情報学の動向——メタデータを主題として

原 正一郎 京都大学地域研究統合情報センター

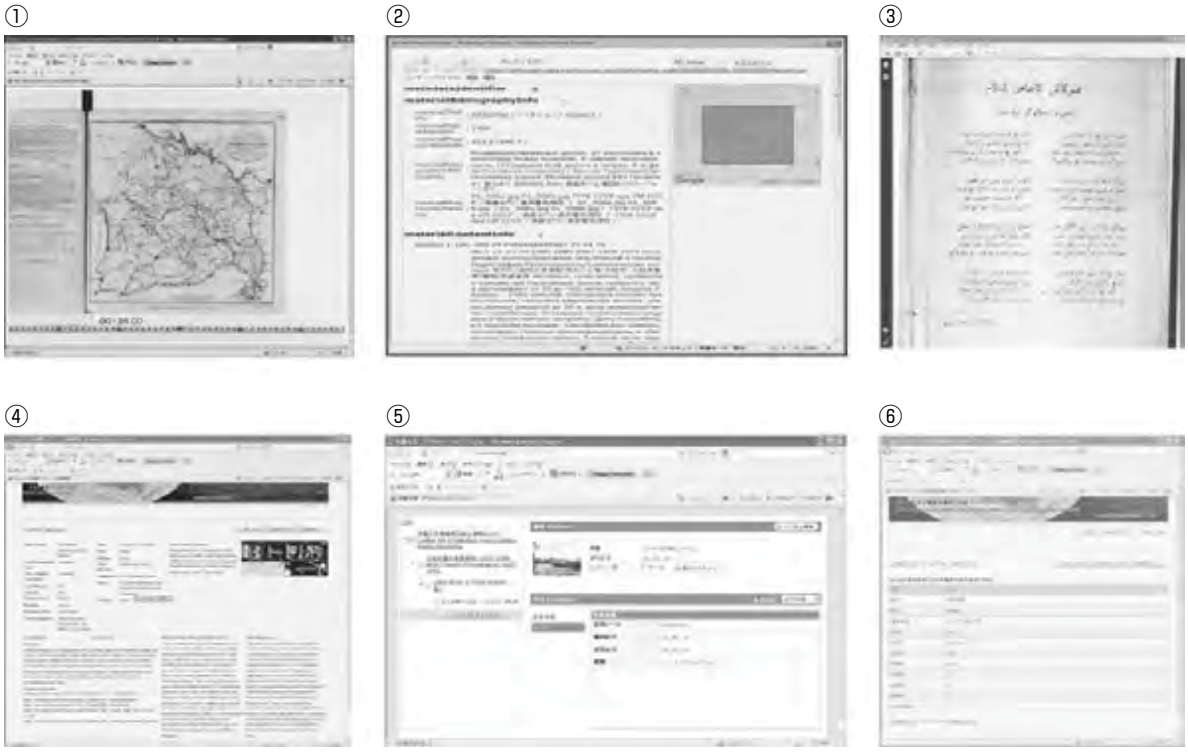
本日はデータベースについてお話しいたします。なおデータベースと言っても色々なトピックスがありますが、ここでは皆さんにはあまり馴染みのない、しかしデータベースを作るためにはとても重要な「メタデータ」についてのお話をします。

地図、雑誌、写真——多様なデータベースに共通するインデックス・データ

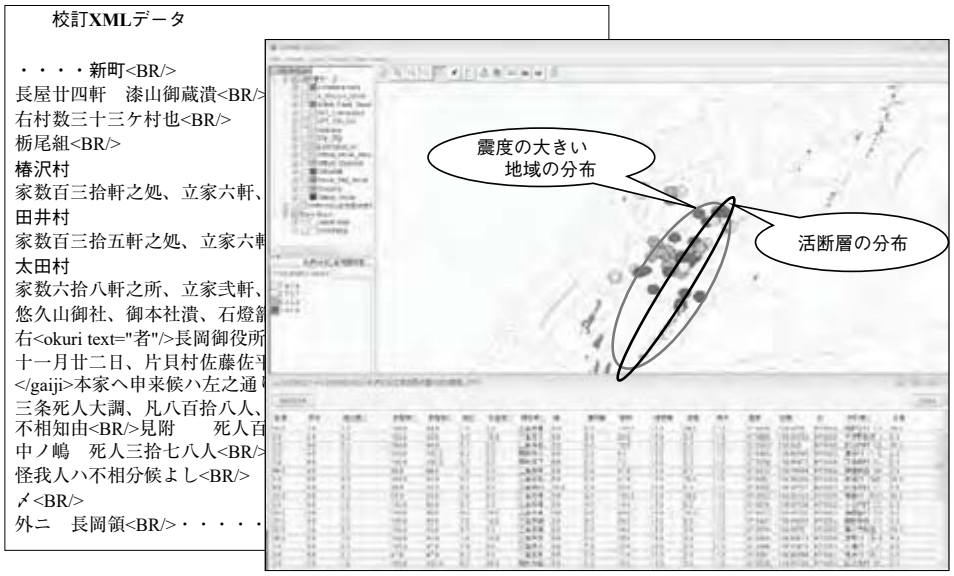
資料10-1に掲げたデータベースは地域研が公開しているデータベースの一部です。これまでにいろいろなデータベースを作ってきました。たとえば、①は英国議会資料に掲載されている古地図のデータベースです。それに対して②は、旧ソ連で作成された現在の地図をデータベース化したものです。③は雑誌記事データベース、④はインド映画のデータベース、⑤は著名な地域研究者がタイで撮影した写真のアーカイブデータベース、⑥は東ヨーロッパの議会選挙資料のデータベースです。

地域研は世界中を対象としているので、このように様々な地域の様々な資料を対象としたデータベースを作っています。内容は様々ですが、これらのデータベースには共通するものがあります。それは、地図や本や写真などを検索するために作られたインデックスというデータです。

資料10-2に掲げたデータベースは、先ほどのデータベースとは少し違っていて、文書の内容そのものをデータベース化したものです。これは1828年に、今の新潟県を中心に発生した地震の記録資料の一部です。どこで家が何件倒れたのか、どこで人が何人が死んだのかといった情報が書かれています。これを表の形に整理して、緯度・経度をつけて、地図上に表示したもの



資料10-1　京都大学地域研究統合情報センターで作成したデータベース



資料10-2　テキストのデータベース――歴史史料から見える災害の例

がこの図です。

この図で、濃いグレーの丸は非常に揺れが強かった場所、薄いグレーの丸は揺れが弱かった場所を示しています。このように揺れのデータと断層のデータとを重ねあわせると、地震がどこを震源として発生したかなど推測することができます。今回の災害マッピングとは別のものですが、災害予測などに利用できるかもしれない、なかなか興味深いデータベースです。

しかし本日はこの話をするわけではありません。このようにいろいろなデータベースを作ってきた経験から、どのようにデータベースを作ったら良いか、ということがテーマです。ですが、それを説明することはなかなかたいへんなので、まずは「このようなデータベースを作ったら使えない」というかたちで説明したいと思います。

■ どのようなデータベースを作ってしまうと「使えないもの」になるか

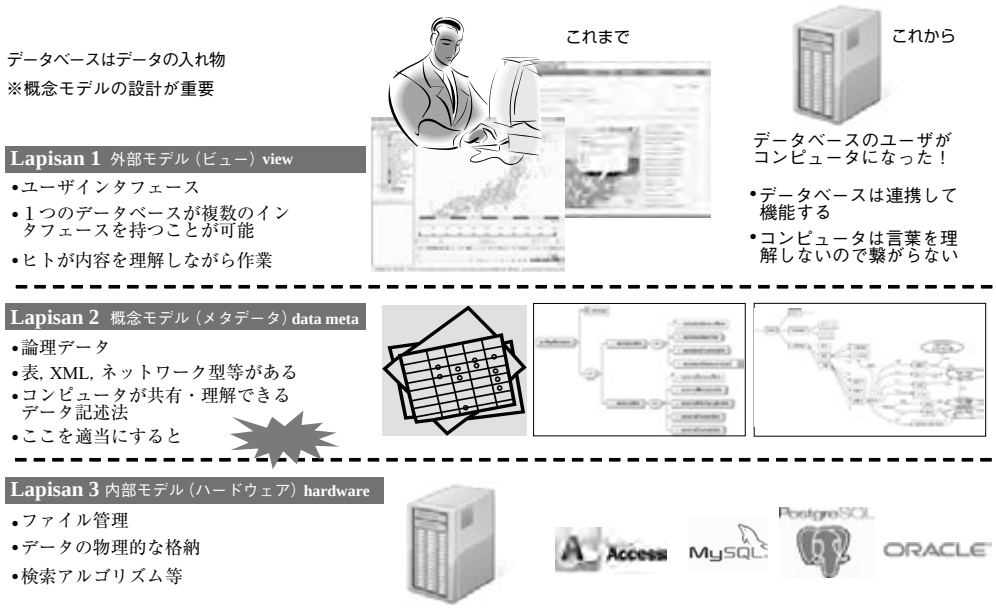
データベースを作るときには、資料の所在を調べたり、資料を集めたりします。それからインデックスをつくります。このインデックスの作り方が良くないと、データベースは使いものになりません。これがデータベースを短命にする最初の極意です。

次に二つ目の極意です。地域研究では写真などをよく撮りますが、そのデジタル化の方法を間違えると、やはりデータベースは使いものになりません。皆さんもデジタルカメラをお持ちだと思います。殆どのデジタルカメラでは、JPEGという形式で画像を蓄えています。JPEGでは撮影した画像データに圧縮という操作を施しています。これによって画像データを蓄積するディスクの容量は節約できるのですが、その代償として画像の情報が落ちてしまいます。つまり元の画像に比べるとボケてしまいます。今は問題ないかもしれませんが、10年後や20年後にその画像データを使おうとすると問題となる可能性があります。つまり、デジタルデータをどのように保存したら良いのかという話と、どのように使ったら良いのかは別の話です。使い方は現在の問題ですが、保存は将来を見越した問題です。

さてメタデータとは何か。メタデータとは、先ほどまでインデックスと呼んでいたデータで、資料の中身についてのデータです。

ちょっとたとえ話をしましょう。ここに水の入ったペットボトルがあります。私はこれを安心して飲んでいますが、それは中身が信用のある会社が販売している水であると分かっているからです。しかしラベルの貼られていないボトルを見たら、皆さんは飲みますか？　おそらく皆さんは飲まないでしょう。このラベルには「これは○○という場所でつくった安全な水だ」と書いてあります。つまりボトルの内容に関するデータです。これがメタデータです。

いまメタデータとは中身についてのデータであると言いましたが、もう少しデータベースに即して言い直すと、どのような内容のデータがどのように書かれているかについてのデータとなります。この作り方が悪いと良いデータベースは作れないということは、このたとえ話から何となく理解していただけたものと



資料10-3　データベースとユーザー、アプリケーションの位置付け

思います。

■ 有効なデータベースの構築にはデータ内容を正確に記したメタデータが必要

次にメタデータの位置づけです。資料10-3の図は情報学の世界でよく使うデータベースモデルです。皆さんがデータベースと呼ぶのは、図の下の部分の「内部モデル」ではないでしょうか。OracleやAccessなどお馴染みの名前が並んでいます。ここは、どちらかというとハードウェアに近い部分です。一方、データベースのユーザは、上の部分の「外部モデル」に関心があります。地図の検索に適した検索方法や、自分の研究に都合の良いユーザインタフェースを実現するアプリケーションプログラムの部分です。

もう想像がつくでしょうか。データベースがきちんと動くためには、下のハードウェア部分と上のアプリケーション部分の中間にある「概念モデル」がうまくできていなければなりません。そして、ここがメタデータの部分になります。

ところで、これまでのデータベースのユーザは人でしたが、最近ではコンピュータのアプリケーションプログラムがデータベースのユーザとなっているケースが当たり前になっています。いわゆるe-コマースなどが典型的な例でしょう。人間の脳はとても柔軟なので、曖昧なデータであっても正しく理解することができます。しかし現在のコンピュータは、それほど賢くはないので、曖昧なデータを処理することが苦手です。ですからコンピュータがユーザとなるデータベースシステムを構築するためには、データの内容を正確に記述したメタデータが重要になるわけです。

資料10-4　問題のあるメタデータ

番号	生年月日	性別	氏名	・・・
0001	1957/10/11	M	大学太郎	・・・

ID	SEX	surname	forename	birthday
1	1	Daigaku	Taro	11 Oktober, tahun 32 Showa

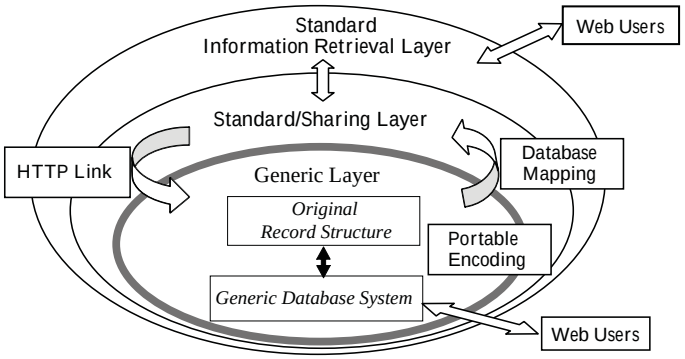
同一内容であるが, 同一処理は適用できない
語彙が異なる／語彙の粒度が異なる／語彙の順序が異なる／値の記述法が異なる

●統合検索・共有化が困難
●定型的／機械的処理が困難
ヒトによる解釈支援が必要
検索やデータ分析の自動化が困難
●メタデータは標準に準拠した方が使いやすい
基本的な語彙や記述法が定義されている
少なくともデータ要素の識別は容易になる
検索・統合・分析等が容易になる

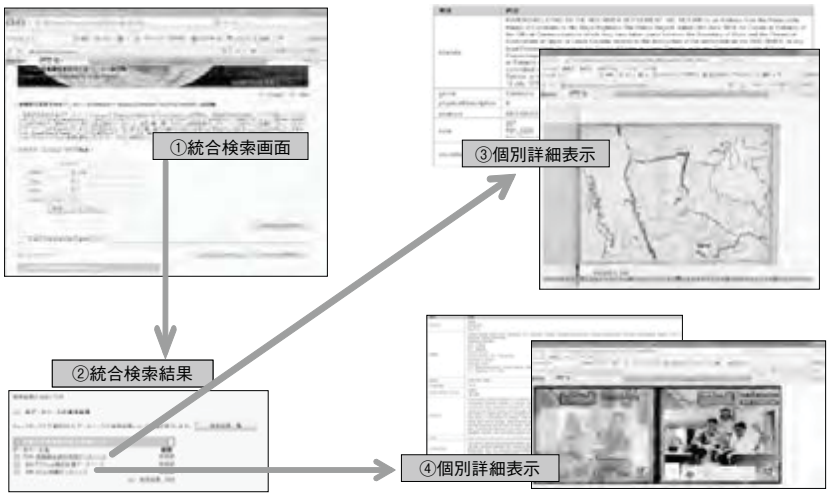
■ メタデータがバラバラではデータの共有が不可能

資料10-4にはメタデータの例を示しました。メタデータで重要な事柄は、データの順序関係と、データを示す名前と、データ要素の内容の書き方の3点です。ここでは表として示していますが、先ほどの3つの条件を満たしてさえいれば、メタデータの表現方法は、どんなものであっても構いません。

実はここに示した2つの表はメタデータの悪い例です。両方の表は同じ内容を表しています。皆さんが二つの表を眺めれば、それぞれの表のどの項目がどの項目に対応しているか、一目瞭然でしょう。でもコンピュータは、たとえば「生年月日」と「birthday」とが同



資料10-5　データ共有化のシステム



資料10-6　地域研の資源共有化システムの検索画面

じデータとは分かりません。仮に同じデータであると分かったとしても、今度は「1957/10/11」と「昭和32年10月11日」が同じ日付であるかが分かりません。あるいは性別とsexを見ると、片方は「M」、もう片方は「1」になっています。皆さんであれば、MはMaleからきているので男性だと想像がつくでしょう。ちなみに「1」というのは国際表示で、「男」を表します。つまりメタデータがバラバラであると、データベースの検索方法もデータ処理もバラバラになり、データ統合や共有を実現することがとても難しくなるのです。使いにくいデータベースは長生きできません。

アチェの震災のあとにいろいろなデータベースが構築されたと聞いています。残念ながら、それらのデータベースのメタデータはバラバラでデータの共有ができないなどの問題が起こっているのではないかと思います。

既存の使い勝手の悪いデータベースを共有化する当面の解決法

使いやすいデータベースを作るためには、メタデータの設計が重要であるということは分かっていただけたと思います。では既に作ってしまったデータベースは使い物にならないのでしょうか。

資料10-5は資源共有化システムと呼んでいる情報システムの仕組みを説明したもので、メタデータの異なるデータベースを統合することができます。この仕掛けを簡単に説明します。たとえば資料10-4の図にあった性別は、上の表では「性別」、下の表では「sex」となっていたため、コンピュータはこれらが同じであることが分かりませんでした。資源共有化システムでは、それぞれのデータベースの項目間の対応情報を保存しています。ですから、あるユーザーが資源共有化システムで性別を検索したとき、資源共有化システムはあるデータベースに対しては「性別」という項目で検索し、別のデータベースに対しては「sex」で検索します。つまりそれぞれのデータベースのメタデータの違いを意識しなくとも、ネットワーク上の多数のデータベースを統合検索することができるようになります。

資料10-6は資源共有化システムの検索画面の例で

す。この画面の後ろには30以上のデータベースが隠れています。資源共有化システムにある検索語を入れると、30以上のデータベースを同時に検索し、その検索語に関連する地図や映画や論文に関する情報を提示します。アチェにおいても、メタデータの異なるデータベースが混在している場合には、この方法が使えるのではないかと考えます。

標準メタデータを探し、組み合わせてデータベースを構築する

ここから先は「こうしたほうが良いだろう」と考えている事柄を短く紹介します。情報の世界は動きが激しいので、「これが正しい」方向と断定することはできません。当面考えられるベターな方向性はこれではないか、という話です。技術な話はしません。

繰り返しになりますが、データベースの寿命が短いのは、勝手なメタデータを使っているためにデータベースの使い勝手が悪いためでした。そうであるならば、既に作成されていて、しかも広く使われているメタデータを皆で使えば良いのではないかということになります。そのようなメタデータを標準メタデータと呼ぶことがあります。

世の中にはたくさんの標準メタデータがあります。本、写真データ、地理データなど、それぞれの分野やメディアごとに様々な標準が用意されています。これはそのような標準の極々一部にすぎません。ですから、データベースを作る際に最初にすべきことは、既に存在している標準で使えるものがないかを調べることです。

もし、ある標準のなかに欲しいデータ要素がなかった場合でも、別の標準の中に使えるものがあれば、それらの標準を組みあわせる方法があります。これをアプリケーションプロファイルと呼び、メタデータの最新の使い方となっています。このような技術を駆使す

資料10-7　標準メタデータ――語彙の共有

Trend Dunia: Datalink di atas jaringan network
世界の趨勢はネットワーク上でのデータリンクと高度利用
そのためには語彙と記述法の標準化が必須
標準メタデータの利用は重要

Dublin Core
WWW上におけるリソースに関する情報を記述：
http://dublincore.org/

MODS（Metadata Object Description Schema）
簡略版XML ベースMARC21：
http://www.loc.gov/standards/mods/

EAD(Encoded Archival Definition)
アーカイブズ用メタデータ: http://www.loc.gov/ead/

GML(Geography Markup Language)
空間データや位置情報を含む各種のコンテンツを記述：ISO 19136:2007

語彙のレポジトリ
Meta Bridge：総務省「新ICT利活用サービス創出支援事業」の一つ
http://www.metabridge.jp/infolib/metabridge/menu/
WordNet http://wordnet.princeton.edu/

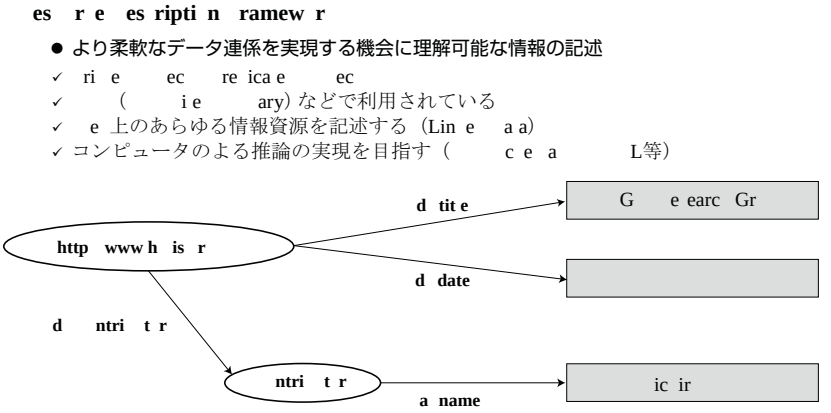
シンポジウム／ワークショップに参加して

アチェ震災情報のデジタル化と共有化

原 正一郎

今回の国際ワークショップにおける一番の印象は、日本語とインドネシア語によるバイリンガルの討論であった。従来の国際ワークショップ、とりわけ理工系の場合は英語が基本である。多くの国や地域から研究者が集まるので致し方ないことではあるが、自分自身を含めて語学スキルがそれほど高くないものどうしでは、微に入り細にわたる意見交換が困難で、もどかしい思いをしたことが多い。日本語とインドネシア語によるバイリンガルな討論が成り立つためには、インドネシア語が堪能な上に関連する学術知識の豊富な通訳の存在が欠かせない。今回は西准教授がその役割を果たしたわけであるが、地域に根を下ろし現地語を駆使して研究を行うという地域研究の底力の一端を垣間見た思いであった。

さて今回はワークショップの合間を縫って、文書館や博物館などを見学する機会があった。国立公文書館アチェ州分室では津波により多くの史資料が流出・水損していた。水損史資料については日本を含む外国の修復保存専門家らによる救助活動が展開されたと聞いたが、どれだけの史資料が失われ、どれだけが修復されたのかはわからなかった。現在、ここには津波災害に関する行政文書が集積されており、目録化が進められている。残念ながら、集積された文書量に比べるとアーキビストの人数が少ないためか、目録作成作業の進み具合はそれほど速くはない印象を受けた。個人的には、どのようなメタデー



資料10-8　Resource Description Framework(RDF):Mendiskripsi data meta――メタデータの記述

ることで、異なるデータベースをネットワーク上で繋げることが可能になります。

柔軟なデータ連携を実現するRDFデータの可能性

資料10-8はRDFというセマンティックWeb技術を簡単に説明した図です。技術的な説明は省きますが、従来のデータベースなどよりも柔軟なデータ連係を実現する手法として注目されており、今後の基盤技術となる可能性が高いと考えています。

この図は世界中で共有されているRDFデータの数です。2011年9月の段階ですが、310億タプルが繋がっています。ある意味で知識の膨大な集積ということができます。将来、アチェの重要なデータもこのようなRDFの形に変換することにより、世界中のほかの災害データとの統合に貢献できるのではないかと期待しています。

タや記述ルールを採用しているのかに興味があったが、短時間の訪問であったので詳しく聞くことはできなかった。ただしジャカルタのアーカイブ専門家や日本の国立公文書館などの支援を受けているようなので、標準メタデータに則した目録が作成されている可能性がある。データベース共有化への可能性に期待している。

津波博物館には、地震や津波の発生メカニズムなどの解説や地震体験装置など、地震および津波災害に関する教育的展示は一通り用意されていたが、充実しているという印象にはほど遠かった。予算不足が主な理由と思われるが、それ以外にモノ資料の収集が進んでいるのか、所有権などの権利問題は解決しているのか、調査・整理が追いついているのかなどの疑問を持ったが詳細はわからなかった。阪神・淡路大震災の史資料管理に多少とも関与した立場からは、災害関連のモノ資料の展示が今度どのように展開される予定なのか、写真や動画や音声資料の収集・保存状況はどのようになっているのか、それらの整理手法やメタデータはどのようになっているにかに一番の関心があったが、見学のみで担当者に尋ねる時間がなかったのは少々残念であった。

今回のワークショップのテーマの一つは震災情報のデジタル化と共有化であった。ワークショップにおける発表や短時間であったが何カ所かの施設訪問を通じ、多くの情報がデジタル化されたものの収集や共有化は進まなかったとの印象を受けた。震災直後から中央政府、各国政府、NGO、NPOなどによる様々な専門的援助活動が実施されたと聞いているが、情報収集・管理についてどのような組織がどのように連携して活動していたのかは不明であり、今後の震災情報の組織化を考える上でも検証する必要があると感じた次第である。