# コレクションの
# デジタル化プロジェクトを開始する

要約:

本文書は、博物学コレクションのデジタル化プロジェクトを計画中の読者に、開始する自信と適切な判断をしてもらえるよう作成された。執筆者は長年にわたるコレクション作業の経験を持ち、読者がこの作業にリソースをコミットする前に、適切な質問を効果的に投げかけ、適切な計画を立てられるように、本文書に専門知識を注ぎ込むことに努めた。

# 第 1 章: 序論

## 本章の目的

デジタル化プロジェクトの開始を決意するのは厄介な作業である。そろそろあなたも「みんながやっている」と恐らくすでに気づいているだろう。 すぐに何か始めなければ、この情報化時代に取り残される怖れがあると不安を募らせているかもしれない。電子データの提供を求める利用者団体、組織の経営陣、 同僚からプレッシャーを受けているかもしれない。あるいは、 あなたが関心を寄せるのは、コレクションがデジタル化されていれば、もう少しうまく対処できると感じる任務を果たす上で、直面する現実的欠陥が主な理由かもしれない。

コレクションの大きさや、デジタル化プロジェクトを開始したい理由とは関係なく、バイオインフォマティクス、システム設計、情報理論についての膨大な学習があなたの関心の的ではないかもしれない。学習したいけれど、どこから手をつけてよいか分からない。この道の途中でコンピュータ科学の別の学位を取得する必要はないとしても、あなたのニーズやリソースに適したソリューションを見つけるために、実際に知るべきことは何だろうか。

ニーズに合った実施可能なソリューションの導入を求めているなら、多くの要素を検討し、数多くの決定をする必要があるという事実を、絶対に回避できない。 多くの人が、着手するとき、「みんながしていることをすればよい」と最初に考えることを、私たちは経験から知っている。それだけで良いなら簡単である。 国産のもの、市販のもの、オープンソースのもの。現在使用できるソリューションは数多くある。状況の性質に合わせて調整したもの、カスタマイズ可能な汎用のもの。実際にはどのソリューションにも何らかの欠陥があり、それが最終的に重大な問題となる。あるものは複雑だがパワーがあり、あるものはシンプルである。あるものはきちんと解説してあり、あるものには説明がない。 市場に出回っているものを知るだけでは十分ではない。というのも、担当者や 研究機関は使用中のソリューションに満足しているか、満足していないか、多くはその中間である。

本文書は、デジタル化プロジェクトを計画中の読者に、開始する自信と適切な判断の手助けのために作成された。私たち、執筆者は長年にわたる大小のコレクション作業の経験を持ち、読者がリソースをコミットする前に、適切な質問を効果的に投げかけ、適切な計画を立てられるように、本文書に専門知識を注ぎ込むことに努めた。本文書の目的は、私たちがデジタル化プロジェクトに着手した時に、欲しかった情報を提供することである。私たちは、具体的なソリューションの販売員や支持者と話し合っても入手できない情報を提供することに努めた。

情報技術も情報科学理論も急速に進化しているので、ニーズに合った ソリューションを見つけることは、動く目標を追いかけるようなものである。わずか数年前に流行した具体的なソリューションが、今では時代遅れで古めかしいものになっている。場合によっては、ニーズにぴったり合うと思われた新しい可能性の兆しも、約束通りの実行力があるかどうかわからない。本文書を読むのではなく、最新の学会の抄録を精読するのに時間を費やすべきだろうか。

私たちは、すべての IT システムについて学んだ、単純な事実を中心に、本文書が時間依存性の低い、具体的なものになるように努めた。例えどんなにパワーがあり、最先端の高価なものであっても、IT システムは包含するデータの品質によってのみ価値がある。現在も将来も、質の高い情報を入力、維持、出力できるソリューションを見つけたいだろう。従って本文書では、保有するリソースと特別な目標のもとで、いかにこれを実施するかについて決定するのを助けることに重点を置いた。 これは本質的に、事前に予測できない多くの分枝をもつ広範なテーマである。いくつかの分野については出版された文書ですでに十分論じられており、その内容を繰り返すより、テーマを完全に理解する必要に

応じて参照するべき適切な資料を紹介する。 本文書で紹介する知識は、私たちが様々なデジタル化プロジェクトに携わる現在の経験の中で生まれた最善のものであり、私たちと同じ立場の研究者の助けになることを願っている。

場合によっては、本文書で具体的なソリューションを推奨しているが、私たちの全般的な意図は、「ベストプラクティス」のセットを提供することではない。私たちの考えでは、具体的な研究機関やコレクションに適したソリューションは、各研究機関に固有の事情に大きく依存するので、単独のベストプラクティスは存在しない。正しい答えを見つけることが、あなたのプロジェクトを成功に導く第一歩である。

## デジタル化の意味

デジタル化とは、情報を電子的形態で保存することである。 私たちの業界の基本情報は、チェックリスト、現場ノート、収集した試料に由来するものか、あるいは出版物、記録文書、他の媒体からの抜粋である。情報の基本単位は、標本館の試料などの物理的対象や、ある一日の森でさえずる鳥類の観察や、一晩の間落とし穴にかかった多くの生物コレクションなどの出来事に関係する。デジタル化の成果は保存され、フォーマット資料、マークアップ資料、スプレッドシート、ウェブページやウェブサイト、フラットファイルやリレーショナル・データベース、地図や GIS システムなど、様々な方法で表現される。

私たちの業界のもう一つの重要な特徴は、デジタル化が、オブジェクトの*画像*を電子的に保存し、オブジェクトに関する*テキスト情報*、あるいはテキストに含まれるオブジェクトからの抜粋の保存を意味することである。これら両方をデジタル化と呼ぶが、前者には「画像化」という用語を使用した。試料の画像化の詳細は、本文書では深く論じていない。地球規模生物多様性情報機構 (GBIF) の文書「生物学的試料のデジタル画像化：ベストプラクティス文書」 (Häuser et al., 2005)は画像化の優れた情報源である。「データベース化」という用語は、テキストベースの情報を保存するプロセスを記述するのに使用した。さらに、多くのデジタル化プロジェクトは、画像とテキストベースの情報の両方を保持する情報システムの構築を伴う。この場合、「データベース化」あるいは「デジタル化」は、情報を保持するこのシステムの構築を意味する。

## 対象になる読者と本章の範囲

本文書は、 デジタル化プロジェクトを始めて開始することに関心のある人を対象にしている。また熟練したデジタイザーにとっても有益であることを願っている。本文書の焦点は、試料ベースの生物学コレクションの管理者を支援することである。それでも管理者は、様々な理由で、デジタル情報の様々な目標に沿って、コレクションのデジタル化を望むことが分かっている。例えば、ある者は、単に所蔵コレクションの電子カタログを構築したいと望み、他の者は、コレクションイベントの詳細な記録を構築しようと努め、さらに他の者は研究機関のプロジェクトベースの研究を支援する、情報システムに、証拠コレクションを統合する方法を模索する。本文書では、具体的な目標の細目に関係なく、デジタル化の取り組みを計画する上で有益であろうと思われる情報を提供する。

デジタル化プロジェクトの規模は、単一の試料をデジタル化する単独の個人から、無数の試料を記録する研究機関の広範なプログラムまで、まちまちである。本文書で論じた課題の多くは、コレクションの規模やデジタル化の取り組みの範囲にこだわらない。いくつかの事例では、取り組みの範囲に固有のガイダンスを提供しようと努めた。

本文書は、データベース設計、コンピュータ技術、博物学に対する、高水準のリテラシーを前提としたものではない。しかし場合によっては 、読者は、ここで論じた概念の速やかな理解を助ける補助的資料を併読するよう指導されるかもしれない。

# 各章の概観

第 2 章「コレクションをデジタル化する目的」は、 デジタル化に取り組む理由と、デジタル化で期待できる利益を簡潔に紹介する。第 3 章「 デジタル化をはじめる前にするべきこと」は、いくつかの主要な部分に分かれる。 はじめの 4 つの項目は、目的の設定と現状分析に関係する。この作業が終われば、プロジェクト導入法の具体的な詳細を検討でき、これは、「適切なデータベースの選択」、「試験的なアクションプランの作成」の章でカバーされる。最後の項目では、計画を実行に移すことを述べる。第 4 章では目で見る情報とコンピュータが見るデータとの対比を論じる。この区別を利用して、データベースに入力し保持するために、システムに保持してほしい情報を操作する方法を説明する。 第 4 章では、標準、データ品質、言語、知的財産権など、他のデータ関連問題についても論じる。 第 5 章では、簡単なカタログから、情報管理システムのモジュール設計まで、データモデルの概念について論じる。 第 5 章では、続いてデータモデルをコンピュータシステムに導入する方法を論じ、導入に関連する基本的な問題を論じる。最終第 6 章では、ニーズに合った具体的なデータベースソリューションを評価し選択する方法を詳細に検討し、助言を与える。第 6 章まで読破した読者は、開発プロセスに適した事業計画やアクションプランの構築に着手する自信が持てるだろう。付録 A と付録 B には、これらの計画を開発するプロセスの簡潔な概要を収録した。

# 第 2 章: コレクションをデジタル化する理由

コンピュータ導入前の博物学コレクションは、骨折って転写される情報の物理的データベースであった (Lane, 1996)。 その後データが発表されない限り、転写は一人の関係研究者がデータを利用するだけである。コンピュータの出現によって、またインターネット経由のデータアクセスの増加に伴って、博物学 コレクションの潜在力を活用する新しい道が開かれた。

デジタル化は、 コレクション、スタッフやワークフロー、潜在的データ利用者にとって大きな利益がある。それでもデジタル化には実質費用がかかるので、厄介なデジタル化を開始する理由を明確に理解し取り組む必要がある。具体的な理由を説明されなければ、管理者や潜在的資金団体は労力や費用を正当化する理由を明確に理解できないだろう。プロジェクトが進行中または完了した時点で、これらの理由がデジタル化 プロジェクトの有効性を評価するベンチマークや他の基準を定めるために使用される。

デジタル化された データの使用に関する広範な考察は、Chapman (2005c)に基づく本*文書*の第 1 章で論じた。デジタル化 プロジェクトを開始する一般的な理由には以下の事柄が含まれる。

**広範なデータの普及**

試料の一次 情報は、一般的に試料シートそのものに記述されたデータだけである。従って、現在試料を所有する人物だけが利用できる。 デジタル化されていないデータを研究機関同士で閲覧するには、関係研究者が個人的に訪問し、あるいは、運搬や学芸員業務など、潜在的に高いコストがかかる、試料の貸し出しを受けなければならない。 デジタル化されたデータは、多くの方法で流布できる。主としてインターネットを使用して、より多くの人がデータを入手し利用できる。

**データを様々な方法で研究することを可能にする**

一度コレクションをデジタル化しておけば、以前は容易でなかった方法でデータを照会できる。 例えば、収集旅行の旅程を追跡できるので、収集者や収集日別にデータを整理できる。 生物分類の科ごとに整理された コレクションでは事実上不可能である。 デジタル化された 試料記録も種の多様性を評価する上で重要な役割を果たす (Meier and Dikow, 2004; Chapman, 2005c)。関連データが優れた

構成のデータベースに記録されていれば、あなたが求めるどんな方法でもこれを閲覧できる。

**学芸員の業務を向上させる**

コレクションのデジタル化は、通常、必要な蔵書の量を削減し、 研究機関の日常業務を支援できる。試料の貸し出し状態を追跡してコレクションの経路を把握できる。間違いが検出され、コレクションの品質が向上する。「喪失した 試料」が発見され (Peterson, 2002) 、試料ラベルに記された用語を統一できる。デジタル化すれば、主に、前の諸点で述べた様々な方法でデータを研究しようとする時、素早く有益なデータの欠如を明らかにできる。デジタル化のような、コレクションの真の奥深さを知ることを可能にする道は他には見当たらない。(Peterson, 2002)

**試料を保護する**

デジタル化するには必然的に、オリジナル試料を参照しなければならない。一度試料を参照しておけば、代わりに試料データが移送されるので、試料を取り扱う必要が削減される。。資料の取り扱いが減るので、オリジナル試料の寿命も延びる。これは特に、基準標本など、掛け替えのないアイテムにとっては重要である。しかし、今でも多くの研究形態がアイテムそのものの物理的検査を求めているので、試料へのアクセスを不可能にし、あるいは制限を加えるものではない。試料の画像を含めてデジタル化しても、試料を取り扱う頻度が少なくなるだけであり、完全に置き換えることはできない。デジタルコレクションは、災害管理の一形態としての機能も果たす。 最悪の事態が起きてオリジナルコレクション が破壊されても、デジタルコレクションは価値あるリソースを提供し続ける。

**未来の転写時間を削減し研究を助ける**

一度試料データを転写しておけば、同じ試料に関わる未来のプロジェクトは転写を繰り返す必要がない。これによって将来のプロジェクトの効率を上げ、費用要求を削減できる。

**研究機関/コレクションのプロファイルを向上させる**

研究機関は、所蔵コレクションだけではなく、幅広い情報源からのデータにアクセスできることに関心がある。 研究機関のコレクションに、より多くのアクセス可能性、従って研究プロジェクトの（財政的またはその他の）リソース改善を求める要求も増している。 多くの新しいプロジェクトは、デジタルデータの充実を推進し、結果として生じるデータにオンラインアクセスできることを要求する。高品質データに対する利用者の評価は、その後のオリジナルコレクションに対する高い評価につながり (Lane, 1996)、コレクションの重要性を高める。デジタル化は、コレクションの大きさ、発達状態、利用法を監視することも可能にする。新しいプロジェクトの財源を求める時、非常に有益である。デジタル化 は、試料データを発信国に送還するとき CBD（生物多様性条約）の要件を満たすこともできる (Meier and Dikow, 2004)。

**研究機関が従来の権限を超えた分野に貢献する能力を高める**

伝統的に、博物学研究所は試料を保存し、研究者の命名法上の研究に寄与してきた。 試料データが利用可能になれば、分類学研究者に提供する利用法だけでなく、新しい関心領域にも提供できる。データは教育現場にも、また研究機関の業績に対する一般市民の理解を深めるためにも利用できる。データ はコレクションの欠落部分を特定するために分析され、未来の収集旅行に役立つ収集の指針を作成できる。数多くの潜在的データ利用法が存在し、 デジタル化された データが利用可能になれば、容易に導入できる。

**法整備**

データの広範な利用可能性は、各国の「公的資金を受けた研究機関の情報入手」に関する法律でますます求められている。

# 第 3 章: デジタル化をはじめる前にするべきこと

## 計画立案は重要である

明快な計画立案は適切な データベース作りに不可欠である。具体的に言うと、どのような規模であれデジタル化 の取り組みを計画することは、それ自体がプロジェクトである。 組織的なプロジェクト管理技術の適用は、具体的なプロジェクトを成功させる可能性を向上させる。 優れた管理原則に従ってデジタル化 プログラムを導入することを強く推奨する。しかし、プロジェクト管理技術の徹底的な議論は本文書の範囲を超えている。 プロジェクト管理のテキストは広く利用可能であり、読者は、プロジェクトに着手する前に参照することを推奨する。本文書では、プロジェクト管理から直接派生する 3 つのテーマを論じる。 それは、投資対効果検討書、アクションプラン、リスク分析の 3 つである。

**投資対効果検討書**は、実現したいことを設定し、 提案された作業を実行することで獲得が期待される利益を明らかにする。また、プロジェクト導入に必要なリソースを評価し、現在利用可能なリソースを特定することも含まれる。 不足しているリソースをはっきり特定し、関連コストをはっきり述べた方がよい。これらの事実を 1 つの文書にまとめることによって、プロジェクトの実現可能性を明快に判断でき、限られたリソース（たとえ 1 人しかスタッフがいなくても） がデジタル化に使用される理由を明確にできる。

**アクションプラン** は、投資対効果を実際に導入する方法を詳述する。これには必要な コンピュータやデータベースの数など、実際的情報が含まれる。スタッフの人数、トレーニング、デジタル化作業を進める方法 (一般に ワークフローと呼ばれる)なども検討する。アクションプランは、リソース不足を賄うための資金はどこで調達できるかを詳述する。

**リスク分析文書**は、 何らかの不具合の対応策の検討を目指すアクションプランの一部である。コンピュータが機能しない場合、あるいはプロジェクトの一部を担う資金が確保できない場合、どうなるかという分かりやすい例を示す。事件が起きた場合のリスクを最小限にとどめる方法も検討する。 定期的にデータをバックアップし、資金調達の別の可能性を模索し、利用可能な予備のコンピュータを準備する。これらすべてがリスク分析文書で検討される簡単なリスク緩和法である。

投資対効果検討書は、1 つのプロジェクトで完了するには大きすぎる目標全体を概説する可能性があるが、それぞれが独自の投資対効果検討書、アクションプラン、リスク分析文書を備えた、いくつかのより小さなプロジェクトに細分化される。これは完全に受け入れ可能な慣行であり、全体的な投資対効果検討書のアクションプランは、個別のプロジェクトを概説するとともに、全体的目標を提供するために個別のプロジェクトを結合させる方法を概説する。こうした作業方法によって、全体的展望を見失うことなく、しばしば長期的で大きな目標を徐々に達成することができる。

計画段階を急がないことが重要である。というのも、データベースが公開されてから、問題を修正するのは困難で時間のかかる作業になるからである。正しい導入を目指す研究機関の計画段階は、優に 6 か月から 1 年はかかる。これは当初やる気をそぐかもしれないが、 要件を正しく理解するのに時間をかけることは、不適切なパッケージが公開された時の落胆を回避するものである。

計画したデータベースが準備できる前に、短期的ソリューションを準備する必要があることもある。この場合、短期的データベースから恒久的ソリューションにデータを移動させる時間も計画の中に含める必要がある。そうしないと、短期的ソリューション が事実上恒久的 データベースになってしまう!

# 目標を設定する

本章では、デジタル化 プロジェクトを開始する理由の一般原則を特定した。正当な一般的理由はあるものの、具体的なデジタル化 プロジェクトを開始する明確な理由を特定した方がよい。本章では、あらゆるデジタル化 プロジェクトが、その存続期間のある時期に解決しなければならない適切な疑問を提起することを目標としている。 多くのプロジェクトは、実際にプロジェクトを開始する前に、これらの疑問のすべてを検討していない。しかしプロジェクトの変更を余儀なくされた時、概して相当量の追加作業が生じる。 本章で特定した疑問を解決しておけば、プロジェクトのリソース、要件、制限に関して、はるかに優れた考えを持てるはずである。

**研究機関の目標と個人の目標**

研究機関全体を対象にしたプロジェクトか、 それとも個人向けのプロジェクトなのか。プロジェクトの規模を認識することは、 プロジェクトが動き始めた時に直面する多くの制約を明確にする。 例えば、個人向けプロジェクトであれば、必要なワークフロー、コンピュータの使用、物理的スペースの要求は、研究機関の全システム向けプロジェクトよりもはるかに小さい。同様に、小規模プロジェクトの場合、研究機関全体に比べて比較的少数の試料しかデジタル化できない。研究機関のプロジェクトの場合、担当スタッフを配置することが非常に重要である。つまり、プロジェクトと研究機関の日常業務との相互作用には細心の注意が必要である。 プロジェクトの目的や手順についてのトレーニングや教育は不可欠である。これらのプロセスを初めから理解していないと、 プロジェクトが成功裏に完了する可能性が著しく低下する。システムの規模が大きいほど、基礎データベースが複雑で厳密に設計されたものになる傾向があり、個々の研究者のニーズに応える新しい領域を臨機応変に創り出す余裕が少ない。

**ソリューションの主要なクライアントは誰か**

潜在的なデータ利用者は数多くいる。試料から作成した初期データの利用者には、分類学者、管理者、研究者、検査助手、収集者、環境問題専門家、NGO、薬理学者、一般市民などが含まれる (Chapman, 2005a; 本文書第 3 章も参照)。必然的に、主要なオーディエンスであるターゲットユーザーの小集団ができる。通常、博物学デジタル化プロジェクトのターゲット・オーディエンスは以下の三者である。

- 具体的なプロジェクトに携わる個人
- 一般の研究者
- 研究機関の学芸員

一人用のシステム導入は通常、非常にシンプルである。その個人に固有の要件は容易に構成することができる。彼らは大抵、データ入力とクエリーシステムの単純な形態を要求するだけなので、論文、植物相、チェックリストなど、単一のターゲット結果が得られるように設計される。しかし、こうしたデータセットは、広く利用可能にするための具体的な取り組みがされない限り、初期プロジェクト以上の使用は限定される。 研究機関レベルでは、すべての記録データが中央ロケーションで利用可能になるという要件を、内部プロジェクトに課すことがしばしば有益である。その結果、他のスタッフが以前に記録されたデータを見つけることができ、プロジェクト全体の重複作業を削減する。

研究者には、 研究機関の内部研究者と外部研究者がいる。これら 2 つの広いカテゴリーの研究者はそれぞれ、データを入手できるある種のインターフェースを要求する。外部研究者は、通常、ウェブサイトを使って情報を入手する。内部研究者も、外部ユーザーと同じウェブサイトを利用するが、追加情報 (分かりやすい例として、戸棚の中の試料の位置)を入手できる方が望ましい。個人研究者だけ

ではなく、すべての研究者にデータを利用可能にすることに専心することは、研究機関レベルで 2 つの有益な結果をもたらす。第一に、研究プロジェクト全体のデータを統一する必要がある。これは同時に長期的なデータ保存の保証を促し、未来の研究者にデータを利用可能にする。資金提供団体はしばしば、情報が広く普及する形態であるという目標をプロジェクトに求める。外部研究者にデータを利用可能にするという計画は、一般にこの要求を満たすことに役立つ。データを広く利用可能にする一つの方法は、地球規模生物多様性情報機構 (GBIF)などのデータ提供者に加入することである。GBIF は独自に特化した データベース接続形態をもっており、これを導入する必要がある。

デジタル化は、特に、試料の貸し出しやデータ入手の分野で、研究機関のベストプラクティスを助けるために利用できる。これを可能にするために、 スタッフは 研究機関内部で利用可能なシステムを必要とするが、地球規模で利用可能なインターフェースは必要ない。

もちろん、多くの他の利用者が、デジタル化 の取り組みによって保存される追加情報を含む、それぞれが独自の特別要件を備えた、データベースにアクセスできることを目指すことも十分可能である。

データベースを提供したいターゲット・オーディエンスが増えるほど、データベースが複雑になることに留意しよう。上記のクライアントの場合、それぞれのクライアント・グループは少しずつ異なる方法でデータにアクセスできることを望む。おそらく、3 種類の異なるインターフェース (データ入力、内部アクセス、外部アクセス)を構築することが求められる。時間をかけて、具体的な要件について、それぞれのクライアント・グループの代表者数名と意見を交わすとよい。あなたの目標がターゲット・オーディエンスのニーズと一致することを保証するのに役立つだろう。

**サポートする言語数**

情報を提示する必要のある言語が増えるほど、データセットやインターフェースは必然的にかなりの程度、より複雑になる。例えば、インターフェースだけでも複数の言語で提示する必要がある場合、データをそれぞれの言語に翻訳しなければならない。少なくともデータベースシステムは、発音区別符など、(英語から見て)独特の文字を処理できた方がよい。

**データの数**

小規模なコレクションの管理タスクは簡単であるが、コレクションの規模が大きくなるにつれて徐々に困難になり、最終的にすべての試料のデジタル化は、かなり長期的目標でない限り、実行不可能になる。 コレクションの特定の部分を対象にすることが、しばしば優れた戦略になる。これは研究機関の即時的要件によって変化するが、通常は生物分類の科や具体的な地理的区域など、簡単に定義できるグループに焦点があてられる。最も重要な試料(大抵は 生物分類のタイプ)を選び、それを中心に デジタル化 に取り組むのも有効な技術である。

デジタル化の取り組みが小規模プロジェクトに焦点を当てている場合でも、 デジタル化するべき試料の数量を把握することは非常に重要である。 この数量は、データベースに情報を記録するのに要する時間を考慮するための基礎数であり、またプロジェクトを設定する際のほとんどの前提となる側面である。実際に、プロジェクトでの見積は保存された試料の実数の半分ほどである場合がある。これは、多くの問題を引き起こし (特に、利用可能なリソース不足) 、プロジェクトが部分的にしか成功しない、あるいは完全に失敗する怖れもある。これが、プロジェクト開始前に試料を評価する唯一の方法であっても、プロジェクトを開始する前にデジタル化する試料の数量は知っておくべきである。

**データ 品質とは何か**

「数量こそがその品質のすべてである」と言われている。できるだけ多くの記録をデータベースに残したいと考えるのは常に自然な願望である。記録の数を明確にできれば、プロジェクトの成否を測る簡単な判定基準になる。しかし、 試料を単純にリストアップしても大部分の利用者には価値がない。適切な補足データがなければ、データを有効にするために相当量の後続作業を実行する必要があると

思われる。

明らかに、個々の試料は、研究機関の観点からは、初めてのデジタル化から完璧に処理することが効率がよい。しかし、具体的なプロジェクトの資金提供団体が、そのプロジェクトと直接関係のない情報の記録まで賄ってくれるのかという問題がある。 資金提供団体が支払いを拒否した場合、作業を完了するために、不足分にソースを合わせることを検討するしかない。それが実行不可能で、大部分の研究機関の利用可能な資金が限られている場合、最も一般的に要求されるデータと、具体的な プロジェクトに特別に要求された固有のデータを記録するというのが、完璧なデータ記録と、現在のプロジェクトの限られたニーズとの、合理的な妥協案になる。 一般に要求される試料 データをまとめると、その大部分が登録番号またはバーコード、収集者、コレクションデータ、コレクション場所、コレクションの決定、現在の決定である。

**データ保存またはデータ解釈**

試料に記録されたデータは、通常、収集者の現場ノートから得られるので、手書きの記録と同様に多くのエラーがある。データが書かれたままに保存するべきか（データに歴史的視点を与える）、あるいは最新の解釈を付与するために訂正されるべきか（スペルミスの訂正や、試料収集以降の政変を反映して国名を更新するなど）という疑問が自然に生じる。 どこかに記録されている限り、どちらの方法も納得のいく慣行であり、この慣行はデータセット全体に一貫して適用される。

オリジナルデータ特有の問題は、分類学的解釈の領域である。根拠のない名前が決定として頻繁に入力されている (よく見られるのは、種名に間違った命名者を挙げている )。2004 年、Meier & Dikow は、*Euscelidia* （ムシヒキアブの一種）のすべての決定の 62～73% が誤認されていることを発見した。従って決して小さな問題でないことは明らかである。

こうしたデータを記録するとき、実際に 2 つのオプションがある。歴史的観点から、このデータはまったく変更せずにそのまま残す。これでは、分類学的研究にとってデータの有用性が低くなるので、データを訂正する強力な根拠になる。あらゆる決定に訂正作業を施すにはかなり時間がかかるので、可能であれば、国際植物名索引（IPNI）など、名称の情報源となる出版物を参考にして新しい決定を採用することを推奨する。

デジタル化のもう一つの潜在的側面は、実際に試料では入手できない有益なデータの追加である。最もよく見られる例はおそらく、地理情報システム (GIS) を利用して位置データを提供することである。そのためには 試料をジオリファレンスしなければならない (試料の収集場所を見つけ、緯度と経度を割り当てる)。 過去 10 年以内のコレクションでは、 GPS システムが提供した経度と緯度が付記されているのが普通である。 しかしそれ以前のコレクションでこのデータが付記されたものは稀なので、追加する必要がある。これは価値のある取り組みだが、追加調査が必要なのでかなりの時間を要する。 すべての利用可能なコレクションの位置データが記録されれば、こうした付加価値データを残して、後日追加した方が良いだろう。そうすればジオリファレンシング専門家は、その 1 つの地域に集中できるので、潜在的利点になる。

**研究機関の既存の慣行を向上させるのか**

試料をデジタル化する重要な潜在的理由は、研究機関の学芸員の業務を向上させることである。しばしば試料そのものからきわめて少ないデータを要求しても (しばしば名前のみ)、研究機関の学芸員の業務としてはかなりの追加データとなる。学芸員の業務を助けるために追加される標準機能は、様々な試料を識別するために使用される固有のバーコードや受入番号である。これによってデータベースの利用者は、他のやり方では容易に分類できない多様な試料を含むコレクションから、 *Quercus robur* （ヨーロッパナラ）の特定の試料を見つけることができる。

**画像化**

圧倒的に多くのケースで、他の方法で容易に記録できない情報を保存するとき、画像が大いに役立つ。 時には画像化が適切でないこともある。特殊な例として、藻類やコケ植物の画像化の価値については議論の余地がある。試料の特徴を識別するには、高い倍率で拡大しなければならない。もちろん、高倍率の画像や試料ラベルの画像も可能であるが、それはリソースの追加のオーバーヘッドである。画像化にはかなりの関連コストがかかるが、通常それは高品質の画像から生まれる利益のほうが上回る。画像化について詳細に論じるのは本文書の目的ではない。その点については ENSCONET (Haüser et al, 2005) が発表した以前の論文ですでに論じている。

## デジタル化できないものを理解する

多くのデジタル化 プロジェクトは、データベースに非現実的な期待を持ったために目標達成に失敗した。これまで本文書では、デジタル化 プロジェクトが提供できる実例を論じてきたが、達成できないことを明らかにするのも重要である (McLeod and Winans, 1991)。本章はデータベース化プロジェクトができることではなく、不可能な目標は除外されることを確かめることについて論じる。

### データベース化は費用を節約できるオプションではない

ある種の活動はデジタル化より効率よく、安いコストで達成できるが、データへのアクセスが増加すると、研究機関へのクエリーも増加する。情報技術を導入すれば、コンピュータ設備とメンテナンスに比例した費用がかかる (コンピュータと デジタルコレクションそれ自体の両方)。 実際に試料をデータベース化する必要があれば、短期費用がかかる。綿密な計画によって、増大した費用の一部は効率節約によって相殺できるが、コレクションをデジタル化することで拡大した能力に見合った、費用増加が生じるだろう。

### コレクションのデジタル化は新たな情報をもたらさない

試料に情報が提示されていなければ、それを作成するための適切な基準を決める追加作業が発生する。試料に不正確なデータが書かれていれば、データベースに誤って入力され、場合によってはスペルミスまで引き継ぐ可能性が高い。 これらの欠陥は、システムが期待通りに作動することを阻み、プロジェクトの失敗につながる。幸い、試料をデータベース化すれば、データの欠陥を容易に特定でき、予防措置を講じることができる。他の記録と比較することによって、結果として得られたデータの利用法を開発すれば、試料から直接得られたデータに、価値のある追加データを加えられる。

### 現在でも試料は物理的に保管され扱われる必要がある

個々の試料に対するリクエストが、デジタルコンテンツの利用可能性のせいで減少しても、データアクセスの増加は、試料に対するリクエストの増加を招くだろう。 どれ程きめ細かい試料画像であっても、画像では記録できない試料の物理的特性がある。

## いつまでにデータベースを利用可能にしたいのか

大部分のプロジェクトは、すでに利用可能なデータの欠陥を埋めるために開発されたので、この質問に対する直観的な答えが「昨日」であっても驚くにはあたらない。特に、研究機関のデジタル化プロジェクトの担当者の場合、実際に目標が相当量のリソースを必要とする。多くの場合、目標は短期、中期、長期の各段階に細分化される。Chapman (2005a) は プロジェクトの目標を以下のカテゴリーに細分化した。

- **短期**　6～12 か月で完了できる作業
- **中期**　約 18 か月で完了できるデータ入力
- **長期**　18 か月以上続くあらゆるプロジェクト

多くの受託プロジェクトには決められた期限があることを考えると、プロジェクトの期限を研究機関のデジタル化目標に設定することは、しばしば実際的である。通常、小規模プロジェクトを運営している

なら、プロジェクトの期限が必然的に最長目標となる可能性が高い。すべてのプロジェクトにとって、実用的な短期目標を立てることは 、適切な割合で進捗状況を確かめる有益な方法である。これは研究の具体的なサブセットを完成させたり、他の研究機関にデータを提供したりする方法でもある (選択したデータ交換方法の実用性をテストするのに非常に有益である)。

個別のプロジェクトは大抵、期限が明確に決められているが、研究機関は大抵、1 つのプロジェクトに資金提供される期間より長い目標を設定する必要がある。目標と期限の設定は、今でも、新規プロジェクトを研究機関の要件に適合させることに役立つ価値ある活動である。

現在進行中のデータベース化作業は必然的に、将来有益になる可能性があり、それを考慮して構成するのが理想的である。そうしないと、 試料データを再入力する必要が生じ、研究機関に費用負担が発生する。しかし、今後 10～20 年以内にどのデータが重要になるか今の時点で判断し、記録するデータを選ぶのは困難である。 大部分の将来保証システムはすべてを記録しているが、これは最もリソース集約度の高いオプションでもある。後段の「データ 品質」の項で論じるように、一般的に必要とされる領域の大部分 を単純に記録し、未来の要件は追加データに記録するのが実際的である。

**未来の要件**

コレクションをデジタル化すると、長期間データを維持する必要がある。 簡単に言えば、データ が維持されていなければ、その妥当性や実用性が低下し、最終的には役に立たなくなる怖れがある (Wheeler, 2004)。 試料コレクションにおける主要な変化の例として、分類学上の名前を決定する分野の進歩が挙げられる。これを回避するために、他のコレクションと同じ方法で、データを維持した方がよい。データセットの妥当性を維持するために、学芸員の業務の一環として、物理的コレクションを追加するのと同様に、データコレクションも追加した方がよい。ジオリファレンシングデータなどの付加価値情報を追加できれば理想的である。そうすれば、データの有用性を高められる。

研究機関は、種子、栽培、ゲノム情報など、試料に関連する追加情報を含む、他のデータベースも所有しているので、自然な流れとしては、これらのデータセットを統合することである。 これは研究機関の所有するコレクション全体への理解を深めるとともに、研究者は以前の利用可能性より大幅に広い範囲のデータにアクセスできるようになる。

技術は常に進歩しているので、新しい技術の統合が常に課題になる。 現在の課題は、収集現場でモバイル・コンピューティングを使ってコレクションを記録することである。完全に構築されたデータベースに、こうした新しい機能を組み込む試みは、データベース構築時に設備を構成するのと比べて、はるかに困難である。将来利用したい機能を考慮することは、たとえ今すぐ導入しなくても、将来的なシステム拡張時の「産みの苦しみ」を緩和するだろう。

# 現在の限界とリソース

目標を設定したら、今度はプロジェクトに提供できる実際的要素を検討する時である。希望のプロジェクトを実際に実行するための現在利用可能なリソースが不十分な可能性は大いにあるが、リソースの不足を確認することで、状況を改善する方法を決定できる。この行為は、アクションプランに定義されている。多くの要素はプロジェクト計画者の目には一目瞭然であるが、ここでは完璧を期して記述しておく。 これらの限界のいくつかは実際にはプロジェクトや研究機関にとってメリットにもなるが、大部分のプロジェクトは、プロジェクトを開始するために追加 リソースを調達する必要がある。

**スタッフの配置**

プロジェクトの成功を確実にするのを助けるために果たすべきいくつかの異なる役割がある。プロジェクトに適任のスタッフがいれば大いに助かるが、例えそれが単にデータベースの利用法や、研究機関の要件に適合した試料の取り扱い方であっても、職員には具体的トレーニングが必要なことが多

い。もちろんこれはプロジェクトに割り当てられた時間を使って行う。追加トレーニングが必要な程度を評価すれば、目標達成のために必要な時間やリソースのレベルが分かるだろう。

スタッフ及び スタッフ配置の費用については、デジタイザーだけを対象にして検討してはいけない。他のプロジェクトと同様に、デジタイザーを管理する必要があり、これには時間と関連費用がかかる。プロジェクトが求める適切なレベルで作業が完了したことを確認するデータ品質検査を含めるのがグッドプラクティスである。 プロジェクトの規模に応じて、これらの役割のそれぞれに 1 人の職員、あるいは 1 つの役職を組み込む必要がある。 5 人以下のスタッフに 1 人の管理者/データ検査者の役職という割合が適切である。さらに データベース管理者の役割も考慮しなければならない。IT 問題に取り組むスタッフも決めなければならない。例えそれがコンピュータを購入し、データベースをインストールし、すべてが正常に機能することを保証するだけでも!

プロジェクト関係者の標準的役割は以下の通り。

**デジタイザー。** プロジェクトを遂行するために利用できるいくつかの人材プールがある。プロジェクト管理者は、作業の大部分を担うスタッフのプールを利用できる。しかし、デジタル化 プロジェクトにデータを追加できる可能性があれば、機会を逃してはいけない。.

**学芸員が通常業務の一部として担当する。** データベースが主に学芸員の業務をサポートするために利用される場合、これは非常に有益なオプションである。研究機関を出入りする試料の貸付情報を保存するには非常に実用的でもある。しかし学芸員チームは、すでにフルタイムの作業を抱えているので、必然的にデータ収集率は、着実であっても、比較的遅い。

**外部の契約スタッフ/契約企業。** 外部のデジタル化チームにデータを渡すのはリスクを伴う。 契約者の約束にもかかわらず、雇用した研究機関の元に成果が戻って来るまで、作業が正しく遂行されるという保証はない。これは余分な手間がかかるという意味である。デジタル化のためにデータを送付する前には対象コレクションの準備があり、デジタル化の後には詳細なデータ検査が必要である。さらに、外部企業に依頼してデータを訂正する場合、たびたび送料が発生して費用がかさみ、試料に変更を施すために要する時間にも課金される可能性がある。無料サービスは稀であるが、研究機関にプロジェクト・スタッフを雇用するよりしばしば安価であり、その上、プロジェクト管理者とデータ品質検査者以外には研究機関内に事務所スペースが必要ないというおまけつきである。しかし、通信回線を介して利用可能なデジタイザーは、博物学コレクションの取り扱いについてトレーニングを受けていることは稀であり、研究機関の内部スタッフと同等に、複雑なデータを解釈することはできない。 データが最近のもので、印刷されたファイルやラベルから容易に引き出すことができる場合は、これは前進するために有効な手段である。

**ボランティア・スタッフ。** 多くの研究機関は、進行中の作業に深くかかわることを希望するボランティアを抱えている。彼らはデジタル化作業の担当者として理想的な候補者と思われるが、これは両刃の剣である。ボランティアは大抵、深くかかわることに熱心だが、自らの自由意思でそうしているのであって、退屈すれば、恐らく何の説明もなく関与を停止する。デジタル化は非常に重要ではあるが、残念ながら研究機関で実行される刺激的な作業と同じとは言えない。ボランティアは、自身のスケジュールに沿って働く意思があり、週日完全就労で試料のデジタル化に取り組む意思はない。必然的に、ボランティアはフルタイム勤務のプロジェクト・スタッフほど作業を遂行できない。彼らは作業用の専用の事務所スペースを必要とする。特に、あまり時間的制約のない小規模プロジェクトなら、ボランティアはコレクションのデジタル化に有効な手段である。 プロジェクトのスタッフ離職率が上昇傾向にあることに留意しよう。これは新しいボランティアを採用し、トレーニングを試みる着実な進捗を阻むものである。

**客員研究員。** 研究機関によって客員研究員が稀なことも普及していることもある。彼らは、自身の

ノートを取りながら、コレクションをいくらでも調査したいと望む。いくつか問題がある。客員研究員は、研究済みのコレクションを調査したいと思わない。彼らは、研究機関が望むデータ品質で入力したいと思わない。ほぼ確実に、研究機関が望む形ではなく、自身のプロジェクトに望ましい形を求める。 これらの傾向から、研究者は研究機関の広範なプロジェクトに向いており、小規模プロジェクトでは非常に限られた有用性しかない。

**プロジェクト・スタッフ。** フルタイムの有給プロジェクト・スタッフを活用することは、通常、特別にトレーニングを受けたスタッフが特異的にプロジェクトに集中する利点があり、一貫性に優れた品質でデータを入力するための、全体として最善の方法である。これは通常、最も費用がかかる リソース集中的オプションであり、トレーニングを受けたスタッフ（通常、博物学の経歴を持つ）に高い報酬を支払わなければならない。トレーニングを受けたスタッフは、自身でデータを解釈できる強みがあり、未熟練スタッフより自立して操作できるので、必要な管理水準が低くなる。

**学生。** 学術機関では、学生がデータ入力に携わる。これは特に、学生の勤労に奨励金が支給される体験学習プログラムの場合、相対的に安価なソリューションである。学生はデジタル化を、博物館界と交流し、学習し、関与することのできる、初級職務と見なしている。学生を利用することは、学術機関の視点からプロジェクトの価値を高め、場合によっては、プロジェクトの作業に取り組む教員に解放的な時間を与える。場合によっては、レベルの高い学生はプロジェクトの先進的な側面を担当できる。例えば、生物学大学院生は、デジタル化する前の品質管理、試料の分類、検査を手伝うことができる。美術科の学生は画像化プロセスの助手として役立つ。 学生の離職は、退屈な作業であることと、学生生活につきものの潜在的娯楽が原因である。

**データ所有者。**研究者、他の研究機関、市販のデータセットは、例えそれが検索リストを追加する基準値の提供だけであっても、既製のデータを提供する際に大きな助けになる。データを使用可能にする方法を制限する、知的財産権(IPR) が問題にならないように、何らかの配慮は必要である。

**データ専門家。**学芸員や様々な分野の専門家は、常に相談相手であることが求められる。プロジェクト期間中に生じる疑問に答える人間が求められるのはほとんど必然である。障害を乗り越える手助けをしてくれる、プロジェクト参加者以外の人物を配置しておく。彼らからプロジェクトに時間を割く同意が得られれば、大きな財産になる。

**技術スタッフ。** 技術スタッフは、コンピュータを設定しに来た IT サポート技師から、使用するデータベース/スプレッドシートを設計した技師にまで及ぶ。彼ら全員が、プロジェクトの重要性について何らかの知識を必要とし、恐らく使用するシステムのメンテナンスを担当するだろう。Microsoft Access より大きなデータベースシステムを使用している場合（恐らく Microsoft Access それ自体でも）、データベースのシステム管理者（シスアド）がきわめて重要な人物になる。彼は、実際のデータコンテンツに責任はないものの、データベースが起動し、作動していることを確認する技師である。

**プロジェクト管理。** プロジェクト管理に関連する 2 つの役割がある。 1 つは、プロジェクトの日常的な稼働に責任のあるプロジェクト管理者、2 つ目は、データ品質の検査とデータ維持に責任のあるデータ管理者である。データ管理者はしばしば、プロジェクト終了後もデータに責任を負い続ける人物である。プロジェクト擁護者を置くことも有益である。 研究機関レベルでプロジェクトをサポートする役割を担う組織内の幹部である。.

## データ入力の手順

一度にデータベース化できるデータ数も検討しなければならない。デジタイザーが一度に操作するデータの合計数はいくつかの影響を及ぼす。特に、プロジェクトに必要な物理的スペースの量や、コンピュータの数など、必要な実際的リソースの数に影響する。データベースに求められる複雑さのレベルや、潜在的にプロジェクトに必要な I.T. インフラにも影響を及ぼす。

以下のようなオプションがある。

**1 人のスタッフが 1 つのデータベースを担当する。**最も簡単なオプションだが、同時にプロジェクトの最終期限は最も遅い。必要なのは、1 人のスタッフ、1 つのデスク、1 台のコンピュータ、デジタル化する試料を並べる十分なスペースだけである。要求されるデータのセキュリティレベルにもよるが、IT インフラを追加する必要はない。しかし、個人のコンピュータが故障すると、それまでに蓄積したすべてのデータセットが失われ、プロジェクトの失敗を招く恐れがある。何らかの形でデータをバックアップすることが不可欠である。

**数人のスタッフが個別のデータベースを使用する。**最初のオプションと同じように、1 人のスタッフのリソースを増加させていくシンプルな形である。しかし、1台以上のコンピュータが故障し、交換する必要がある可能性があるので、データを保護する必要が増大する。ほぼ確実に、プロジェクト終了時に個別のデータベースを1つのデータセットの形に統合するための追加措置が必要である。つまり、数人のスタッフを使うことでコレクションのデジタル化に要する時間を短縮するのである。担当するスタッフの人数という要因（1人のスタッフが作業を担当する場合と比べて）によって時間が短縮される。その所要時間に、プロジェクトの設定に要する時間と、プロジェクト終了時のデータ統合に要する時間を加えるのである。大量のデータが失われるリスクは、最終的なデータセットをいくつかの部分に分割することで若干軽減される。しかし、データを再入力する必要がないように、データの個別の部分をバックアップすることは依然として推奨される。これによって、デジタル化に携わるスタッフ数に比例してプロジェクトのオーバーヘッドが増大する。

**数人のスタッフが同じデータベースを分担する。** このオプションは、数人のスタッフが同じプロジェクトを担当するメリットを組み合わせたものである。しかも、プロジェクト終了時にデータを統合するオーバーヘッドもない。バックアップするべきデータベースは1つしかないので、データ保護もはるかに容易である。これを実現するためには、ある種の IT ネットワークを配置しなければならない。また、プロジェクトのリソースに負荷を加える可能性がある。

**コレクションの大きさ**

コレクション全体の大きさを知ることは、すべてをデジタル化するのに要する時間の長さを推し測るのに役立つ。コレクションの実際の大きさは過小評価されがちなので、正確な評価は重要である。また、新しい試料の入手率の現実的評価を含めることも重要である。これも大きさの評価に含めるべきだからである。

**データへのアクセスは何らかの制約を受けているか**

各国は、固有の動植物を開発することで得られる、潜在的財産に対する認識を深めるにつれて、利用できる情報の拡散や利用を保護する姿勢を強めている。その結果、生物多様性条約（CBD）などの協定が生まれた。これらの協定は、試料が収集される前に許可を求めるよう各国に要求している。その基本合意書（MOA）は、データを閲覧できる人物を含む、様々な制約を課している。制約は 1 つの研究機関を対象とするが、研究機関の 1 部門だけに課されることもある。データを共有するときは、知的財産権 (IPR) に配慮しなければならず、時にはデータを公開する前に、研究機関に法的契約への署名を求めることもある。二国間の法的契約に合意するのは困難であり、大抵は、他国で制定された法律に基づく法的拘束力のある契約への署名を弁護士が拒否し、多くの知的財産権（IPR）契約は、覚書（MOU）として成立する。契約に法的強制力があるかどうかに関わらず、データ公開時にはこれらの文書に配慮しなければならない。これは、公開したウェブサイトから許可なくデータを採り入れてはならない理由でもある。発信元のウェブサイトはデータを出版する許可を得ているが、あなたは希望通りにデータを使用する許可を得ていないので、法律に違反することになる。少なくとも、大量のデータを使用する方法について、出版元の研究機関に説明しなければ、出版元の信頼を裏切ることになる。

独自のデータの一部を一方的に隠したいと望む場合もある。例えば、IUCN（国際自然保護連合）のレッドリストに登録されているような希少種が、商業利用の対象になることがある。 地理的位置（特に、GPS リファレンスによる可能な限り正確な位置情報）などのデータを公開すると、野生生物の集団に計り知れない被害をもたらす恐れがある。そのため、固有の動植物集団を保護するために、このデータを公開しない、あるいは、非常に広い範囲の地理情報だけを公開するという選択ができる。この最後の点には根拠がある。その場所で得られた試料や他のコレクションからデジタル化した複製など、他の形式ですでにデータが利用可能な場合、固有種の破壊や違法な利用を確実に困難にする方便として優れた選択である。現在は国際基準として合意されていないが、保護必要データに対する標準的なアプローチの構築に向けて若干の前進は見られる (Chapman and Grafton (2008)に基づく本文書第 6 章参照)。プロジェクト導入の際には、同僚と相談して、良心の導きに従って行動するべきである。

**研究機関は既存のシステムの使用を要求しているか**

研究機関がすでに中央データベースを構築している場合、他者がすでに提供したデータベースにデータを追加するよう研究機関が要求するのは不合理ではない。すぐに使用できる有益なデータや、既存のデータ入力システムを保有できるメリットがある。しかし、データの解釈法や記録法は制約される。この場合、プロジェクトのデータは個別のシステムを使って記録した方が良いが、データは、プロジェクト終了時に、中央データベースに容易にインポートできるフォーマットで、研究機関に提供されることが重要である。これは、研究機関が求める標準的データ要件の基準がデータベースに引き継がれることを意味する。ほとんどの場合、所定のデータベースは実際にプロジェクトの遂行を容易にするので、研究機関の基準に従うのが通常、最善策になる。

**レガシー データはあるのか (電子または紙)**

既存の研究データは、前項で述べたように、中央データベースに組み込まれないまま、すでに研究機関に存在する。このデータは、許容できる水準までデータ品質を高めるための時間、あるいは選択したシステムに適応させるための時間が必要であるが、きわめて短時間に大量のデータの構築を可能にする。既存の紙のシステムは、しばしば容易にデジタル化でき（特に、タイプスクリプトの場合）、ほとんどデータ解釈する必要がなく、外部デジタル化の強力な候補である。データの正確さと、記録されたデータが作成されてから試料に記された注釈の両方を検査するために、レガシーデータに関して記録されたデータと物理的試料を照合することに時間をかけることは価値がある。

**物理的要件**

デジタル化は真空空間では実行できない。 プロジェクトに携わるスタッフやボランティが作業する場所が必要である。検討案件は以下の通り。

**デジタル化を実行する場所。** プロジェクトの作業場所として 3 つの基本的なオプションがある。

- **収集場所でデジタル化する。** 試料のすぐそばで作業できるメリットはあるが、通常、作業スペースが狭いので、一度に作業できるデジタイザーは数人である。コレクションの人気が高いと、デジタイザーはたびたび作業を中断され、作業量の低下を招く。

- **デジタル化専用の場所を設置する。** 上述の課題は除去されるが、研究機関内に事務所スペースが必要である。しかし、デジタイザーグループは作業に専心でき、中断されることもない。デジタル化する場所にコレクション試料を移動させることで生じる問題もあるが、慎重な計画によって大抵は解決できる。

- **全く別の場所でデジタル化する。**デジタル化する場所が別の建物にある場合、さらに困難である。 試料の移動には通常、特別慎重な取り扱いが必要だからである。この場合、コレクションの試料を画像化し、画像を使ってデジタル化するのが、移動によって生じる問題を解決する有効な手段になる。

これらすべてのオプションは、選択した場所の要件に適切な配慮を払う限り、過去のプロジェクト導入に利用されて成功を収めている。

**既存の IT インフラ。**技術要件には適切な配慮を払うべきである。日常業務にはノート型パソコンより大型コンピュータ (PC または Macintosh) が望ましい。すでに配置されている場合は、データベースの選択に影響するだろう。しかし、収集旅行に出かける時は、現場で観察を記録できるノート型パソコンが実用的である。試料を画像化する予定なら、カメラかスキャナーが必要である。インターネットや、デスク周辺を越えた別の場所と接続したい場合は、例え電話やモデムだけとしても、ある種のネットワークが必要である。現在所有しているものや、必要になるものを検討し、記録しておこう。

**プロジェクトの最終期限はすでに決まっているのか**
すでにプロジェクトに着手しているなら、プロジェクトに影響する最終期限に向けて作業を進めているだろう。恐らくこの期限を十分意識しているだろうが、特に、プロジェクトの遂行に他のスタッフの協力が必要な場合、最終期限を記録することは他のスタッフにその重要性を明確に示すのに役立つ。

**研究機関の外で作業する予定があるか**
他の研究機関に出向いたり、現地でデータ収集するなど、正規の勤務地の外で作業する場合は、プロジェクトに特別要件が課される。試料データを紙に記録し、後日転写することはできるが、ノート型パソコンや、場合によっては個人用デジタル補助装置（PDA）など、ある種の携帯端末を利用すれば、はるかに実用的である。試料を直接データベースにデジタル化したい場合は、以下の前提条件の 1 つ（以上）をクリアする必要がある。

携帯が可能

データ入力を処理でき、メインデータベースにインポートできるモジュールがある

ウェブベースのデータ入力システムがある

最後のオプションは、フィールドワークの現場では滅多に実用化されていない。どこにいても適切なインターネット接続ができることを前提にしているからである。 強力な衛星通信網の発達した今日の世界でさえ、居住地の外で作業する際に（時には居住地内でさえ）、長時間の連続接続は滅多に実用化されていない。

**資金調達**
資金提供源をすでに持っているか。また現実的にどの団体から資金提供を受けるのか。

**変化する意欲はあるのか**
研究機関内にデジタル化作業のサポートがない場合、プロジェクトの導入ははるかに困難である。プロジェクトの擁護者として行動するキーパーソンを置くことが非常に有益である。また研究機関内に支持者を集めることも役に立つ。このようなサポートがなければ、大規模なデジタル化プロジェクトを成功させることは困難である。

# 投資対効果検討書の試案を作成する

投資対効果検討書を作成すれば、はっきりした形で目標や制約を明確に設定することができる。検討書は、プロジェクトが開始する価値のある理由を他者に提示するものである。デジタル化プロジェクトを開始運営する有益な段階の 1 つは、研究機関の他のスタッフにプロジェクトを承認させ、関与させることである。これはしばしば、データ品質を向上させることが可能な専門家を対象とし、彼らにプロ

ジェクトの要素を検討する時間を提供する意欲があるということを意味する。学芸員チームがプロジェクトの重要性を認めた場合、適切な試料を入手するタスクは、はるかに容易になる。プロジェクトの完了に向けてリソースを提供する意欲のあるスタッフがいれば、プロジェクトが成功裏に完了する可能性が高まり、資金提供を引き寄せる可能性も高くなる。以下のような、実際的な質問に答えるために、目標と制約の単純な声明以上のことをするべきである。

**このプロジェクトを実行して何が得られるのか**

研究機関や世界中のその他すべての地域にとって、プロジェクトが実行する価値のある理由を説明すること。

**プロジェクトは実現可能か**

プロジェクトが実用的で達成可能であると信じることが重要である。すべての議論を尽くしたと認識することが、それを大いに容易にする。多くの大規模プロジェクトは、この問いに「いいえ」と答えざるを得ない。この場合、ここで概説した目標は、投資対効果検討書全体の目標に適合した、小規模で短期の、実用的なプロジェクトに細分化する必要のある、創造的目標にするとよい。この場合、プロジェクトグループ全体を対象にした投資対効果検討書を作成し、それぞれの小規模プロジェクトにも詳細な投資対効果検討書を作成することが推奨されている。

**目標は限界を超えているか**

プロジェクトが実現可能な場合、プロジェクト達成に必要な追加リソースを評価する必要があるだろう。数少ない非常に幸運なプロジェクトには、プロジェクトを達成するための十分なスタッフ、時間、機器が前もって揃っている。しかし、大部分のプロジェクトはどこかから追加リソースを調達しなければならない。 ここで、必要になると思われる追加リソースを書き出しておこう。そのリソースの調達方法の正確な詳細もアクションプランに含めるべきである。

**限界を克服する**

問題点を検討し、プロジェクトを開始するのが妥当だと判断したら、いよいよ行動方針の概要をまとめることができる。これをアクションプラン (あるいは、プロジェクトをいくつかの段階に細分化することを選択した場合は、複数の詳細なアクションプラン)に詳細に記述する。 以下の質問に答えを見つけることが、既存のリソースを補完するのに役立つだろう。

**業務慣行の変更によってプロジェクトに取り組む時間を確保できるのか。** あらゆる問題を解決する普遍的な答えとなる業務慣行の変更を期待してはならない。スタッフが現在のやり方で作業しているのには理由がある。また、業務慣行の単純な変更により、デジタル化の目的を修正するほど大量のリソースを確保することは滅多にない。 しかし、プロジェクトの特定の部分について専門家の援助を仰ぐことができる十分な時間を稼ぐことは可能である。業務慣行の変更は、デジタル化プロジェクトに必要な物理的スペースを確保するのに役立つ。これは、組織の他のスタッフに広く受容される大きな恩恵をもたらす部分である。というのも、スタッフは、業務時間に対する新しい要求に直面して、円滑な慣行を変更せざるを得ないことに不信感を抱くのが普通だからである。

**近隣の研究機関の援助は期待できるのか。**近隣の研究機関と共同でデジタル化に取り組む協定を結ぶことは、十分可能である。より多くの有益なリソースを活用でき (Snow, 2005)、同じような規模の研究機関と導入費用を分担できる。

**誰がプロジェクトに資金提供するのか。**国内・国際両レベルの多くの資金提供団体が存在する。例えば、GBIF、アンドリュー・ウイリアム・メロン財団、ゴードン＆ベティ・ムーア財団などは、毎年多くのプロジェクトに資金提供している。営利団体から資金を調達することも可能である。すべての資金提供団体が実際に提供できるよりはるかに多くの申し込みを毎年受けているので、資金調達の機会については現実的に考えることが重要である。資金提供団体の要件を慎重に調査する

ことがきわめて望ましい。それに基づいて申込書を作成するとよい。

**プロジェクトを個別のユニットに細分化するべきか。**プロジェクト全体が実現可能と信じるなら、この質問に「いいえ」と答えても驚くにあたらない。しかし資金調達は、大規模プロジェクトより小規模プロジェクトの方がはるかに容易なので、この質問はもっと慎重に検討した方が賢明である。 この場合もやはり、組織の他のスタッフに投資対効果検討書を公表することが、計画案の実行可能性を迅速に評価するのに役立つだろう。

**概念実証は可能か。** デジタル化プロジェクトの試験を実施することは、非常に有益なエクササイズになる。試験によって、プロジェクトの計画に役立つ多くの実際的な情報が得られる。 常に概念実証を実行できるとは限らないが、大いに推奨される。

**付録 A** のチェックリストは、検討した議論をまとめ、投資対効果検討書の作成を容易にするのに役立つ。プロジェクトの基礎的前提を疑問視することは常に有益なので、あえて時間を取って投資対効果検討書をじっくり検討し、他者の建設的なフィードバックを受けるとよい。当初からプロジェクトを支援する意思のない同僚からのフィードバックが最も厳しいが、いろいろな意味でこれが最も有益である。結局、すでに実行する価値があると認めている人を説得するのは容易であり、疑問を持つ人を説得する方がはるかに有益で、満足感がある。他者の意見を十分に考慮して、投資対効果検討書を改訂すれば、実行可能な投資対効果検討書であると十分確信できるだろう。

# データベースソリューションを選ぶ

投資対効果検討書が完成し、同僚の合意が得られたら、目標を達成する方法を詳細に検討するべき時である。投資対効果検討書の作成にかなりの時間を費やしたとしても、まだ急いで実施段階に進まないことが重要である。プロジェクトのこの段階で貧弱な決定をすると、この先数年間に影響する。決定事項を試験する、少しの時間と忍耐と意欲があれば、最終的にプロジェクトが始まったとき、円滑に運営できるという報いがある。本章と第 6 章を実行した後で、もう一度ここに戻って計画案を練り直したいと考えた場合、長期的に見るとプロジェクトに良い結果をもたらすので心配ない。

適切なデータベースを選ぶことは、様々な要因を考慮しなければならない難しい決断である。これは第 6 章「特定のデータベースソリューションを決定する」で深く論じる。利用者が使用するデータベースについて意思決定するときは、この第 6 章を読み直すことを推奨する。

# アクションプランを立てる

投資対効果検討書を作成し、使用するデータベースを選択したら、プロジェクトの実施方法について多くのアイデアがあるはずである。プロジェクトを実施したとき、これらのアイデアがうまく機能することを証明した方がよい。投資対効果検討書で提起された問題を考慮して、プロジェクトのアイデアを分かりやすく書き始めよう。そして、以下の諸点についてアイデアをチェックし、提起された問題を考慮して、必要に応じて計画を変更する。

ここで論じる問題の多くは相互作用する。そこで、少なくとも 2 回リストを詳細に検討するのが実際的な予防措置である。そして、前に提起された問題に答えるリストを詳細に検討して、変更された部分を確かめるのである。このタスクを首尾よく完了すれば、可能な限り強固な、達成可能なプロジェクトであるという確信が持てる。

**選択したソリューションは、目標、制約、リソースに適合するのか**

適合する場合、プロジェクトを完了する絶好の位置にいる。しかしほとんどの場合、リソースとソリューションの間に不足があり、それを埋めるために追加リソースが必要になる。 ここで示す質問の多くは、

その不足を詳細に述べ、ソリューションの基準コストを提供するのに役立つ。これが入手できれば、資金提供者を探し、利用可能なリソースに適合するようソリューションを修正することもできる。

**ソリューションは未来の要件に対処できるのか、対処できない場合はどうするのか**

もちろん、変更ニーズに適応できる余地をソリューションに残しておくのが望ましい。あらゆる可能なソリューション向けの設計を試みることは、短期的には非常にコストが高く（例え長期では有益だとしても）、簡単に法外な価格になる。

**必要な スタッフ数**

必要なスタッフ数は、コレクションのデータベース化をどの程度急ぐかに若干左右され、現在の物理的スペースの総量によって制限される。最終期限と、データベース化したいコレクションの規模を考慮して、働かせるスタッフ数を判断する必要がある。必要なスタッフ数は、少なめに見積もらず、多めに見積もる方がよい。というのも、多めに見積もればプロジェクトがより速く進むが、少なめに見積もればプロジェクトが失敗する可能性がある。ここでは慎重にバランスを取らなければならない。プロジェクトに過剰な費用がかかると、資金調達が困難になり、計画が無駄になる。必要なスタッフ数を判断する最も効果的な方法は、概念実証を行うことである。計画したワークフローを使って、かなりの数の本物の試料を実際にスタッフがデジタル化するのである。 他の研究機関やデータベース販売業者が宣伝する推定速度を使うときは十分に注意しよう。非常に多くの場合、研究機関や企業は実際よりも良く見せようとするからである。 非常に大まかな経験則から言って、解像度の高い画像を使って、きわめて詳細にデータベース化する場合、1 人のデジタイザーが 1 週間に 100 件の試料を処理する速度だと推測できる。画像化しない場合は、1 週間に 200 件の試料、高水準のデジタル化ではなく画像化もしない場合は 1 週間に 300 件の試料を処理できる速度である。これらの数字は控え目に見積もっているが、プロジェクトの目標を達成し、あるいは上回れば、感謝される。その時、これらの数字が目標達成に役立つだろう。スタッフの休日や病欠を考慮に入れることも忘れてはならない。 これらはプロジェクトの存続期間中にデジタル化できる試料の全体量に影響するからである。

**スタッフのトレーニング法**

システムの使い方をスタッフに教える作業は時間とリソースがかかる。スタッフは、ワークフローの実施方法、コンピュータやデータベースソフトの使い方、さらに最も重要なことは、一貫して、効率良く、正確に、試料をデジタル情報に変換する方法を学ぶ必要がある。データをシステムに取り込むだけでは十分ではない。「優れた」データでなければならない。これは、作業に適した人材が選ばれ、作業方法のトレーニングをきちんと受けたときにしか起こらない。

**構築または導入に要する時間**

すべてのプロジェクトに、ある程度のリードタイムがかかる。しばしば 6 か月、あるいはもっと分かりやすく資金提供者が見つかるまでと言ってもよい。しかし、コンピュータの購入や設置、作業場所の確保、スタッフの補充など、すべて合わせるとかなりの時間がかかる。リードタイムを概算し、プロジェクトの実施段階の期間にそれを加算する。専門職員を使う大規模プロジェクトの場合、スタッフの雇用に最短でも 2 カ月、しばしばそれ以上かかる。この時間は以下のように細分化できる。広告制作と掲載に 1 週間、応募の到着、人選、面接の設定に 2 週間。面接に 1 週間。最後に、合格者が試用期間としてトレーニングを受けるのに 1 か月。 これらのタスクを実行する別のスタッフを雇用する必要があれば、さらにもう 1 か月かかると見込んだ方がよい。そうでない限り、大抵はこの試用期間中に、物理的要件を所定の場所に整えることができる。 作業の遅延を考慮し、未処理のタスク（例えば、プロジェクト報告書の作成）を完了するために、プロジェクト終了時に少し時間の余裕を持たせるのも賢明である。データをウェブサイトで公開する予定があり、デジタル化の取り組みのあとでウェブサイト開発を行えばはるかに円滑にできる。これをしない場合、ウェブサイトデータに変換するデジタル化に予想外の展開が生じ、プロジェクトが存続期間を越えて遅延する恐れがある。ウェブサイト設計は研究機関の IT インフラに大きく依存して変化するが、ゼロからウェブサイトを設計する場合、大量のデータを 1 人のプログラマーが処理すると通常は約 4 カ月かかる。

**購入するべきもの**

新しいスタッフを補充する場合、恐らく、新しいデスク、椅子、コンピュータ、電話、それにオフィス環境に必要なあらゆる備品が必要になる。データセットを保存するサーバーや、外部とのリンクを購入する必要がある。画像化するときの手段はどうするのか。この場合もやはり、特定の試料コレクションにとってのベストプラクティスに従うことである。その多くを研究機関が準備して使用できることが望ましいが、プロジェクトを始める前からそれを期待してはいけない。

**費用**

ご承知の通り、デジタル化プロジェクトの運営費用は安くない。 隠れた費用を含めて研究機関によく確かめよう。海外で作業する場合の旅費や生活費などを予算に計上するのを忘れてはならない。

**ワークフロー**

言い換えると、デジタル化を実施するとき、スタッフにとって最も効率の良い方法である。以下の要因が ワークフロー計画に影響する。

**デジタイザーの人数。** 人数が多ければ、特定の職務を専門的に担当させることができる (画像化、データベース化、ジオリファレンシング、品質検査などが考えられる例である)。しかし、専門化した担当部署がデジタイザーの選択した職務でなければ、スタッフは退屈し、稼働率が下がるので注意しよう。

**試料の収集と返還。**この業務は、通常一括処理で行われる。従って、1 日分または 1 週間分の試料は、試料の大きさに応じて、瞬時に収集できる。収集作業の場合、必要に応じて、それぞれの試料を収集することもできる。デジタル化するために移動させた試料を、学芸員スタッフが必要になり、あなたが持っており、学芸員スタッフの意向に同意する場合、学芸員スタッフが見つける方法を検討しよう。昆虫や粗雑な取り扱いによって試料が損傷するのを防ぐ前処理が必要かどうかを検討し、それに要する時間を計画案に盛り込むこと。.

**試料の取り扱い。**適切な試料取り扱い手順を文書化するのも得策である。 デジタイザーが、特定のコレクションを取り扱うトレーニングを受けた学芸専門家の経歴を持っている可能性は低いからである。

**試料がコレクションから無くなる期間は。**他のプロジェクトから試料を要求された場合、試料を利用できない期間を最短にとどめることが奨励される。

**場所。** スタッフが主要コレクションとは別の場所で作業する場合、デジタイザーの元に試料を移動させる困難さと所要時間を考慮しなければならない。この場合、画像からデジタル画像とデータベースを移動させた方がよい。

**データ 品質。** 品質は、記録しようとするデータに関して考慮するべきである。例えば、公認の規格に照らして検査したかどうか。手描きの記録は判読が困難で、デジタル化プロセスにかなりの時間を要する。Chapman (2005b) は、後日エラーを訂正するより、正確なデータを入手する方がはるかに安くつくと指摘している。

**元データに価値を付加する。** デジタル化されたデータに価値を付加する方法はいくつかある。例えば、出版された規格と比較する。一般に認められたリストと照合してデータを解釈する (手書きの収集者は、ほとんどの場合この方法で解釈しなければならない。彼らの署名はデジタイザーを混乱させるためにデザインされているように見える)。緯度や経度などの有益なデータを付け加えることもできる。これを実行する時間を考慮し、特定のプロセスを記録しておくこと。

**画像化。** 試料を画像化する場合、その作業をどのようにワークフローに組み込むつもりなのか。試料を記録する前か後か。専門家がこの作業を実行するのか、あるいはタスクを分担するのか。

**データの順序。**試料ラベルに記されたデータの順序に従ってデータベースに入力するのは効率的だと思われる。しかし、データベースに表示されるデータの順番の領域は、ラベル上の順番と一致しない。また、ラベルが完全に一貫したフォーマットであることは滅多にない(もし完全に一貫しているなら、幸運である!)。経験豊かなデジタイザーは、独自の望ましい作業方法を開発するが、データベースに試料データを入力する最も効率の良い方法について、分かりやすい指導を与えるトレーニングを受けたときである。

**データ検査。** データ品質を検査する時間を必ず考慮することは、計画した基準に一致する。無作為のデータサンプルの単純な検査や、必要なら詳細な検査によって、すぐに発見できる単純なデータ入力ミスでも、他の高品質な作業を簡単に台無しにする。そのための 1 つの方法は、データを表にして閲覧することである。データは分類され、類似した入力は同じグループに集められ、さらに容易にミスを発見できる。Redman (1996)は、5% 以下のエラー率が期待でき、これらのミスははるかに簡単に訂正できる。 単に元データに返還すればいいだけのことである、と指摘している (Chapman, 2005b)。

**手順を重複できるのか。**デジタル化のいくつかの段階を同時に進めることは可能である。個別の試料では難しいが、いくつかの試料を同時に処理することは可能である。単純な例として、1 つの試料をスキャンしながら、別の試料をデータベース化し、同じデジタイザーがすべてを制御する。

**スタッフの欠席はワークフローにどのような影響があるか。** スタッフ の主要メンバーが休日または病欠の場合、デジタル化手順のその他の部分も停止することになるだろうか。小規模チームではこれは避けられない。しかし大抵は、緊急配置によって問題を回避することが可能である。

**計画に障害はあるか。**綿密な計画によって円滑なプロセスが可能になる。しかし、プロセスの各段階に割り当てる正確な時間は慎重に計算しよう。仮説例を挙げる。1 人のデジタイザーが 1 つの試料をデータベース化するのに 10 分かかる。それをジオリファレンスするのに 3 分かかる。この時、1 人のデジタイザーから 1 人のジオリファレンサーに引き渡されるなら、ブロセスに一切遅延はない。つまり、ジオリファレンサーはフル稼働である。もしもジオリファレンシングを実施する各スタッフに 4 人のデジタイザーがいた場合、障害があると言える。ジオリファレンサーは、次の 4 束が整うまでに、3 束の記録しか処理できない。その結果、30 分ごとに 2 束のジオリファレンシング未処理分が残る。この未処理分の処理方法を事前に検討することによって、プロジェクト運営中の時間が節約できる。

以上のように、デジタル化の実際のプロセスを実施する際、記録するべき多くの事柄がある。プロセスを実施したら、デジタル化スタッフのためのユーザーマニュアルとして、それを詳細に記録することを推奨する。それはスタッフトレーニングをかなり容易にするだろう。

## プロジェクトが終了したとき何が起きるのか

デジタル化が終了してもデータは存在する。 故に、データに起きることの概要を計画に含めるべきである。データベースは別のプロジェクトでも利用できる。場合によっては、既存のデータセットにコンテンツを付け加えることもできる。データの妥当性を確保することも重要である。できればプロジェクト終了後も妥当性が保たれることが望ましい。データの更新に特に責任のあるスタッフを配置することでこれが実現できる。あるいは研究機関でこのデータを広く利用可能にしておけば、学芸員スタッフの日常業務の中で妥当性が維持できる。 あらゆる新しいデータベースの妥当性の維持に、学芸員の就労時間の最大 0.3 正規職員 (FTE) がかかる (Snow, 2005)。

**ソリューションは適切な水準の データ 品質を提供できるのか**

受け入れるデータ品質の水準を決定しよう。出版された基準の利用は、データ品質を大幅に改善し、データ入力を容易にできる。基準は、多くの分野でシンプルなドロップダウンリストとして利用できる。しかし、特に古い試料には注意が必要である。試料に使われている用語は、もはや現在の命名法と一致していない。国名はこの問題の顕著な例である。この場合、元の国名を記録する能力は非常に有益である。しかし、必然的にデータベースの複雑さは増大する。多くの基準が存在し、時には複数の基準が同じテーマを対象にしている。しかし、結局は 1 つの基準が別の基準を含んでいる限り、どれを選択しようが全く問題ない。

**人的要因の計画**

データベース化は、コンピュータシステムだけではできない。人的要因も無視してはならない。デジタル化にはトレーニングが必要である。前述したプロセスだけでなく、実際のデータ入力システムのトレーニングもある。複雑なデータの解釈を学ぶには、時間と適切なトレーニングが必要である。お粗末なトレーニングは、かなりの割合のデータエラーの原因になりかねない (Chapman, 2005b)。デジタル化の経験を積むにはやはり時間がかかる。その間スタッフは最適な効率で作業していない。そこでプロジェクト期間中にデジタル化される現実的な試料数を計算するとき、これを考慮すること。

**コレクションのデータベース化に要する時間は**

データ入力には 2 通りの基本技術がある。詳細なデータ入力とは、正確さと検索の容易さを最大限に高めるために、リスト、データ検査、適切なデータ構造化を最大限に活用し、データを慎重に入力することを意味する。 これは最善のデータ品質を実現するが、多くの時間を要する。高速データ入力とは、迅速かつ容易にデータを入力する能力を意味する。しかしこれは、データ検査の削減や、しばしばそれほど高度に構築されたデータではないことを意味する。これはデータ入力エラーの増加を招き、従ってデータ品質を低下させる。後日このデータを訂正するには通常、リソースの深い関与が必要である。前述の 2 通りの極端な選択肢から選ぶ必要はないが、受け入れ可能な両者のバランスを取るのは容易ではない。また、個別のプロジェクトに受け入れ可能な方法が、研究機関のシステムに適切とは限らない。 試料の品質は大幅に変化するので、試料ごとにデータベースに要する平均時間を設定するのは難しい。これは、記録されるデータの品質水準や、受け入れ可能な正確さのレベルによっても変化する。この場合もやはり、デジタル化率を現実的に評価する実際的な方法は検査しかない。品質保証 (QA) もデータベース操作に必要不可欠の重要な要素である。早い段階でエラーを訂正でき、恐らく組織の長期的費用を削減できる。

**取り組みに優先順位をつける**

デジタル化の取り組みにとって考え得る最良の結果は、間違いなく、コレクション全体の利用可能なすべてのデータをデジタル化することである。しかし、これを完了するにはかなりの時間がかかり、プロジェクトに大量のリソースを集中的に投入する必要がある。 手持ちのリソースを最大限効果的に使うために、投資対効果検討書で論じたように、取り組みに優先順位をつける必要がある。簡単に要約すると、特に重要な試料や特定の科や種を対象にして、デジタル化する試料の総数を削減するのである。主要なデータ領域に焦点を当てて（通常、コレクションの基礎情報や命名情報）、保存するデータの量を削減することもできる。もちろん、目標を達成し、あるいは実際的なプロジェクトを構築するために、両方の技術を組み合わせる必要があるだろう。この場合も、長期的な考察が功を奏すだろう。データを長期的に利用可能にするつもりなら、短期的には有益であっても、保存する試料の総数を最大限にするより、試料ごとに最大量のデータを保存する方が研究機関にとっては有効である。

**危機管理計画/リスク分析**

プロジェクトが計画通りに進まないとどうなるだろうか。間違った方向に進みそうなものを検討し、それに備えることが、何が起きても混乱を最小限に抑えて、プロジェクトを運営することに役立つ。 リスク

分析はプロジェクト管理ツールである。危機管理計画の導入方法を明確に決定するために、読者は適切なプロジェクト管理テキストを参考にすることが望ましい。検討するべきいくつかの単純な事柄を以下に示す。

**スタッフの死亡または長期欠席。**デジタル化率が低下したらどのように対処するのか。新しいスタッフを雇用する、あるいは不足分を補うために目標を修正する必要があるかもしれない。 他のスタッフだけで作業を完了できるように、プロジェクトの期限を少しだけ延長することもできる。これを依頼するのは政治的に難しい。しかし、大部分の資金提供団体は、問題が不可避的であれば、やむを得ない問題に対して寛容である。

**必須デジタル化速度を達成できない問題にどのように対処するのか。**これは深刻な問題である。プロジェクトには多くの試料があり、一定の期間内にデジタル化しなければならない (これがない場合は、現実的な目標を設定する価値が十分ある。これによって業績を評価できる)。これは、ガイドラインとして設定するべきデジタル化率を意味する。スタッフをトレーニングするには時間がかかる。従って、早急にデジタル化率の達成を期待してはならない。目標が現実的であれば、 (通常のスタッフの欠席を考慮しても) 最大進捗度を上回るだろう。目標を達成できない場合は、作業プロセスを見直すべきである。回避できる遅延がないかよく見極め、あるいは関与するスタッフの追加を検討する。プロジェクトの予算が許せば、有給残業は一つの選択肢である。サービス残業はプロジェクトの運営がうまくいっていない兆候である。明らかに必要なリソースをプロジェクト計画案に盛り込んでいなかったのである。目標を達成しても、デジタイザーに次のプロジェクトを続ける意欲がない場合、トレーニングを受けたリソースを失うことを意味する。より多くのデジタイザーを確保できない場合は、目標を現実的なものに修正するまで、デジタル化する試料の数を減らすしかない。

**コンピュータが故障するとどうなるのか。**いずれにしても、業務委託契約を結んでコンピュータを取り替えるしかない。しかし、これには時間がかかることに注意しよう。小規模プロジェクトでは滅多に実現しないが、予備のコンピュータを準備する資金があれば理想的である。また、故障による遅延をカバーする時間もプロジェクト期間に盛り込んでおこう。

**バックアップ戦略。**ハードウェア障害からデータを保護する方策は、アクションプランに必ず含めることを強く推奨する。1 台しかないコンピュータが機能しなくなり、データセットをリカバーできなければ、プロジェクトは失敗である。これは、資金提供団体が寛容になれる問題ではなく、将来的にさらなる資金を引き出せる可能性がなくなることを意味する。

**悪意のあるデータ改ざん。** これは滅多にない事件であるが、外部から不正侵入される可能性のあるオンラインシステムを使用した場合に発生する。データベース管理のある種の構造を持った単純なパスワードのセキュリティが役に立つ。 最後にデータを編集した人物などの基本情報を追跡することも、品質保証プロセスの助けになるだろう。

## ソリューション/導入の文書化のためにするべきこと

特に、新しいスタッフのトレーニングなど、実行したことを文書化すれば、プロジェクト期間中の作業がはるかに容易になる。正確な文書化は、複数のスタッフが担当するデータ入力プロセスに一貫性を持たせる。最初のプロジェクトを振り返り、そこから学ぶことが可能になり、その経験を未来のプロジェクトに生かすことができる。しかしながら、文書作成には時間がかかり、作成するスタッフがプロジェクトの他の職務を兼任する場合、それ自体が障害になりかねない。事前に適切な文書化を計画することは、ワークフローの円滑な進行を図り、スケジュールを守るのに大いに役立つ。

## プロジェクトを評価する

プロジェクトの成否を計るためにどの指標を使うか検討しておこう。作業が完了した試料の数やその

所要時間だけを考慮せず、品質の高さも含めよう。また、スタッフの士気の高さも含めよう。というのも、有効なプロジェクトは未来のプロジェクトに役立つトレーニングを受けたスタッフを生みだす可能性があるが、彼らがその後研究機関で働くことを望まなければ、貴重なリソースの損失である。

**ソリューションは投資に見合う高収益があるのか**

本章の冒頭で述べたように、実際のプロジェクト実施中に、解決すべき様々な問題のすべてが、調和して機能することを確実にするために、アクションプランを何度も見直した方がよい。この段階が完了したら、プロジェクト全体、とりわけリソース要件を検討しよう。費やした努力に見合う価値のある成果があるかどうかを検討する必要がある。これは資金提供団体が使用する基準である。綿密に計画されたプロジェクトの答えは、断固とした「はい」でなければならない。確信がない場合は、実行する予定の仕事量を見直し、デジタル化する予定のコレクションの最重要部分だけに焦点を合わせ、他の部分は次のプロジェクトに先送りするとよい。

最終的に、資金を引き寄せる可能性の高い、綿密に計画されたプロジェクトになり、目標達成が可能になる。すべてを詳細に書き終えたら、リソースと資金を確保し次第、プロジェクトの運営が可能になる。今からするべきことは、全体をまとめて考えることである。

# プロジェクトを運営する

ようやくプロジェクトを実際に実施できるスタート地点に到達した。恐らく、目に見える成果はないものの、多くの作業がすでに終わったように感じるかもしれない。 しかし、見返りは開始直後からのプロジェクトの円滑な運営である。この期間中、プロジェクトが実際に始まる前でさえ、まだ実行するべきことが数多くある。しかし、すべてがプロジェクト開始に向けた実務である。

**仮説をテストする**

これが最初のプロジェクトなら、デジタル化のワークフローやデジタル化速度に関連する仮説が有効かどうか分からない。前段で論じたように、概念実証がこれを確かめる唯一の方法である。（選択したシステムの専門技術を見るのに十分な）およそ 100 試料の小規模なプロトタイプのプロジェクトを実施すれば、アクションプランを改善するために利用できる貴重な見識が得られるだろう。実際的なシステムの専門知識が得られれば、デジタル化に伴う問題をかなり深く理解できるので、デジタル化スタッフに対するトレーニングの質を高めることができる。従って、プロジェクト・スタッフの管理者はこのタスクを実行することが強く推奨される。プロジェクトの一環としてデータベースを構築または修正する場合、これがデータ入力の第一段階を形成する。この時、作業をしっかりと監視し、プロジェクトの初期段階でワークフローを修正する準備をする。この修正が遅くなりすぎると、プロジェクトの損失時間を取り戻せないことも十分あり得る。

**資金を調達する**

研究機関のものでも、外郭団体のものでも、すべてのプロジェクトには費用がかかる。十分に開発されたプロジェクトは、6〜10 週間で適切に準備できる (Snow, 2005)。適切な財源を見つけることは、本文書で実際的に詳細に論じることのできない問題である。利用可能な資金提供団体は、国によってまちまちであり、また研究対象のコレクションの厳密な種類によって異なるからである。投資対効果検討書とアクションプランを適切に書き上げることだけが、プロジェクト提案書が成功する可能性を高める。従って、本章の提案に従えば、資金を獲得するのはおそらく簡単である。

**データベースを構築する**

選択したデータベースに応じて、データベースを修正または、場合によっては作成するための時間を確保しなければならない。モジュール設計プロセスの導入を推奨する（後段で論じる）。システムの各部をテストすることができ、あるいは、データベースの他の部分が開発されれば、実際に使用することもできる。

**スタッフを雇用する**
スタッフが着任するまでに 2 か月かかることを覚えておこう。つまり、資金が確保できたら、まず初めに取り組むべきタスクである。

**文書化を開発する**
トレーニング・マニュアルやデータベース設計などの実用的文書は、未来の作業にとって非常に重要である。適切な文書は、経験豊かなスタッフの育成や、データベースシステムの維持を容易にする。プロジェクトが適切に始まる前に、優れた文書化の開発に時間をかけよう。

**適切な事務所スペースを整える**
スタッフには作業するための場所が必要である。従って、スタッフが作業を始めるまでに事務所スペースの準備を整えよう。普通のオフィス用デスクと椅子の上や、上の空間に必要な専門家用の機器を設置する十分なスペースを確保することを忘れてはならない。

**機器を購入し設置する**
事務所スペースと同様に、スタッフが着任するまでに機器も準備する必要がある。 機器が到着するまでスタッフを待たせるのはリソースの無駄遣いである。

**スタッフをトレーニングする**
研究機関に着任して即座にデジタル化の作業に取り掛かる準備が整っているスタッフはいない。 最低限でも、スタッフに研究機関の作業手順を教え、恐らくデータベースの使い方をトレーニングしなければならないだろう。以前に作業手順の文書化に費やした時間がここで戻って来る。これによってスタッフは、かなり早く、効率よく作業を始めることができるだろう。

**デジタル化を開始する**
ようやく、プロジェクトの主要部分を開始できる。

**プロジェクトを継続的に監視する**
恐らく、リスク分析に取り組んで対応できるように備えた、予想外の出来事に注意しよう。プロトタイプのプロジェクトを実行して準備していても、うまくいかないこともある。短期的目標を使い、また通常業務での成果を見直すことで、予想外の問題に素早く対処し、それに適応して打開することができる。.

**プロジェクトを見直す**
成否はともかく、プロジェクトが終了したら、アクションプランで提示した成功基準に照らして見直すこと。何が功を奏し、何が改善できるか検討しよう。このプロジェクトから学んだ知識を次のプロジェクトに生かすことができる。

最初のプロジェクトが完了したら、次のプロジェクトのテーマを検討する時である。デジタル化プロジェクトに関係する問題の経験を積んだので、今回は容易である。恐らく、今は経験を積んだスタッフと利用できるデータベースも手元にある。ここでもまた、投資対効果検討書の作成または更新や、新しいアクションプランの作成に、時間をかけることを推奨する。時間の経過に伴って、新しい技術や機会が利用可能になるからである。

# 第 4 章: 情報を整理しデータを表現する

情報が先か、データが先か。「情報」を定義するときデータという語を使用し、「データ」を定義するとき情報という語を使用する。鶏と卵のような堂々巡りをする議論である。Losee (1997) は、この点について詳細に論じている。幸い、これに関連する重要な定義はそれほど多くない。 ただし、以下の 3 つの概念は明確に理解しておこう。

関心のあるオブジェクトについての知識を、私たちは **オブジェクト情報**と呼ぶ。

**参照情報と補助的情報**を利用して、オブジェクト情報の品質を向上させ、充実させる。

コンピュータは何らかの方法で、オブジェクト情報、参照情報、補助的情報を保存し、提示できる。そのために使用するものを、私たちは**データ**と呼ぶ。

# オブジェクト情報 (目で見る情報)

すでに既存のデジタルフォーマットの情報でも、デジタル化されるのを待っている情報でも、入手した情報は、2 つのカテゴリーの 1 つに分類される。一次情報か二次情報である。一次情報は、オブジェクトを特定し、あるいはオブジェクト（すなわち、試料に付随するラベル、タグ、現場ノート）から直接取られたものに分類される。これらは、見解に関係なく（第一に正しい情報である限り）、決して変わらないと分かるものである。二次情報は、オブジェクトを記述または分類するとき、あるいはオブジェクトを他の種類の情報と関連付ける時に使用される。二次情報は、特定の時点での知識を反映する。従って、時間の経過とともに、見解の変化や知識の増加によって変化する。

**代表的な一次オブジェクト情報**

**オブジェクト識別子**– 識別子は、単一のオブジェクトを一意的に分離する。受入番号、バーコード、LSID 、あるいは試料に固有の価値を与える他の方法であり、*一意識別子*型である。

**コレクションイベント** – 収集者、コレクション番号、収集日から成る。 それぞれ、*人物*型、*識別子*型、*日付*型である。

**コレクションイベント場所** - オブジェクトが収集された場所の記述（*テキスト*型）。 座標（例：経度（*GPS* 型）と緯度（*GPS* 型））、土地所有または土地管理（*テキスト*型）、標高（*数値*型）を含むものと含まないものがある。

**コレクションイベント方法** – オブジェクトの収集方法に関する記述。

**オブジェクトの記述情報**– 生体試料の収集者の記述 (*テキスト文字列*型)。試料作成についての情報 （*テキスト文字列*型）。収集者または試料作成者による試料に含まれる注記または備考（*テキスト文字列*型）。

**環境情報** – 試料が収集された場所の特徴の記述。例えば、生息環境（*テキスト文字列*型）、植生（*テキスト文字列*型）、関連種（*テキスト文字列*型）、自然地理 (*テキスト文字列*型)。

**ドナー情報** – 個人ドナー (*人物*型)、研究機関 (*場所*型)、詳細な連絡先 (*場所*型) 、寄贈に適用さ

れる特別条件（*テキスト文字列*型）。

**参考図書** – 表題、発行日、ジャーナル名、書誌情報、著者から成る。

**代表的な二次オブジェクト情報**

**地理情報 (空間)** – 政治地理。例えば、国、州、地区、コレクション場所のジオリファレンス。

**分類学上の名称と命名法上の名称**- 収集者が命名した初期名称、その後の複数の名称または修正名を含む。命名者の氏名と命名した日付を伴う。名称は*分類名*型、命名者は*人物*型、日付は*日付*型と見なされる。

**保存場所**- 研究機関 (*場所*型) が試料の所有者なら、バーコードまたは受入番号（*一意識別子*型）、研究機関内の保存場所（*場所*型）。様々な所有者や保存場所の時系列の足跡を含むこともある。

**分子**– 収集時に記録されることは滅多にないが、次第に普及している。証拠試料を添付した DNA サンプルや DNA 配列と関連付けるのが優れたプラクティスである。*配列*型と見なされる。

**ステータスマーカーとラベル**– 関心のある他の情報と関連するオブジェクトについての情報をすべて集める受け皿。保全ステータスマーカーは、オブジェクトが何らかの指定を受けていることを示す。例えば、希少、絶滅寸前、絶滅危惧、感受性など、通常はその名称から（すなわち、希少種）指定されている。型ステータスは、特定の命名規則に従った型として試料を指定する。トランザクション・マーカーは、試料と複数の貸借が関連付けられる。その他のマーカーは、試料とプロジェクトや出版物が関連付けられ、コレクション全体のサブセットの一部としてマークし、特定の使用適合性を指定し、特別な配慮をマークする。

**備考欄/注記** – オブジェクトについて、あるいはオブジェクト自体の一部ではないオブジェクトに関連する情報についてのコメントをすべて集める受け皿。 例えば、「瓶は少し水漏れしているようだ」、「ID が正確かどうか分からない」、「収集者の氏名は判読不能だが、A. Smith のコレクションだと思う」、「コレクションからこの 試料を見つけられない」など。– Felisa Jones 5/1/1999

この潜在的情報をすべて記録すると、短期的に高いコストがかかる (Armstrong, 1992)。しかし、長期的利点は、同じ作業を二度と繰り返さなくて良い点である。一次情報を部分的に記録すれば、短期的にはコストが安いが、長期的にはさらに高騰する。というのも、追加データが必要になった時、デジタル化ワークフローの一部をもう一度実行するために、追加リソースを使う必要があるからである。大規模な研究機関では、この活動だけでデジタル化プロセス全体の 3 分の 1 から半分を占めることがある。その結果、試料ごとのデジタル化の正味コストが、かなり膨れあがる。

システム内で取り扱うこの種の情報は、目的に沿って決定され、それぞれの詳細度はそれに応じて変化する。原則として、一次情報はできるだけ完全に記録するように努め、二次情報は選択して記録することを確認しよう。

# 参考情報と補助的情報

デジタル化プロジェクトが、関心のあるオブジェクトと直接的に関連する情報の記録だけに制限され、それぞれのオブジェクト情報を自由に入力するとは考えられない。**補助的情報**とは、オブジェクト自体と直接的に関連しない、デジタル化プロジェクトの一環として管理する、すべての電子情報のことである。 オブジェクトについて入力する価値を制限するために補助的情報が使用された場合、それが **参考情報**と考えられる。

補助的情報は、特徴の有無を指定するために使用する 2 つの価値、オブジェクトのラベルとして入力する価値のリスト (例えば、雄、雌、雌雄同体、不稔) 、すべての領域が出版物に関連するようなかなり複雑な情報セットなどと同様に単純である。補助的情報はしばしば、参考情報についての情報を伴う。例えば、参考情報がオブジェクトを所有する研究機関の名称の場合、補助的情報には学芸員の氏名や彼女の連絡先が含まれる。参考情報が、試料の連邦ステータス指定の場合、補助的情報には、そのステータスが連邦公報で公開された日付、その参考図書、適用対象の個体群が含まれる。

**代表的な補助的情報**

命名法

地理学

形態学

人物名

プロジェクト

研究機関

出版物

トランザクション

# データ (コンピュータで見る情報)

データについて重要な点は、 情報を正確に表示することと、入力したものと同じ情報を入手できることである。試料情報は大抵、基本単位と呼ぶ場所に保存される。 これらの基本データユニットは、関連データを一緒に保存でき、情報をデジタル表示する無制限の方法と結合できる、概念ツールである。後段の各章で様々な例を挙げる。実際にこれを実行した例や、これを考慮してデータベースソリューションを選択した例がある。

いくつかの 共通データの基本単位を以下に挙げる:

人物 – ファーストネーム、姓、イニシャル、肩書き

分類名 – 階級、別称

場所 – 緯度、経度、高度、地名、住所

日付 – 日、月、年、世紀

配列

染色体数

参考図書 – 表題、書誌情報、発行年、発行元

**基本的なデータタイプ**

どのようなデータタイプが最良のデータかを知ることは、データベースソリューションを構築または選択する上で重要な段階である。基本的に、3 種類のデータがある。数字、文字、日付である。 最初に考えたより高い水準でデータベースを作動させる方法を選択する。

**テキスト/文字列領域**は、データタイプの中で最も単純であり、初心者でも迅速に簡単にデータ入力ができる。通常、目で見えるものが、入力したものである。利用者は文字と数字の両方で打ち込めるが、キーボードで数字を打ち込んでも、コンピュータは、正確な数字領域の中で、同じ方法で処理しないことに注意しよう。

- **プラス面:**
  生データが入力した通りに見える
  並べ替えが文字に期待される結果を示す
  図表形式が使いやすい
  出力データをフォーマットしやすい
  コピー＆ペーストが期待通りに機能する

- **マイナス面:**
  テキスト領域は数字領域より多くのスペースをとる
  文字列が長い場合、テキストベースの参照テーブルが非効率的で遅い
  同じことを何度も繰り返して打ち込まなければならない
  入力したときとは異なる目的でデータを利用すると、通常、データが変化して、予想外の結果を招く
  数字が関与し、広すぎる領域に置けない場合、並べ替えが予想外の結果を招く

パラメータ:
*長さ* – これは、領域内に収容できる文字数を示す。長さを短く設定すると、データを切り詰める。テキスト領域にデータをインポートした場合、すべてのシステムがこの現象が発生したことを示さず、カット＆ペーストに問題が生じやすい。長さを長く設定すると、並べ替え機能に深刻な影響を与え、 場合によっては、この機能が働かない。

*パディング* – 領域内の実際のデータ量に関わらず、特定の長さの領域を設定したテキストフォーマット。これはデータベースのサイズを大幅に膨らませ、これが要求された場合は十分注意すること。

**メモ領域** は、テキスト領域の 1 つの形式であり、限定された制限があるので、単独で言及される。この種の領域には大きさの制限がなく、従って、利用者は詳細に記述できる。特に、生息環境などの記述領域として有効である。しかし、潜在的サイズが広いので、並べ替えできることは滅多になく、検索も難しい。ボールド体やイタリック体の文字表記などのテキストフォーマットを処理できない。キャリッジリターンは不安定である。控え目に利用し、頻繁に検索するデータに利用しない方がよい。

**数値領域**は、数字だけの領域であり、テキスト領域よりさらに単純である。しかし、しばしば単独では意味がなく、テキスト領域より制限が多い。数値領域には 2 種類ある。浮動小数点領域はデシマル表記が可能であるが、整数領域は整数しか表記されない。

- **プラス面:**
  数字領域は保存スペースをあまりとらない。
  参照テーブルを効果的に利用できる
  データ入力システムに基づくフォームを容易にコード化できる
  データの細分化を容易にする

- **マイナス面:**

通常、テキスト領域を要求し、限定するので、データ処理に 1 つではなく 2 つの領域を使わなければならない。
テキストの処理に参照テーブルを使わなければならない
データの細分化が多く、エクスポートデータが複雑になる
図表形式のデータが読みにくい

パラメータ:
*長さ* – 領域内に入力できる最大数を決定する。通常、数字を保存するとき要求されるビット数で示される。

*基数* – 数字は 10 進法で保存する必要はない。例えば 16 進法も可能である ( 8 進法)。コンピュータとプログラマーは、好んで使用するが、利用者には解釈が難しい。可能であれば、10 進法だけを利用するとよい。

**日付/時間領域** は、修正数字領域であり、注意して利用した方がよい。すべてではないがほとんどが「1」という既知の開始日からの特定の日付を示す整数で保存される。この整数を組み込みフォーマットに入れて、様々な方法でこの日付を表示する。しかし、様々なパッケージが様々な開始日を使用する。従って、開始日が同じで、データをエクスポートまたは転送するとき表示される数字ではなく、日付を転送していることをチェックする必要がある。

**ブール領域**は、「はい・いいえ」、「真・偽」、「有・無」、「である・ではない」などのニ者択一を反映するために設計された。いくつかのソリューションでは、入力される実際の価値はラベル上のものと同じであるが、通常、ラベル上の上記の二者択一の選択肢、あるいはチェックボックスを元に、基礎領域に「1」か「0」が入力される。 ブール領域のデフォルト値やヌル値には注意が必要である。例えば、特定の押し花試料が無花果であることを示すために「不稔」領域を使用する。特定のソリューションでは、その領域に何も入力されていない場合、「ゼロ」に変換され、「不稔ではない」と示される。あるいは、「1」しか入力できない場合、価値が何も入力されない「ヌル」では「不稔ではない」と識別できない。最後の可能性は、「1」、「0」、「ヌル(入力なし)」の三者択一で保存されるケースである。しかしこの場合、「はい」か「いいえ」を入力した後で領域をクリアするのが難しい。

**BLOB (バイナリ・ラージ・オブジェクト) 領域またはコンテナ領域**によって、画像、音声、ビデオ、文書、他のバイナリデータなどのファイルをデータベースに保存できる。特定の BLOB 領域に保存できる最大ファイルサイズを事前に決めておかなければならない。もちろん、データベースにファイルを保存すれば、データベース・ファイルのサイズが劇的に拡大し、パフォーマンスに影響を及ぼす。 いくつかのソリューションでは、ファイル自体ではなくリンクを保存できる。この場合、データベース・ファイルはそれほど大きくならないが、リンクファイルやデータベースの移動によって、リンクが破壊されることもある。

計算領域

測定値の領域から実際に計算し、すでにデータベースに入力した情報を処理するケースがしばしばある。いつものように、これを実行する方法は数多くあるが、以下の事柄を心にとどめておくとよい。

1. 公式とそのパラメータの変更頻度
2. 計算に使用する測定値の変更頻度
3. 結果の最終的な使用目的

公式やパラメータが比較的安定している場合、データベースに領域を作ることを考慮する価値がある。

その領域にコードを手動入力（手で打ち込む）しなくても、計算結果が保存される。

測定値が変更される可能性がある場合は、データベースにその変更を記録する方法を検討する必要がある。しばしばリンク表が役に立つだろう。

結果が広く利用される予定で、元データに価値が付加される可能性がある場合は、 他の人が利用できるようにデータベースに保存するのが恐らく価値があるだろう。そうでない場合は、精通した環境で計算した方がよく、データベースに保存する必要はない。あるいは、最終的な表示の一部として計算してもよい。 計算用データベースではなく、パッケージを利用するのも間違いではない。

関数
データベースで領域を計算し、公式を使用すると、関数と接触する。この時データタイプが重要である。ほとんどの場合、テキスト領域で使用できる関数は、数字領域では使用できない。逆もまた同様である。日付/時間領域はテキスト仕様で表示されるが、現実には数字なので、加算・減算ができることを覚えておこう。方程式や公式を書くために関数を使用すると、データ入力時間をかなり削減できるが、ロジックをコード化し、組み立てるための投資が必要である。

全体として、コード化が可能ならコード化しよう。すべての入力は常に同じ方法で計算されるので、一度コード化操作をするだけで済む。もちろん、プログラミング技術がなく、新しい言語を学ぶ時間もない場合は、別の人物のコード（本物のプログラマーが実行する方法）を再利用するとよい。他の方法もすべて失敗した場合、手動入力すれば、一応答えは求められる。

**特殊文字とエンコード**
特殊文字には、発音区別符、アクセント、数学記号、非ラテン文字が含まれる。データが様々な元の言語で入力される場合に重要であるが、一貫して使用しないと紛らわしい。 例えば、Wurdack と Würdack は同一人物だろうか。ゼロからデータ入力する場合、データ入力スタッフがそれを使用するかどうか、根気強く取り組むかどうかを*事前に*決めておいた方がよい。

エンコードとは、プログラマーが特殊文字を解読する方法である。様々な方法があるので、特にレガシーデータをインポートする必要がある場合に、データベースソリューションが使用する方法を知る必要がある。

両テーマは以下のサイトで論じている。
http://www.nada.kth.se/i18n/iab-charsets/terminology.html

# データ表示とストレージフォーマット

大部分のデータベースソリューションは、4 つの方法を組み合わせてデータを表示する。データ入力表示、ストレージフォーマット、エクスポートフォーマット、マルチメディアディスプレイの 4 つである。これらは、行、すなわち表、あるいはデータ入力ボックスの画面、すなわちフォームとして利用者に表示される。

これらの違いを利用者に説明することは有益なエクササイズであり、 大きな混乱を回避できる。特に、データ入力スタッフが、プログラマー自身ではなく、他のオーディエンスにデータを表示する方法の責任者でもある場合に有益である。

**データ入力表示**
データ入力表示は、データベースにデータを入力する場所である。1 つのデータ入力システム（特に、利用者自身が構築したシステム）がすでに存在する場合、新しいデータ入力システムを導入すること

は、しばしばデジタル化プロジェクトの成功を左右する。それは不可能ではないが、プロジェクト開発の初期段階からこれらの人間が作業に関わることが重要である。しばしば既存のシステムが「予測できない」形で動作する正当な理由である。ロジックを学ぶ姿勢をなくしてはいけない。見逃したことを教えてくれるだろう。 見事なソリューションを、必要とあれば 2 回、説明するだけでよい。常識的で、それほど速度を落とさない限り、同意するだろう。もちろん、何も存在しない場合は、データ入力スタッフを評価し、正確さを落とさずに、効率を最大化するものを選択しなければならない。

基礎データベースの構造がしっかり構造化され、細分化されるほど、データ入力表示はフォームベースのソリューションになることが多い。データ入力はしばしば、これらのシステムではより時間がかかるが、データ検査は非常に厳しい。入力速度を速めるために、いくつかのシステムでは、高速データ入力 (RDE) 表を利用している。インポートはできるが、データ検査はそれほど厳しくない。これらの中からどれを選ぶかは、データ入力を担当するスタッフ次第である。

データ入力表示は実際にはデータを表示/提示するものではなく、データ入力を能率的にするべきであることを理解し説明することが重要である。データ入力システムを仕様書に含める場合、このことを心に留めた方がよい。格好よく見せるためだけに過剰に設計に凝るのはやめた方がよい。 柔軟性を持たせ、できるだけ多くのタイプの利用者が、基礎データベースの正しい部分にデータを変換できることがさらに重要である。誰にでも適した単一のデータ入力ソリューションは、決して手に入らない。

例えば、データベースが SQL サーバーに保持され、研究機関内での日常的なデータ入力用の MS アクセスのフロントエンドや、外部データ入力用のウェブフォームを持っているようなケースである。

**ストレージフォーマット**

ストレージフォーマットとは、データベースにデータを保存する方法である。適切なデータ入力表示を持っている場合、利用者はストレージフォーマットに関与する必要がない。しかし一般に、データモデルが複雑であるほど、ストレージフォーマットが、リンク表、記録 ID、オブジェクトを使用することが多くなる。この場合、実際のデータを見ることで着想することはほとんどない。もちろん、単純なフラットデータモデルでは、ストレージフォーマットはデータ入力表示と同じになる。

**エクスポートフォーマット**

エクスポートフォーマットは数多くあり、データベース以外のパッケージでデータを分析する必要がある場合、ソリューションを選択する決定的な基準になる。この場合、データベースは作業領域ではなく、一次データの保存場所になる。

**マルチメディアディスプレイ**

マルチメディアディスプレイは、様々なオーディエンスを対象にし、特定のものを強調するために、データを操作する場所である。これは、非常に頻繁にデータ入力と混同されるエリアでもある。データが適切なフォーマットに適切に入力された場合は、ほぼ満足できる形で表示できる。従って、データを入力する方法と場所と、表示するものを、明確に区別することが重要である。

デジタル化プロジェクトのこの部分は、最も少なく割り当てられるリソースと日限とともに、最後まで残されることが多い。 これは外部オーディエンスがプロジェクトを判断する方法であり、基礎データベースそのものと同等に重要である。リソースが不足した場合は、独自の資金調達ラインやスタッフ配置を伴う単独のプロジェクトとして、これを検討する価値がある。

## 標準

**標準とは何か**

標準とは、公認団体が承認した文書のことであり、一般に繰り返し使用される、製品や関連プロセス

や製造方法の規則、指針、特徴を提供する。最もよく見られるのは、コレクション・データベースや情報管理システムの構築法を教える、コレクション・データベース標準がどこかにあるという誤解である。そういったものは存在しない。様々なコレクション・データベースや他の情報管理システムが使用されており、共通の標準データモデルが根底にある訳ではない。

しかし標準は存在し、コレクション・データベースや情報管理システムの設計を含む、生物多様性情報科学活動に影響を及ぼす。ここでは、これらの標準を 4 つの大きなカテゴリーに分類した。

**データ交換標準。**この標準は、転送プロトコルあるいはトランスポート・プロトコルとしても知られ、情報ソースに関係なく、情報を交換し、合成できるように、情報を整理し、フォーマットするために使用される。コレクションデータのデータ交換標準として最も一般に知られているのは、標本館情報システム＆データ交換プロトコル (HISPID) (Conn 2000)、生物学コレクションデータアクセス（ ABCD ）( http://www.tdwg.org/activities/abcd/ ) 、ダーウィン・コア（ DwC ）( http://www.tdwg.org/activities/darwincore/ )である。データ交換標準は、ヘッダー、領域、タグ、データを整理する要因を提供する。ABCD と DwC は、XML スキーマで表現される。ABCD は階層構造を持ち、 生物学コレクション情報をモデル化する包括的で詳細なフォーマットになることを目的とする。DwC ははるかに単純なフォーマットを持ち、一般に有用な「最も重要な情報」の交換を容易にするために設計された。

**標準データセット**は、ある種の情報の「管理用語」を定めるために使用される。これらは、索引リストや参照表の基準として使用された場合にきわめて有益である。標準データセットの例として、ブルミット＆パウエルの『植物名の命名者』(1992) が挙げられる。植物名の命名者略称の標準として、植物命名法国際会議に承認されている。別の標準は ISO 3166 である。国名や主要な下位区分（例えば、州や郡/県）をコード化する地理標準である。これらのコードは、地理的行政単位に価値を制約するとき非常に有益である。データセットが「標準」だと言うことは、必ずしもある種の情報が利用可能な唯一の選択だという意味ではない。例えば、連邦情報処理規格 (FIPS) 10-4 には、ISO3166 の国名コードとは異なる 2 文字の国名コードのオプションリストがある。データセットが「標準」だと言うことは、必要性に完璧に、あるいはうまく適合するという意味でもない。例えば、FIPS 10-4 も ISO 3166 も、標本ラベルに示されたような国名のドロップダウンリストにこれらの標準を使用した場合、イングランド、ウェールズ、スコットランドの入力には問題がある。

**ベストプラクティス文書**は、手法やプラクティスの統一を助ける指針であり、通常、組織や団体に厳しく検査される。例えば、米国哺乳動物協会の哺乳動物の自動データ処理の文書化規格 (McLaren et al. 1996) や、Arthur Chapman が地球規模生物多様性情報機構 (GBIF)のために作成した文書がある (Chapman 2005a, 2005b, 2008)。

**技術標準**は、前述の 3 つのカテゴリーに当てはまらない標準の用語をすべて受け入れる包括的な受け皿。通常、技術標準は、データの交換、表示、操作ができるシステムの設計や導入に影響を及ぼす。ソフトウェア開発者は、インターフェースのサポートを構築し、製品やサービスにエンコードを設定するために技術標準を使用する。例えば、TDWG 情報検索アクセスプロトコル (TAPIR) は、XML ベースの要求を転送するための HTML の使用法を特定し、分散型データベースの数字やタイプを保存した、アクセス構造化データ (すなわち、ABCD や DwC フォーマットのデータ) に対応する。別の例として、オープン GIS ウェブマップサービス (WMS) 導入仕様書がある。多様な情報ソースから届く情報の地図状の表示を構築し、表示するのに役立つ。

標準は、情報の共有や解釈のための共通の言語、規則、プロトコルを提供する(Conn 2003)。標準を理解し使用することは、情報システムの品質を向上させ、開発を合理化し、システムや情報と他のシステムや情報との相互運用性を高める。他方、数多くの標準があるので、現状に合った、最も目的に沿った標準を選択できる、標準に精通した高水準の専門知識が必要である。

**標準化団体**

標準化団体とは、標準を開発し、維持する団体である。次第に、相互に影響し合い、有意義な方法で標準をリンクする方法を模索している。数多くの国際的標準化団体があり、地域的広がりや国家的広がりで活動している団体はさらに数多くある。

標準化団体

**生物多様性情報標準 (TDWG):**
http://www.tdwg.org/

**オープン GIS コンソーシアム (OGC):**
http://www.opengeospatial.org/

**国際標準化機構 (ISO)**:
http://www.iso.org/iso/en/ISOOnline.frontpage

リスト
http://www.consortiuminfo.org/
http://bubl.ac.uk/link/i/internationalstandards.htm

さらに、オンラインリソースを提供し、この学会での標準の役割の理解を助ける有益な出発点となる会議やワークショップを開催する、非標準化団体がある。

**地球規模生物多様性情報機構 (GBIF):**
http://www.gbif.org

**博物学 コレクション保護協会 (SPNHC):**
http://www.sphnc.org

**自然科学 コレクション連合**:
http://www.nscalliance.org

**及び、多数の分類学会.**

# データ品質

**データ 品質とは**

データ品質は、すべて相対的であることを覚えておこう。「使用適合性」(Chrisman 1983)、「潜在的価値」(Dalcin 2004) 、「無欠陥」(Redman 2001) などの用語は、すべてデータ品質を説明するために使用される。確かに、これらの用語はすべて、データが少しは役に立つかどうかの指標と考えた方がよい。しかし結局、要約すると、希望する作業にデータを使用できるか、持っているものを他者に説明できるか、まったく違う目的で別の人物が使用できるか、ということである。

Chapman (2005a) は、データ品質はデジタル化プロセスのあらゆる段階で重要な役割を果たすものであると述べた。新しいデータに生じる問題を防ぎ、既存のデータのエラーを訂正できるので、決定的に重要である。現有データと新規作成するデータの品質を評価する単純な方法は、以下の項目を検討する Redman (2001)のリストを使用することである。

入手可能性

確度

完全性

他の情報源との整合性

妥当性

包括性

詳細度

解釈の容易さ

これらの品質は、デジタル化プロジェクトの規模に関わらず妥当である。従って、目標に沿って、各項目に対処する方法を決めておくことが重要である。作業制限に沿って各項目に優先順位をつけてもよいが、アクションプランには盛り込んだ方がよい。

**新しい データを入力する**

もちろん、ゼロから作業を始める場合、持ち時間内に手に入れたい結果を得るために、記録する必要のあるデータだけを取ればよい。 しかし、プロジェクトの目的にとって知る必要があり、文書化する必要のある最も重要な部分を選択する。

高品質の新しいデータを作成する一つの方法は、索引リスト、ドロップダウンリスト、管理用語を利用することである。これらは、データベースの特定の領域に 1 つ以上のオプションを選択する基準になる、統一されたデータ/用語のリストである。データ価値はすでに検査してあるので、これらのリストを使用すると、データの確度を高めるメリットがある。ただし、間違ったオプションを選択する危険性も拭えない。検索条件を追加する階層検索は、入力をより正確にするもう一つの有益な方法である。索引リストを利用できるときは利用しよう。エラー検査に費やす時間を削減できる。しかし、プロジェクト開始時に検討するべき、索引リストで統一されたデータセットを使用するには制約や障害がある。第一に、データ入力を開始する前に、それらを入手し、フォーマットし、そして恐らく拡張しておかなければならない。これは、タイミングよくデジタル化を開始できるかどうかに影響する。第二に、これらの標準データセットは、システムにインポートした後で、変更または更新される。既存のデータの他の部分にこれらの変更を組み込んで調整するのは必ずしも簡単なプロセスではない。

信頼できる利用者にはデータ入力を許可し、他の利用者には入力を制限するという、あまり厳密でないやり方で索引リストを利用するという選択肢もある。

非常に複雑な索引リストの利用は、大抵のデータベースソリューションにおいて、プログラミングのオーバーヘッドを増加させ、データベースの構造そのものを複雑にする。

**既存のデータをインポートする**

ほとんどの場合、プロジェクトの目標達成に役立てるために、再度目的を持たせ、組み込み、統合し、構築したい既存のデジタルデータセット（レガシーデータ）がある。別のケースでは、ワークフローのシステム外でデータセットを作り、その後それを取り入れることもある。これらのデータを新しいシステムに移動したい場合も、あるいは既存のデータベースシステムに機能を付加したい場合も、まず最も重要なことは、持っているものと、作成された理由を理解することである。データセットが、何も操作を加えられずに、システムから別のシステムに転送されることは滅多にない。データ品質の実質的評価の第一原則は「目的」である。持っているものと、データセットの結合方法が、このデジタル化プロジェクトの目標に役立つと分かれば、目標達成に必要な追加情報を判断できる。

レガシー システムには独自の規則があり、データはデータベースに保持されているので、ソフトウェアがデータベースとして機能しないことを覚えておこう。その目的と実際のコンテンツを決定するために、それぞれの表と領域の評価に費やす時間が、長い目で見れば時間の節約になる。レガシー データは必ず浄化しなければならない。このタスクにかかる時間を過小評価してはならない。

Maletic and Marcus (2000) は、データクリーニングを以下のように定義した。

- エラーの種類を定義し決定する
- エラー事例を検索し特定する
- エラーを訂正する
- エラー事例とエラーの種類を文書化する
- 将来の類似エラーの発生を削減するために データ入力手順を修正する

これについては、『データクリーニングの原則と方法』(Chapman, 2005b)に基づいて、本文書第 4 章でさらに詳しく論じたが、以下の諸点に注意しよう。

領域名

領域名は様々に誤解されやすい。

**ケース 1:** 異なる学問分野が、完全に異なる概念を表すのに同じ用語を使用している。動物名のデータセットで「有効な」という用語は、植物名のデータセットと同じ意味ではない。従って、両方のケースに正しく使用されても同意ではない。

**ケース 2:** 領域の中身が領域名と全く関係がない。これは、必要なデータを入力するシステムがないため、あるいは利用者が領域名を理解していないために生じる。

**ケース 3:** 領域がその使用目的を変更した。「日付」と記された領域は、当初試料の収集日を記入するための領域であったが、2 人目の利用者は命名日の領域だと考えた。従って、データそのものは正しく見えるが、実際には 2 つの異なる情報片にすぎない。

列/ 領域配列

適切に設計されたデータベースなら、これは問題にならない。 しかしすべてのデータベースが適切に設計されているわけではない。

**例:** スプレッドシートには 2 つの領域がある。1 つは命名日を記録し、もう 1 つはタイプ検証日を記録する。左側に「命名者」や「検証者」とそれぞれ記されていれば、領域の意味が分かる。しかし、どちらの領域にも「日付」と記されている。

スプレッドシートは、データ要素と列を特定するためにセル参照を使用するので、これは完全に有効である。しかし、これらをデータベースパッケージにインポートすると、領域名が不一致なため、問題を経験する。運が良ければ、コンピュータの指示で名前を変更するオプションが与えられ、最悪の場合は、領域に自動的に名前が記入される。インポートする前に、列に特徴的な名前を 記入しておいた方が良い。

「行」対「記録」

データベース表で、行は記録を表現し、それぞれの記録は何か（例えば、試料、人物、出版物など）に固有のインスタンスを表現する。それぞれの記録は、データが追加されていなくても存在する領域番号を含む。 それぞれの記録は、同じデータまたは潜在的データを持つ。テキスト文書やスプレッドシートを元にしたデータは、必ずしもこの方法で整理されない。データは、インスタンスごとに 1 度だけ繰り返される、ヘッダーとともに階層的に整理される。

**ケース 1**: データベース表の列は、分割したラベル情報を保持するために使用された。下表の記録 1 と 5 には、データベースの最初のいくつかの領域から間違ってパースされた、元文書のヘッダーが記入されている。

| ID | バーコード | 収集者 | コレクション番号 | 場所 | 名前 | 命名日 | 国 | 収集日 | 利用者 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | The | Adam | Smith | コレクション | 1960-9 | | | | |
| 2 | 98987 | Smith,A. | 90 | BM | S. aph. | 2/6/1969 | エクアドル | 1969 年 5 月 | yy |
| 3 | 98988 | Blogg,B | 1 | KEW | B. perr. | 5/12/2006 | イギリス | 1908 年 6 月 | xxx |
| 4 | 98989 | Anon. | 306 (Smith,A と推測) | NY | E. sup. | 8/8/1971 | フランス | 1701 年 9 月 | xxx |
| 5 | The | Richard | Spruce | コレクション | | | | | |
| 6 | 10001 | Spruce,R. | 5040 | BM | M.aus. | 1/1855 | エクアドル | 12/1852 | Yy |
| 7 | 10002 | Spruce,R. | 5041 | BM | M.apr. | 1/1855 | エクアドル | 10/1851 | yy |

テキスト文書やスプレッドシートでは、1 つの記録から次の記録に、情報が一貫性なく繰り返される。

**ケース 2:** 記録 4 の行は、コレクション番号のセルに、後に挿入されたコメントを含んでいる。

フォーマッティングとデータタイプ

データベースでは、特定の領域にあるデータはすべて同じデータタイプである。ただし、テキスト文書やスプレッドシートからインポートする場合、これは当てはまらない。例えば、元文書の日付列は、ほとんどが mm/dd/yyyy フォーマットのデータで追加されるが、時おり、「1/11-13/2001」や「Spring 196?」といった価値のセルがある。これらはデータベースに正確に反映されない。さらにひどいケースでは、「05/01/2001」という基礎日付をフォーマットして「May 2001」といった形で表現する。インポートは適切にできるが、正しい収集日情報を正確に反映していない。

領域または類似領域のコンテンツが一貫してフォーマットされていると思い込んではいけない。編集できない索引リストを元に入力されない限り、これは当てはまらない。例え、索引リストがあっても保証の限りではない。必要なデータセットのすべての価値をフォーマットし直すのに要する時間を過小評価してはならない。 今手に入るのは自動解析スクリプトだけであるが、それらを使いこなすのに要する時間について現実的に考えよう。ある部分では手作業も必要になる。

データセットを合成する

様々な元の目的を持った、様々なシステムの、様々なフォーマットの、1 つ以上のデータセットを持つことができる。これは完全に有効であり、しばしば合成するのが望ましいとしても、様々な元の焦点を持った、1 つのデータセットの制約が、他のデータセットの実用性を害することに留意する必要がある。合成したデータセット全体の品質は、それに含まれる最も品質の低いデータセットより悪くなることはないが、 最も品質の高いデータセットより良くなることもないことに注意しよう。データ品質は、実際にはデータセットを統合するより、リンクした方がより良く保護される。また、データセットの統合が、初めの印象ほど単純でないのも事実である。

**例 1:**
データセット 1: 特定種の複合体の遺伝的浮動を期待して、個体群生態学者が作成した試料データセット。

データセット 2: 受入台帳。

データセット 1 の 1 つの領域は「配列?」と呼ばれ、 証拠試料を伴う DNA 配列があることを示すために「y」が使用される。ヌル値は DNA 配列がないことを示す。

データセット 2 にはこの領域はない。

配列が存在しないことを明確に示すために、データセット 1 で「n」を使用するように変更せずに、2 つのデータセットを合成した場合、ヌル値が 2 つの意味の 1 つである可能性を持ち、利用者に伝わらない。従って、この領域の品質は低下したことになる。

**例 2:**
データセット 1:サンゴ藻の分類処理

データセット 2: サンゴ藻のタイプカタログ

データセット 1 は、特定の分類群と、タイプを示す試料についての情報を記録する。「分類群」と呼ばれる領域は、当該分類群の基底部を保存するために使用される。

データセット 2 は、現在はタイプフォルダに入った標本館のすべての試料を記録する。現在の試料名を保存するのに「分類群」と呼ばれる領域を使用する。

ここでは、表そのものの性質としてはきわめて正確に、2 つのまったく異なる情報片を記録するのに同じ領域名が使用された。しかも各列は類似していない。これらのデータセットを統合するのは可能であっても簡単ではない。

## 言語

データベース入力スタッフや一次ユーザーは、データベースでデフォルト言語を使用するのが適切である。データベースでは 1 つ以上の言語で要求に応える必要がある。ラテン系言語のような共通言語で一般用語を使用する場合以外は、複雑化の大きな要因である。データはそれぞれの言語で個別に記録されるだけでなく、適切なバージョンでデータを表示するには、適切な手順の導入が必要である。データの保守はさらに複雑である。2 つのバージョンで記録を更新しなければならない。データの自動翻訳は可能であるが、常に正確な翻訳ではなく、従って、データの信頼性を低下させる。

データベースは多様な言語を保持できるが、保守問題がデータ入力の実用性を上回る。少なくとも、非ラテン系言語だけの場合、適切なエンコードの使用は必須である。

## 知的財産権

知的財産問題は、広大で複雑であり、本章の範囲外である。

しかし、あらゆるデジタル化プロジェクトが認識し、できる限り広範に対処すべき問題である。知的財産権（IPR）は、デジタルリソースを作成するために、情報やデータを使用するプロジェクトに影響し、さらにそのリソースをターゲット・オーディエンスに普及させる方法にも影響を及ぼす。研究者や研究機

関が研究のために定期的にデータセットにアクセスし、利用するとしても、当初の目的以外で、元のままの形で (ネット上でも紙媒体でも)実際に出版する権利はない。

原則として、常にデータセットの元の情報ソースを見つけ出すように努め、最大限の知識を使って実行したことを文書化し、常に使用許可を得、常に情報ソースに感謝しよう。研究機関の指針や地方の法律を確認しておこう。規則は国ごとに異なるものである。

開始する場所は以下の通り。

**GBIF**
http://www.gbif.org/News/NEWS1174645079

**アメリカ**
http://usinfo.state.gov/products/pubs/intelprp/index.htm

**イギリス**
http://customs.hmrc.gov.uk/channelsPortalWebApp/channelsPortalWebApp.portal?_nfpb=true&_pageLabel=pageLibrary_ShowContent&id=HMCE_CL_000244&propertyType=document

# 第 5 章: データモデル

## 序論

第 4 章を読み終えた読者は、恐らく、記録したい一次オブジェクト情報と、二次オブジェクト情報、および、デジタル化プロジェクトの一環として保持したい補助的情報の識別プロセスを開始したことだろう。コンピュータがデータを見る方法や、様々な使用目的に沿ってデータを表示するためのシステムの構築法について、何か考えているだろう。今こそ、コンピュータシステムを使って、この情報をデータとして整理し処理する方法に取り組む必要がある。これがデータモデルのテーマである。

データモデルについて、まずは単純な事例であるカタログについて論じよう。 そして、最初の整理概念として、関心のある**基本単位**を紹介する。 その後、基本単位とデータモデルの**焦点**が、モデル設計において考慮すべき重要な点となる、より複雑な状況に論を進める。これらの諸点は、情報管理システムのモジュール設計という、より複雑な次のテーマの理解を助ける。

データモデルは、データを入力し、閲覧し、操作し、他者の利用可能にするシステム導入の基盤である。次に、導入システムの構造を含む、データモデル導入の基本概念を論じる。これには、データを使用可能にするために、モデルにできることと、導入そのものにできることを区別することの重要性も含まれる。その後、導入された情報管理システムで、情報をデータとして処理する方法に影響を及ぼす、データモデルの複雑さに関連する主要な概念に取り組む。

## 単純なデータモデルの基本単位

コレクションをデジタル化するとき、現実に利用する最も単純なデータモデル「カタログ」についての考察を始めよう。**カタログ**とは、オブジェクトをまとめてリストの形で表現したものである。カタログは、リストの中のオブジェクトについての記述情報を含む単純なリストとは異なる。基本単位の役割を実証するために、思考実験をしてみよう。関心のあるオブジェクトは、無脊椎動物コレクションの節足動物である。このコレクションを担当する学芸員の目標は、研究機関が保有しているものを内外の人間に分かるように、所蔵品のカタログを作成するだけである。ここまでは単純で簡単に見えることだろう。し

かし、この単純な例でさえ、数多くの異なる基本単位の選択肢があり、データモデルを整理する様々な方法がある。

**基本単位としての個体**

第 1 のケースでは、基本単位としてピンで留めた昆虫の個体を選択する。電子カタログがコレクションの正確な表現となるように、それぞれの記録やそれぞれの昆虫は、固有の識別子を持た**なければならない**。これらの識別子は 2 つの機能を果たす。1 つ目は、データベースのそれぞれの記録を固有の試料と一致させる。2 つ目は、他のすべての試料と区別して、それぞれの試料を識別する。このデータモデルでは、すべての情報が試料と結び付いていることに留意しよう。ただし、識別子が記述情報と見なされ、データモデルに対してある程度選択的である場合を除く。 例えば、カタログ番号さえあれば、例えその番号に関連する分類名がなくても、試料をカタログに載せることはできる。.

**劣化したデータ モデルの基本単位**

第 2 のケースでは、コレクションは主にピンで留めた昆虫である。しかし、クモのコレクションはアルコール瓶漬けで、アザミウマや他の小型の無脊椎動物はスライドとして保有している。数百、数千の個体がアルコール瓶の中やスライド上にあり、各個体に識別子を割り当てなければ、実際には役に立たない。そこで、それぞれのピンや瓶やスライドに固有の識別子を割り当てる。論理上の基本単位は、今や個体ではなく、 ピンや瓶やスライドのコレクションが「動物標本」である。それぞれの基本単位の動物標本が、実際に関心のあるオブジェクトのコレクションを表現するという意味で、このデータモデルは**劣化して**いる。劣化したモデルは、記録する情報とその解釈に影響を与える。例えば、スライドに付けられた分類名は、スライド上の各個体の同一性を表現し、あるいはすべての個体に共通（すなわち、それらはすべて同じ科の構成員）の、最も低い下位分類を反映したより上位の分類を示す。個体総数などの総計領域や集計領域、あるいは性別リスト、瓶で示された発達段階などがある。

もちろん、劣化したモデルは、情報を個体あるいは集合基本単位の中の個体のサブセットと関連付けたいときに、最大の矛盾を示す。 例えば、1 個の瓶に入った 2 匹のクモだけに明確に基づいてタイプを発表したいとき、あるいは、魚類標本用の広口瓶に入った 1 匹の魚だけの DNA サンプリングをするときなど。しかし、この情報をデータベースに記録する方法や、動物標本を伴うこれらの事象を表現する方法は何か。答えは単純ではない。注記領域やリンク表を使用し、あるいはジルタグのようなサブユニットマーカーを使用し、あるいはサブユニットを発展させる必要があるだろう。従って、例えば 2 匹のクモを取り出して、別々の瓶に入れ、新しい固有の識別子をつけるのである。

例えば、コレクションが非常に大規模で、できるだけ速く所蔵品を知ることだけが目的のとき、それぞれの試料にタグがなければ、目的達成には時間がかかりすぎるので、それぞれの試料にラベルを付けることにする。この場合、分類名に基づいてカタログを作成しようと考えがちである。カタログは、試料総数を伴う基本単位としての固有名のリスト、あるいは記録した主要な記述情報としてそれぞれの名前を付した動物標本のリストになるだろう。別の選択肢は、基本単位としてドローアを使用することである。コレクションの数多くのキャビネットの中には数多くのドローアがある。従って、それぞれのドローア、キャビネットの中のドローアの記憶場所、各ドローアの中の試料の分類学的同一性、そして恐らく、各ドローアの中の試料や動物標本の数だけを記録してはどうだろうか。このやり方が目的に叶う場合は、実行するとよい。

劣化したモデルの問題点は、定義によれば、潜在的サブユニットについての情報を排除することである。目的にとって情報が不必要な場合は、劣化したモデルの使用は必ずしも「間違い」ではない。しかし、将来的に、既存のソリューションがニーズに合わないと判断したとき、劣化したモデルは別の基本単位に基づくモデルに容易に変換できないことを覚えておこう。例えば、ドローアの記憶場所やドローアごとの総数に基づいてシステムを開発する場合、システムに基づいて個体や動物標本を変換すると、ゼロから新しいシステムを始めるのと実質的に同じ時間とリソースがかかる、

# データモデルの焦点

カタログより複雑なデータモデルでは、モデルに用いられた基本単位や、モデルの焦点を検討しなければならない。前段の節足動物コレクションのピンで留めた試料コレクションを例として使う。モデルに基づく試料では、基本単位は試料 (または動物標本) であり、これがデータベースの焦点でもある。分類名やコレクション情報などの他の情報は、基本単位試料の特性として付加される。しかし、特に節足動物コレクションでは、1 つの有機体を収集するとき、同じコレクションイベントで、大量の他の有機体も同時に収集されることを知っている。従って、恐らく、データモデルの焦点をコレクションイベントそのものにした方が効率的である。この場合、それぞれのコレクションイベントに固有の識別子を与え、例えば、現場にいた人間、日時と場所、理由など、大量の情報をそのイベントに関連付ける。そして、コレクションイベントの概念的特性として「何を収集したか」を試料に関連付ける。

これら 2 つのモデルの違いは決定的である。例えば、最初のケースでは、それぞれの試料は固有の識別子を持っているが、焦点がコレクションイベントであれば、これは当てはまらない。コレクションイベントは、一意的に識別された多数の試料にリンクできる。あるいは、劣化した試料情報 (1,200 体のアザミウマ、200 体のゴキブリなど) にリンクできる。あるいは収集されたものと、個別に識別された試料の記述リストを合成して保有できる。

コレクションの節足動物のカタログに基づく、試料の単純なケースなら、コレクションイベントについての情報を入力することができ、逆にシステムに基づいてコレクションイベントを入力することもできることに注意しよう。その違いは、この情報をデータモデルのデータに変換する方法である。最初のケースでは、コレクションイベント情報は、それぞれの試料を自由裁量で入力し、あるいは、それらが同じイベントで収集された場合は、次から次へと記録を切り貼りすることができる。しかしどちらのケースでも、特定のコレクションイベントについて詳しい情報を与える十分なシステム情報がない。 コレクションイベントの情報を分類することは、どちらのケースでもあまり助けにならない。というのもそれぞれの試料記録についての情報がいくぶん異なる形で入力されているからである。 他方、この場合の焦点は試料なので、試料記録を追加するために、コレクションイベントについて大量に入力しなければならないという負担はない。

データ モデルに関して、単純なカタログより複雑な次の段階は、 識別された基本単位、単一の焦点、及びこの基本単位についての追加記述情報を含むものを伴うシステムである。この記述情報は、1 つ以上のリストを含むが、これらのリストは焦点情報の多値特性として処理される。焦点は、データ入力の効率を高めるために使用され、デジタル化プロジェクトの目的に沿っている。焦点は、システムに情報を記録し保持するときの優先順位を決定する。焦点はまた、システムから生まれそうな生産物の種類を決定する。

以下に、共通の情報製品や、私たちの学問分野の目的、学会で使用する情報との関係、それによって生成した情報システムの焦点のいくつかの実例を挙げる。

**動植物相**

これらは**分類群に基づく**生産物である。焦点は、生物分類の上位の集合に含まれる分類群と、言外の地理的範囲（例えば、北アメリカの植物相、クイーンズランド州の哺乳類）である。

*分類情報と命名情報*、現在一般に認められた分類群名、異名、この動植物の仲間に名前を適用したことについての議論を提供する。

*地理 (空間) 情報:*各分類群の分布情報を提供する。

*出版情報:*各分類群または各分類群名に関する参考文献情報を提供する。

各分類群の*記述 データ*

執筆者または引用機関の*備考と注釈*

各分類群の代表的または実例となる*画像*

この動植物の裏付けとして、あるいは別の面からある分類群がこの動植物に含まれると実証した、執筆者が観測し収集した試料を含む*証拠情報。*

*分類学キー:* 他で使用されていない三次情報の実例を提供する。

地図

**プレゼンス・チェックリスト**

**場所に基づく**チェックリストのデータベース。焦点は、地理上の区域と、そこに存在することが何らかの方法で記録された分類群である。

*地理(空間) 情報:*関心のある分画された区域における分布情報を提供する。

*分類情報と命名情報:*分類群情報それ自体を提供する。

*出版情報:*観察結果の参考文献情報、あるいはある場所にある分類群が存在することに関する他の考証文献を提供する。

*ステータスマーカーと存在の記述子* (例えば、移動集団、定住集団、歴史的動植物、短命な動植物、希少動植物、偶発動植物、一般動植物)

執筆者または引用機関の*備考と注釈*

動植物の種類の豊かさなどの*要約情報*

**ステータス・チェックリスト**

**保護状況に基づく**分類群リストまたは分類群集団リスト。あるいは逆に、1 つ以上の出典に裏付けられた保護状況を伴う分類群リストまたは分類群集団リスト。

*分類情報と命名情報:*分類群情報それ自体を提供する。

*出版情報:*特定の分類群や分類群集団に特定の状況を適用することに関する参考文献情報を提供する。

*保護状況マーカーと記述子*(例えば、希少種、絶滅寸前種、絶滅危惧種)

執筆者または引用機関の*備考と注釈*

**コレクションノート**

通常は、**コレクションの数に基づいた**、あるいは直接基づいた**コレクションイベント**。

*収集者とコレクションイベント:*収集した人物、収集した日時、およびコレクションの識別子を記録する。

*地理(空間)情報:*収集地点の広大な場所の詳細を提供する。

*分類情報と命名情報:*収集者の命名を提供する。

*環境情報:*収集が実行された具体的な現場の情報を提供する。
*記述情報*も含まれる。

収集者の現場ノートの中身の*画像*

**試料カタログ**

**試料に基づく、**植物標本集のようなカタログ。

*収集者とコレクションイベント:*収集した人物、収集した日時、およびコレクションの識別子を記録する。

*地理(空間)情報:*収集地点の広大な場所の詳細を提供する。

*分類情報と命名情報:*収集者の命名とその後の追加命名を提供する。

*環境情報:*収集が実行された具体的な現場の情報を提供する。

必要に応じて、*記述情報*も含まれる。

*ドナー情報:*コレクションイベントの後、試料が直接研究機関に届いていない場合、寄贈する個人や研究機関の詳細な情報が必要になる。

試料の*画像*

**トランザクションの文書化**

焦点は、オブジェクトを取得し、移動し、貸与し、交換する**トランザクションイベント**である。

*オブジェクト識別子:*トランザクションに関連する特定のオブジェクトを示す。

*コレクション管理:* トランザクションに関与した人物、その人物の連絡先、およびトランザクションの条件を記録する。

*分類情報と命名情報:* 現在の名前と、試料返還時に付けられた新しい名前を提供する。

*ドナー情報:*コレクションイベントの後で、試料が直接研究機関に届いていない場合、寄贈する個人や研究機関の詳細な情報が必要になる。

*制限:*試料に課された制限の説明。

*出版情報:*トランザクションされた試料に基づいて制作された記録文書または文書の参考文献を提供する。

## 目撃情報と観察情報

野鳥の目撃などの**観察情報**。

*観察者と観察イベント:*観察した人物、観察した日時、観察の識別子を記録する。

*地理(空間) 情報:*収集地点の広大な場所の詳細を提供する。

必要に応じて、*記述情報*も含まれる。

*分類情報と命名情報:*観察者の命名を提供する。

*環境情報:*大いに役立つが、経度と緯度は特に参考になる。

*画像:* 観察を記録する。

*出版情報:*観察に基づいて制作された記録文書または文書の参考文献を提供する。

## プロジェクトの文書化

### プロジェクトに基づく

*収集者とコレクションイベント:*収集した人物、収集した日時、およびコレクションの識別子を記録する。

*地理(空間) 情報:*収集地点の広大な場所の詳細を提供する。

*分類情報と命名情報:*収集者の命名とその後の追加命名を提供する。

*環境情報:*収集が実行された具体的な現場の情報を提供する。

必要に応じて、*記述情報*も含まれる。

プロセスと手順の情報: プロジェクトが試料を管理する方法の詳細。

*制限:*プロジェクトの一環として、試料に課された制限の説明。

*承認:*プロジェクトの一環として、コレクションを実現するために取得した許可についての詳細。

他のプロジェクト関連データ

*出版情報:*プロジェクトの一環として、制作された記録文書または文書の参考文献を提供する。

上記のリストに記されなかった他の可能性はあるものの、デジタル化プロジェクトを開始する**主な**理由は、ほぼこれらのカテゴリーのどれかに該当するだろう。しかし、何としても実現したいことは、これらの様々な活動の多くを 1 つのシステムで処理することだろう。所蔵品のカタログ作りを目指していながらも、プロジェクトを支援し、トランザクションを管理し、収集者の現場ノート情報を追跡することも望んでいる。どのように 進めて行くつもりだろうか。

これらの活動のいくつかは、カタログについて前述したような単純なデータモデルを、単に拡大するだけでも処理できる。例えば、試料が貸与されているか、また誰に貸与されているかを記録するために、いくつかの領域を増やすことができる。保護状況や、コレクションが実行されたプロジェクト名を記録するために領域を増やすことができる。しかし単純に考えると、実行するべきことは、データを整理し、様々な目的に沿って様々な方法でデータを閲覧できるような、総合的な**情報管理システム**を開発することである。モジュール設計のデータモデルが必要になるだろう。

# データモデルのモジュール設計:情報管理システム

**情報管理システム (IMS)** によって、保存した情報を 2 つ以上の焦点から閲覧できる。例えば、特定の情報管理システムでは、具体的な試料記録に焦点を合わせ、 試料を獲得したコレクションイベントについての詳細な情報にアクセスできる。また二者択一的に、コレクションイベント記録に焦点を合わせ、イベントに関連して収集したそれぞれの試料についての詳細な情報に進むことができる。情報管理システムの設計は、必然的にモジュールの設計である。この実例の場合、試料または動物標本を基本単位とする 1 つの試料モジュールがあり、独自の基本単位(コレクションイベント)を持つ、明確に分離したコレクションイベントのモジュールがある。 情報管理システムによって、試料とコレクションイベントの両方に焦点を合わせることができ、モジュールに焦点を合わせた特性の一種として、他のモジュールからの情報を処理できる。

情報管理システムでは、それぞれのモジュール記録が基本単位で構成され、大部分が各基本単位に固有の識別子を導入し、その基本単位(特性)についての追加情報もある。特定の特性データは、そのモジュールの領域に直接入力され、あるいは参照リストや他のモジュールから選択される。コレクションイベントモジュールの以下の例でも分かるように、固有の識別子は、モジュール設計のシステム内で重要になる。コレクションイベントの基本単位は、日付、場所、収集者の独自の組み合わせである。この基本単位を反映するモジュールに含まれる唯一のデータ は、コレクションイベント ID である。

1 つの焦点をもつ単純なデータモデル

試料データベース
試料 ID、試料の名前、収集日、収集場所、収集者、備考

分類学上の名前、参考文献リスト
分類群名

収集者、参考文献リスト
収集者の名前

3 つの焦点を持つモジュールデータモデル

コレクションイベントモジュール
コレクションイベント ID
収集日
収集場所
収集者

試料モジュール
試料 ID
分類群名 ID

コレクションイベント ID
備考

分類法モジュール
分類群名 ID
分類群名
科
属
種
出典
タイプ基準

収集者、参考文献リスト
収集者の名前

モジュールデータモデルを使った情報管理システムは、 学芸員のタスクや情報配信を助ける力強いツールである。また、高性能データモデルの開発は非常に時間がかかる。 しかし、データモデルが詳細であるほど、管理するべきデータが多くなり、データの相互作用も多くなるのは明らかである。例えば名前リストなど、ある種の情報が外部の情報ソースからインポートされ、あるものは組織内で付加される場合に、命名法モジュールを管理する方法は何か。更新した名前リストをインポートし、それを試料や組織内の他の命名情報と一致させる方法は何か。

# データモデルを導入する

データモデルの設計には 2 つの基本的なアプローチがある。スプレッドシートなどのフラットファイル・アプローチと、より複雑なリレーショナル・データベースモデルである。それぞれのメリットとデメリットを以下で検討する。

**フラットファイル/スプレッドシート**
フラットファイル/スプレッドシートには、簡単さというメリットがある。領域名（科、属、収集者、収集者の数など）を決めるだけで、データ入力を開始できる。短時間で設定できることがフラットファイル設計の決定的なメリットである。同様に、何かを忘れたとき、追加領域を増やすのも非常に簡単である。データを紛失した場合も、特定の領域に無理やりデータを入力する必要はない。これはリレーショナルデータベースシステムには当てはまらない。従って、通常、スプレッドシートを利用したデータ入力には時間がかからない。ただし、データ品質が犠牲になるので、データの訂正が必要になり、プロセス全体を遅延させる。

しかし、いくつかのデメリットもある。概してデータ検証のフォームがなく、すべてのデータがテキスト（数字も！）で入力され、検索要求が難しくなる。1 つの領域にいくつかの価値（例えば、命名履歴）を入力するようになっている場合、すべてのデータを 1 つの領域に入れ込むのが非常に困難であり、その結果、データ検索が困難になる。関連情報だけを表示するように、検索要求し、検索条件を追加するのは簡単ではないので、データ調査はさらに困難である。

スプレッドシートでは、ドロップダウンリストなどの単純な設計制約を導入し、特定のデータタイプを保持するために領域をフォーマットし、ある程度これらの問題に対処することが可能である。しかし、利用者がこれらの制約を回避したいと望んだ場合は、比較的容易にできる。ドロップダウンリストの保守に問題が多いのも事実である。要するに、ファイルに保存したいデータが複雑であるほど、リレーショナル・データベースを使用する方がより適切になる。

**リレーショナル・データベース 管理 システム (RDBMS)**

リレーショナル・データベース管理システム (RDBMS) は、単純なスプレッドシートと同じような方法で作成できる。表を作成し、名前やデータタイプの領域を決める。データ記入が必須の領域を設けることもでき、その領域にデータがなければ記録が拒否される。設定が終われば、表を開いてデータ入力を開始できる。この方法は、スプレッドシートのアプローチと同じ問題を生じ、設定に少し時間がかかるが、入力データの制御レベルはいくぶん高くなる。

しかし、RDBMS システムに可能なことはこれだけではない。検索リストを作成でき、キーと呼ばれる領域によってリンクした追加の表を使って、この検索リストを維持できる。キーには 2 種類ある。1 つは *一次キー*—であり、定義したい価値に通常連番の固有の価値が割り当てられている。2 つ目の *外部キー*—は、探しているデータの一次キー—を保持するためだけに作られた領域であり、一次キー—とともに主表で使用される。一次キーと外部キーが様々な表と表の関係を作るので、ここに RDBMS の「リレーショナル」部分が入る。同様に、多値領域の問題は、キー領域でリンクした独自の領域を持つ別表を作成することによって解決できる。

RDBMS システムを利用すれば、はるかに高度な方法でデータクエリーが可能である。データ検索の条件を追加して、関連情報を見つけ、必要に応じて更新するのも容易である。これによって RDBMS アプローチは、スプレッドシートよりはるかに強力になった。

この追加機能の代償は、システムの複雑さの増加である。本文書では、RDBMS システムという深いプールにほんのつま先を浸すほども深入りすることはできない。従って、読者は、開始する前に、IT スタッフに相談し、システムのデータベース設計を研究することを強く助言する。これは必然的に、プロジェクトに要する設定時間をかなり延長するが、データの確度の上昇はしばしば価値がある。ここでも RDBMS が複雑になるほど、使いやすい情報提示が一層困難になる。Microsoft Access は、表のユーザー表示に検索機能をとり込むことができるが、多くのシステムにはこの機能がなく、適切なユーザーインターフェースの設計は、追加要件になる。

**オブジェクト指向およびオブジェクト ・リレーショナル・ データベース**

オブジェクト指向およびオブジェクト・リレーショナル・データベースは、徐々に利用可能性を高めている。これらは、RDBMS オプションと同じようなメリットとデメリットを持ち、リレーショナル・データベースと同じ方法で検討した方がよいだろう。

## データモデルソリューション対プログラミングツール

データモデルは、データを保存する方法を簡単に表現したものと理解しよう。データモデルは、プロジェクトに必要なデータを反映するが、実際にデータが入力され維持される方法は反映しない。データの最新性を保持し、できるだけ最新のデータにしておくためには、ファインドやリプレースシステムなどの追加ツールが必要になる。単純なスプレッドシート・アプローチでは、現在使用中のシステムにすでに組み込まれている。これは、RDBMS システムにもある程度当てはまるが、大抵は使いやすいフロントエンドをシステムに付け足した方がよい。これは一般にデータ入力、クエリー、検索のフォームを内蔵している。通常は、領域情報の計算ツールを含み、結合データを提示する。従って、読みやすいフォーマットで提示される。システムの動作方法を明確に理解し、データ作成・維持の様々なタスクに必要なツールを判断できるスタッフを、組織内に配置することが重要である。これは書き留めておこう! この人物は、システムの管理者または専門家の立場で行動し、直面する問題の解決を図る。

## 複雑さ

データモデルの複雑さの議論の中で述べたように、詳細度や純度が高くなるほど、データモデルが一

層複雑になることを考慮しなければならない。

**多値領域**

単独の領域にいくつかの別々の実体を表示するのは珍しいことではない。収集者がその好例である。単独の領域は、主要な収集者を保持でき、あるいは以下のように収集者集団を保持することもできる。

- T. Wajima, S. Yoshizawa & T. Kitayama

これは、検索するときは特に使いやすいわけではない。そこで、以下のように、別々の価値に分割して記入したいと考えるだろう。

- T. Wajima,
- S. Yoshizawa
- T. Kitayama

価値の数が少ない場合は、領域を「収集者 1」、「収集者 2」、「収集者 3」と呼び、フラットファイルの別の領域にこれらの 1 つ 1 つを保持することが可能である。しかし、これでは、試料の大部分に使用できないほど多くの領域が必要になり、すぐに間に合わなくなる。1 つの記録が「収集者 10」の領域を必要とした場合、例えそこには決してデータを記入しなくても、すべての記録に「収集者 10」の領域ができる。必要な収集者の数だけを保存する別の表にデータを記入した方がよい。これらに、一次キーと外部キーをリンクさせる (外部キーは収集者表に保存される)。下記のダイアグラムに 2 つの記録を示した。試料キー1の記録には 3 人の収集者、試料キー2 の記録には 1 人の収集者しかいない。前の試料の 3 人の収集者の中の 1 人とたまたま同じ人物である。

| 試料表 | |
|---|---|
| 試料キー | 名前 |
| 1 | Codium latum |
| 2 | Mastocarpus yendoi |

| 試料収集者表 | | |
|---|---|---|
| コレクションキー | 試料キー | 収集者名 |
| 1 | 1 | T. Wajima |
| 2 | 1 | S. Yoshizawa |
| 3 | 1 | T. Kitayama |
| 4 | 2 | T. Kitayama |

**細分化**

収集者名の問題は、名前を集めたときにはっきり表れる。T. Kitayama の例で見ると、以下のように記されている。

- T. Kitayama
- Kitayama, T.
- Kitayama
- Kitayama; T.

- T. Kitayaama

同じ名前でも多くのバリエーションがある。もっと多くの名前のバリエーションが入力されるので、データを一貫して表示することができなくなる。この問題を解決するために、細分化と言われるプロセスで、名前をいくつかの部分に分割する。例えば、以下の表ではイニシャルと姓に分けた。

| 収集者表 | | |
|---|---|---|
| 収集者 ID | イニシャル | 姓 |
| 1 | T. | Kitayama |

この情報は、一貫した計算領域を作るために、以下のように操作される。

- “T.” + “ ” + “Kitayama”
- “Kitayama” + “; ” + “T.”

これによって、データを様々なフォーマットで一貫して表示できる。入力には追加時間がかかるが、データの確度を向上させ、すべてのプロジェクトデータに共通のフォーマットを強制する。このような細分化を、ニーズやワークフローに応じて適用するとよい。しかし、細分化してもスペルミスは訂正できないので、イニシャルの例で見た問題のすべてを解決するものではない。これは、検索リストを使用し、データを検査して、解決しなければならない。

**正常化**

ページトップの収集者表を見直してみよう。表の中に「T. Kitayama」が 2 回繰り返されていることに気づくだろう。細分化の例で見たように、多くの記録の中に繰り返しが多いと、スペルミスが生じる可能性が高くなる。ミスの可能性を削減するために、データモデルは正常化というプロセスを実行する。つまり、何度も繰り返される共通データの領域を取り出して、別の表に入れる。通常は、これが一般に認められた価値の検索表になる。冗長になるまで、繰り返しデータを正常化することは可能だが(例えば、収集者のイニシャルを切り離して別のリストに入れることは可能だが、それで得られるものは何もない)、かなりの現実主義に基づいて正常化を適用することが推奨される。1 つの表の中の 1 つの価値を変更すれば、多くの記録のエラーを訂正できるので、正常化によってデータの維持は容易になる。行を訂正することによって、あるいは記録を変更して、検索表に正しい入力を指摘することによって、スペルミスを取り除くことは可能である。

キーを使ってもう一度練習してみよう。単独の価値が記入された領域では、これは簡単である。下記のダイアグラムでは、2 つの試料があり、タイプはタイプ表のリストで定義されている。この場合、2 つの試料はアイソタイプと定義されている。

| タイプ表 | |
|---|---|
| タイプキー | 名前 |
| 1 | タイプ |
| 2 | アイソタイプ |
| 3 | クレプトタイプ |

| 試料表 | | |
|---|---|---|
| 試料キー | タイプキー | 名前 |
| 1 | 2 | Codium latum |
| 2 | 2 | Mastocarpus yendoi |
| | | |

収集者の例でも同じ原則が当てはまる。2 つの表をリンクするキーを使って、収集者を抜き出し、独自

の表を作成する。最初の実例と全く同じ結果が示される。

| 試料表 | |
|---|---|
| 試料キー | 名前 |
| 1 | Codium latum |
| 2 | Mastocarpus yendoi |

| 収集者表 | |
|---|---|
| 収集者キー | 収集者名 |
| 1 | T. Wajima |
| 2 | S. Yoshizawa |
| 3 | T. Kitayama |

| 試料収集者表 | | |
|---|---|---|
| 試料収集者キー | 試料キー | 収集者キー |
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 1 | 3 |
| 4 | 2 | 3 |

これによって数字リストだけの表ができる。一見して理解しにくいが、実際にこうなる。複雑さのレベルがここまで来ると、この複雑さを隠すために、使いやすいフォーマットでデータを表示するフォームの使用が必要になる。

データを分離すればするほど、適切な結果を得るために様々な表を正確につなぐ必要があるので、クエリーが難しくなるのも事実である。表を正しくつなぐ困難さは、本文書の範囲外であり、利用者は、このレベルの詳細設計を始める前に、データベースソリューションを詳細に研究することが強く推奨される。最後に、データが多くの表に広がるほど、コンピュータが情報を処理し、表示するのに長い時間がかかる。

**多値領域、正常化、細分化を合成する**

上記 3 つの例を組み合わせると以下のような 3 つの表になる。

| 試料表 | |
|---|---|
| 試料キー | 名前 |
| 1 | Codium latum |
| 2 | Mastocarpus yendoi |

| 収集者表 | | |
|---|---|---|
| 収集者 ID | イニシャル | 姓 |
| 1 | T. | Wajima |
| 2 | S. | Yoshizawa |
| 3 | T. | Kitayama |

| 試料収集者表 | | |
|---|---|---|
| 試料収集者キー | 試料キー | 収集者 キー |
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 1 | 3 |
| 4 | 2 | 3 |

見て分かるように、これはきわめて複雑である。しかし、システムが複雑であるほど、システムが実行していることの詳細を理解する人物に頼ることが多くなるという事実を表している。また、どのように、

いつ、利用者に提示されるかは、データベースを正しく作動させる努力の舞台裏で決まることが多い。

# 第 6 章: 特定のデータベースソリューションを決定する

本章では、データベースを選択するときに考慮するべき様々な分野について簡単に考察する。使用するデータベースを選択する前に、前段の各章を検討することを強く推奨する。何をするにしても、データベースが研究機関の IT 能力に適合することが第一である。実際に使用できないデータベースを選ぶことは、恐らくプロジェクトの失敗を保証する最短ルートである。大規模な IT インフラ要件のデータベースが必要な場合、IT システムを拡張するためのリソースも含める必要があり、プロジェクトの予算は大幅に高くなる。

どのデータベース設計を選択しても、人的要因を忘れてはならない。最重要とは言わないまでも、選択するデータベースシステムの重要な側面はユーザーインターフェースである。データ入力がやりにくければ、必然的にデータ入力に時間がかかり、スタッフは作業環境に満足できないだろう。 データベースが使いやすければ、より広く受け入れられ、データ入力がより速く進むだろう。これが、実際の試料に適切な検査をせずに、データ入力システムを評価するときのきわめて難しい側面であり、選択したデータベースの実際の検査が重要視される理由である。

個人のコレクションから本格的なマルチコレクション用の研究機関のデータベースまで、幅広い規模の数多くのコレクション・データベースがすでに市販されている(Berenson et al, 2003)。 幸い、データベースの価格も幅広い設定になっており、プロジェクト資金に適したものが見つかるだろう。しかし、大部分の市販のデータベースはかなり汎用化されているので、特定のニーズに合わせてデータベースを適応させる方法を検討する必要がある。データベースが、ほとんど要件に適合するが、厳密には適合しない場合、ニーズに合わせてシステムを若干修正するために、しばしば業者を雇うことがある。完全に新しいデータベースを構築する必要がある場合、最も速く、最も安いオプションではないが、実行したいことを厳密に設定する能力が与えられる。

設計を選択するとき、現実世界のデータがどれほど複雑か忘れてはならない。単純な例として、コレクションが作成された日付の記録がある。単純に考えると、日付は日と月と年だけの単純なものである。しかし、収集者がひねくれた考えで物事を複雑にし、月と年だけ、年だけ、範囲の広い日付などを記録する。例え日付が「5/9/1815」と記録されていても、「1815 年 9 月 5 日」を指すのか、あるいは「1815 年 5 月 9 日」を意味するのか。これは収集者への言及によって明らかになるだろう。しかし、追加研究が必要である。データ処理については第 4 章で論じた。

使用するために選択した設計は、データモデルと呼ばれる。多くの研究機関の基礎データモデルは、選択した市販のデータベースパッケージによって決定する。独自のデータベースを設計すれば、独自のデータモデルを設定することができる。データモデルは、研究機関のデータベースの外にデータを広める手段としての役割を果たすので、きわめて重要である。データモデルについては第 5 章で論じた。また、プロジェクトの要件であれば、研究機関のデータベースにデータを移動させる役割も果たす。現在の主流は ABCD[1] や Darwin Core[2] などのデータ交換スキーマであるが、数多くのデータモデルが存在する。独自のシステムを設計する場合は、これらのスキーマの検討が、考えをまとめる助けになるだろう。

データの普及はもう 1 つの問題である。市販のデータベースパッケージは、外部からデータにアクセスできる(通常はインターネット)メカニズムを備えている。これは要件に適合するのか。あるいは、別のアプリケーションを作る必要があるのか。別のアプリケーションは独自の要件にきちんと合った設

[1] http://www.tdwg.org/standards/id/81/参照
[2] http://digir.sourceforge.net/schema/conceptual/darwin/core/2.0/darwincoreWithDiGIRv1.3.xsd 参照

計ができるというメリットがあるが、同時にそれに関連する追加コストがかかる。外部のアクセスを可能にすると、特にオンラインデータ入力を利用可能にするつもりなら、それ自身がリスクをもたらす。いわゆるハッカーによるネット攻撃は、データ公開時の大きなリスクであり(Morris, 2005)、ネット攻撃に備えて適切な措置を講じた方がよい。

候補のデータベースを選択したら、実際にプロジェクトを開始する方法を検討し始めよう。これについては次章で論じる。

# どのデータベースを使うのか

平たく言えば、既存のソフトウェアパッケージをどこかで入手するか、自分で構築するかのどちらかである。どちらのアプローチにも長所と短所がある。完璧なソリューションはないということを覚えておこう。うまくいくものと、もっとうまくいって欲しいものと、まったくうまくいかないものとがある。しかし、適切に計画すれば、ほぼニーズに適合し、時間と資金とリソースの有利な投資となるソリューションを手に入れられるだろう。

**既存のパッケージ**

既存のパッケージは、市販かオープンソースのどちらかである。有償のリソースであれば、市販のパッケージの前払金と、場合によっては継続的な代金がかかり、適切な費用を払えば動作を速くできる。設計作業のすべてまたは多くはすでに済ませてある。オープンソースパッケージでも同じことが当てはまるが、費用は安く、無償のこともある。どちらの場合も、技術サポートと文書を入手することが重要である。アクティブさの程度は変化する。現在何を実装しているか。将来は何を導入する予定か。どのくらいの頻度で新しいバージョンが発売されるか。プログラムの次のバージョンに移行するのに何が関与するのか。他の誰がプログラムを使用しているのか。将来的にいつまでプログラムのサポートが受けられるのか。しかし最も重要なのは、望み通りの、期待通りの動作を実行しているのか。

既存のパッケージを使用する主なメリットは、使用契約を結ぶ前に評価できることである。使用契約を結ぶ前に、実物説明や参考文書を研究できる場合は、できれば、また少なくとも試してみよう。 特定のパッケージの販売員は、ソフトウェアの限界や欠点や欠陥に焦点を当てていないことも心に留めておこう。プロセスの評価段階で、できる限りこれらを自分で発見する必要がある。パッケージが文書に書かれていないことを実行すると思い込んではいけない。関心のある機能が期待通りの方法で実際に動くことを確かめよう。

セールストークとは関係なく、既存のパッケージが状況に必要なものと厳密に適合することは絶対にない。 あまりにも簡略化され、期待するあらゆる種類のアクティビティをサポートできない。あるいは、あまりにも複雑すぎ、管理・維持するには、利用可能なものより多くの専門技術やリソースが必要である。まったく不要であるが、プログラムを使うためだけに保持し交流する必要のある特徴を持っている。あるいは、ただ単に異なる焦点を持っているだけかもしれない。当初写真コレクションをデータベース化するために設計されたパッケージを、試料コレクションをデータベース化するために使用することはできる。しかしおそらくこれが最善の方法ではない。

既存のパッケージは常に、フレキシビリティ・ソリューション、カスタマイズ・ソリューション、アドホックソリューションが評価される。一方で、パッケージは、閲覧、報告、検索などの能力を内蔵したデータモデルの抽象システムを表現する。このシステムを使用するには、多額の開発費が必要であり、恐らくプログラミングとその利用は、システム全体を新たに設計するより若干節約になる程度である。 他方、システムはひどく柔軟性に欠け、状況の詳細に合わせてわずかな修正もできない。あるいは、事前に予測しなかったかなりの追加費用を払わなければ修正できない。これは特許で厳しく守られた市販のシステムでとりわけ顕著である。.

これら 2 つの両極の間にアドホック・カスタマイズのサポートがある。例えば、多くのパッケージには、当初の設計で予期しなかったデータに名前を付けて使用するプレースホルダー領域がある。しかし、この領域は、入力し、検索し、入力を制御し、エクスポートし、閲覧することさえ容易ではない。さらに、これらの変更は基本データモデルあるいは設定データモデルと互換性がなく、解釈が不明瞭である。

既存のパッケージを批評し、詳細にリストアップするのは本文書の範囲を超えている。GBIF は、一般に流通している既存のコレクション管理ソフトやデータ保存ソフトのソリューションの調査を委託した (Berendsohn et al 2003) 。これが良いスタート場所になるだろう。他の出発点は以下の通り。

- **生物学コレクションデータの TDWG 下位区分: 生物学コレクションの管理ソフト**:
  http://www.bgbm.org/TDWG/acc/Software.htm
- **GBIF のソフトとツールのリンク集**:
  http://www.gbif.org/links/tools
- **デジタル分類法: オープンソース生物多様性情報科学のウェブ リソース :**
  http://デジタル taxonomy.infobio.net
- **コレクションをデータベース化している検索可能な標本館のリスト:**
  http://www.cals.ncsu.edu/plantbiology/ncsc/type_links.htm

将来的には完全な形で作成することも可能だが、これらのサイトに現在リストアップされたものよりはるかに多くのプログラムが使用されている。現状で使用するパッケージを開発中であれば、関心のあるパッケージの現在の利用者に相談し、類似した他のコレクションをチェックして何を使っているか参考にするのが賢明である。

**独自のソリューションを構築する**

データベース化のニーズが比較的単純なものであれば、あるいはリソースが不足気味であれば、独自のソリューションを構築する価値がある。数百あるいは数千の試料しかないコレクションなら、1 日足らずで構築できるフラットファイルあるいはスプレッドシートソリューションでカタログを作成できる。簡単な下準備をし、品質管理問題を検討し、ABCD や DwC スキーマ (前章「標準」を参照)を調査すれば、データの潜在的利用者にも大いに役立つ、システム構築をすぐ始められる。独自のソリューション構築は良い出発点でもある。慎重に思慮深く実行すれば、後日既存のソフトウェアパッケージにデータを移動することができ、あるいは将来的にニーズやリソースへのアクセスが変更したとき、修正を施してソリューションを拡大できるだろう。

あるいは、より高度なソリューションを構築するニーズ、リソース、専門知識へのアクセスを保有しているかも知れない。独自のソリューション構築は、具体的なニーズに合わせてソリューションを調整する柔軟性を与えるだろう。コンピュータ、プログラミングの専門知識、実行する時間があれば、ゼロから独自のソリューションを実際に構築するのは、希望にぴったり合わせられる最善の方法である。しかし、精巧な情報システムを設計し構築するとき、それに要する時間と労力の量を過小評価しがちであることに注意しよう。もう 1 つのソリューションは、既存のパッケージをニーズに合わせてカスタマイズすることである。 希望に十分近く、希望通りに十分カスタマイズできるオープンソースパッケージが手に入ったら、これも満足できる方向に進むだろう。

ほとんどの場合、「自分で構築する」のは、実際には、データベースや情報管理システムを構築する人間あるいはスタッフを配置するという意味である。状況にもよるが、意欲のある、十分熟練した大学院生に構築させる単純なものから、多額の投資をするものまである。あるいは、民間企業と契約して、その現有スタッフとリソースを使ってソリューションを設計し導入する、あるいはそのスタッフを直接雇用することもできる。あるいは、研究機関の現有 IT スタッフにプロセスの開発と構築という職務を任

せることもできる。この方向に進む場合、2 つの重要な留意事項がある。1 つは、開発者に希望とニーズを伝える能力が、獲得する品質を大きく左右するという点である。情報管理システムに求めるものを明確に述べることができない場合、あるいは開発者にこれを伝達する時間がない場合、希望のものを構築するための適切な専門知識を見つけるのに苦労する。2 つ目は、ソリューション開発者が、生物学データや博物館コレクションデータの性質を十分理解していることが非常に重要である。データベースプログラミングの専門知識は絶対的に不足している。 理論的には非常によくできているが、作業で使用するデータの現実や、ソリューションが強化するべき現実世界のワークフローに直面したとき、破綻するソリューションを何度も目にした。ソリューションが破綻すると、忠実なデータベース開発者は、しばしば、彼らが構築したデータシステムに適合するデータに変更しようと、説得を試みるだろう。別のケースでは、彼らは、常に意図したほど完璧ではない、あるいは時間と労力と資金を浪費しただけで決して完成しない、見事なデータベースソリューションのデータ問題を「解決」しようと努めるだろう。

Morris (2005) は、生物多様性情報科学の分野に限ったリレーショナル・データベースの設計と導入に関連する多くの問題を論じている。これは明らかに、計画段階で参考にし、開発スタッフにも推奨する価値のある文書である。

自分で構築するソリューションのメリットは、結果の透明性が高く、市販製品を使用するより容易に修正できることである。これは常に当てはまることではない。専門知識と開発方法にもよるが、自家製ソリューションには、他のソリューションに比べて多くのブラックボックスがある。しかも、時間とともに必要になるメンテナンスやアップグレードの方法を事前に検討しておかなければならない。新しいオペレーションシステムが現れたときまだ使えるのだろうか。ソースコードは利用可能で、アクセス可能なのか。バージョンをコンパイルするだけでよいのか。最後に、文書作成のためにどのような努力をすればよいのか。データベースについての文書がまったくないか、ほとんどない場合、経験豊富なスタッフが離職したとき、新しい職員にソリューションの使用法を教える方法をしっかり検討しておかなければならない。

## 優れたデータベースソリューションの特徴とは

市販のパッケージやオープンソースソフトウェアを調査し、あるいは独自のソリューション構築を検討する場合も、提案システムを評価する基準を理解することが重要である。最終的に、「優れた」ソリューションは、ニーズに適合し、うまく機能するものである。以下の質問リストは、どのソリューションが自社に合っているかを評価する基準の設定に役立つことを目的としている。

**実際の費用は**

費用には、開発の初期費用、ハードウェア費用、ライセンス、メンテナンス、アップグレード、追加ソフトウェアの要求などが含まれる。システムの正常な稼働を保持する専門技術の入手費用も含まれる。

**安定性**

優れたデータベースソリューションは、3 つの点で安定している。1 つ目は、現在進行中の作業や、圧倒的件数の「近日中に発生する作業リスト」も含めて、ほとんどの目的を果たすという点である。提案ソフトウェアに実行可能なことと、実際に実行していることを対比して明確にしておこう。 特に、どのようにワークフローに影響するかを含めて、どのように修正やアップグレードを実施するかを明確にしておこう。また、変更した、あるいは追加した特徴を検査する方法も評価しておこう。実在のサンプルデータで予備テストをするのか。あるいは、最善の結果を期待して実在データを入力したとき、変更や追加がうまく作動するか検査する必要があるのか。

2 つ目は、ソフトウェアがバデッグ済みかという点である。プログラムが強制終了するまでの堅牢さはどの程度か。特に関心があるのは、プログラムが強制終了したときのデータの破損度を判断すること

である。今後プログラムにバグが見つかった場合、どうすれば解決できるのか。 パッケージの一部としてデータをバックアップするのか。あるいは外付けのバックアップシステムが必要なのか。

3 つ目は、コンピュータアーキテクチャ、オペレーティングシステム、データベースプログラム、ネットワークプロトコル、プログラミング言語などが将来変更されたとき、プログラムがどんなサポートを受けられるのかという点である。技術は急速に変化しており、データベースソリューションが現在のフォームで存続できる期間には限りがあると考えるのが現実的である。できれば年単位で考えたいところだが、ソリューションによっては、数か月で時代遅れになり、最悪の場合は構築された時点で時代遅れという可能性もある。

**優れた文書化と技術サポートはあるのか**

高性能の精巧な情報管理システムでも、使用法が理解できなければ、何の役にも立たない。文書を作成して、ソリューションに実行可能なことと、その使用法を記録するべきである。指導書も含まれるだろう。技術サポートには、セットアップ時の問題、使用法、潜在的バグへの対応と報告などが含まれる。技術サポートの入手方法、利用可能な程度、応答時間、初期費用や導入後の費用の金額などを必ず調査しておこう。コンピュータハードウェアからバックエンドデータベース、フロントエンド、導入に至るまで、パッケージ全体の文書化とサポートのニーズをよく検討しよう。何らかの形でデータモデルを閲覧できる方がよい。

**期待通りの性能があるのか**

きわめて単純な機能を命令したとき、コンピュータが不安定で、延々と待たされることほど苛立つことはない。パフォーマンスが遅い理由は数多くある。コンピュータまたはサーバーのパフォーマンスが遅い（処理速度が遅い）、ハードドライブが満杯である、あるいはメモリー容量が一杯であるなど。あるいは、コンピュータのパフォーマンスは十分速いが、プログラムが遅いだけということもある。いくつかのルーティンは他の仕事に比べて本質的に効率が悪いので、貧弱なプログラミングによってパフォーマンスが遅くなることがある。いくつかのソリューションは、記録変更履歴や検証プロセスなどのバックグラウンドアクティビティを実行しているので、それがパフォーマンスを遅くする。いくつかの導入には、パフォーマンスを低下させないために、例えば、手動でインデックスを再構築し、データベースを「バキューミング」するなど、定期的なメンテナンスが必要である。 ソリューションが、その機能の一環としてネットワーク接続を可能にし、あるいは要求する場合、特定のソリューションを実行しているかどうか判断するときは、ネットワークの信頼性と速度に取り組む必要もある。

いくつかのパフォーマンス問題は、システム内のデータ量の規模に依存する。例えば、特定のソリューションでは、2,000 件の記録しかない時はよく働くが、記録が 200,000 件あると、検索、インポート、報告、エクスポートに延々と時間がかかる。

**学習曲線とは何か**

箱から出してそのまま使えるプログラムはほとんどない。プログラムを適切に使用するには、トレーニングを受け、参考文書を研究しなければならない。一般に、プログラムの機能が多いほど、使用するための学習曲線が長くなる。しかし、学習曲線に影響を及ぼす他の要因がある。いくつかの設計は、他と比べてより直観的である。ボタンに依存するナビゲーションや機能性、メニュー項目、キーストロークコマンドは、どのように学習曲線に影響を与えるのか。データ検索やデータ閲覧はどの程度簡単なのか。報告書を作成するために検索を行い、スクリプトを書くには、SQL の知識が必要なのか。特定のプログラムの機能性と、すでに知っているプログラムの機能性は、どの程度類似しているのか。

長い学習曲線は、例え参考文書が優れていて、提案プログラムが最終的にニーズに合った妥当なソリューションであっても、欲求不満の元である。提案ソリューションを評価する上で最も難しい部分は、例え「優れた」プログラムでも、それを使用するための学習に要する、時間と労力を費やす価値があるかどうか判断することである。

**システムに最初にデータを追加する方法とは**

すべての必要なデータをシステムに入れる方法を前もって考えておく必要がある。 通常、私たちの主要な焦点は、試料情報であるが、コレクションイベント、プロジェクト、出版物、命名法、これらと他の種類の情報との組み合わせにも関心がある。この種の情報を入力する場合、他の種類の情報を入力する前だろうか。例えば、試料を入力する前に、収集者名や分類群名のリストを入力する必要があるだろうか。これが必要な場合は、これらのデータをどこで入手するのか。システムからデータが提供されるだろうか。システムは、どのような種類の情報を参照基準に使うだろうか (すなわち、ドロップダウンリストの価値のスタティック型リスト)。どれが修正可能なのか。修正する場合は、誰がどのように実行するのか。既存の情報ソースからどのデータが入手できるのか。どのデータを自分でコンパイルする必要があるのか。

**データを入力・編集・閲覧・削除する方法とは**

これら 4 つの機能が、システムの初めから終りまでに、どのように発生するか慎重に検討しよう。例えば、データ入力は一方向で行われるが、既存記録の編集は別々に、あるいは別のデータ表示で行われるなど、別々に発生するのか。誰がデータを削除できるのか。偶然に削除されるのはどの程度簡単なのか。情報が入力され、変更され、1 回で削除されるとき、モジュールとモジュールの間のリンクはどのように保持されるのか。どの情報を有効な記録に入力する必要があるのか。不完全な情報しかなかったら、どうするのか。多くの試料で常に不変の、データ入力を助けるショートカットはあるのか。固有の組み合わせでデータを閲覧するのはどの程度簡単なのか。

**プログラムを容易にナビゲートできるのか**

情報の多重表示や、多機能モジュールを伴う最新のリレーショナル・データベースを利用できれば、非常に高性能である。ナビゲーションの容易さは、使いやすさの重要な特徴である。データ入力画面は、直観的タブオーダーが可能であり、通常、領域から領域に容易にフローできるべきだ。通常、特定の記録に対して、少なくとも読みやすいサイズのフォントで、単一画面に適合するより多くの入力・閲覧領域があるので、より完全な記録情報を見るためにどのようにナビゲートするか調べておこう。単一記録と多重記録リストの表示の間をどのように移動するのか。モジュール同士がリンクしているとき、あるいはモジュールが個々に独立しているとき、モジュールとモジュールの間をどのように移動するのか。例えば、試料記録を見ているとき、分類名、収集者、場所、コレクションイベントについての詳細な情報にどのようにアクセスするのか。データ入力機能と、ラベル作成やトランザクション記録モジュールの間をどのようにナビゲートするのか。データベースが単純な、直観的な、あるいは実用的でさえあるナビゲーションシステムを持っている場合、プログラムは日常業務の満足度を高めるだろう。不格好な、あるいは見るからにランダムなナビゲーションは、うまく行っても学習曲線を大幅に延長し、最悪の場合は生産性を著しく低下させ、頭痛の種になる。

**ソリューションがデータ 品質を改善する方法とは**

スプレッドシートより高度なほとんどすべてのソリューションには、記録の品質を改善する特徴が期待される。これはいくつかの領域のドロップダウン選択を必要とする。計算領域では、再入力しなくても様々な方法でデータを使用することができ、領域レベルの検証は、入力データが領域の最小限の期待に適合していることを保証する。

あるいは、プログラムが現状より高いデータ品質を期待する場合、品質の悪いデータを入力できるだろうか。例えば、収集日の領域が「10/32/1964」の入力を阻むかもしれない。また、収集日が「summer 1964」であれば、これを入力できるだろうか。あるいはアドホックソリューションを使わざるを得ないのだろうか。優れたプログラムは、微妙なバランスをとり、データ入力エラーを却下し、あるいは注意を促すのが望ましい。しかし、品質の悪いデータを無効にする、別の入力方法を使った方がよい。

データ品質のもう 1 つの側面は、内蔵検証プロセスと品質管理プロセスのサポートに関連する。品質の悪いデータを識別し、データ入力後の検査のために回収できるのか。管理検査や専門検査の対象になる記録を選び出せるのか。特定のデータ入力担当職員とともに、特定のデータ入力問題に管理者が関与できるのか。管理者は、例え変更されたとしても、いつ記録が変更されたのかを判断し、どの情報が変更に関与したのかを即座に判断できるのか。

**どのようなインポート機能があるのか**

データがデータベースに直接入力され、あるいは情報が外部ソースからインポートされるという期待は、プログラムごとに大きく異なる。相当量のレガシーデータがある場合、どのようにプログラムに取り込んでいくのか。インポートする前にどのようにフォーマットするのか。インポートは、新しいソリューション構成時の初期機能としてのみ取り組まれるのか。あるいはプログラムは、将来的に新しいデータのインポートを容易にサポートするのか。いくつかのプログラムはさらに先に進んでいる。インポートされたフォーマットにプログラムの外からデータ入力することを模索している。あるものは、収集者用のスタンドアロンデータ入力モジュールを提供する。こうした状況で、インポートデータやインポート可能なデータのデータ品質問題への取り組みを検査することが重要である。インポートする前に取り組まなければならないのか。そうだとすれば、メインプログラムの外にデータ品質の評価を助けるツールはあるのか。インポートログを使ってインポート中に品質評価に取り組まなければならないのか。あるいは、エラーリポートをデータ問題に取り込むのか。システム内に特定のインポートセットに属す記録を区別するマーカーがあるのか。システム内に品質を改善するツールがあるのか。あるいは、エラーが多すぎる場合、取り除くツールはあるのか。

**どのようなエクスポート＆報告機能があるのか**

エクスポート＆報告機能も、プログラムごとに大きく異なる。プレパッケージの報告は、一般に予想される出力を容易にするために含まれるが、固有の報告を出力する機能も持つべきだ。必要なときは紙に印刷し、あるいは印刷機能が電子出力用にも使用できる場合に、報告をフォーマットする方法を検討しておこう。 Pdf、html、xml、エンコード選択をはじめとする、出力選択肢のサポートを検討しておこう。 UTF-8 、ISO-8859-1、ACII テキストでデータをエクスポートできるのか。エクスポート中に使用するエンコードを見分けることができるのか。

ウェブベースのデータアクセスに対するサポートは、現在の多くのデータベース利用者にとって重要である。極端な例として、データ入力から一般人の表示アクセスまで、データベースとのインタラクションのすべてまたは大部分が、ウェブインターフェースを経由する。一体型ソリューションは、セキュリティとパフォーマンスに大きな影響を及ぼす。あるいは、メインのデータベースは、別のサーバーでウェブ表示するための適切なフォーマットにデータをインタラクトまたはエクスポートできる。これはいくつかの問題を解決するが、ウェブ上の存在感の開発・維持に関連する追加要件に取り組む必要がある。

データベースが同じ目的のデータベースを連合したノードとして機能するような、自動的または半自動的データ取得を可能にすることが、今日の関心の的である。そのためには少なくとも、ABCD や Darwin Core などの既知のスキーマにエクスポートするための機能が必要である。これはおそらく、DiGiR や TAPIR などのデータポータルに対するサポートや、現在のリフレッシュした互換性のあるデータ表示を維持するために必要なスクリプトにも関連するだろう。いくつかのプログラムは、こうした機能をプログラムにパッケージし、あるいはこれらの拡張機能を自分で開発するためのサポートを提供する。

リレーショナル・データベースからデータをエクスポート＆報告することは、基礎データモデルが複雑になると次第に複雑になる。つまりエクスポートに時間がかかるようになる。いくつかのプログラムからフラットファイルの Darwin Core エクスポートを準備するだけで、記録数が比較的少ないときでも、24 時間もの処理時間がかかる。これはまた、大急ぎで報告やエクスポートを生成する時に、単に気をくじく実行困難なタスクになるを意味する。定型の報告やエクスポートは有益であるが、新しい報告

＆エクスポートの生成や、実行に必要なツール、スクリプト、言語などに関連する詳細事項は必ず学んでおこう。

**ソリューションは、ネットワークや多重アクセスにどのようなサポートができるのか**

次第に一般的ではなくなってきているが、いくつかのソリューションは、一度に 1 人のローカルユーザだけが使用する 1 つのコンピュータに存在する。高度なソリューションは、複数の利用者を追跡し、あるいは異なるクラスの利用者に異なる権利を付与できる。大規模プロジェクトは、複数の同時アクセスをサポートすることを目指している。例えば、1 人以上の人間が同時にデータ入力できる。あるいは、データ入力は別の場所で実行され、コレクション管理がこれを使用して、事務所からのローンを処理する。 一般に、1 つが「本物のコピー」で、他のコピーが他のスタッフたちに配布され、検索や報告や他の機能を利用するなど、データベースの複数のコピーが流布するのは良い考えではない。

**可能であれば、どのようにカスタマイズするのか**

データベースを完全に自分で構築しない限り、ソリューションにはある程度の「ブラックボックス」がある。つまり、どのように動作するか分からない機能がある（ただし、機能するものとして）。実行した方法で機能する理由が分からないこともある。恐らくトレードオフが関与しているのだろう。 望んだ方法で X が機能した場合、恐らく Y はそれほどうまく機能しないか、まったく機能しない。プログラムの実際の欠陥や、プログラムを正しく使うために身につけるべきことを理解するために、評価期間や学習曲線の大部分を費やすことになる。

いずれにしても、ニーズが変化し、あるいはニーズの理解が深まった時点で、プログラムが機能する方法をカスタマイズする必要がある。特定のワークフローのニーズに合わせて若干微調整する必要がある。レガシープロトコルと互換性のあるドロップダウン価値が必要になる。より多くのリソースが追加可能になったとき、まったく新しい機能のモジュールを追加する必要がある。可能なレベルまで、事前に確認しておいた方がよい。こうした変化に対して提案ソリューションはどれほど柔軟になれるのか。カスタマイズを可能にするどのような内蔵機能を持っているのか。領域を特定の表示で再編成できるのか。領域間でタブオーダーを変更できるのか。ある種のタスクに新しい表示を作れるのか。それが可能ならその方法は。基礎的データモデルを修正できるのか。あるいは、比較的静的で不変のモデルとインタラクトする方法だけを変更するのか。

特注プログラムとオープンソースプログラムには、正反対の問題があるが、プログラムの機能を邪魔する変更をうっかり起こしがちである。データ入力担当者がグローバルナビゲーションの基盤になるスクリプトを変更し、あるいは表示を削除できるのか。

## 自分に合った焦点と機能があるのか

プログラムは見事に機能するが、自分のニーズに合っていなければ、優れたソリューションとは言えない。第 5 章の考察を受けて、将来を見越したプログラムは、自分にとって重要なことに焦点を合わせた方がよい。芸術作品、建築、人類学的アイテムを含む、あらゆる種類の博物館オブジェクトを管理するために設計されたプログラムは、ピンで留めた昆虫コレクションには適さないかもしれない。適切なカスタマイズをしなければ、記録したい情報と何の関連もない領域をナビゲートして、データ入力時間の多くを費やしてしまうだろう。最悪の場合、任務を果たすために、必要な、期待する方法でアクセスし、検索することができないデータモデルに情報を入れるために、すべての時間を費やすだろう。

試料やコレクションの詳細情報を記録できるプログラムをすでに見つけたかもしれないが、本当に必要なのは、様々なプロジェクトのすべての詳細と、これらのプロジェクトに伴って作成した文献を追跡できるものである。プログラムが望み通りにプロジェクトに焦点を当てていない場合、すべての努力が期待したより小さな成果しか挙げないだろう。

既存のソリューションを評価し、あるいは独自のソリューションの構築を準備するときは、入力し、保持し、システムに保有したいデータを列挙しよう。自分の情報システムの焦点は何か。複数の焦点があるのか。関心のあるオブジェクト (第 4 章参照) の中で、一次オブジェクト情報とは何か。二次オブジェクト情報とは何か。一次オブジェクト情報の補助的情報とは何か。参照情報として何を利用するのか。

高度な情報管理システムが保有する、あるいは保有したい機能は数多くある。いくつかの共通機能と、私たちの学界が関心のあるデータ問題を以下で論じる。システムの特質、複雑さ、焦点にもよるが、これらは、必ずしも他の方法より良くも悪くもない、様々な方法で管理される。評価期間中あるいはシステム開発計画中に、これらをどのように管理するかを詳細に検討することが重要である。

**命名情報を管理する**

命名情報は、データベースシステムのデータに変換するのが信じられないほど難しい。分類名は、ある意味で、1 つのラベルであり、あるいは試料の特性である。例えば、X は分類群「Y」の試料である。しかし、これは解釈の対象でもある。ラベルに記された名前があり、 スペルが正しかったり間違っていたり、出典や階級が一貫したフォーマットであったり一貫していなかったり、そして、期待される、または正しいフォーマットと一致するように浄化された名前の「一般に認められた」バージョンがある。

分類名をパースするのは必ずしも簡単ではない。ラベルに記された名前は、様々な分類学的表現で示される。例えば、「未確認の節足動物」や「Rosa alba subsp. alba forma angustifolia」など。いくつかの分類群では、「種名」にだいたい二名法（属名＋種名）や三名法（属名＋種小名＋亜種小名）をよく使っている。 しかし他では、特に植物でははるかに複雑である。例えば、亜種と変種が同じ階級に入ることもあり (すなわち、Rosa alba subspecies alba or R. alba var. alba) 、あるいは別の階級に入ることもある (R. alba ssp. alba var. ternata)。交配種や栽培種の分類群は、命名法の出典に含めるとき、さらに複雑さを加える。

厳密には名前ではない情報がラベル名に記されていることもある。例えば「Rosa fulva (新種か?)」や、「好意的に見て Rosa Alba.に一致」 など。こうした情報を入力できるシステムもあり、できないシステムもある。入力できる場合は、付随修正キーを使わずに、同じ名前に関連付けるのは簡単ではない。

次に、分類学的階級の問題がある。ラベルに記された名前は「界」までのリンク名の階級内の同定を表現する。特定の階級内の種の配置は、解釈の対象であり、一般的に同じ種に複数の階級が適用される。これらの階級は、必ずしも同じ階級構造に従っていない。それがひどく悪い選択ではない場合、大部分の分類学的階級制度において、常に配置が不確かな分類群があり、その結果、次の上位階級にリンクされないまま放置される。

異名は大きな問題である。名前と名前は階級制度によって互いに関連付けられるが、部分的または全体的に同じ価値である。異名を管理すると、より多くの名前をシステムに保持するだけでなく、両者の関係を保持し、それらを操作するメカニズムを開発し、現在認められた一貫したステータスをシステム全体で識別し、適用する。関連する問題は一般名である。システムによって、それぞれの学名と結び付いた 1 つ以上の一般名を使用できるものと、使用できないものがある。 これらを維持するには追加の管理負担がある。

特定のソリューションが命名法を管理する方法は、潜在的に適用できる名前がどれほど多いかという期待に関係する。ドロップダウンリストは、アイオワ州の哺乳動物のデータベースについてはうまく機能するが、北アメリカの節足動物についてはまったく機能しない。

タイプ試料は、試料のタイプの名前がラベル名やその分類群の現在一般に認められている名前と同じであり得る、また同じでない可能性があるという点において、追加の複雑さを加える。

これらすべての考察は、命名情報を管理する完璧なソリューションを獲得あるいは開発することが現

実的でないということを明確にした。その代わり、ニーズを比較的うまく管理でき、異常な状況を管理する十分な柔軟性を持ち、極端に重い管理負担をかけないソリューションを探そう。

**命名の変更を追跡する**

ラベル名は解釈の対象であり、後日改訂されることがある。後日専門家が試料を見直し(「命名」または「注釈」)、あるいは発表済み原稿やデータ処理(「命名の更新」)に成文化された通り、新しい分類名の解釈を適用する。特定の名前は、単なる間違いや、データ入力中にミスタイプされた可能性があり、その場合、訂正する必要がある。様々な理由から、試料に付けられた名前の変更を追跡し、変更日と変更者を追跡するのは有益である。変更された理由を追跡するのも有益である。例えば、誤字の訂正と、専門家の命名を区別する。すべての命名を追跡するか、あるいは最初の名前と最新の名前だけを追跡するのが有益である。

**ラベルやタグを作成する**

コレクション管理プログラムはしばしば、試料のラベルやタグを作成する。これは通常、植物標本によく見られる。サイズが小さいので、ピンで留めた節足動物のラベルはしばしば、省略形や、すべての情報の短縮形で示される。例えば、データベースの情報から自動的に作成するのが難しい場合など。コレクション管理プログラムがラベルを作成できる場合、ワークフローとのインタラクションは慎重に検討した方がよい。例えば、新しいラベルがデータベース化された試料と結び付いたとき、新たにラベルを付けた試料と、まだデータベース化する必要のあるラベル付き資料とをはっきり区別しておく。

**データ管理とトランザクションを追跡する**

データ管理情報は、試料が現在置かれている物理的場所を示す単純な情報である。入念な処理段階を経由して、コレクションの中に最終的に配置されるまでの、オブジェクトの詳細な足跡を示す。適切な識別子を保持することや、処理段階を通してそれを関連付けることから複雑さが生じる。受け入れは、最初はすぐに識別できる個別のオブジェクトを含むが、多くの場合、オブジェクトがコレクションの中に配置される前の、領域確保、ソーティング、そして後日の詳細な識別に関与する。

トランザクションは、ローン、交換、および同定のために送信された試料の追跡に関与する。トランザクションの追跡は、担当者、研究機関、方針、文書についての情報を伴う。いくつかのシステムやモデルでは、トランザクションには、ホスト研究機関内で 1 つのコレクションから別のコレクションに試料を移動させること (例えば、ティーチング・コレクション)、ホスト研究機関の贈与、売却または紛失した試料の記録をつけることが含まれる。

**保護必要データに印をつける**

様々な理由で、記録に保護必要印をつける必要がある。例えば、希少種である場合や、商取引生物コレクションのおそれのある場合や、進行中の研究の証拠物件となる場合である。記録に保護必要印をつけることは、データ制約ポリシーの開発や導入を可能にし、続いてデータ制約ポリシーは保護必要データについて何ができるかを決定する。データ制約ポリシーは、記録全体、記録内のある種の情報タイプ、あるいは記録レベルと領域レベルのデータアクセスの組み合わせに適用される。

記録に保護必要印をつけることは、オブジェクト記録表に保護必要チェック領域を入れるだけの単純なものであるが、精巧なシステムは注記を記録するメカニズムを持っている。記録ごとの複数の保護必要マーカー、記録に保護必要の印をつけた人物、記録が保護必要扱いされる理由とその期間などの注記である。

記録に 1 回に 1 件保護必要印をつけることは、特に記録セットの規模が大きい場合は、集中的な処理となる。 記録を保護必要基準に一致させるのは、現在の保護必要情報を保持するのと同様に困難であり、マッチングプロセスそのものも困難になる。例えば、連邦が指定する絶滅寸前種や絶滅危惧種のリストにあるすべての試料に印をつければよいが、これらの指定はしばしば分類群ではなく生

物集団に適用される。大部分のデータモデルでは、この詳細を保存するのが困難である。リスト上の名前から試料ラベルに記された異名まで、保護必要を転送するのは問題がある。

**記録の変更を追跡する**

理想的なシステムは、何を変更したか、誰が変更したか、いつ変更したか、なぜ変更したかを、それぞれの領域で追跡する。これは一般に、システムに保有するデータを実質的に膨らませるので、非実用的なソリューションである。保持する責任のある記録の「作成日時」や「更新日時」領域は、良い出発点である。分類群名の変更や命名は、情報が異なるので、別のシステムに入れた方がよい。命名者の名前を入力するスタッフとは異なる方法で、命名者は名前を変更する。

Darwin Core のようなデータ交換標準では、それぞれの記録の「最終更新日」領域を見たいと期待する。しかし、エクスポートされた更新日が、エクスポート領域セットではなく、領域の更新で始まる場合、この領域の解釈はさらに複雑になる。

**ジオリファレンシングを可能にし支援する**

ジオリファレンシングとは、場所記述をマッピング可能な表現に変換するプロセスである (Chapman and Wieczorek 2006)。レガシーデータのジオリファレンスを可能にするので次第に有益になっている。(Beaman et al. 2004)。ジオリファレンスがコレクションイベントの実行場所についての仮説を表現することを認識することも重要である。そのため、特定のコレクションイベントや試料について複数のジオリファレンスがある。様々な方法 (例えば、MaNIS や BioGeomancer プロトコル) を使用し、様々な検査レベル(すなわち、初期出力や、専門家や収集者による審査の結果)を使用して表現される。正式のジオリファレンシングの結果と、当初試料から届いた座標情報とを見分けるためにも有益である。

**収集日を管理する**

収集日は情報管理システムで問題を起こしやすい。というのも常にある一日を正確に示すことはないからである。単一の日を記入する領域に入力を強制するシステムは、実際にデータが示すものではなく、より正確な体裁を求めるので避けるべきである。テキスト領域で収集日を管理すると、情報を文字通りに入力できるが、結果として生じるデータは日付として有益ではなく、同じ日付の入力価値に幅ができる (例えば、Aug. 23, 1976 と 10/23/1976)。スラッシュで区切られた日付には、曖昧さという別の問題がある。どちらの数字が「日」を表し、どちらが「月」を表すのか。Morris (2005) が、日付問題について詳しく論じている。一般に、優れたソリューションには、単一の日付、幅のある日付、文字情報（すなわち、「Spring 1976」)などが入力できる多くの領域が必要である。データに日付表示が可能な場合は、少なくとも「年」情報を記入できるとよい。

**地理管理ユニットを管理する**

国名、州/県、郡/地区などの領域のある地理管理ユニット（GAUs）を記録するのは簡単そうに見えるが、必ずしもそうではない。GAUs は、例えば「Soviet Republic」を分割して変更し、「Rhodesia」という地名を変更し、最悪の場合は地理的範囲も変更する (ニューメキシコ州バレンシア郡は 1978 年に新たにシボラ郡と縮小したバレンシア郡に分割再編された)。GAUs は、一般に使用される様々な地名を使用し (すなわち、「United States」と「U.S.A.」) 、通常、様々な言語によってまちまちである。場所情報は、これら 3 つの領域 (すなわち、イギリス、ハワイ島、グリーンランド) の1つに記入するのが難しいユニットである。地方の行政支配権は、その地方そのものとは異なる (すなわち、マルティニ―ク島)。いくつかのコレクションは、どの行政単位にも属さない場所から来ている (例えば、「太平洋上ハワイ島の南方 700 マイル地点」。行政単位の境界となる地物から来るものもある。例えば、2 つの州を分ける川や、2 つの郡を分ける尾根など。

GAUs を管理するソリューションは、この情報を別々の領域に入れず、一般的な場所領域に入れる。しかし、通常、これはあまり満足できない。GAUs は一般に検索基準として使用され、場所領域へのフリーエントリーを可能にするが、はるかに多くのキーストロークが必要であり、タイプミスの元になる。

限定された地理的範囲から来るコレクションは、これらの問題のほとんどが比較的少ない。所蔵品が世界的なものであれば、GUAs を管理する方法に注意を払うことが重要である。

**評価するべき他の特徴と問題**

上記は、潜在的パッケージや開発計画がニーズに十分適合するかどうかを評価するとき、詳細まで注意するべき特徴とデータ問題の一部である。同様の詳細な注意が必要な他の分野を以下に挙げる。

収集者の名前と収集者団体

場所、生態学的記述、関連種、他のコレクションイベント情報によるデータの分類と合成

形態学と試料作成情報

観察記録

画像

プロジェクトと証拠情報

出版物と文献

研究機関と コレクションのメタデータ

セキュリティとアクセス

# 付録 A: 投資対効果検討書

**コレクションをデジタル化する理由**

データを広く普及させるため

様々な方法でデータを研究することが可能になる

学芸員の業務を向上させる

試料を保護する.

未来の転写時間を削減して研究を支援する

研究機関の組織目標に適合する

研究機関が伝統的な権限を越えた分野に貢献する能力を強化する.

# 目標を設定する

**研究機関の目標と個人の目標**

研究機関の目標　　多くの人がデータ入力し、広範な試料に対処できるデータベース

個人の目標　　データ標準に適合し、個人の試料に対処できるデータベース

**ソリューションの主要なクライアントは誰か**

具体的なプロジェクトに携わる個人

一般の研究者

研究機関の学芸員

その他

**サポートする言語数**

処理する言語数が増えるだけ複雑になる

**データの数**

記録の数はデジタル化に要する時間と、情報を保管できるデータベースのスケールに影響する。

**データ品質とは何か**

記録の種類:

コレクションデータ

分類学的情報

格納場所

生息地情報

初期記述

試料そのもの

**データ収集またはデータ解釈**

データ収集　試料に書き記されたデータを記録する

データ解釈　不正確に命名された試料など、エラーを訂正するためにデータを修正する

**研究機関の既存の慣行を向上させているのか**

デジタル化が研究機関の学芸員を支援する方法を文書化する

**画像化**

画像化の対象

画像の入手方法

画像の精度

画像を保管するフォーマット (JPG、TIFF など)

格納場所

アクセス方法

**デジタル化できないものを理解する**

データベース作りは費用を節約できるオプションではない

コレクションのデジタル化は新たな情報をもたらさない

現在でも試料は物理的に保管され扱われる必要がある

**いつまでにデータベースを利用可能にしたいのか**

短期　6〜12 か月で完了できる作業

中期　約 18 か月で完了できるデータ入力

長期　18 か月以上続くあらゆるプロジェクト

**未来の要件**

現在のプロジェクトが終了した後、データベースはどうなるのか

**スタッフの配置**

**誰がデジタル化を担当するのか**

- 学芸員が通常常務の一部として担当する
- 外部の契約スタッフ/契約企業
- ボランティア・スタッフ
- 客員研究員
- プロジェクト・スタッフ

**何人のスタッフが同時にデータベース作りを担当するのか**

- 1 人のスタッフが 1 つのデータベースを担当する
- 数人のスタッフが個別のデータベースを使用する
- 数人のスタッフが同じデータベースを分担する

**専門家の支援はあるか**

- はい　　プロジェクトははるかに円滑に進む
- いいえ　　プロジェクトに役立つ専門知識がどれほど重要かを検討する

**適切な専門知識は利用可能か**

- データ所有者
- データ専門家
- 技術スタッフ
- プロジェクト管理

**制約**

**データへのアクセスは何らかの制約を受けているか**

- はい　　未公開の領域/試料は何か、またなぜ未公開なのか
- いいえ　　データに制約がないための結果を検討する

**研究機関は既存のシステムの使用を要求しているか**

- はい　　プロジェクトにどのように組み込まれ、何らかの制約になっているかを記録する
- いいえ　　データベースは適切なデータ標準とともに選択されるべきである

**レガシーデータはあるのか（電子または紙）**

- はい　　プロジェクトにどのように組み込まれ、どのようにデータ品質を検査できるのか
- いいえ　　データ品質標準を設定し、合理的な品質保証を提供できる

**プロジェクトの最終期限はすでに決まっているのか**

- はい　　試料のデジタル化に実際にどの程度の期間をかけるか優先順位を決め、次に計画にどの程度時間をかけるが逆算する。時間が足りない場合は、プロジェクトの延長を要求することを検討する
- いいえ　時間をかけてプロジェクトを適切に計画する

**研究機関の外で作業する予定があるのか**

- はい　　データベースに及ぼす影響を文書化し、適切な旅費をリソース要求に落とし込む
- いいえ　データベースを選択できる裁量を増やす

## 物理的要件

**デジタル化を実行する場所**

- 収集場所でデジタル化する
- デジタル化専用の場所を設置する
- 全く別の場所でデジタル化する

**既存の I.T. インフラを文書化する**

- 適切なデータベースを選択すれば、より大きなセキュリティが得られる

## 結論

**プロジェクトは実現可能か**

- はい　　計画を実行する方法を検討し始める
- いいえ　実際的になるまで計画を練り直す

**目標は限界を超えているのか**

- はい　　アクションプラン作成時に以下のオプションを検討する
  - プロジェクトに取り組む時間の制限を外して業務慣行を変えられるか
  - 他の近くの研究機関の援助は期待できるか
  - プロジェクトの資金提供者は誰か
  - プロジェクトをいくつかの段階に細分化した方がよいか
- いいえ　アクションプランを書き始める

# 付録 B: アクションプランの論点

**どのデータベースを選ぶのか**

**データベースソリューションを選ぶ**

- 市販のソリューション
- オープンソース
- 修正した市販のソリューションまたはオープンソース
- カスタムメイド

**構築または導入にどのぐらい時間を要するのか**（プロジェクトのリードタイムも含む）

**リソース**

- 何人のスタッフが必要か (デジタイザー、管理者、他のスタッフ)
- スタッフのトレーニング方法
- 購入する必要のあるものとは
- 請求する予算額は

**ワークフロー**

- 試料の収集と持ち帰り
- デジタル化の場所
- データ品質
- オリジナルデータの価値を高める
- 画像化
- データの整理
- データの検査
- 手順が重複できるか.
- スタッフ不足がワークフローに与える影響
- 計画に障害はあるか

**人的要因**

- スタッフの喪失や長期的スタッフ不足
- トレーニング

**危機管理計画/リスク分析**

- 問題になった場合、求められるデジタル化率を達成できない問題にどう対処するのか
- コンピュータが故障した場合、どうなるのか
- バックアップ戦略
- 悪意のあるデータ改ざん
- ソリューション/導入を文書化するために何をするか
- 考慮するべき他のリスクはあるか

## 結論

**ソリューションは適切な水準のデータ品質を提供できるのか**

- はい　　データは現在のプロジェクトにとって完璧であり、他のプロジェクトにも有益である
- いいえ　　以下のことは可能か:
  - データを改善するためにリソースを増やす
  - 対象になる試料の合計数を減らし、残りのデータをの質を高める時間を増やす
  - 未来のプロジェクトでデータ品質を改善できるのか

**選択したソリューションは目標や限界やリソースに適合しているのか**

- はい　　プロジェクトの導入を進めてよい
- いいえ　　ソリューションを改善する

**ソリューションは未来の要求に対応できるのか**

- はい　　容易にデータを維持し拡大できる
- いいえ　　現在のプロジェクトには問題ないが、未来のプロジェクトでは問題になる

**ソリューションは投資に見合う高収益があるのか**

- はい　　計画の導入方法を検討し始める
- いいえ　　実際的になるまで計画を練り直す

**コレクションのデータベース化にどのぐらい時間がかかるのか**