用範囲を絞り込めば、バイオ GUID は十分実現可能と考えられます。またバイオ GUID の議論は国際塩基配列データベース (<http://www.insdc.org/>) とも協調して進んでいくことでしょう。

関連プロジェクト Barcode of Life (BoL)

GUID は生物オブジェクトに人為的に識別子を付与する枠組みです。これに対して、塩基配列を生物種の識別子にしようという試みが、Barcode of Life (<http://www.barcodinglife.org/>) です。GBIF 自身のプロジェクトではありませんが、GBIF は BoL の国際コンソーシアムのメンバーとなって、相互に密接な連携を保っています。

塩基配列によるバーコード付けは、比較的短い塩基配列によって生物種を特徴付けようとする技術です。したがって、特定の生物種内では共通し、生物種間で変異が大きい部分配列を選択することになります。動物標本の場合は、ミトコンドリア DNA の COL1 をターゲットとして配列決定とデータ蓄積が始まっています。BoL の成果は、分類学者だけでなく生物試料を同定しようとする研究者技術者全てにとって有用な技術です。この技術は、特定の生物集団の専門家でなくても利用でき、従来の手法では同定が困難か不可能であった生物試料の同定も可能とします。バクテリアの場合は、16sRNA の配列が分類と同定に決定的な役割を果たしており、種のレベルでのバーコード付けはされています。植物については、BoL は適切な分子あるいは分子群を検討中です。

BoL のホームページによれば、2006 年 2 月の時点で DDBJ/EMBL/GenBank に 26,398 件の BoL エントリーが登録され、17,834 種についてバーコードが付与されたとのこと。

BEATLES, "Long and winding road"

GBIF の規模は一体どこまで広がるのでしょうか。昆虫の場合は「およそ 95 万種が知られている一方で、推定 800 万種も未知種が生息している」と言われています(参考文献)。したがって、GBIF の ECAT にしても、GUID にしても、BoL にしても全生物種を網羅するまでには遙かな道のりです。しかし、歩みを止めなければやがては目標を達成することができ、その時、新しい世界への扉が開かれることでしょう。

遺伝研の GBIF ノードは、国内外の研究機関や研究グループと手を組んで、一歩ずつ前進していきます。

【参考文献】

Watson, R.T., Heywood, V.H., Baste,I., Dias,B., Gamez,R., Janetos, T.Reid, W., and Ruark, G. (eds.)

“Global Biodiversity Assessment–Summary for Policy–Makers”,

Cambridge University Press (1995)