資料知識のカタロギング方法論：情報多様性と標準

鯨井　秀伸

木村定三コレクションの受入により美術館のコレクションは質量共に変化した。そのカバーする地域および時代は広範なものとなり、いわば、ひとつの機関で複数の機関のコレクションを含むようなものとなっている。受入時から基本的な目録作成までの過程は池田氏の報告に述べられている。そこに記されているように、資料の記録記述については新たに異なった形式のカードがいくつか用意されたが、それは資料の分野によって記述形式が異なるからである。こうした異種で多様な記述形式に対応したドキュメント形式をどのように統合的に記録するかが、カタロギングとその活用の要点となる。現在カタロギングの対象となる資料の概略が把握され、個々の資料が属す研究分野の情報カテゴリーの詳細に関する情報を収集している段階であり、それらに基づいた概念化によって既存のカタログ方式の中での処理を行っている。ここではこうした多様で異種の資料群のカタロギングについて考えてみたい。情報は異なった機関においてさまざまな仕方で構成される。この観点から、国際博物館会議のデータ・カテゴリー CIDOC IC（Information Category）は、「構造」に対抗して「コンテンツ」の標準化に向けた努力を表していたにもかかわらず、美術館、考古博物館、自然史博物館、遺跡や記念建造物、保存修復、分析研究所、保全などを同時に満たすような「ひとつのサイズで全てに合う」標準あるいはデータ・スキームはないのである。ドキュメンテーションは証拠を伴った記録という意味であり、アナログ資料の蓄積整理と管理とは基本的な事柄だが、現在のデジタル化に対応する必要もあり、デジタル化しないかぎり、グローバルな利用という情報流通にのれない状況も実際問題として生まれつつある。「Being Digital」という言葉はこの事実を表している。行政文書を含めて多くのアーカイブが現実的に「生まれながらのデジタル」となってきており、従来の「記録」や「文書」の意味をも変えつつある。これは、近年アーカイブ学が新たに取り上げられるようになった理由のひとつである。こうした現状をふまえ「記憶機関（Memory Institution）」のひとつにおける文化遺産の記録という視点からカタロギングについて考えてみたい。

文化遺産情報は大別して二つの基本的な要素から成り立っている。ひとつは機関としてのコレクション管理的機能を持つ管理情報（Administrative Information）であり、通常この機能が一般的に受け入れられよく知られている。もうひとつはいわば研究情報（Research Information）に属するもので、文化遺産が属す学問分野によってその特質が異なるが、それら研究情報は多くの場合テクスト情報あるいはイメージ情報として提供され、こうした情報の中で、主にテクスト情報から文化遺産に特有の情報要素を抽出したものは研究情報とみなしうるものである。イメージ情報や他のメディア情報についても、コンテンツ管理の情報技術により管理されるようになってきている。この情報は書誌情報あるいは文書情報として他の機関により索引が提供されている。これら二つの情報は互に関係し合い補完し合っており、管理情報単独では典拠に欠ける情報とならざるをえないのである。例えば最も基本的な5W1Hの各情報要素を考えてみても、それら各要素は互に関連しあっているのだが、「それは何か」を示す「What」に相当する「品名（資料名、タイトル等）」は、研究情報による裏付けがなければ単なる呼称でしかなく、他領域への参照が不可能となる。この要素は言語文化によって性質を異にするようで、日本語文化圏においては所与のものとみなされる傾向にあり、普通名詞や固有名詞的に扱われ、欧米の概念「Title／Name」とは異なるようである。この「Title／Name」の中の要素の「Name」が日本の場合に対応するのだろうが、「Title」の方は芸術に特有の概念である。この場合例えば「メレアグロスの死」がこれに相当する。この「メレアグロスの死」というTitleは、イメージの「記述」「同定」「解釈」を通して得られる、ある媒体の上あるいは中に表現された「主題」の要約のことなのであって、物の名称と言うわけではなく、強いていえば「主題」の簡略化した呼称なのである。したがって、何が表現されているかは主題研究の展開に従うことになり、その結果によって「タイトル」は変更されることになり、わが国で一般的に受け入れられている「所与」の名称ではないのである。同様のことは、異なった議論になりここでは略すが、「作者名（創作者名、制作者名、制作団体等）」「制作年」などの要素についても言えることである。こうした「人」「時」「場所」「出来事」は互に密接に関連しており分離することはできない。こうしたいわば名付けの方法・仕方は、文化圏により異なり、その方法の基点を記録する者の文化圏に置くか、記録される資料の文化圏に置くかで全く異なったものとなるであろうことは容易に想像されることと思う。この問題は情報処理技術の展開と共に文化情報的情報交換・共有を行う研究者によって、言語文化の問題としても気づかれ始めているものだが、近年では情報多様性（Infodiversity）上の課題として考えられるようになっている。

いわゆる「記憶機関」からの文化コンテンツは典型的に典拠が確かなものであり、流通している情報の中でもその質は高く、教育や専門的研究に有用であって、広範な人々の魅力を引きつける力を持っている。しかしこれらの機関によって管理運営される文化情報は、構造や内容において異種のものである傾向がある。これらの機関は博物館や図書館、文書館であるが、各機関内においてもそのコレクションに応じて異なった情報の管理運営が行われていることが多い。これら機関の管理運営者による情報は、概念的なオーバーラップが目立つにもかかわらず、情報多様性を認め活かしつつ異種の文化遺産情報を統合していくことが目標となっている。このとき情報の概念化と共にメタデータ概念が注目を集めるようになってきているのである。コレクションおよびそのコンテンツ記述は以下のような要因に応じて変化する可能性を持っている。

1） コレクションのタイプ
2） 機関管理者のアプローチ法
3） 対象となる学問分野
4） 記述の粒度
5） 記述の詳細のレベル
6） 記述的データ構造
7） 記述的データ内容の値

これらの要因は客観的な情報要因であるばかりではなく、研究者・管理者のマインドそのものでもあるということが重要なことであって、この多様性のため、全ての博物館、図書館そして文書館のニーズに合うような単一の記述的スキーマも考案されなかった。その代わり過度のローカル化されたドメインやアプリケーション固有の標準や実践がコレクションのドキュメンテーションのために発展してきたのだった。記述情報の多様性は、自然界のエコシステムにおける生物多様性（Biodiversity）に比較しうる。情報多様性は多様でダイナミックな情報エコシステムの自然で適切な反応なのである。しかしながら、博物館、図書館、文書館間を横断する記述スキームにおけるこうした相違は個々のアプリケーションにとって必要なものではあるが、文化情報資源の情報が流通するコンテクストにおいて、クロス・ドメインの発見や相互利用性の妨げとなるものでもある。幸なことにこの相違は、第一義的にはデータ構造とシンタクスのレベルにある。顕著な概念的オーバーラップは記憶機関によって使用される記述スキーマの間に存在するのである。要素的概念、オブジェクト（対象資料）、人、場所、出来事（事象）およびそれらの間の内的な関係性はほとんどが普遍的なものなのである。異種の情報源を横断するアクセスを提供するための伝統的妥協案は、初期資源の発見の目的のために広範で普遍的なセマンティクス（意味論）により全てを単純なスキームにマッピングすることであった。こうしたシンプルな記述は —— これは時に「資源発見メタデータ（resource discovery metadata）」と呼ばれるが —— 研究者が資源を発見し評価するのを助け本来の形式における豊かな記述へと導くものであるが、クロス・ドメイン資源発見へのこのアプローチは確かにいくつかの利点を持っている。なぜならそれは潜在的にシンプルなキーワード検索よりより適格な検索結果を生み出したからである。例えばダブリン・コア・メタデータ・エレメント・セット（DC）は本来的に資源発見のために意図されていた。これは他の多くの目的のために利用されてきたが（それは誤利用であると主張されることも多い）、このアプローチには多くの欠点があったのである。第一に情報源記述（source description）は、資源発見スキームという広範な普遍的意味論へといわば「沈黙」させられているのが現状であるため、アクセスは最小公分母に縮約されており、それは記号論的不可逆的等価物という、大規模データベースを横断する洗練されたクエリーや検索精度に対する適切な援助は提供しないかもしれないからである。第二に、さまざまな意味論的妥協がなされなければならないため、マルチプルで豊富なデータ資源を、矛盾なくよりシンプルなスキームにマッピングすることは非常に難しい（現実的に欲求不満的であり単調で退屈であることについては言及しないが）ものである。第三にユーザが単純な資源発見メタデータを使って関連する資源を発見できたとしても、より豊富で利用価値のある情報源にアクセスできるかどうかは保証されないからである。結果として、コレクションとその管理団体である機関は、ユーザの期待には及ばないであろう単純な記述によってしか表されたり判断されたりしないのである。

意味論に基づいた概念化によるモデルは、異種の文化遺産情報への意味をもった統合されたアクセスを提供するという、上記の問題に対する代替の解決策を提供するのである。意図するところが博物館情報と書誌情報とを統合しようとすることにある場合、この情報多様性はさらに重要なものとなる。博物館情報と書誌情報の統合の利点はエンドユーザにとって計り知れないものがある。所在の場所、絵画の制作年あるいはイコノグラフィー、その作品の研究、アクセス可能な複製物ないしはデジタル化された画像そしてその作品への参照を含む仕事上のツールなどの情報を検索することが可能となる。ここにおいて、情報多様性はそれ自体悪と言うわけではなく、反対に善であると主張することができる。同じ標準を利用するために、異なった目的や異なった必要性をもつ異なった機関を無理に従わせることは不可能であり望ましくないものであるから、情報統合は非直接的手段を通して、それぞれ単一の機関内において改善のレベル以上に高度なレベルにおける標準化を通して、さまざまに構成された情報システムから共通の意味論的な価値を抽出することができる仲介ツールを利用しなければならない。

仲介モデルは、オントロジーに基づくか、あるいはより控えめな表現をすれば、意味論的参照モデルに基づくものである。意味論的参照モデルは、あらゆる相互関係のある「クラス」あるいは「エンティティ」の言語的に総合的な（synthetic）視点を提供してくれる。その存在は所与のドメインにおいて認識され、その間およびその中に存在する関係性の適切な意味を定義する。国際博物館会議の国際ドキュメンテーション委員会のSIGが開発したCRMはそうした意味論参照モデルのひとつである。これは1996年以来開発が続けられ昨年2006年にISO 21127としてISO標準として発行された。CIDOC CRMは、システム間のデータ交換の共通基盤、また統合クエリー・ツールのための基準、仲介システムのデータ・スキームとして機能する。CRMは博物館情報に焦点を合わせているが、それを書誌情報および博物館情報統合のコンテクストにおいて使用することが可能であると証明されている。博物館資料に対して、CRMにおいて定義されている記号の意味関係の多くは書誌情報の記述にもまた有効である。主要な難点は、実は書誌記述が図書館に実際に所蔵されている個々のアイテムによって体現される「出版物」の抽象的概念（notion）に焦点を合わせているのに対し、博物館の記述が物質的で、個々の「ユニーク」な対象（Object）に関連する

という事実にある。CRM SIGはFRBR（書誌レコードの機能要件）モデルを表現することによってこうした二つの視点を調和させようと努力している。これは書誌情報とそれをCRMへ嵌め込むためのIFLAによって開発された概念である。2005年のICHIM会議の時点において、この概念を公的に利用可能にしたドキュメンテーションはなかったため、フランス国立図書館の開発グループは、FRBRとCRMの調和を待たず、CRMモデルのみを使用してSCULPTEURモデルを開発した。FRBRはER理論を採用しているが、これはエンティティ・リレーション・モデルといい、博物館資料の知識集約と目録作成には当初から使用されてきた理論だった。2000年代になって概念化が重要な理論的背景となってから、文化遺産を初めとして図書館資料にも応用されるようになったのである。こうした志向性は文化遺産全般のカタロギングの基盤に同じような資料把握の方法論が採用されるようになった、あるいはそうした方法を取らなければ各資料（object）を理解し認識し管理できなくなってきていることの明らかな証拠でもある。

データ統合は現代の情報システムの展開に鍵となる問題である。ウェッブの指数関数的進展やデータベース管理システムの広範な利用は、多様で多量の情報源泉を一体として相互に連結させるための必要性を前面に押し出している。異種で自立的なデータの源泉に一貫したアクセスを提供するためには、複雑なクエリーと統合メカニズムが計画され実行されなければならない。異種のデータベース統合に必要不可欠なことはマッピング・プロセスである。そこでは、二つのスキームのマッピングを、スキーム１の各インスタンスをスキーム２のインスタンスへの変換に対する十分な仕様として図１に示したような意味をもつものとして定義する。

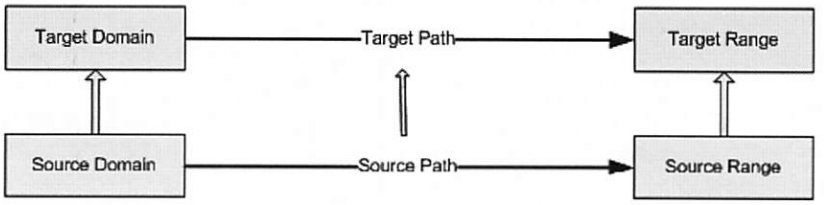

図１　基本的マッピング・スキーム

この定義は特定のインスタンスとは独立でなければならない。マッピングは変換の特定化に従うことによってのみ、スキーム１の全てのインスタンスをスキーム２のインスタンスへ変換する自動化アルゴリズムを実行することができなければならない。この作業において、目標とするスキームは、ノード、リンク、プロパティ、マルチISAやマルチ・インスタンス化を使用するオントロジーあるいは知識再現モデルであると仮定する。目標とするスキームはリンク上の濃度制約を強制しない、したがって濃度に関する異種性を扱う必要がないのである。さらに、ソース・スキームがエンティティ・リレーション性あるいはXMLモデルを使って記述可能であると仮定する。マッピングの定義は、ITの専門家の助けを借りて、手動で作成されるかドメインの専門家により半自動で制作されると想定される。適当なエンコーディングにおけるあるオントロジーの適当な部分は、目標とするスキームとして利用されあるいは解釈することができる。このリポートにおいては、文化遺産ドキュメンテーションにおいて使用される暗黙のあるいは明示的な概念と関係性を記述するための諸定義と形式的構造を提供するCRMを、目標とするスキームとして利用する。CRMは文化遺産からの情報のために創られたにもかかわらず、それは他のドメインにおいても同様にモデル化するのに適しているのである。

複合的な条件やケースが想定されるため、マッピング・ルールを作ることは簡単なプロセスではない。任意の諸スキーム間のマッピングを定義することの問題は、全てのケースで解決できるものでもない。さらに諸スキームのインスタンスはソース・スキームの意図した意味に従わないかもしれず、それによって例外ができる可能性もある。しかしながら所与のドメインや想定した目標スキームの特定の特質においては、異種性のケースは通常限定されたものである。

こうした問題を克服するためには、最も共通したケースをカバーするマッピング・メカニズムを定義し、要求に応じてメカニズムを拡張していかなければならないのである。こうした最も共通したケースは、博物館コレクションや、考古学データやモデルから集められた。こうしたドメインにおける最も共通したケースはまた同様に一般的に共通したケースを反映すると考えられるのである。

マッピングのメカニズムは、ドメインの専門家がそれを理解し利用し少なくともそれを検証できるほどまでに直感的なものでなければならない。そうするためには、当該のアプリケーションのコンテクストにおいて、ソースと目標とするスキームとの間の異種性の最も共通のケースを注意深く吟味しなければならない。

モデルはXMLやリレーショナル・データベースからRDF-OWLや類似のフォーマットで定義されたオントロジーにマッピングを図るために使われる。さらに、このモデルは単純なオブジェクト指向のデータベースにも使用しうるものでもある。リレーショナル・データベースについては、そのスキームをセマンティック・モデルとして解釈するものである。これを実行するため以下のことが必要となる。

１）　テーブルやコラムをエンティティとして解釈する

２）　完全なレコードをエンティティのインスタンスとして解釈する

3） フィールド名をリレーション関係と同時にエンティティとして解釈する
4） フィールド内容をエンティティのインスタンスとして解釈する

各フィールドは、ソース・パスと呼ばれるシーケンスでクラス＝ロール＝クラス（c-r-c）として解釈され、スキーム全体はc-r-cのスキームに分解され、各c-r-cは個々にマッピングされて目標となるスキームとなるのである。適切なマッピングを得るために、以下のことを定義しなければならない。

1） ソース・ドメインとターゲット・ドメインの間のマッピング
2） ソース領域とターゲット領域との間のマッピング
3） 妥当なソース・パス
4） 妥当なターゲット・パス
5） ソース・パスとターゲット・パスとの間のマッピング
6） ある場合には、同様のインスタンスを共有するパスを結合する必要がある

これまでのところCRM以外にこうした定義を提供するマッピング言語はなく、そうした定義の内のいくつかは手順コードによって明瞭に定義されると想定されている。

中間ノードの導入

ソース・パスにマッピングされるべき全パスを適正に定義するため中間ノードが導入されなければならない。これは、ソース・モデルにおいては、中間ノードに情報が格納されない場合、大きなパスを単一のリレーションにコンパクト化して省略するのが通常であるからである。

複合短縮

住所や種の名前、等位語句などで主に観察されるのが複合短縮であり、ソース・スキームにおけるいくつかのクラスがひとつの識別子の部分となる。例えば、住所（Address）には、番地、市、地域などが含まれているのである。

入れ子への並列

ソース・スキームにおいて、クラスAが他のクラスBiに関係し、そうした関係性が、AとBiの関係というよりむしろ、結果的にいくつかの関連するクラスBi間の関連性を意味する場合である。

中間入れ子への並列

前者のケースのみの拡張である。ソース・スキームにおける二つの関係性で同じ中間ノードをもったターゲット・スキームにおけるパスにマッピングされるのである。

マッピング・ルールにおけるクラスの同じインスタンスが複合マッピングに現れるということを定義するための一般的メカニズムが必要となる。また同様に、マッピングが依存する以下の単純な条件を定義するメカニズムが必要となる。

1） あるインスタンスのアトリビュートは用語や定数に等しい
2） あるインスタンスのアトリビュートはもうひとつの用語bによって包含された用語aである
3） アトリビュートのあるインスタンスは存在するかあるいはしない

こうした最も共通した条件のケースは、CRMを使ったMIDASのマッピングによって確証されている。MIDASはイギリスのイングリッシュ・ヘリティジのFISHによって開発された記念碑目録のマニュアルでありデータ標準である。以下に現在提案されているマッピング注釈フォーマットを図示する。

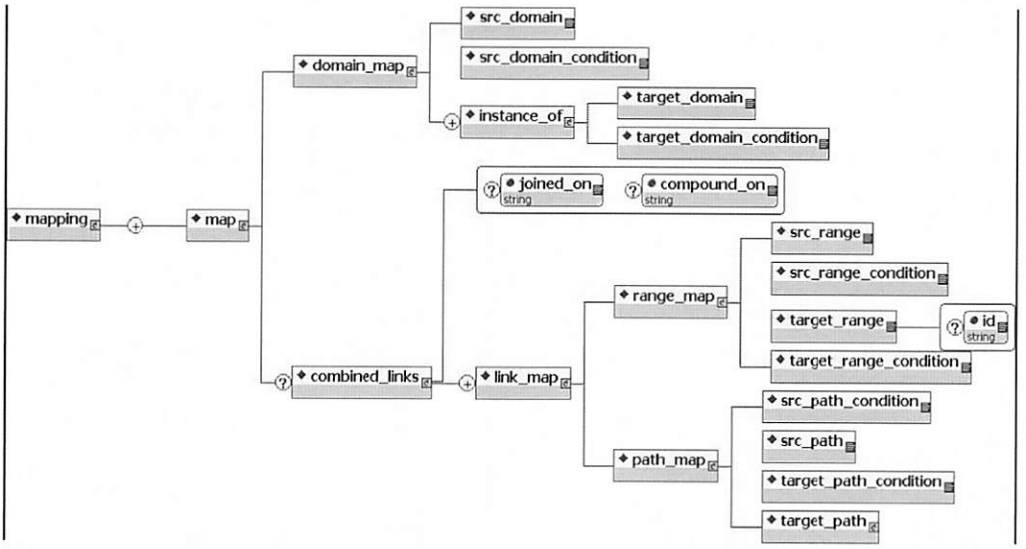

図2　マッピング・ファイルのDTD概略図

この小論は文化遺産情報における仲介システムあるいはデータ変換システムを構築し始めるための必要なマッピング・フォーマットと全体的プロセスを明らかにするために、現状の展望を記したものである。詳細なマッピングにはグラフィック・ツールが必要であるが、そのための枠組みを簡略にまとめたものである。

参考文献

1. N. Crofts, M. Doerr, T. Gill, S. Stead and M. Stiff, "Definition of the CIDOC Conceptural Reference Model," 2003.
2. L. Dempsey, "Scientific, Industrial, and Cultural Heritage: a shared approach. A research framework for digital libraries, museums, and archives," Ariadne, issue 22 (Dec.) 2004.
3. T. Gill, Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model," 2004.
4. P. L. Boeuf, Ch. Lahanier, G. Aitken, et al., "Integrating Museum & Bibliographic Information," in ICHIM, 2005.
5. H. Kondylakis, M. Doerr, and D. Plexousakis, Mapping Language for Information Integration, ICS, 2006.