インベントリー

# 農業環境情報カタログサイト NIAES VIC (Virtual Inventory Complex)

Open data catalog site NIAES VIC (Virtual Inventory Complex)

大澤　剛士

Takeshi Osawa*

## １．はじめに

情報技術の発展に伴い、様々な分野においてデータベースが構築されるようになった。特に環境情報データベースは、自然環境、防災、社会科学等、様々な分野において近年ニーズが急激に高まり、その数、量ともに著しく増加した。環境情報データベースの構築は、現在では国際的な取り組みとなっており（例えば GEOSS; Global Earth Observation System of Systems< http://www. earthobservations.org/geoss.php >、2015 年 12 月 21 日参照）、日本も積極的に参画している（例えば DIAS; Data Integration & Analysis System <http://www.diasjp.net/>, 2015 年 12 月 21 日参照）。

これまで農業環境技術研究所（以下、農環研）では、農業環境に関わる様々なデータセットの整備、データベース化を進めてきた。これらを受け、平成 23 年度より始まった第三期中（長）期計画において、「全国的な土壌、気象、生物、土地利用、衛星画像、農業統計などの農業環境情報を一元的に提供できる農業環境情報統合データベースを構築する。」という内容が提示され、筆者が中心となって統合データベースの構築という研究課題が開始された。本稿は、2015 年に農環研が公開する農業環境統合データベース NIAES VIC (Virtual Inventory Complex)および、その構築に至るまでの研究内容を概説する。なお、本稿においてデータベースとは、同じ性質のデータセットをデータベースマネジメントシステムにおいて一元化したものと定義し、電子化した巨大ファイル等はデータセットと呼ぶ。

## ２．NIAES VIC の概要

第三期中（長）期計画の成果物として公開する NIAES VIC は、データカタログサイト、つまり、これまで農環研で整備、公開してきた各種データセット、データベースを検索し、アクセス可能にするシステムである。検索結果として表示されるメタデータ（データベースの内容を説明するもの）は全てオープンデータとし，可能な範囲でデータ本体もオープンデータとしている。オープンデータとは、データを公開するだけでなく、再利用、再配布も保証した自由な利用を促進するという考え方で（大澤ほか 2014）、オープンデータに適合するライセンスとして、しばしば Creative Commons CC-BY（クリエイティブ・コモンズ・ジャパン http:// creativecommons.jp/ 2015 年 12 月 21 日参照）が利用される。NIAES VIC においても、標準ライセンスとして CC-BY 国際 4.0 を採用している。カタログサイトのシステムは、日本政府のデータカタログサイト DATA.GO.JP（http://www.data.go.jp/ 2015 年 12 月 21 日参照）と同じ CKAN というオープンソースプラットフォ

*農業環境インベントリーセンター

Natural Resources Inventory Center

インベントリー，第 13 号，p18-22 (2016)

ーム（http://ckan.org/ 2015 年 12 月 21 日参照）によって構築され、既存の行政オープンデータとの親和性を最大源高める工夫をしている。同時に、公開するデータベースの一部には API (Application Program Interface)が付与されており、その利用方法がメタデータに記述されている。API を使うことで、インターネットを経由して直接データベースにアクセスできるため、データをローカル環境にダウンロードすることなく利用することが可能になる（大澤ほか 2011, 2012）。これにより、インターネット上で複数のデータを組み合わせたり（大澤ほか 2012：http://agrienv.dc.affrc.go.jp/農業環境情報データセンターgamsDB）、他のリソース、例えば地図と組み合わせたり（大澤ほか, 2011: http://habucollection.dc.affrc.go.jp/オサムシ科標本情報閲覧システム）、モデル計算等の解析を行うこと（http://soilco2.dc.affrc.go.jp/ 土壌の CO2 吸収「見える化」サイト 2015 年 12 月 21 日参照）が可能になる。引用した各種システムは、この API を利用して各種サービスを提供しているものである。NIAES VIC によって、データにアクセスできる API を発見し、それらを自由に組み合わせて新しい価値を生み出すといった「マッシュアップサイト」が農業環境の分野においても促進されると期待できる。この API を一元化し、公開することが、新しいデータベースの統合と筆者は考えている。次項では、データベース統合の考え方について述べる。

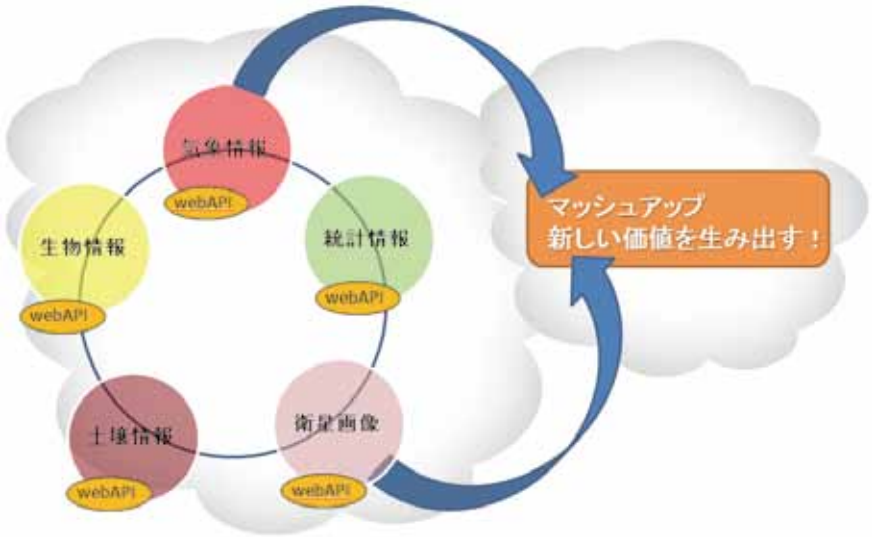


**図１：マッシュアップの概念図。インターネットの雲の中にある各種データベースに API を取りつけることで、雲の中で自由にデータベースを組み合わせ、新しい価値を作り出せる。**

## ３．仮想的な統合

データの『統合』と言うと、多くの方は、データ記述フォーマットを同一にし、同じシステムに様々なデータを一元化することを想像するのではないだろうか。これは、分野を同じくする研究グループが似た形式のデータを共有する場合や、レコード数が数千程度までの、人間が見て認識できる程度のデータサイズを扱う場合は有効な手段である。しかし、多くの専門分野が異なる人々が集まった研究グループにおいて、まるで性質が異なるデータを一元化し、扱うレコード数も数万、数十万という膨大な分量になってくると、必ずしも有効な手段ではなくなってくる。理由は単純で、人間の認識能力を超えてしまうからである。つまり、利用者がデータベースに格納されている内容を把握しきれなくなってしまうのである。特に研究者は、自分の研究分野どころか、自分が扱っている形式以外のデータには極めて疎い場合が多く、これもデータ統合を困難に

している。人間の認識能力を超えた巨大「統合データベース」は何をもたらすのだろうか？本来データベースは、人間が楽をするために構築されたものである。データベースがあることで、必要なデータを改めて収集する必要がなくなり、大量なデータセットから必要なデータのみを抽出する、利用しやすい形に加工するといった効率化を実現してくれる。しかし、認識能力を超えたデータベースを前にした場合、人はどうするかというと、答えは単純で、利用しなくなるのである。無理に使おうとしても、その中身を知るだけで疲弊してしまう。とはいえ、こういった型のデータベースが無意味ということはない。データベース管理担当者（Information Manager）のような専門知識を持つ技術員がデータベースの管理を担い、研究者の要望に対して的確にこたえてくれるような体制であれば、巨大データベースは研究遂行に大いに貢献してくれるだろう。だが、現実的に、少なくとも環境科学の分野において、こういったデータベース管理担当者は皆無に近い（大澤ほか 2013; 2014; 大澤・岩崎 印刷中）。管理者不在のままデータベース構築プロジェクトが乱立した結果、不良債権化してしまったデータベースが多々眠っているという事実は、目をそらしてはならない問題である。

こういった背景を勘案した結果、筆者はデータを物理的に統合することは、少なくともデータベース管理担当者を持たない農環研にメリットをもたらすことはないという結論に達し、別の統合方法を検討した。その結論が、API を利用した仮想的な統合である（大澤ほか 2011; 2012; 大澤・神保 2013）。つまり、分野や性質の異なる個別データベースは独立させたまま、データをインターネット上へ配信する API を実装し，由来の異なるデータをインターネット上で仮想的に統合するという考えに至った。詳細は大澤ほか（2011; 2012）や大澤・神保（2013）に詳しいが、データベースに API を付与することで、データベースの独自性、例えば研究分野やファイル形式、記述フォーマット等に縛られることなく、全てとインターネット上で横断利用することが可能になる。この考え方は、ありとあらゆるデータセット、データベースをインターネットという雲の中に置いてしまい、必要に応じて必要なものだけにアクセスするという、クラウドコンピューティングの考えに基づいている（図 1）。

この技術をデータベースに応用すること、つまりデータベースに API を付与することで、先述のマッシュアップサイトが容易に構築できるようになる。なぜなら、既に全てのデータリソースはインターネット上に存在しているので、システムを構築する際に改めてデータベースを設置する必要がなくなるのである。これまでのデータベースシステムの多くは、システム専用のデータベースを設計し、データを格納し、システムを廃止知る際には基本的にデータベースそのものも廃止してきた（大澤ほか 2011）。しかし、API を使うことで、システムが廃止になってもデータは常に再利用可能になり、さらには同じデータベースを使って複数のシステムを同時に構築することも実現できる（大澤ほか 2011; 2012）。筆者らは実際にこの技術を利用して先述の各種データベースを横断的に組み合わせたシステムの構築を実現し、その有効性を確信することができた。これらの検討およびシステム開発を通し、農環研における農業環境情報統合データベースは、一つの巨大なシステムを構築するのではなく、個別データセット、データベースは独立していても問題ないが、それらを横断利用するため、インターネット上で「仮想的に」統合することが最適解という結論に達した。その成果が、データベースに API を付与し、それをメタデータに記述すること、それらの一元化および検索を実現するカタログサイト NIAES VIC である。

## ４．ライセンス問題

データセットおよびデータベースに関するライセンスとは，基本的に利用条件を記したものである。具体的には，利用に所有者の許可を必要とする，商用利用は不可能といった利用の条件や許諾方法等を明記し，該当データセットやデータベースに付与するものである。だが、データセットやデータベースを公開する際に、ライセンスをどうするかは、知的財産権に関する知識を持たない管理者にとって悩ましい問題である。一般にライセンスをはじめとした知的財産の問題について研究者は軽視しがちであるが、多くの研究はアイディアと共にデータに基づいていることを考えると、研究活動における最も重要な一部を占めると言っても過言ではあるまい。近年、データのねつ造や剽窃が明るみになり、論文の取り下げや学位の取り消しが頻発したことは記憶に新しい。国内の競争的資金に応募する際、倫理教育を受講していることが必須要件となった。一部の心無い研究者によって引き起こされた問題という一面はあるものの、知的財産の問題は、研究者側が知らなかったでは済まされない問題になっていることは認識しておかなければならない。

繰り返しになるが、データセットやデータベースのライセンスは本来データ所有者が策定するものである。しかし、必ずしも専門家ではないデータ管理者が必要な条件を過不足なく策定することは容易でないこと、データセットやデータベースごとに独自のライセンスが設定されることはデータの利用性を損なう可能性があること等から、外部団体が作成した標準ライセンスを利用するという考え方が広がってきた。この標準ライセンスの代表が Creative Commons ライセンス（以下 CC ライセンス）である（クリエイティブ・コモンズ・ジャパン http://creativecommons.jp/licenses/ 2015.12.21 参照）。CC ライセンスについて本稿では詳しく解説しないが，現在データおよびデータベースに付与されるライセンスとして最も一般的なものの一つと言ってよいだろう（大澤ほか 2014）。CC ライセンスは表 1 にある条件マークを組み合わせて設定するもので、一般的に図 2 で示した 6 種類がよく知られている。その中にはオープンデータライセンスと呼ぶべきものもある。具体的には CC BY、CC BY-SA はオープンデータライセンスと呼べるものであり，出典を明らかにする限り，自由な利用と再配布を許可している（大澤ほか 2014）。

表１：クリエイティブ・コモンズで利用される基本的なマーク

| CCライセンスの条件 | 表示マーク |
|---|---|
| 表示（Attribution） | |
| 非営利（Non Commercial） | |
| 継承（Share Alike） | |
| 改変禁止<br>（No Derivative Works） | |

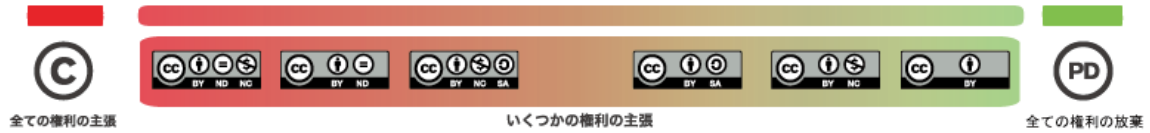


図２：クリエイティブ・コモンズでは、表で示したマークを組み合わせ、主に６つのライセンス表示がされる。

NIAES VIC では、先述の通り検索結果として表示されるメタデータは全て CC-BY 4.0 国際が付与されたオープンデータであり、出典を明記する限り再利用、再配布が可能となっている。これはすなわち、農環研がどんな情報資源を保有しているかについてオープンにすることを意味する。この姿勢は、主に税金を原資として研究に取り組んでいる国立研究開発法人として当然と言えるだろう。ただし、実際のデータセットやデータベースのライセンスについては、個人情報の問題

等を勘案し、データを整備した研究者に委ねることにした。とはいえ、可能な限りオープンデータ化し、多くの方々に利用してもらえる形で公開できるよう努力していきたい。

## ５．今後の展望

データベースシステムの構築は比較的容易だが、それを維持管理し、持続的に運用していくことは簡単ではない。この維持管理をどうするかを考えていくことは重要な課題ではある。しかしその反面、そればかりに囚われていては、データベースの本質を見失う。最も重要なのはコンテンツ、データベースにおいてはデータであり、システムは、それを利用しやすくするだけのものにすぎない。最後にこのことを強調したい。システムのライフサイクルは IT の急速な発展もあって非常に短く、10 年後も同じ仕組みが通用することは考えにくい。しかし、データそのものは 50 年後、100 年後も利用可能である。実際、インベントリーセンターがこれまで整備してきたインベントリーの中には、100 年以上前のものも含まれている。残念なことに、いわゆる「箱もの」としてシステムのみ作成し、コンテンツが充実することなく亡霊のように存在するシステムも少なからず存在している。NIAES VIC も数年のうちに陳腐化してしまうかもしれない。しかし、それが格納する各種データセット、データベースは 50 年後、100 年後も利用可能な形で維持していくことこそが、基盤情報を担う研究者として最も重要視すべきことと筆者は考えている。2016 年 4 月をもって組織が統合され、インベントリーセンターはなくなるが、「農業環境インベントリー」自体は 50 年後、100 年後も維持するように、今後も研究に取り組んでいきたい。

## 引用文献


1) 大澤剛士・岩崎亘典（2016）：環境科学分野における研究データのオープンデータ化の現状と課題. 環境情報科学 44-4:35-40.

2) Osawa, T., Kadoya, T., Kohyama, K. (2015)：Agricultural land use 5- and 10-km mesh datasets based on governmental statistics for 1970 - 2005. Ecological Research 30(5):757.

3) 大澤剛士・神保宇嗣・岩崎亘典（2014）：「オープンデータ」という考え方と、生物多様性分野への適用に向けた課題. 日本生態学会誌　64(2): 153-162.

4) 大澤剛士・神保宇嗣（2013）：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-. 統計数理 61:217-231.

5) 大澤剛士・鎌内宏光・細矢剛・伊藤元巳（2013）：LTER、GBIF における国際的な生物多様性データベースの動向と日本国内の課題 -国際ワークショップ参加報告-. 日本生態学会誌　63(2):269-273.

6) Osawa, T.（2013）：Monitoring records of plant species in the Hakone region of Fuji-Hakone-Izu National Park, Japan, 2001-2010. Ecological Research 28(4):541.

7) 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.

8) 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(2):231-241.


## 問い合わせ先


農業環境インベントリーセンター　大澤　剛士

電話：029-838-8272, e-mail: arosawa@affrc.go.jp