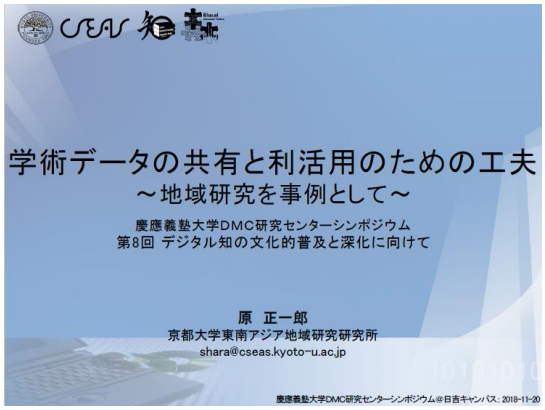
Title	学術データの共有と利活用のための工夫
Sub Title	
Author	原, 正一郎(Hara, Shōichirō)
Publisher	慶應義塾大学デジタルメディア・コンテンツ統合研究センター
Publication year	2019
Jtitle	慶應義塾大学DMC紀要 (DMC review Keio University). Vol.6, No.1 (2019. 3) ,p.6- 28
JaLC DOI	
Abstract	
Notes	特集 DMC研究センターシンポジウム第8回「デジタル知の文化的普及と深化に向けて」メタデータ再考 開催日時 : 2018年11月20日 (火) 14:00 ~ 17:30 開催場所 : 慶應義塾大学日吉キャンパス西別館1 講演
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO32002001-00000006-0006

講演

「学術データの共有と利活用のための工夫」

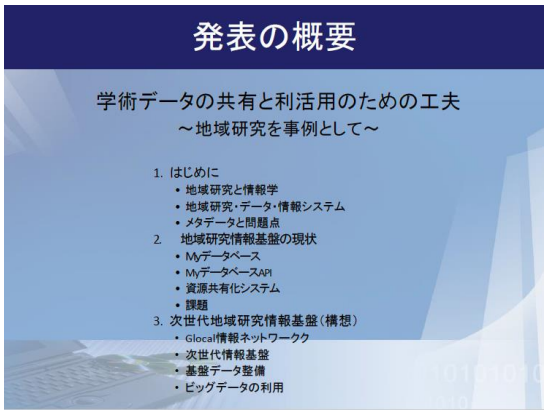
原 正一郎

（京都大学東南アジア地域研究研究所教授）



ご紹介にあずかりました、京都大学東南アジア地域研究研究所の原と申します。この発表では、東南地域研と呼ばせていただきます。

本日お話しする内容は、このようになっております。

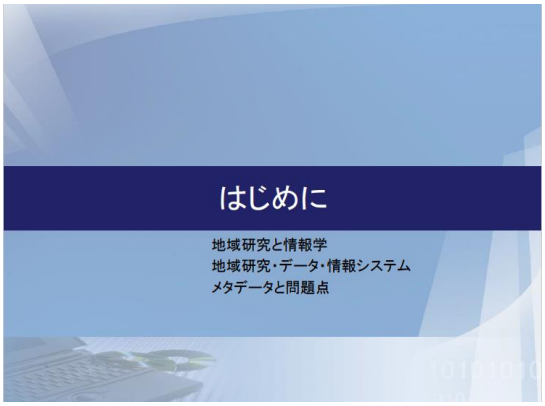


最初に、地域研究とは何か、地域研究を支援する情報システムとはどのようなものか、そこで考慮しなければならない事柄についてメタデータを中心にお話をします。この発表の基礎となる事柄です。

続いて、東南地域研が公開している 3 つの情報ツールについて、具体的に説明します。この発表の中心です。

最後に、オープンデータやビッグデータに対応した、新しい情報プラットフォームの開発についてのお話をさせていただきたいと思っております。

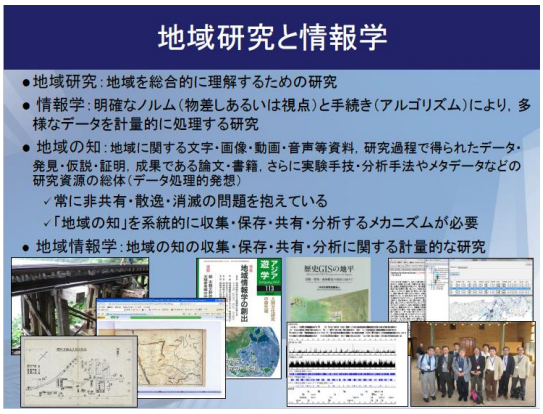
と言いましても、規模の小さい人文系の研究所の、その中の小さな情報センターの話なので、それほど大がかりなことはできません。ですが、手を広げ過ぎていて、どれもこれもが中途半端な状態になっております。



まず、「地域研究」や「報学」や「知」について、整理しておこうと思います。

言うまでもないことですが、地域研究は複合領域です。実際、東南地域研の構成員は、歴史・文化人類・経済・医学・農学・生物などを専門とする研究者です。ですから、地域研究の定義は研究者ごとに違って

います。ここに示したのは、あくまでも私の定義ですが、「地域を総合的に理解する」研究領域としています。もちろん、「地域」とは何処かあるいは何か、「総合的に」とはどのようなことか、どうしたら「理解」したと言えるのかなど、曖昧この上ないのですが、研究領域を厳密に定義する意義があるとは思えないので、ざっくばらんにこの程度に考えています。



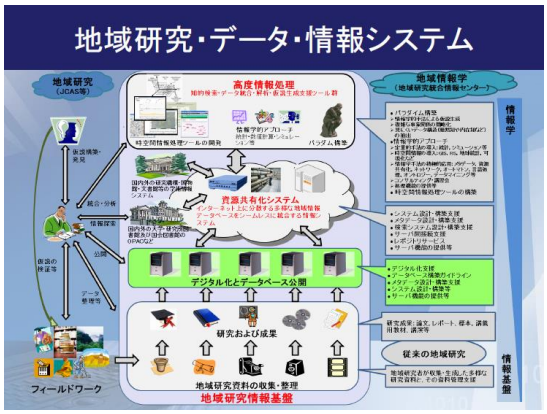
一方、情報学ですが、「明確なノルム（物差しあるいは視点）と手続き（アルゴリズム）により、多様なデータを計量的に処理する研究」と定義しています。これも、なかなか突っ込みがいのあるところですが、地域研究と同じく、情報とはなにかを突き詰めていくと分からなくなってしまうので、ステレオタイプ的ですが、人文社会科学の定性的アプローチに対比するという意味で、計量性と再現性を強調しています。つまり、同じデータと同じ方法を使えば、誰でも同じ結論を再現できる、と言うことです。もちろん、情報学で扱うデータが全て計量的とは言いませんが。

地域の知ですが、これはもっと曖昧です。ここでは「地域に関する文字・画像・動画・音声などの資料、研究過程で得られたデータ・発見・仮説・証明、成果である論文・書籍、さらに実験手技・分析手法やメタデータなどの研究資源の総体」を知としています。データ処理的発想とでもいいましょうか。コンピュータを前提とすれば研究資源はデジタルデータとなりますが、ここにはアナログデータが入っても良いかなと考えています。

最後に、地域情報学ですが、地域研究に関する知を組織化して利活用を支援する取り組みと考えております。たとえば、左下の写真は「戦場にかける橋」の場所ですが、このようなフィールドに足を運んだり、その右の写真は地図ですがさまざまな資料を集めたり、それらをデータベース化したり、分析するためのツールを作ったり、右側の写真のように成果を国際学会で発表したりする活動をしています。

これは、地域研究を支援する地域情報学

を説明するために、いつも使っている図です。地域の知を情報の流れとして、東南地域研で開発した情報ツールと必要な情報技術を整理したものです。



地域研究ですから、全てはフィールドワークから始まります。都市や村落や森林などいろいろなフィールドに滞在して、観察したり、インタビューしたり、資料を収集したりします。それらの研究資源を研究室に持ち帰って、整理・分析して、論文や書籍を執筆します。ここで研究は終了しますが、私はこれを「従来型の地域研究」と呼んでおります。

一方で、収集した研究資源、例えば写真や動画などは、再現のきかない、その時・その場所でしか収集できない資料なので、その意味で貴重です。さらに、地域で集めた資料や研究成果は地域や一般に還元すべき、広く学術コミュニティーで共有すべき、あるいは執筆した論文や書籍の証拠として保存すべきなどといった、社会的・学術的な要請が強くなっています。そのため、研究資源のデジタル化とデータベース公開は、地域研究においても必要不可欠な要件となってきました。これが、東南地域研でもデータベースの構築と公開を積極的に進めている理由で、そのために開発したものが「地域研究情報基盤」です。

詳細は後にしますが、「地域研究情報基盤」において、研究資源のデジタル化と公開の機能を担う情報ツールが、Ｍｙデータベースです。東南地域研のほとんどのデータベースは、Ｍｙデータベースにより構築されています。

ところが、データベースの数が増えてくると、どの資料がどのデータベースに入っているか分からない、一つずつデータベースを検索するのは面倒だということで、データベースを統合する仕掛けが必要になってきました。そのために開発した情報ツールが「資源共有化システム」です。これも後で詳しく説明します。

さらに、統合されたデータの高度情報処理を進めるために、いくつかの分析ツールを試作しています。

このようにして得られた研究成果は、地域にフィードバックされます。これで地域の知が一巡します。この地域の知の循環のプロセスを支援するＭｙデータベース、資源共有化システム、分析ツールなどの情報ツールの総体を地域研究情報基盤、その開発に関わる情報学を地域情報学と呼んでい

ます。こういうとカッコは良いのですが、どれも完成の域には達しませんし、新しい技術が次から次へと現れてくるので、どの技術を導入するか、どの段階で古い技術を入れ替えるかなど、戸惑いが多いのが実際のところです。

さてデータベースを構築するうえでメタデータは欠かせません。このメタデータですが、「あるデータにアクセスするための要約というか Proxy となるデータ」と言われています。自分なりに翻訳すると、ある対象物や現象の特徴を何らかの視点あるいは基準で計測して定型的に要約したデータとします。たとえば、人物が対象なら、生年月日や住所や地位などのデータの集まりになりますし、同じ人物でも健康が対象であれば、血糖値やコレステロール値などのデータの集まりになります。つまり物理的には同じ対象物や現象であっても、視点や切り口が違えば、違うメタデータとなります。

メタデータについて、もう少し掘り下げたいと思います。筑波大学の杉本先生の説明法が面白いので、それを拝借します。

さて、お茶の自動販売機のディスプレイが左側のボルトの写真のようだとします。ここでボトルの中身をデータ、自動販売機をデータベースと考えます。このお茶は何でしょうか？　私は「おーいお茶　濃い茶」が欲しいのですが、右側のようなラベルがなければ正しいボトルを選ぶことができません。メタデータとは、このラベルのように中身について記述した事柄です。

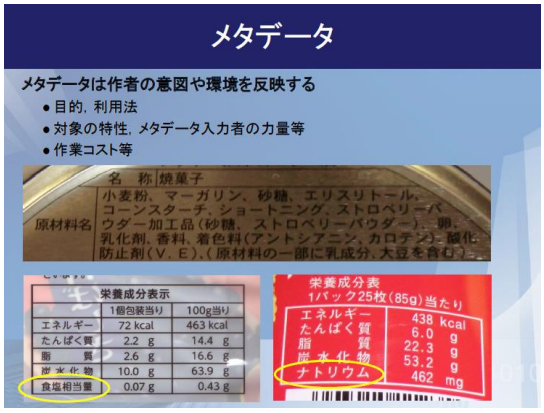
つまり、ラベル「おーいお茶　濃い茶」がなければ、ボトル「おーいお茶　濃い茶」を選ぶことができない。言い換えると、ラベルというメタデータがなければ、正しいデータにアクセルでないことになります。メタデータの書き方は、今のところは問わないけれど、これがなければデータベース検索はできませんということです。

このメタデータですが、作者の意図や環境を反映したものとなります。先ほどの人物データベースと健康データベースは、対象物は同じヒトであっても、目的によってメタデータの内容が異なるという例です。この視点は、学術データベースにおいて重要です。後ほど説明いたします。

それから、あまりいい言い方ではないですが、メタデータの質は作成者の力量に左右されます。どんなに精緻なメタデータを設計しても、必要な情報を読み取る目がなければ、正しいメタデータを作成できないからです。たとえば、医療データを検索す

るためのメタデータがあったとしても、必要な語彙や医学知識あるいは症状を正しく認識できる技術などがなければ、使えるメタデータを作成することはできません。

それから、作業コストにもメタデータ作成において重要な要素です。たとえば完璧なメタデータを 1 件作るのに 1 万円かかるとします。この場合、メタデータの質は高いもののデータ量の少ないデータベースで我慢するか、メタデータの質は多少劣るけれどもデータ量の多いデータベースとするかなど、悩ましい決断が必要になります。



もう少しメタデータの中身を見てみましょう。先ほどの、「作者の意図を反映する」ことに関係しています。これらの食品表示ラベルもメタデータの例です。このメタデータを例にした理由ですが、私にとっては死活問題だからです。私は大きな病気をしまして、食塩の摂取量が 1 日 6 グラムに制限されています。食塩 6 グラムはどのぐらいか、カップラーメン 1 個で 5 グラムといったら、一食あたりの食塩量がどれだけ少ないか想像つくと思います。なので、スーパーなどで食材を買う時に、「食塩相当量」や「ナトリウム」は、私にとってはとても重要なデータなのです。ですから、下の 2 つの栄養成分表は私の意図に合ったメタデータですが、上側の食品表示ラベル役に立たないメタデータということになります。

次に「メタデータの書き方は、今のところは問わないが」に関係しますが、下の 2 つの栄養成分表を改めてご覧下さい。これらは「内容的」には同じですが、記述法が違っています。まず左側は一包あたりあるいは 100g あたりの値で、右側は 85 g あたりの値です。また左側は食塩相当量ですが、右側はナトリルウム量です。メタデータの記述法が多様だということがお分かりと思います。

実際のデータベースを例にします。この上下 2 つの仮想データベースの内容は同じですが、項目名、データ型、記述規則、粒度が異なっています。

まずデータ項目ですが、上が日本語であるのに、下は英語となっています。これが同じであることはヒトにとっては当然ですが、コンピュータにとっては別物です。

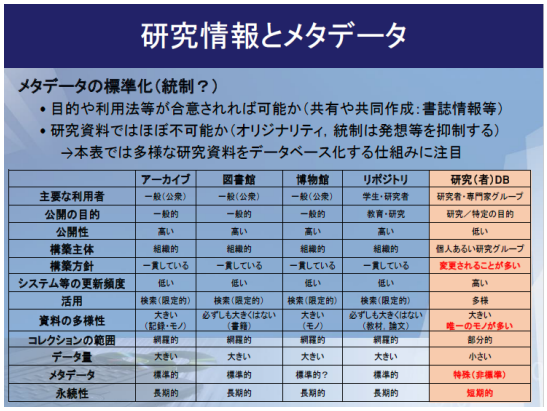
データ型について、上の識別では「0001」ですが、下の ID では「1」です。「イチ」ということでは同じかもしれませんが、「0001」は文字型、「1」はおそらく整数型で、違うデータ型です。

上の性別の「M」と下の SEX の「1」は記述規則の違いです。ちなみに、JIS（JIS X0303（性別コード））では男は「1」、女は「2」となっています。上の誕生日の「昭和 32 年 10 月 11 日」と下の birthday の「1957/10/11」も記述規則の違いです。

上の氏名は「大学　太郎」で、下は surname が「ダイガク」で forename が「タロウ」で、粒度が異なっています。これらのように、内容的の同じメタデータでも、記述法は違っていることが分かります。

では、異なるメタデータをどうしたら良いでしょうか。一つの手段はメタデータの統制あるいはメタデータの標準化です。みんなで同じメタデータを使いましょう、ということです。データサービス機関の間で標準についての合意が形成できるなら、良い解決法だと思います。図書館の機械可読目録（MARC :MAchine-Readable Cataloging）はその典型例です。日本では、国立国会図書館の JAPAN MARC や国立情報学研究所の NACSIS-CAT などがこれに該当します。

一方、研究分野におけるメタデータの統制あるいは標準化は、必ずしも容易ではないし、良い方法とも思えません。なぜかというと、分野が違えば語彙が異なります。あるいは、同じ語彙であっても意味が違うこともあります。それから、研究ではオリジナリティーを重視しますから、対象物が同じであっても、メタデータが異なるのは当たり前です。標準化や統制は、発想の自由を阻害します。新しい発想のもとで作られるメタデータは、それまでとは全く違う構造となる可能性があります。そのような理由で、研究分野のメタデータの統制や標準化は不可能であり有用でもないと考えています。

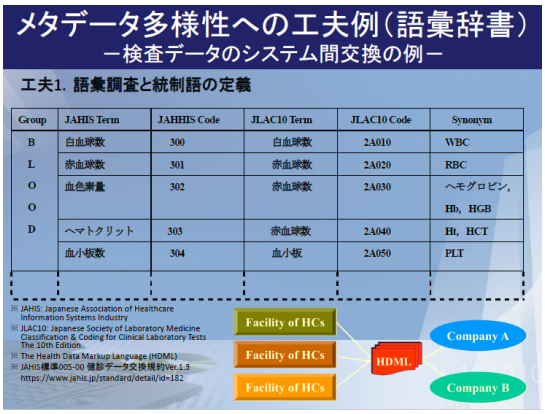


ここまで述べたことを表にまとめてみました。図書館のメタデータは標準化が進んでいます。アーカイブ・博物館・レポジトリについても、標準規約の制定は進みつつありますが､導入は進んでいないようです。これらのいわゆる機関データベースに比べると、「研究（者）DB」の様相はかなり異なっています。主要な利用者は研究者あるいは研究グループなどの個人あるいは小規模なグループです。そのためもあり、公開性は必ずしも高くはありません。個人データなどが含まれている場合は、もちろん公開できません。目的も研究の遂行に特化していますから、メタデータの構造はデータベースごとに異なっていて、これをメタデータの heterogeneity と呼ぶことがあります。

データ収集法やデータ構造や利用法についても、図書館や博物館は一貫していますが、研究者は変化します。今日、私はデータサービス側の立場で話をしているので、メタデータやアプリケーションの変更はありがたくないのですが、反対にデータを利用する研究者側の立場のとき、朝令暮改はよくあることです。

このような多様なデータをどうやって管理・運営するかというのが、これまでの課題でした。ついでですが、もう一つ、今、直面してる問題は、多様なデータの長期的な保存と利活用をどのように実現するかということです。

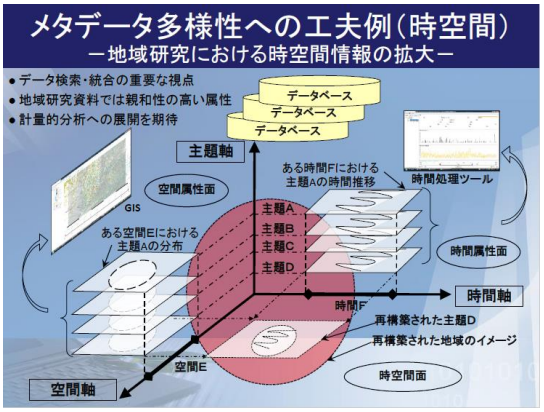
研究メタデータの特徴を明らかにしたところで、異なるメタデータを統合するあるいは繋ぐにはどうしたら良いかという話に進みたいと思います。



皆さんが健康診断を受けると、血糖値は幾つなど書かれた検査結果票を受け取ると思います。検査項目名には、いろいろな書き方があります。「血色素量､ヘモグロビン、Hb、HGB」といった具合です。これらが同じであることを専門家は知っていますが、コンピュータはそうはいきませんから、このままでは、情報システム間でデータを統合したり分析したりすることはできません。そこで、検査項目語彙を辞書にまとめて利用しようとしたのが、この研究です。

ここでは、主要な検査ラボにアンケートを送って、利用している語彙や単位などの関連情報を集めて同義語集を作り、同義語に代表語と ID を付けました。こうすることで、情報システムの間で交換する検査データの項目名については、対応できるようになりました。後で述べる「資源共有化システム」は、この方法を応用しています。

では、検査項目名が交換できれば問題が解消されるのかと言えば、そうは問屋が卸さないわけです。情報システム間で交換された検査値を利用する場合、先ほど述べた整数や文字といったデータ型以外にも、単位や計測法や記述規則などの情報が必要となります。このように検査値を記述する上で考慮しなければならない事柄、データの意味ということが出来ると思いますが、それらを属性としてまとめたのがこの表です。この研究では、これらの属性を検査値と一緒に交換することで、異なる情報システム間でのデータ利用の可能性を試みました。



異なるデータを統合する方法として、データ項目名とデータ属性の交換以外に、時空間データの利用可能性についての研究を進めています。例えば書籍資料のデータベースであれば、資料名や著者名や出版社名などに注目した検索やデータ統合が可能です。しかし、遺物や発掘された壷のようなものには、書名や著者名などはありません。一方、これらの対象物に関するデータベースの検索や統合を行う上で、壺などが利用されたと想定される時期や発掘された場所といった時空間データは有効です。多様な史資料を対象とする地域研究において、このような時空間データは親和性の高いものと言えます。また、時空間情報をグレゴリオ暦や緯度・経度のように数値化できれば、計量的な取り扱いも容易になります。

この図は、時間と空間を考慮した、我々の情報モデルです。主題軸は書誌データベースの書名や著者名や主題に相当する文字的な情報です。同じ主題を持つ史資料を検索するあるいは集める際に利用します。

それに対して、空間軸は対象物の場所に関連する情報です。緯度経度や地名などによって、史資料を検索する、集める、あるいは空間的関係を分析する際などに利用できると考えています。時間軸は空間軸と同じ機能を時間で実現しようとするものです。

このモデルを地域研究的に解釈すると、この 3 次元的な仮想空間内のデータ点の塊が、ある地域の全体像に対応すると見なせます。また、この塊を主題・空間・時間あるいはそれらを組み合わせた面で、ちょうど CT のように輪切りにして、可視化・分析することは、地域を多様な視点で観察することになると考えています。これは、情報学による地域研究という新しい可能性を示すものと期待しています。あるいは、そのようなことができたら面白いと考えています。

以上が前触れで、ここからは、東南地域研の情報基盤の紹介です。大きく分けて 4 つの機能を実現または実現しつつあります。データベースを「作る」機能、「使う」機能、「共有する」機能、「使う」機能です。

データベースがなければ、データ公開も、共有も、使うこともできませんから、「作る」機能は最優先です。そのために My データベースという情報ツールを構築しました。

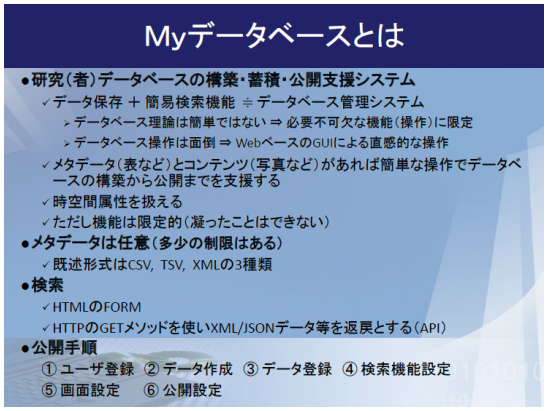
ところで、ユーザーインターフェースや可視化アプリケーションなど、データベースを「使う」機能も重要です。Ｍｙデータベースは誰もが簡単に使える情報ツールですが、機能は制限されていて凝った使い方は出来ません。そこで Application Programming Interface あるいは API とも呼ばれる機能を公開しています。API はプログラムでＭｙデータベースを使うための仕掛けです。研究に適したアプリケーションは、自分で作って下さいということです。

次に「共有する」機能です。Ｍｙデータベースで多くのデータベースが作られていますが、どの資料がどのデータベースに入っているかは、検索して見なければ分かりません。といって、一つ一つ検索するのも面倒なので、データベースを共有化して一

括検索できる情報ツールも構築しました。それが「共有する」機能です。

さらに「使う」機能として GIS ツールなども構築していますが、今日、この話はいたしません。

それでは、「作る」機能、「使う」機能、「共有する」について、詳しく説明します。



Ｍｙデータベースはデータベースの公開を支援するツールです。もしデータベースを自力で構築するなら、サーバーを購入して、OS やセキュリティ環境などを整えてから、データベースソフトウェアの設定や、アプリケーションを作成する必要があります。また、基本的なデータベース理論や検索言語の習得なども必要です。残念ながら、データベースの構築は特に人文社会学研究者にとっては、敷居の高い作業です。

しかしながら、人文社会学においても、データは EXCEL などのツール使って作ることが当たり前になっています。そこで、EXCEL レベルのデジタルデータからデータベースを作れるように工夫した情報ツールがＭｙデータベースです。

技術的な説明はしませんが、ID を必要としない、データ型やデータサイズを気にしない、データ項目名の記述制限を緩やかにしているなど、データ作成の制約を減らす工夫をしています。また、データベース作成の作業は、コマンドではなく、GUI によって直感的にできるようになっています。さらに、地域研究用なので、地図あるいはタイムラインを使ったデータ検索機能や表示機能をはじめから用意しています。

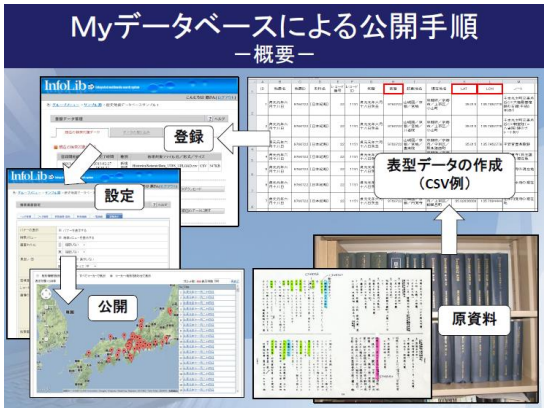
ということで、Ｍｙデータベースのメタデータスキーマは任意です。データフォーマットは、CSV、TSV、XML の 3 種類です。スキーマは任意と言っても、多少の制限はあります。たとえば XML の場合、データ構造は入れ子になっている、つまり整形式でなければなりません。また CSV あるは TSV の場合、下部構造を持たない、つまり第一正規形を満たしていることなどです。

余談ですが、EXCEL と ACCESS の違いを理解している人文社会学研究者は多くないようで、かなり手を入れないと使えない EXCEL データを持ち込まれる方もいらっしゃいます。

Ｍｙデータベースの利用法は、HTML フォーム、つまり通常の Web 検索ページによるか、API を使ったアプリケーションによるかの 2 種です。

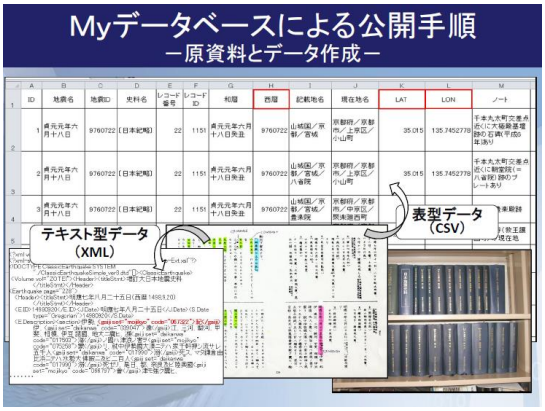
Ｍｙデータベースを利用する前に、ユー

ザー登録の申請をしてもらいます。なお、ここからの説明では、Ｍｙデータベースを作成する研究者をユーザーとします。管理者は、Ｍｙデータベースのシステム上に一定の作業領域を確保した上で、ユーザーがデータベースへアクセスするためのURIとIDとパスワードを発行します。

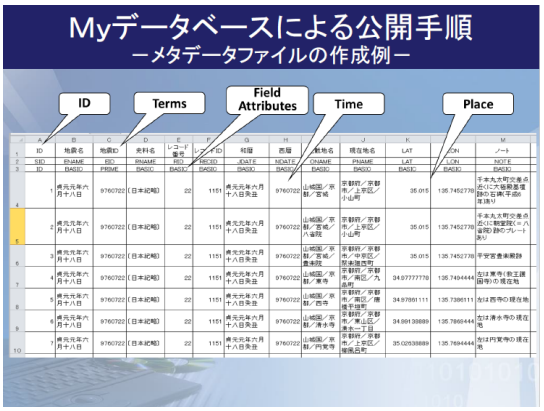


その後は、この図に示す手順でデータベースを作成します。

ユーザーは資料に基づいてメタデータを作りますが、多くの場合はテーブル形式のファイルとなっています。ここまで出来ていれば、発行されたURIを使ってＭｙデータベースを呼び出し、メタデータや画像データなどをアップロードします。さらに幾つかの設定､例えば検索するデータ項目や、表示するデータ項目、データベースの利用権限などを設定したら、データベース作成の準備完了です。データが少なくてエラーがなければ､「構築ボタン」を押して数分以内にデータベースの構築が終了して、直ちに公開できます。



もう少し詳しく説明します。これは文字資料の例です。ユーザーは、図の上側のような内容に関するメタデータをテーブルとして整理したり、左下のようにXMLを使って本文を翻刻したりします。また、この例では内容に関連する場所と時間も推定しているので、それらはLAT、LONと西暦としてメタデータに追加されています。なおＭｙデータベースにおいて、緯度と経度の表記はWGS84測地系に基づく十進表示、時間はグレゴリオ歴によるyyyymmddとなっています。



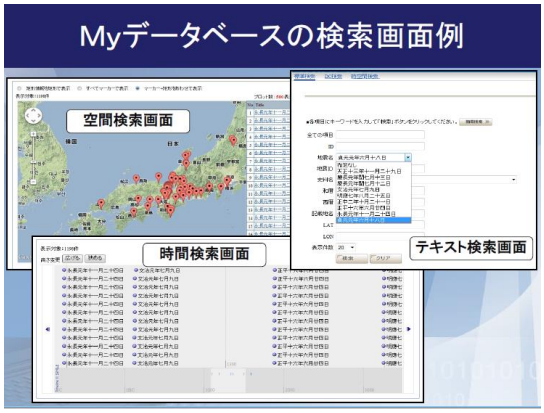
これは、Ｍｙデータベースに登録したメタデータの例です。上の3行分はＭｙデータベース独特の書き方で、1 行目はデータ

項目名、2 行目は他言語によるデータ項目名、3 行目は次に説明するデータ属性です。メタデータをＭｙデータベースにアップロードする段階で、属性の行は不要です。

データ項目名を二つ用意した理由は、例えばメタデータがタイ語で記述されているとデータ項目の意味が分かりにくいので、それを日本語や英語などの別の言語で記述する必要があったためです。地域研究データベースならではの工夫です。

メタデータをアップロードしたら、データ項目ごとに属性を設定します。データ属性ですが、どの項目が検索対象なのか、どの項目が表示対象なのか、どの項目が緯度経度なのかなどを設定します。少し高度な設定としては、データの値を参照してプルダウンメニューやスイッチボックスを作成したり、階層的な選択メニューを作成したりもできます。

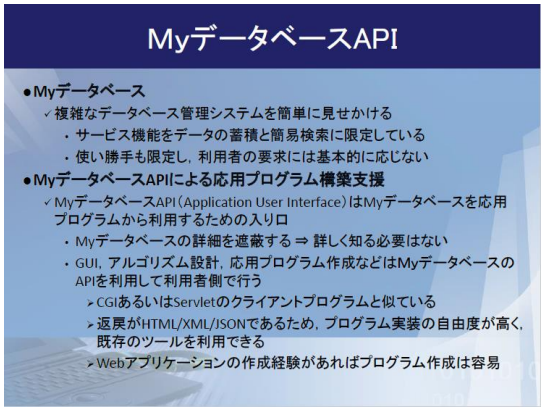
さらに、これとは別の画面ですが、データベースへのアクセス制限を設定したりします。一通りの設定が終了したら、データベースの構築を開始します。



これは、できあがったデータベースの検索画面の例を示しています。右上は通常のテキストによる検索画面ですが、一部ではプルダウンメニューを使っています。空間情報があれば、左上のように、地図画面を使った検索もできます。さらに、時間情報があれば、下側のようなタイムラインを使った検索もできます。

現在、50 個ほどのデータベースを公開していて、構築中や実験を含めると 100 個ほどのデータベースがＭｙデータベースにより構築されています。

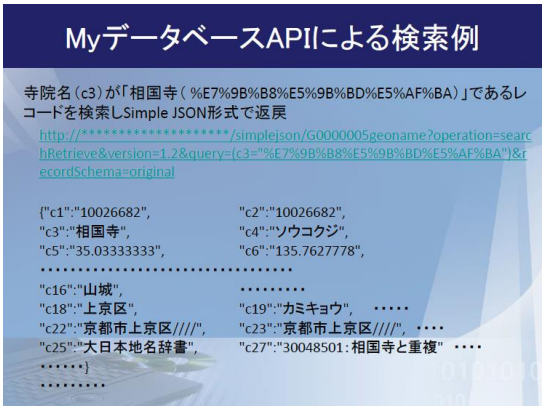
これらは、公開されているデータベースの一例です。



このように、Ｍｙデータベースを使うのは簡単なのですが、システム自体はけっこう巨大で複雑です。それを、情報系の教員 2 名だけで管理・運営しています。そのため、ユーザーからの個別の要求にはとても対応できません。というわけで、よほどの理由がない限り、Ｍｙデータベースの機能拡張に応じることはしていません。また、ユーザーのアプリケーションをＭｙデータベースに組み込むことも認めていません。システム更新によるＯＳやミドルウェアのバージョンアップへの対応や、セキュリティの責任分界などの問題があるためです。

その代わりに API を公開しているので、ユーザーは自分でアプリケーションを作ることができます。残念ながら、アプリケーションの自作は簡単ではありません。実際、少し昔であれば、Ｃ言語や、データベースシステムが提供しているライブラリや、インターネットのセッション管理などの技術や知識がなければ、アプリケーションは作れませんでしたから、専門業者に委託するしかなく、コストもかかりました。

API を使えばアプリケーションを作る負担は大幅に減ります。それでも、人文社会学研究者にはまだ敷居が高いかもしれません。ですが、Web プログラムの簡単な技術があれば、データベースシステムの中身を知らなくともアプリケーションを作ることができますから、外注してもコストはかなり低く押さえることができます。



これはＭｙデータベースの API の例で、地名辞書データベースから「相国寺」というお寺の位置情報を JASON という書き方で取り出しています。このように、API の検索式も検索結果も単純かつ定型的なので、アプリケーションプログラムの作成はかなり楽になります。

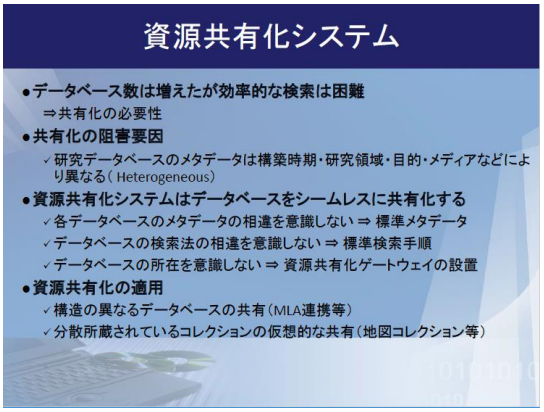
東南地域研の同僚が作った API アプリケーションの例を二つお見せします。

まず左側です。地域研究の論文などでは地図を多く使いますが、紙面に掲載できる枚数には制限があります。また、ズームインやズームアウトもできません。彼の工夫は、紙面には最小限の地図を掲載し、それ以外の地図は QR コードを使ってＭｙデータベースから参照しようとしたことです。QR コードにスマホをかざすと、京大出版会のサーバーが My データベースを検索して地図データをスマホに表示します。表示する枚数に制限はないし、デジタルデータなのでズームインやズームアウトも容易です。

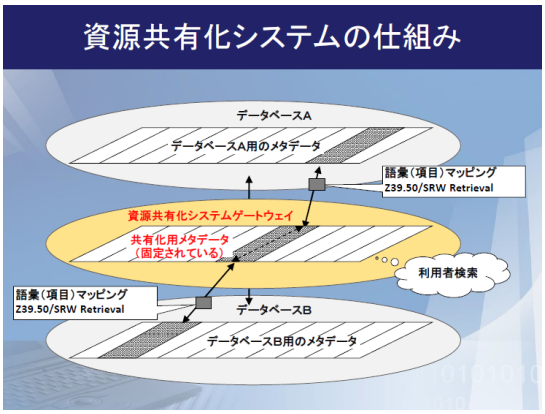
このアイデアの応用は広くて、別の研究書の例では、舞踏の写真の脇に QR コードをつけて、ここにスマホかざすと、動画や音楽が出てくるようになっています。

右側の例は、いわゆるマルチメディアデータベースです。右側はマレー語原本の画像で、画像データベースからの出力です。左側はそれを電子翻刻したデータで、テキストデータベースからの出力です。API を使って別々のデータベースを検索して、Web 上で合成しています。

これら以外にも GIS と組み合わせたアプリケーションなども作成されており、ユニークなものが、これからも作成されると期待しています。



次に資源共有化システムです。前にも述べましたが、Ｍｙデータベースで作られたデータベースは、ユーザーの研究目的が異なるので、同じメタデータ構造を持ったものはありません。一方、Ｍｙデータベースの検索者にとっては、どのデータベースに何が入っているのか分かりません。かといって、一つ一つ検索するのは面倒です。資源共有化システムは、Ｍｙデータベースで作られたメタデータ構造の異なるデータベースを統合検索する仕掛けです。実際は、東南地域研以外のデータベースも統合検索できるようになっています。



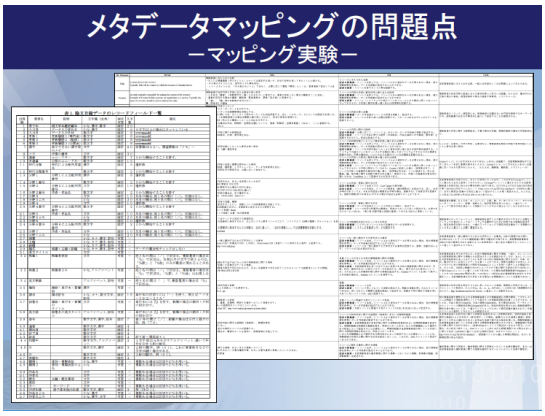
資源共有化システムの仕掛けは単純で、個別データベースから独立した中立的なメ

タデータを利用します。これを共有化メタデータと呼んでいます。現在の共有化メタデータは、Durbin Core と Metadata Object Description Schema（MODS）の 2 種類です。

Ｍｙデータベースで作ったデータベース、図ではデータベース A と B を、資源共有化システムに登録する際に、各メタデータのデータ項目と、共有化メタデータのデータ項目の関連を定義します。これをマッピングと呼んでします。

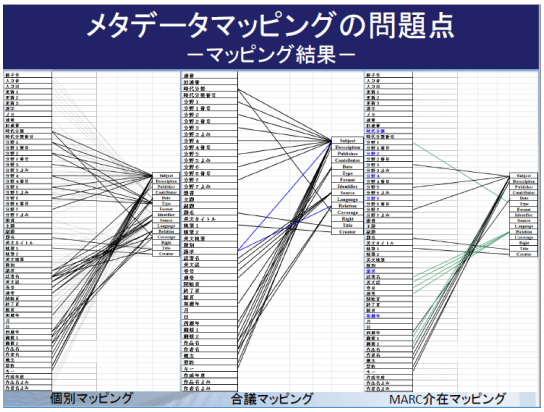
検索者は、個別のデータベースではなく資源共有化システム、図の資源共有化システムゲートウェイにアクセスします。検索には共有化メタデータの項目を使うので、検索者は個別のメタデータを気にする必要はありません。

資源共有化システムゲートウェイは検索命令を各データベースに渡しますが、その際に対象データベース用の検索命令に変換します。資源共有化システムゲートウェイと各データベースは、Z39.50 または SRW という情報検索プロトコルで自動的に通信します。そのため、検索者は個別データベースの所在や検索式を気にする必要はありません。



言うまでもないことですが、資源共有化システムの検索性能は、各メタデータと共有化メタデータのマッピングに異存します。

そこでマッピングの実験をしました。この図の左は、私が以前に勤務していた国文学研究資料館の論文目録データベースのデータ項目の説明文、右側は Dublin Core の説明文です。この説明に基づいて、論文目録から Dublin Core へのマッピングを試みました。被験者は、国文学研究資料館の目録担当図書館員と教員の 5 名です。



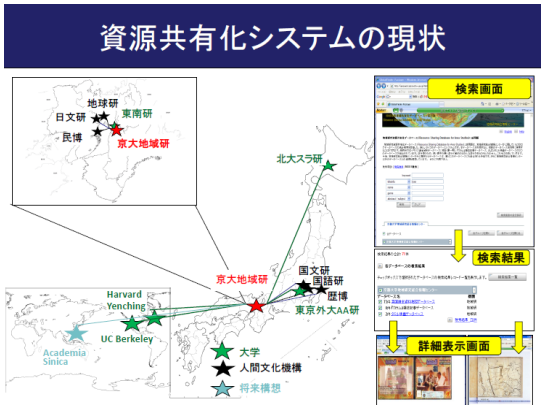
左側は各被験者のマッピング結果をまとめたものです。できる限り全項目をマッピングしようとしていた者もいれば、できるところだけをマッピングした者もいるなど、マッピングの程度はさまざまでしたが、そ

の結果もバラバラでした。

中央は、個別マッピングの結果を受けて、5 人で検討したマッピング結果です。「これでなければならない」ではなく「これなら妥協してもよい」というレベルの合意です。個別マッピングよりも収斂した印象を受けますが、論文目録から Dublin Core への対応が 1 対多になっている項目があります。

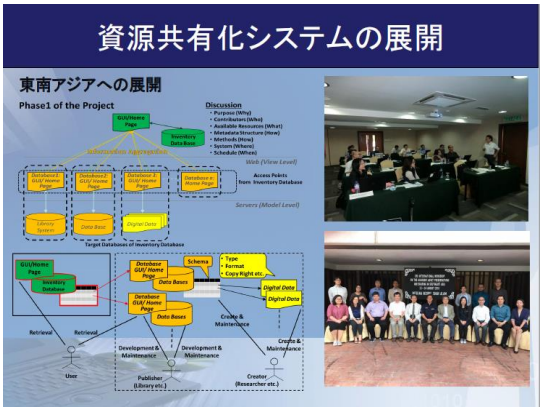
右側は MARC へのマッピングを介した結果です。目録関係者なので MARC はご存知でした。MARC から Dublin Core へのマッピングは、既存の Crosswalk を使いました。その結果、マッピングされない項目が増えた反面、マッピングはかなり整理されました。

マッピングがどれだけバラツクのかを調べる実験だったので、バラツキの違いが検索精度にどれだけの影響を与えるのかについての検討は行いませんでした。しかし、想定した以上にバラツキが大きかったので、資源共有化システムの評価は、「統合検索できないよりは良いけど、検索精度は余り良くない」というところと考えています。



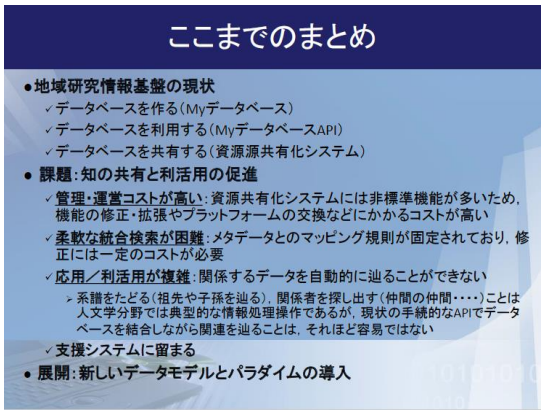
資源共有化システムの現状は、この図に示した通りです。東南地域研の外部のデータベースとして、人間文化研究機構の百数十個のデータベースのうち、地域研究に関連する国立民族学博物館、総合地球環境学研究所、北海道大学スラブ・ユーラシア研究所図書館、東京外語大学 AA 研、UC バークレー・東アジア図書館、ハーバード・イェンチンライブラリーの統合検索を実現しています（本発表後、ハーバードについては、情報システム変更により統合検索を停止中）。

このような成果を受けて、資源共有化システムの対象を東南アジアの大学図書館などに広げようとしていますが、進んでいません。東南アジア諸国の情報基盤のレベルがあまりにも大きくて、日本と遜色ないあるいは日本以上の国から、基本的な図書館情報システムすら整っていない国まで様々です。この状態で資源共有化システムを導入するのは困難と判断しました。



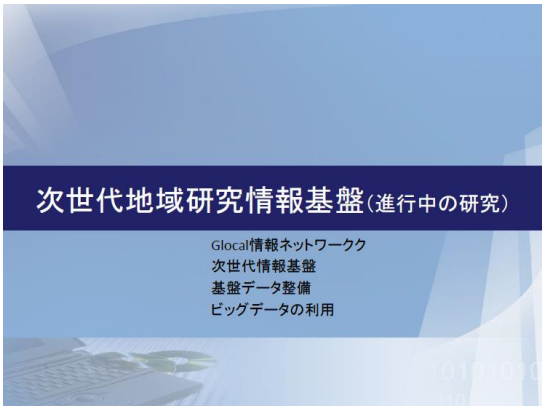
ところが、この議論を進めている中で、お互いがどのようなデジタルデータを持っ

ているのか知らないということが明らかになりました。たとえば東南地域研では三印法典などの資料価値の高いタイ語データベースを公開していますが、現地の人は知りませんでした。そこで、共有化はいったん諦めて、デジタルデータのインベントリシステムを共同で作ることにして、このスライドのような国際 Workshop などを開催しています。



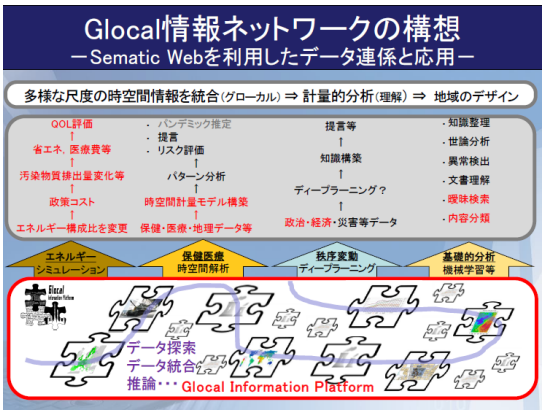
ここまでのまとめです。データベースを「作る」、「使う」、「共有する」ための地域研究情報基盤はだいたいできました。

ただし、幾つかの問題があります。というのは最初の設計が 1999 年でほぼ 20 年前です。よく 20 年も動いてきたなと思っています。ですが、古い技術や非標準の技術を多く使っているので、管理・運用コストがかかります。また、メタデータマッピングが固定されているので融通が利かない。それから、ただ検索するだけで終わってしまって、それ以上のこと、例えば関連する Web データに繋げることが難しい、といった問題があります。つまるところ、今のオープンデータやオープンサイエンスに対応することができません。



そこで、新しい地域研究情報基盤の構築に着手しました。これを Glocal 情報ネットワークと呼んでいます。Glocal は、global と local を合わせた jargon です。

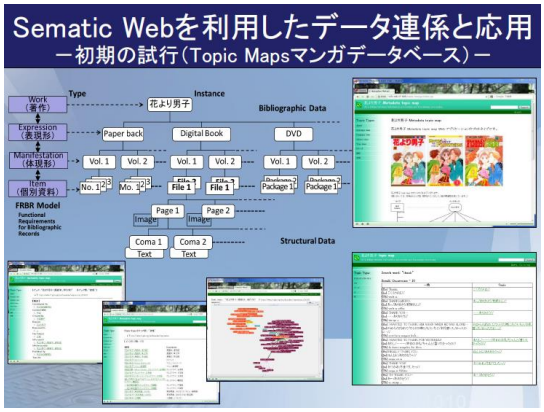
地域をいろいろな尺度、たとえば地球レベルや国のレベルで俯瞰したり、村のレベルで詳細に検討したり、時間についても長い時間で見たり、短い時間で見たり、時間の移動で見てみたりと、そういうことが柔軟にできるような情報基盤にすることを意図した名前です。



Glocal 情報ネットワークでは、大学内のデータベースとインターネット上の膨大な

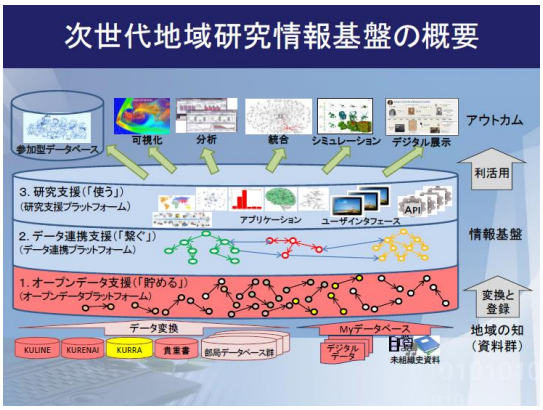
データのシームレスな結合を目指しています。多くの情報がインターネットにあって、これを使わないと地域研究も進まない状況になっているためです。たとえば巨大ハリケーンや巨大地震などの大規模災害が発生した場合、最新の情報はテレビなどのマスメディアではなくインターネット上を流通しています。大統領選挙のような政治情報や、パンデミックなどのような医療情報や、気象などの環境データでも同じです。

このような Web 上のビッグデータと我々のデータベースを効率的に繋げる、反対に、我々のデータベースを Web からより広く利用して貰えるためには、データ連係のための新しい技術的な枠組みが必要です。今のところ、Semantic Web 技術、特に Resource Description Framework（以下、RDF）による Linked Open Data（LOD）が回答ではないかと考えています。



少し脇道にそれますが、Semantic Web の機能を実装する枠組みとして RDF が有名ですが、最初のころは Topic Maps を使っていました。RDF よりも高度なデータの関連づけができることと、研究グループに Topic Maps の専門家いたためです。

この図は、Topic Maps を使って漫画のデータベースを構築した例です。面白いデータベースなので、これを話すだけで 1 時間はかかってしまうのですが、要は、漫画情報を本気で記述しようとするとネットワーク状になってしまうので、普通の図書目録など記述しきれず、Semantic Web の手法が必要になったということです。



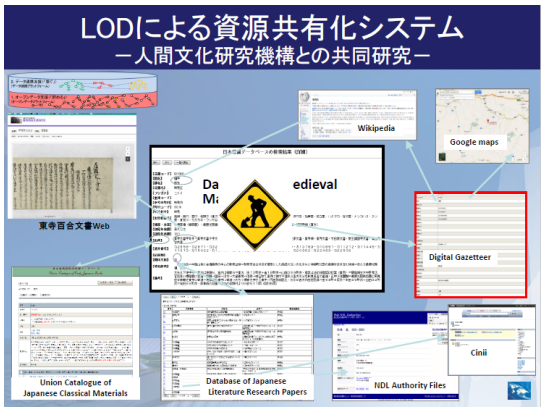
これは Glocal 情報ネットワークの構造と現状を表した図です。3 層構造となっています。

一番下の層が RDF データベースの実体です。M y データベースに蓄積されているメタデータを RDF トリプルに変換して蓄積する作業を続けています。

これらの RDF トリプルの主語と述語の語彙には、元のデータベースで使っていたデータ項目名を流用しています。語彙がバラバラなので、このままでは統合検索など

に利用できません。語彙の関係を記述したオンロジーが必要となり、それを実装するのが第 2 層です。これまでの資源共有化システムの機能を再現する場合、各メタデータのデータ項目名と共有化メタデータのデータ項目名の関係を RDF で定義します。あるいは図書館の視点でデータを統合するならば、MARC を中心したオントロジーを定義します。もちろん、独自のオントロジーを定義しても構いません。これにより、これまでの資源共有化システムよりも柔軟なデータの統合が可能となります。

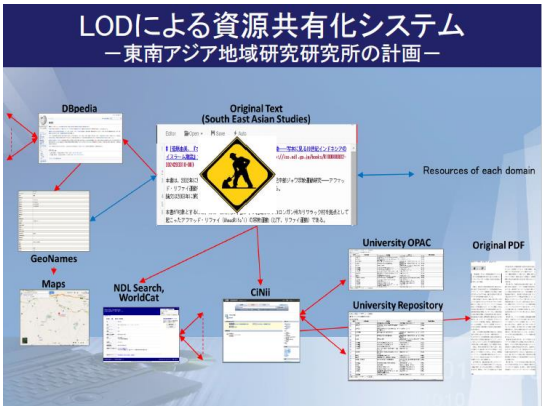
これまでに、第 1 層と第 2 層を少しずつ作り始めているところです。第 3 層は応用プログラムで、実装は進んでいません。この第 1 層と第 2 層を公開すれば、我々のデータベースをより広く利用して貰えるようになると期待しています。



その実験を人間文化研究機構と共同で行っています。これは、国立歴史民俗博物館の荘園データベースのあるレコードを起点として、関連する論文や地図データベースを辿れるように工夫した情報ツールです。例えば荘園の所在地名を使って地名辞書を検索し、地名辞書から荘園の緯度・経度を取り出して、地図上に荘園の位置を表示します。つまり、RDF で記述されている情報の所在 URI を参考にして、関連する情報を次々と自動的に繋いでいきます。

これまでキーワード検索は検索結果を表示して終わりでした。ですが、この情報ツールでは、検索をきっかけとして、関連していそうな情報をどんどん繋げていくので、新しい発見やヒントを得る上で有用な研究ツールなることが期待されます。

ただし、すべてのデータに URI を付けなければならないので、データ作成は時間と手間のかかる大変な作業になっています。



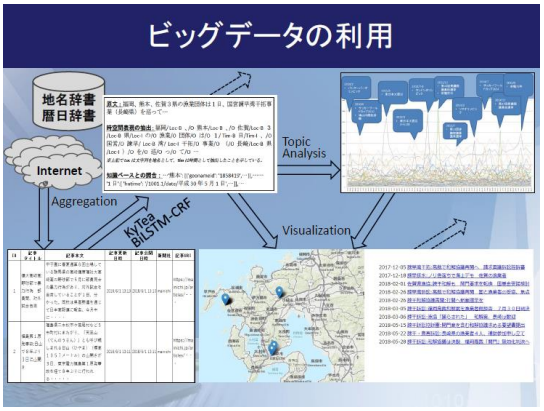
これは東南地域研で進めている別の実験です。論文検索システムは検索された論文が終点だったのですが、ここでは論文を起点にして、関連する情報へリンクします。残念ながら、これも工事中です。

先ほど地名を緯度・経度に変換すると言

いましたが、そこではデジタル地名辞書を利用します。現在の地名については国内外でフリーの Web サービスが幾つかあって、それらを使うことができますが、日本の歴史地名についてのフリーのデジタル辞書はありません。ないと困りますが、かと言って誰も作ってくれなかったので、我々で作ってみました。

幾つかの史資料を使っていますが、最初のデジタル地名辞書には大日本地名辞書を使いました。辞書の見出し語が地名なので、その内容を読んで場所を推定し、それを緯度経度に変換するという作業を繰り返しました。

この手順を応用して、延喜式の式内社名、明治につくられた旧 5 万分 1 地形図の全地名などを緯度・経度に変換した結果、これまでに約 37 万件の地名を収容したデジタル地名辞書を作りました。日本の歴史地名のフリーデータとしては、最大規模ではないかと考えています。このデータは、人間文化研究機構からダウンロードできます。



最後に、データベースとは違うのですが、進行中の 2 つの研究について手短にお話しします。

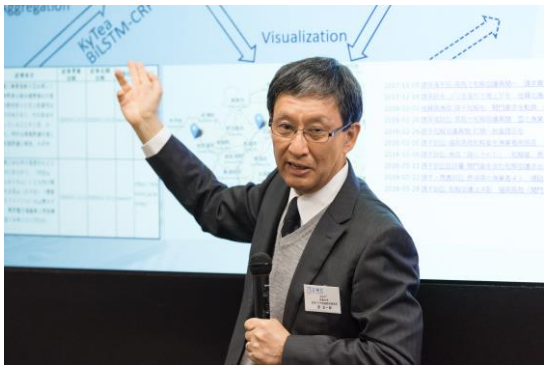
Glocal 情報ネットワークでは、大学内のデータベースとインターネット上の膨大なデータのシームレスな結合を目指すと言いましたが、これまでに説明したのは、自分達のデータベースの Linked Open Data 化についてでした。

現在進行中の研究の一つ目は、Web ビッグデータについてです。ここではソーシャルネットワークデータを使って、地域の状態を可視化することを目指しています。ソーシャルネットワークデータを使った地域研究の例はいくつもありますが、多くは、災害のような特定のデータ収集と分析が目的で、データ収集には、キーワード検索を非常に巧みに使っています。

ところが、この研究では地域の状況を俯瞰することが目的なので、特定のキーワードでデータを選別することはできません。そこで、ここではソーシャルネットワークデータを機械に分類させるという方法をと

っています。実験段階なので、ソーシャルネットワークデータと言いながら、実際は Web 上の新聞記事を利用しています。

手順は、この図に示すように、新聞データを定期的に crawling して、広告などの不要部分を取り除いて、記事テキストデータを作成します。次に記事のテキストを単語に分解します。その際に地名や時間名に関する語彙を抽出して、地名についてはデジタル地名辞書に当てはめて緯度・経度に変換します。単語単位に分割されたテキストにトピックモデルという手法を使って、内容を 200 種類のトピック言い換えれば主題に機械的に分類します。最後に、記事を地図上に可視化します。これによって、地域を 200 の観点から俯瞰することができるようになります。字句解析とトピックモデルの処理は機械学習の応用です。



これまでの地域研究のように、少数の文献やデータを緻密に読み解くのではなく、ざっくりであるが全体的に眺めるというアプローチです。最近 Distant Reading という言葉が流行っていますが、それに類した方法と考えています。必ずしも正確とは言えない大量のデータを含めて分析することで、地域をより全体的に俯瞰できるようになり、新しい研究のヒントが見えてくる可能性があれば、素晴らしいかなと考えています。

最後に、研究データの長期保存はいま喫緊の問題です。検討を始めましたという段階なので、笑い話としてお聞き下さい。

メディア変換、いまはハードディスクやクラウドを使っているので、ほとんど問題ないです。ですが、フロッピーディスクなどの昔の媒体に入っているデータの変換は大変です。何といっても、装置がなくなりつつありますから。

ワープロ専用機なんて若い方は殆ど知らないでしょうね。このころに作成された文書データは、今となってはほとんど使えないと思いますが、プレインテキストでよければ、何とかなるかもしれません。

ゲーム、これは実は簡単だと思っていたのですが、専門家にお聞きしたら、非常に難しいということが分かりました。プログラムはエミュレータなどを使えば何とかな

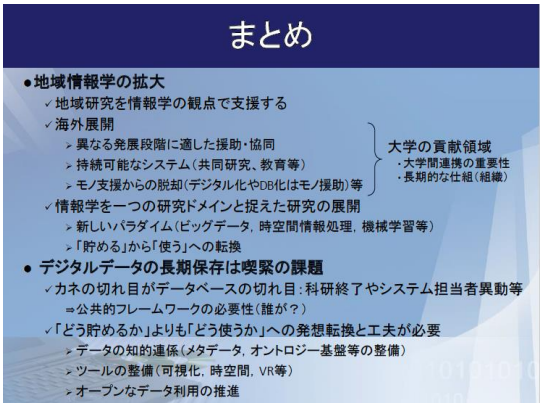
りますが、問題はジョイスティックなどのハードウエアです。設計図を保存しておいて、必要に応じて作るしかないようなので、これは意外に難しいです。

肝心のデータベースです。もしデータベースソフトがなくなってデータしか残っていなかったら、上の図のようなバイナリデータを分析するしかありません。かつて、この経験をしたことがありますが、本当にたいへんな作業です。

次の図は、データだけが読めた場合です。何のことか分かりません。

次の図ですが、これだと血液検査データかなと、何となく分かる気がしますが、実際には使い物になりません。まともな研究者だったら絶対に使わないでしょう。なぜかというと、例えば、最初の項目は GOT で単位は U/L と分かります。ちなみに GOT は肝臓の機能を調べるものです。しかし、どんな計測法を使ったのか、どんな人が検査の対象だったのか、どのような状態で血液を採取したのかなどが明らかでなければ、比較研究に使うことができません。

先ほど言ったように、研究データの長期保存が喫緊の問題となっていて、各大学では図書館やメディアセンターなどが中心となって検討しています。ですが、話を聞いていると、ビットデータを安全に蓄積することが話題の中心になっていて、データを利用する際に必要となる情報についての議論はあまりされていないし、その重要性を述べても、なかなか理解されていないという印象を受けています。適切な標準メタデータが定義されれば良いのでしょうが、当面は、機器の説明書、実験ノート、論文、あるいは手書きのメモなど関連する文書情報をデータ本体に混載させるしかないのかもしれません。



メタデータの話からだいぶズレてしまいましたが、まとめです。

ここでも最後にちょっと書いたのですが、研究データの長期保存の問題は、どう保存するかよりも、どうやって使い続けるかを考えたほうが良いのではないかと考えています。そのためには、先ほど述べたデータのオープン化とデータの関連付けを続けて、色々な経路から古いデータに辿り着ける、あるいは発見できるようにすることが重要になると思われます。そのためのオープン環境あるいは情報インフラの整備を進めていく必要があり、そうなってくると、メタデータのあり方がとても重要で、ことによったら、もう一度再考しなければなら

いのかもしれません。

まとまりのない話になってしまいましたが、以上です。ご静聴、どうもありがとうございました。

原 正一郎（はら しょういちろう）

1957 年、千葉県生まれ。京都大学東南アジア地域研究研究所教授。専攻は情報学。共著に『歴史 GIS の地平』（勉誠出版）、『ナチュラルコンピューテーション 1』（パーソナルメディア）、翻訳書にM・ジェームズ『人工知能 BASIC』（啓学出版）など。