# 第 1 回 SPARC Japan セミナー2018

「データ利活用ポリシーと研究者・ライブラリアンの役割」

# 地域研究画像デジタルライブラリにおける データベース協働構築の実際

丸川 雄三（国立民族学博物館）
石山 俊（国立民族学博物館）

## 講演要旨

国立民族学博物館では、世界各地で撮影された写真や動画を集積するデータベース「地域研究画像デジタルライブラリ」の構築を進めている。当事業は科研費による学術研究支援基盤形成事業であり、公募によって選ばれた科研費プロジェクトの代表者とともに、写真のデジタル化および撮影内容のデータ化を実施する。そのためデータベースの構築においては各プロジェクトの研究者と密接に連携する必要がある。実際に支援を担当する発表者の立場から、データベース協働構築における情報システムと支援業務のあり方や課題等について報告する。

**丸川 雄三**

国立民族学博物館人類基礎理論研究部准教授。2003年、東京工業大学大学院博士後期課程（計算工学専攻）修了。博士（工学）。東京工業大学精密工学研究所助手、国立情報学研究所連想情報学研究開発センター特任准教授、国際日本文化研究センター文化資料研究企画室准教授を経て、2013年10月から現職。専門は連想情報学による文化情報発信手法の研究。これまで手掛けた主なサービスは、『文化遺産オンライン』、『国立美術館遊歩館』、『想 -IMAGINE 早稲田大学演劇博物館』など。

**石山 俊**

国立民族学博物館地域研究画像デジタルライブラリープロジェクト研究員。専門分野は、アフロ・ユーラシア乾燥地域研究および同地域における文化人類学的研究。主な調査地は、サハラ・オアシスおよびサハラ南縁乾燥地域（サーヘル・スーダン地域）。

●丸川　本発表では、研究者が調査で撮影した写真をライブラリとして活用可能な形にする支援事業「地域研究画像デジタルライブラリ」の実際について紹介します。

国立民族学博物館は 1977 年に開館し、昨年で 40 周年を迎えました。国立民族学博物館は、構想の段階から、研究施設を持つ博物館という形でデザインされています。初代館長の梅棹先生が「博物館ならぬ博情館」と語られているように、もちろん標本資料があってのことですが、情報を重視しているという点が非常に特徴的な博物館です。そのようなコンセプトを受けて、国立民族学博物館では、標本資料のデータベースの他に、研究活動の中で収集した研究資料、あるいは研究者自身が撮りためた調査写真のデジタルアーカイブズを公開しています。こういう活動を長年にわたってこつこつと進めています。

## メタデータ付与

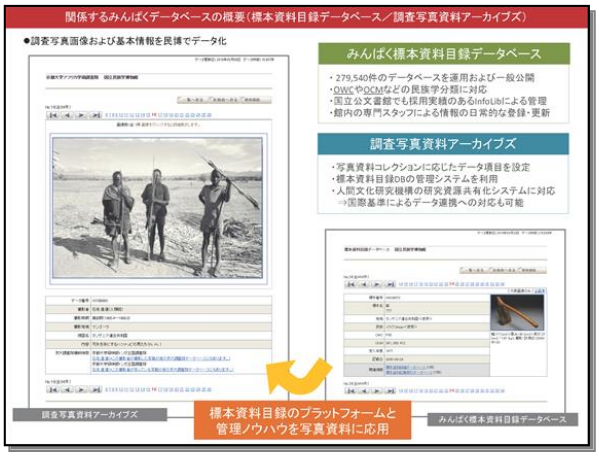
民博が地域研究画像デジタルライブラリの支援事業を実施する背景には、そのような永年の蓄積と実績があるということです。この事業は現在科研に採択されている研究者を対象に、その調査写真をデジタル化し、メタデータを付与し、データベースとしてネットワーク上で共有できるようにする、研究活動の支援をおこなうものです（図 1）。プロジェクトが採択されると、写真を研究者から提供していただきます。あまり整理されていない、ライブラリという形にはなっていない写真のまとまりとしてお預かりすることもあります。代表的なものは、ポジフィルムで撮影された調査写真です。それを整理し、デジタル化し、そこにメタデータを付与していく作業をおこないます。

国立民族学博物館では既に、支援事業という形ではないのですが、研究者の調査写真を図 2 のような形でデジタルライブラリとして公開しております。さらに右下の画面は国立民族学博物館の標本資料データベースです。35 万件ある資料をほぼ全てデータベース化して公開しています。標本資料にメタデータを付ける段階では、OWC や OCM という、イリノイ大学が開発し公開している HRAF（Human Relations Area Files、フラーフ）の分類を付与する作業をおこなっています。

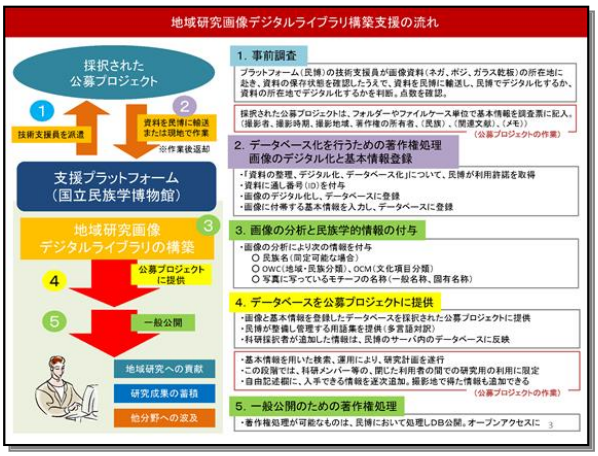
OCM は Outline of Cultural Materials の略で、日本語では「文化分類」と呼んでいますが、例えば農業で使っているものには「241（耕作）」、漁業で使っているものについては「226（漁撈）」、というように、資料の種類に応じて系統的に番号を付ける作業に 30 年来取り組んでいます。今回の地域研究画像デジタルライブラリの整理においても、この OWC や OCM を付与してきたノウハウが生かせるのではないかということです。

また、OCM、OWC という専門的な枠組みではなく、もう少し分かりやすい用語の例もあります。図 3 は展示場で公開している調査写真ですが、右側にあるような民族、カテゴリ、撮影地というような比較的狭い範囲の関連する用語を事前に整理して統一した上で、約 1,000 枚の写真にそれぞれ付与しています。この例ではさらに写真の撮影地を地図上にプロットする作業もおこなっています。

以上の実績も踏まえ、今回の支援事業では、OCM と OWC を写真に付与しています。また、ライブラリとして写真の整理をしやすくするため、コレクション




（図 2）




（図 1）


（図 3）

ごとにそれぞれ独自の用語集を整理して提供しています。当てはめる用語が決まれば、その対訳をつくることで、多言語化が比較的少ない労力で実現できることにもつながります。

## フィルムのデジタル化

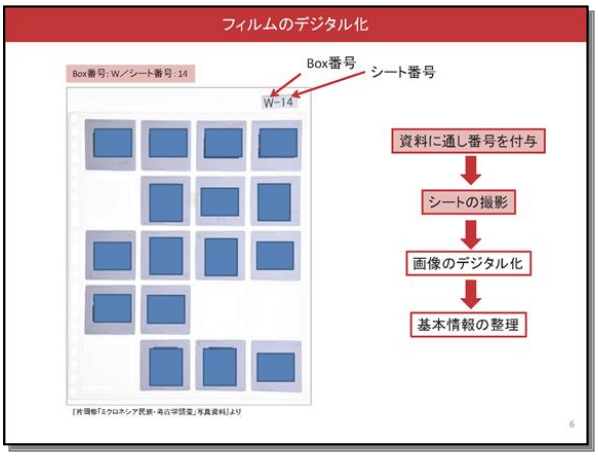
次はフィルムのデジタル化について説明します。ポジフィルムでしたら、まずコマ（スライドマウント）を特定するための番号を入れます（図 4）。写真資料の現物に対する整理をするということです。

それからスライドマウントをシートごと撮影します。さらに写真のデジタルスキャンをおこないます。現在は業務用のフィルムスキャナはほとんど製造中止で手に入れることが難しく、プロジェクトでは SlideSnap Pro を活用しています（図 5）。このスキャナはスライド上映用の装置を改造したもので、デジタルカメラを取り付けて使います。スイッチ一つでスライド送りとシャッターを同期させることができ、慣れれば比較的高速に高精細なスキャンが可能です。

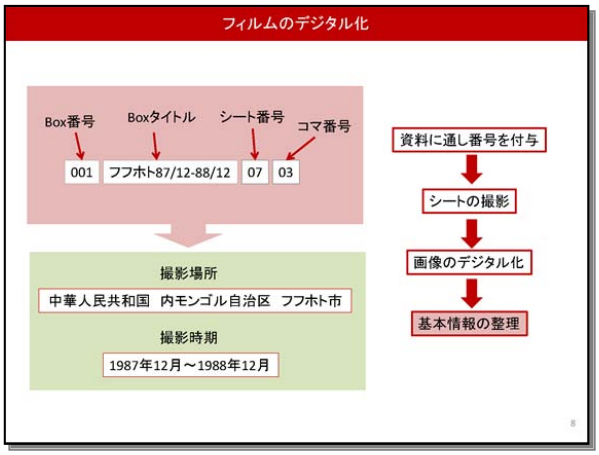
場合によっては研究者が自分のための整理として手書きなどで記入している文字情報もテキストとして取り込みます。略号なども多く読み取りはたいへんな作業なのですが、本人からのヒアリングもおこない、テキストに落とし込んでいくという作業をします（図 6）。

## データベースの構築

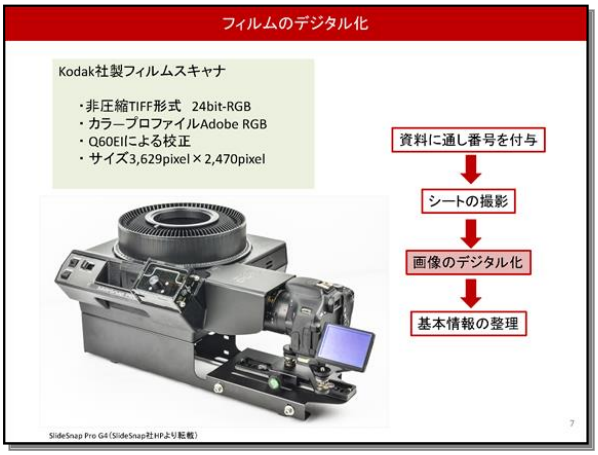
最後にメタデータの付け方です（図 7）。おおまかに五つほどのカテゴリに分けています。最初は写真の説明です。撮影者本人でなければ分からない情報として、まずは写真のキャプションを基本としています。提供時の資料などから分かる情報についてはこちらでなるべく埋めていくのですが、どのような意図で何を



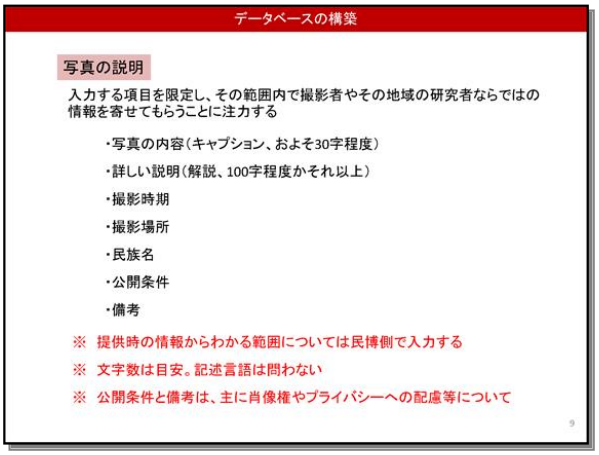

（図 4）




（図 6）




（図 5）




（図 7）

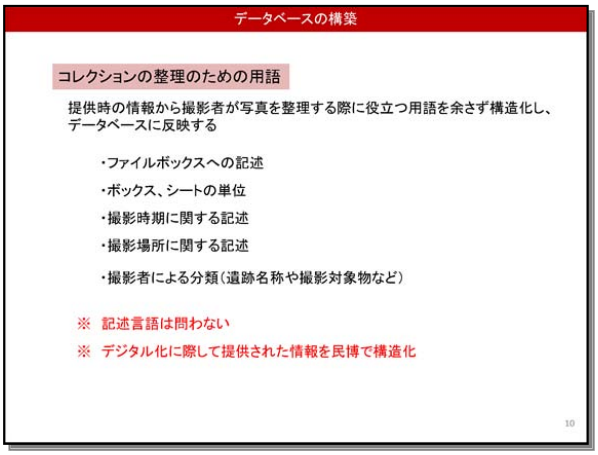
撮影したのかという部分は、やはり研究者自身に入れていただく必要があります。

二つ目は、写真整理のための用語です（図 8）。スライド写真などは、多くの場合 20 コマほどを納めたシートが一つのボックスに 20 枚程度入れられているのですが、そのようなまとまりは閲覧性が比較的良いものです。ぱっと取り出して、開いて、さっと見て、何枚目に何が写っているかということがすぐに分かります。ですがデジタル化をおこない、全体をまとめてフラットにしてしまうと、その使い勝手が失われてしまいます。そこでそのような元々の使い勝手の良さをデジタルにも落とし込んでいくためのデータを付与します。物理的な整理の情報もなるべく落とし込んで、そのまとまりで、例えばきちんとシート単位で取り出すことができるようにするということです。

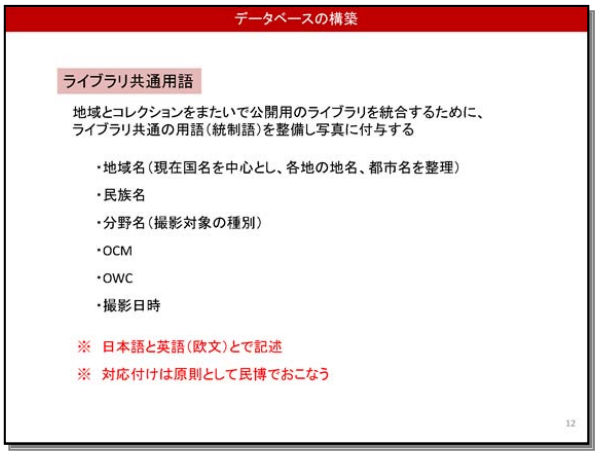
三つ目は、撮影、機材、フィルム、デジタル化に関する情報です（図 9）。撮影機器に関する情報、フィルムの種別やメーカーなどです。デジタルカメラで撮影した写真については Exif というメタデータからいろいろなことが分かります。これらの情報を適切に取り出します。

四つ目は、ライブラリ共通用語です（図 10）。地域名や民族名、あるいは研究分野によっては撮影意図や撮影対象に合わせた用語を付けます。冒頭で説明した OCM や OWC もここに含まれますが、それ以外の用語についても、コレクションをまたいで共通に利用できるよう意識しながら整理します。

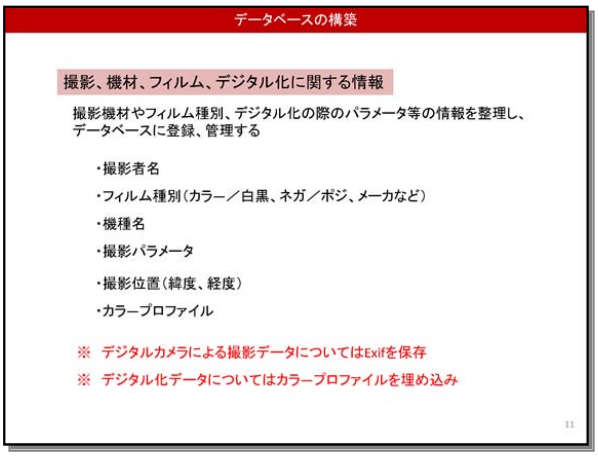
五つ目は、写真の検索時にファセットを自動的につくるためのものです（図 11）。撮影者が入力したキャプションから単語を抽出するなど、自動的なインデックスの更新を検討しています。



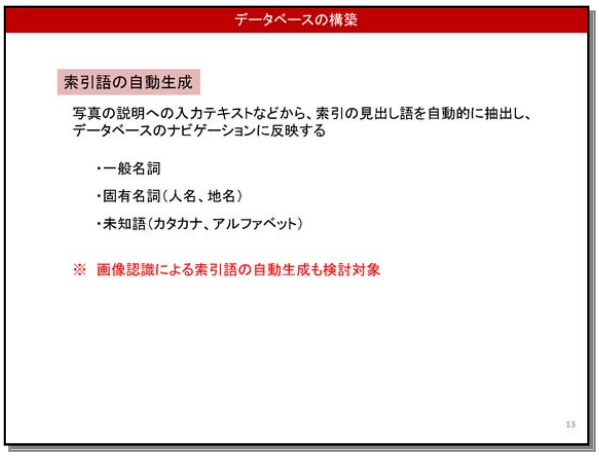

（図 8）




（図 10）




（図 9）




（図 11）

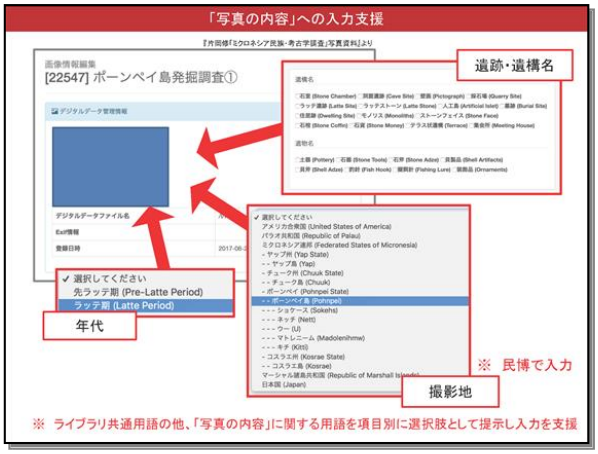
これらのデータを入力し修正するためには、編集用のインタフェースも重要です。図 12 の左が現状の画面ですが、今後はこれを右の構成図のようにブラッシュアップする予定です。先ほど物理的な使い勝手という話をしましたが、スライドを収納したシートの見栄えをある程度踏まえたような見せ方、配置ができるようなものも検討しています（図 13)。入力支援として用語を選択肢として提供することなども可能です（図 14)。

以上の要件や技術要素を組み合わせつつ、写真資料からデジタルアーカイブを構築し、さらにはライブラリとして公開し活用可能な形にしていくという作業に地道に取り組んでおります（図 15)。

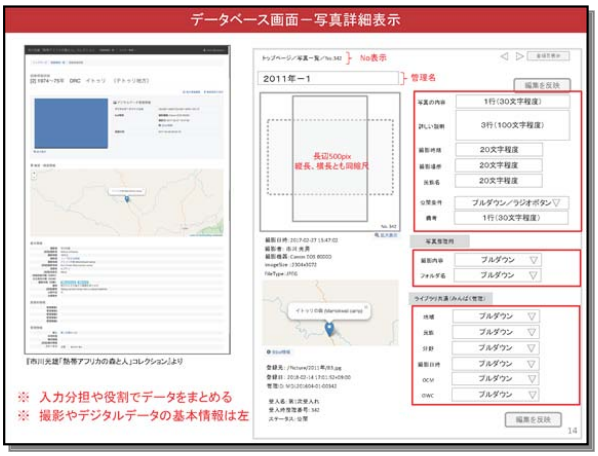
## ライブラリへの展開

本事業の目的は研究者に対する支援ですので、関係者を越えた公開は必須ではないのですが、国立民族学博物館としては、差し支えのない写真についてはいずれライブラリとして公開する予定です（図 16)。一般の方にとっても非常に興味深いものになるのではないかと思います。

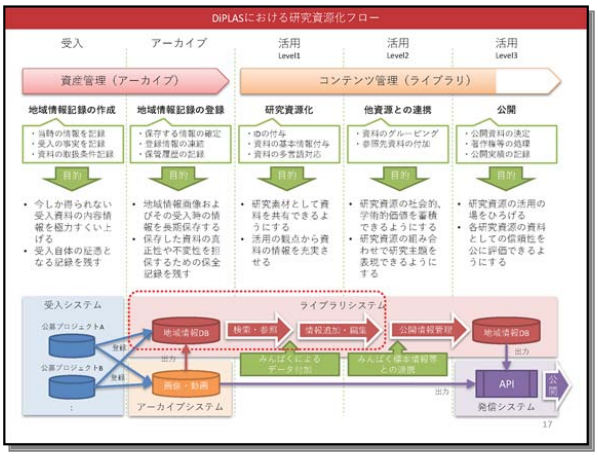
後半は石山に代わります。




（図 14）




（図 12）




（図 15）


（図 13）


（図 16）

●**石山**　今、丸川からデータベースに関して説明がありました。当初プロジェクトの設計としては、データベース上で情報を入力していただくということをメインに考えていたのですが、実際にやりはじめてみると、Excel にまず入力して、そこからデータベースに流していこうという動きが出ています。

簡単に自己紹介をさせていただくと、本日の参加者の中で私が一番デジタルやオープンサイエンスに弱いのではないかと思っています。私の専門領域は地域研究・文化人類学で、アフリカの乾燥地を中心とした生業文化や生活文化の研究をしていて、ひょんなことから 2017 年 10 月より、この地域研究画像デジタルライブラリプロジェクトに従事しています。地域研究者は、若い方はデジタルに慣れ親しんでいる方が多いのですが、キャリアの長い方になるとアナログの方が結構いるのではないかと思います。私はどちらかというとそのような方の立場に立って支援をして、うまくデータベース化につなげていけたらいいなという気持ちで関わっています。

## Excel を用いたメタデータ入力

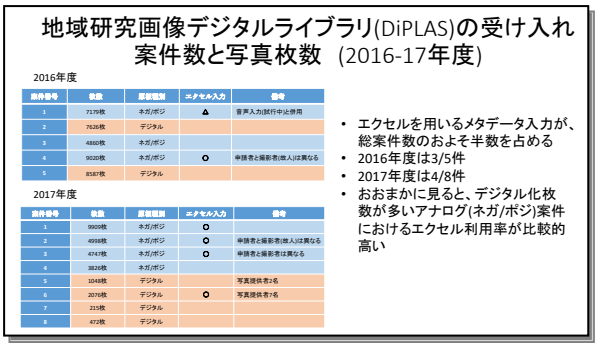
図 1 が昨年度・一昨年度の採択プロジェクトで支援を決定した案件、枚数、原板種別の一覧です。ほとんどの場合は 1 名の方が撮影者ですが、場合によっては科研プロジェクトで数名の方が撮影者になっている場合もあります。ざっと見ると、およそ半分で Excel を用いたメタデータの付与作業が進行しています。デジタル化の枚数が多くて、原板がネガ／ポジの人が、Excel を使っている傾向があるのではないかと思っています。

Excel によるメタデータ付与作業が始まった背景には、まず、文系研究者のデジタルやデータベースに対するハードルがあります。慣れ親しんでいない中で、何千枚分のデータを入れろと聞いた瞬間のプレッシャーはすごいと思います。それをいかに和らげて、みんなにやってもらうようかということがあります。また、デジタル化写真の枚数が多い、撮影者が故人である、撮影者がアナログ世代ということもあります。科研の代表者が同地域の先人の写真のデータベース化を要望するという案件もあり、その場合は、撮った方は既に亡くなって情報を付けられないという状態です。

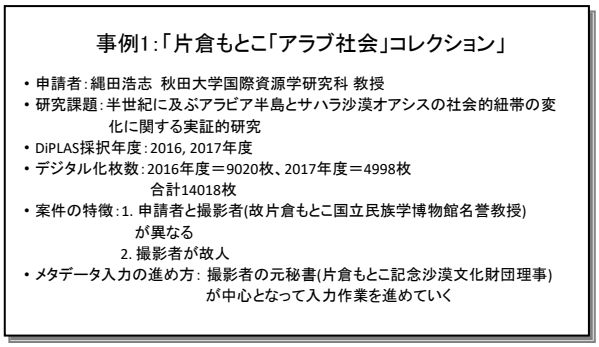
このような課題を抱えながら、Excel によるメタデータ付与作業が始まってきたのです。

## 事例１：「片倉もとこ『アラブ社会』コレクション」

事例を二つほど紹介したいと思います。一つ目が「片倉もとこ『アラブ社会』コレクション」というタイトルの案件です（図 2）。申請者は、科研の代表者、秋田大学の縄田先生です。科研の研究課題は「半世紀に及ぶアラビア半島とサハラ沙漠オアシスの社会的紐帯の変化に関する実証的研究」です。この課題は、実は今年度もそうなのですが、2016 年度と 2017 年度で約 14,000 枚のデジタル化が進んでいます。案件の特徴は、申請者と撮影者（故片倉もとこ国立民族学博物館名誉教授）が異なる、撮影者が既にお亡くなりになっているということです。そのような写真に一体どうやってメタデータを付ければいいのかということで、




（図 1）



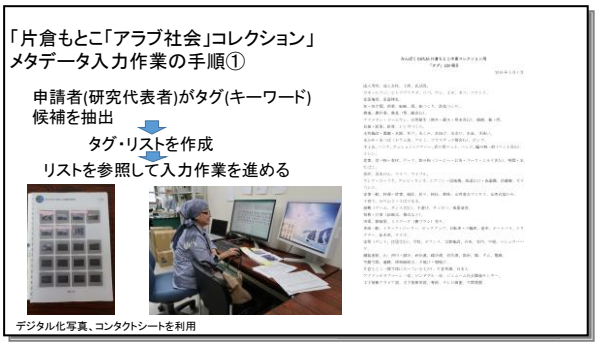

（図 2）

撮影者の元秘書の方にいろいろな情報を参考にしながら作業を進めていただいています。

その写真が図 3 です。このようにして作業を進めています。まず申請者、つまり秋田の縄田先生が写真をざっとくまなく見て、タグ（キーワード）候補を 120 ぐらい抽出します。この中から元秘書だった方がその写真に合うタグ（キーワード）を Excel に入力していきます。

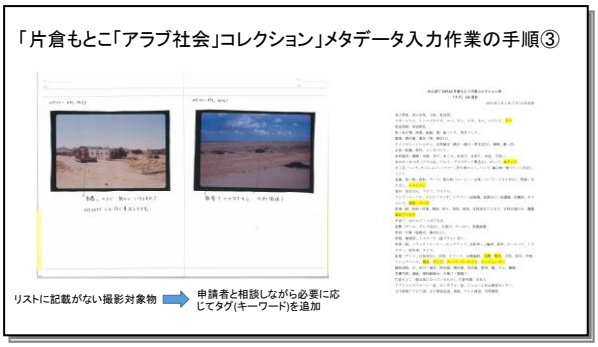
その際、片倉先生が残したフィールドノートやメモも参考にします。それから、入力者の方の記憶なども思い出しながら、どんどん情報を入れていくという作業を地道に続けています（図 4）。

ただし、その途上で最初に想定したタグ（キーワード）に当てはまらない写真が出てくるのです（図 5）。そのときは写真を印刷して、例えば「『重機』のタグを新しくつくりますか？」と、申請者（科研の代表者）とみんなでやりとりしながら、では増やそうかというので黄色マーカーをしたところは増えたり、そのような一発で決まらない往復作業が出てきています。




（図 3）


（図 4）

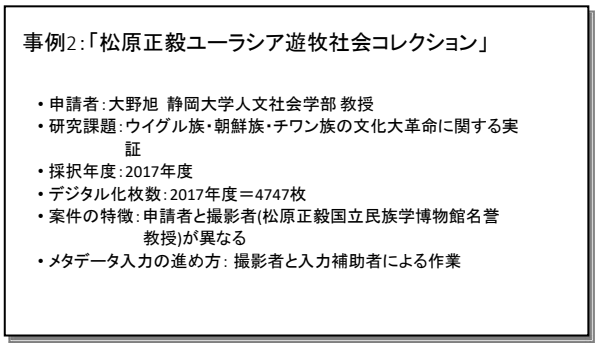



（図 5）

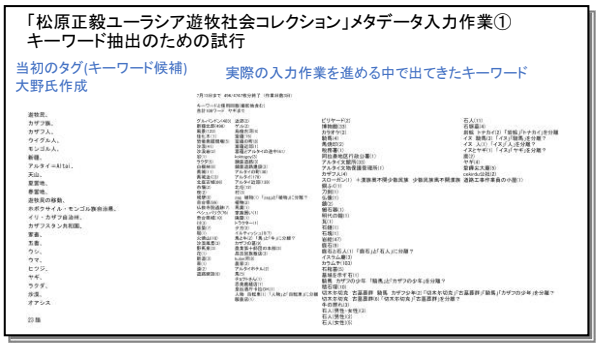
## 事例２：「松原正毅ユーラシア遊牧社会コレクション」

二つ目の事例は、「松原正毅ユーラシア遊牧社会コレクション」です（図 6）。この案件も申請者と撮影者が異なる写真です。採択年度は 2017 年度で、5,000 枚弱です。この案件の場合は、撮影者である松原正毅先生と、実際に Excel に文字を打ち込んでいく入力作業者がペアになって作業を進めています。

図 7 の左は、当初のタグ（キーワード）候補です。最初に申請者の大野先生が「こんなものかな」と作成した 23 個のワードです。ただし、実際に松原先生が作業を始めると、どんどんタグ（キーワード）が増えていきました。4,747 枚中 496 枚時点で、100 いっているか、いっていないかぐらいです。まだまだタグ（キ




（図 6）




（図 7）

ーワード）は増えていくと思うのですが、これをデータベース上で検索しやすいようにするために、1 回またこれを整理する作業も必要になってくるのではないかと思われます。

実際の Excel の画面を見ると、地名がいろいろ書いてあります（図 8）。最初、タグ、タグ、タグ、キャプションと入力していくのですが、フィールドノートを見ながら作業しているので、どんどん羅列して、項目が乱れてきてしまって、これも途中で整理しなければいけないという課題も残っています。

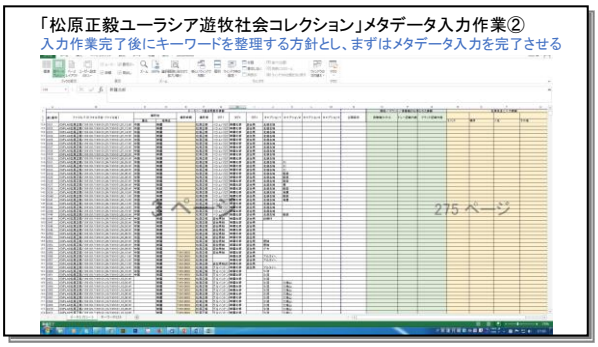
図 9 のような感じで、先生がノートを見ています。作業補助者は内モンゴル出身の総研大の研究生の方で、現地の言葉も分かり、中国語も分かり、松原先生が言った言葉をすぐ Excel に反映させていくことができています。

このプロジェクトの第一義的な目的は、科研研究への支援です。ただ、実際にいろいろな作業をする中で、今までにない新しい面白い副次的効果が見えてきました（図 10）。松原先生は国立民族学博物館の名誉教授でいらっしゃるので、実際に学生、総研大の院生の指導はしないのですが、こういう機会が生まれたことによって、研究コミュニケーションが成り立ったり、松原先生の本が中国語に翻訳されて、この楊海英先生というのが申請者の大野先生なのですが、「中国語の本が余っているから欲しい人がいたらどうぞ配ってください」と言って、留学生に配られたりしています。

事例としては一つで、そんなに大きな広がりにはなっていないのですけれど、私自身、こういうデジタル化の作業もすごく重要だと思いますが、このような副次的な効果がいろいろなところで出てくると思うのです。そのようなものも面白い結果になってくるのではないかと思います。

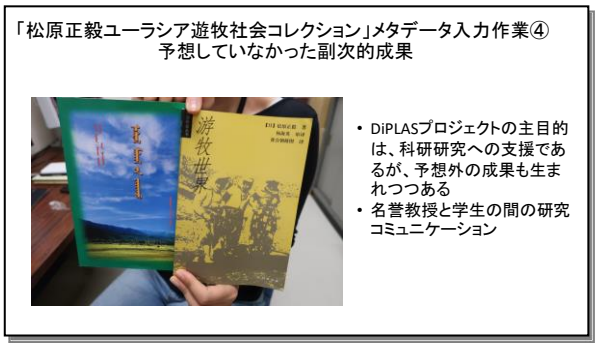
## まとめと課題

やはり一番ハードルが高くなっているのはメタデータ付与作業です。何千枚、人によっては 1 万枚オーバー、それをどうやったらプレッシャーを和らげて作業を進めることができるか。ただ「データベース画面があります。やってください」でやってくれる方ももちろんいると思いますが、それではなかなかスタートできない場合は、やはり何らかの支援、その場を設ける、作業補助者に来ていただくといったことも必要だと思います。




（図 8）


（図 9）




（図 10）


（図 11）

最初にタグ（キーワード）候補を設定して、それにのっとってデータを入れていくということを想定しているのですが、実際にやってみると対応できない部分があって、増やさなければいけません。減らすということはまずありません。ですから、どうしても何往復かやりとりしながら進めていく必要があります。私にとって、デジタル化はかなり先端技術のイメージがあったのですが、そのような従来のフェース・ツー・フェースのコミュニケーションが今の段階では重要だと思っています。

このようなことが果たして、地域研究あるいは文化人類学分野の画像・デジタルデータ固有の問題なのか、あるいは他の分野の方々と共有できる何か課題・問題があるのかどうか、これは私の情報ではよく分からないのですが、考えていく必要があるだろうということです。

地域研究者あるいは文化人類学者は、一つの地域・人々・コミュニティと何年も、場合によっては何十年もかけてじっくり付き合っていく方が非常に多いです。ですから、オープンサイエンス、デジタルデータという中での地域研究の位置付け、どうやったらうまく回っていくかということは、実際にこうやっていろいろ作業をしながら考えていければと思っています。

私はこの事業に関わりながら、科研のメンバーで提供支援を受けるという側にも立っています。図 11 の左は、先ほどの片倉もとこ先生の写真です。1970 年のオアシスの灌漑農地です。右は、2018 年 5 月に行ったときの写真です。全く同じ場所ではないのですが、農業が廃れていることによって、社会、いろいろなものが変わってきている、そのような研究の一つのきっかけになる、大変有用な写真だと思います。私の個人的な研究も含めて紹介させていただきました。

**●林賢紀**　ありがとうございます。それでは質疑を受けたいと思います。

**●フロア 1**　国立国会図書館の職員です。お二人のどちらかは分からないのですが、データ公開を前提としていない研究手法ゆえの苦労かなと感じました。研究手法自体を変える方向に、支援する立場としてコミットすることはあり得るのでしょうか。

**●丸川**　写真をデジタル化して、いろいろな形で活用できるようにすることで何が起こるかというところはこれからなのですが、支援事業で、これ自体が研究ではないということで、そこまで積極的に踏み込んで情報学的な研究テーマをそこに当てはめてというアプローチはしていません。むしろライブラリが公開された後で、分野横断的に他の分野、他のサイエンスの方がこの写真のデジタルライブラリをある種のデータとして活用することでイノベーションが起こるということは期待されているのではないかと思っています。

**●フロア 2**　国立極地研究所の図書館員です。ある程度分野の知識を持っている方でなければ入力補助者は務まらないということを前提にされていると思いますが、補助者の作業時間は 1 日どれくらいなのか、あるいは 1 週間単位でこのぐらいの画像をやってというような幅を設けていらっしゃったら教えてください。

**●石山**　基本的にその地域のことを分かっている方に補助者を今のところお願いしていますが、松原先生の事例のように、「これ」と言って、それを日本語や英語で入力するだけなら、必ずしもその地域に精通した方でなくても大丈夫ではないかという気はしています。

作業時間はまちまちなのですが、午前 11 時から昼 1 時間休憩を挟んで 5 時まで、これを週 1 あるいは週 2 のペースでやっている場合が多いです。それは松原先生の案件の場合です。うまく進むときはその日のうちに 200 枚ぐらい、情報が分からないとか、時系列が狂っている場合、50 枚以下という日もあります。ですから、5 時間作業でアベレージ、大ざっぱに言って 100〜200 ぐらいと考えていただければいいと思います。

●**フロア 2**　そこは正確さを重視していて、枚数などの制限を設けてやっているわけではないということでよろしいですか。

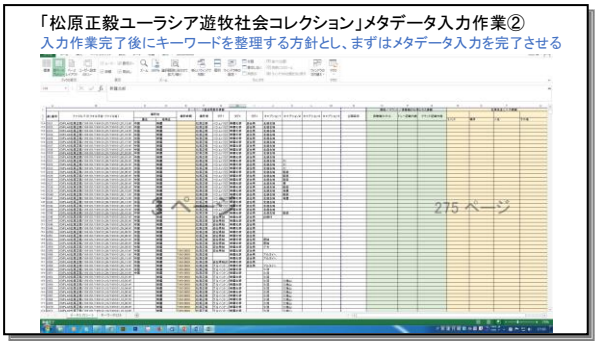
●**石山**　枚数の制限については、結構疲れる作業なので、作業をする方がやめると言うまではやっていただければいいかなと思いますし、片倉先生の案件の場合は、入力者の方が自宅でも少しやったりなど、いろいろなパターンでしています。枚数が多い場合はなかなか終わりが見えてこないということも今の課題です。

●**林賢紀**　はい、ありがとうございます。
もう一つ質疑の時間がございますが、他に何か聞いてみたい、あるいは質問・コメント等いかがでしょうか。はい、どうぞ。

●**フロア 3**　関東学院大学の職員です。松原先生の「ユーラシア遊牧社会コレクション」メタデータ入力作業②（図 8）を見ると、松原先生のコメントを入力する欄が Excel 上で見えるのですが、独自に追加するテーブルというか、ローカルの入力はどれぐらい許容されているのでしょうか。

●**石山**　文字数などですか。

●**フロア 3**　際限がなくなってしまうので、ローカルで、例えば松原先生のコメントを取るところは何個までというような制限を設けていらっしゃるのか。




**（図 8）**

●**石山**　丸川さんからあった話では 30 とか 100 とか。文字数のマックスはありましたか。

●**丸川**　目安です。

●**石山**　目安ですね。今のところ松原先生の方にはそのような制約は課していません。またいろいろデータをいじっていく段階になって、長過ぎるというものが幾つか出てきたらお願いする場合はありますが、今はほとんど自由記述に近い感じでお願いしています。

●**フロア 3**　その項目は、検索のターゲットにはされているのですか。

●**石山**　それもそこから検討したいと思うところなのですが、例えば 100 字の記述の中で重要な単語があると判断した場合は、それをまたキーワードとして別に抜き出すという可能性はもちろんあります。