第1章 メタデータと名前

多様なものごとや情報をうまく扱うためには、その主要な特徴を記した情報、すなわちメタデータを利用するのが効率的です。メタデータは「何の、どんな特徴が、どんな値を持つか」、すなわち主語―述語―目的語の関係で、ものごとのある面を表現します。この3つの名前を情報の送り手と受け手が相互に理解できなければ、メタデータの共有は成り立ちません。メタデータの先駆者である図書館の目録を手がかりに、文書のメタデータの要件を考え、そこで用いられる名前についての基本を整理します。

1.1　メタデータの誕生

多数のものを管理し、必要に応じて探し出せるようにするためには、ものをそのまま扱うのではなく、名前を付けて特徴を記述整理する、つまり目録あるいはカタログを作成して、どこにどんなものがあるのかを把握します。コンピュータデータの管理や検索に用いるデータは「メタデータ」と呼ばれますが、広い意味では目録もものに関する「メタデータ」と考えられます。その歴史は紀元前までさかのぼり、使い勝手をよくする工夫が今日まで続いています。

1.1.1　紀元前2世紀

ファロスの港を見下ろす王宮内に「真面目な関心に値する限りあらゆる人々の著作を収集しよう」と建てられたその図書館は、紀元前2世紀には蔵書が50万にも達しようとしていました。この膨大な書物を一点ずつ調べたり探したりしていてはとてつもない時間がかかりますから、有効に利用するためには、個々の書物の概要をまとめ、どんな内容のものがどこにあるかを記し、探し出せるよう分類整理した目録が不可欠です。

> 彼の採った基本的分類法は主題によるもので、修辞学、法学、叙事詩、悲劇、喜劇、抒情詩、歴史、医学、数学、自然学、雑学という分類であったことがわかっている。それぞれの主題のもとに著者たちがアルファベット順に配列された。それぞれの名前のあとには短い書誌的記述と、その作者の作品についての批評が続いた。（『古代アレクサンドリア図書館』*）

＊エル＝アバディ

カリマクスが作成したアレクサンドリア大図書館の目録『ピナケス』は、120巻の規模に及んだと言われます。50万点分の情報を圧縮し、主題、すなわち内容のテーマ別に整

理したこの目録は、まさに知識のカタログでした。「過去の遺産の、完璧な研究と理解に基づく最高水準の学問」と呼ばれるアレクサンドリアの知的世界も、書庫に収められた情報源への手引きとなるこの目録なしには花開かなかったかもしれません。

＊北嶋

＊バトルズ

図1.1

古代アレクサンドリア大図書館のホールで学者たちがパピルスを読む様子を描いた、19世紀ドイツの版画。

1.1.2　カタログの役割

私たちの周りには、図書館の目録のほかにも保管庫に収めた品々の台帳や製品カタログなど、さまざまなカタログ（目録）があります。これらのもともとの役割は、実際に現物を確かめに行く代わりに、目録を見て在庫を把握できるようにすること、つまり「何があるのか」「それはおよそどんなものか」を記録・閲覧する機能です。

一方、現代のカタログは存在の確認だけではなく、多くのものの中から必要な品や資料を見つけ出すための検索手段としての役割が重要です。製品カタログは、多彩なラインナップを見るのに加え、評判になっている特定の製品を探したり、自分の予算と目的にかなった品を選ぶためにも使えなくてはなりません。さらに必要なものが見つかったら、現物がどこにあるのか、「どうやって入手できるのか」についての情報も不可欠です。

これらを整理すると、「カタログ」には、次の役割が求められることが分かります。

- 識別：どんなもの（情報）があり、それが具体的に何を指すのかを、他のものと区別して示す。
- 記述：もの（情報）そのものを見なくてもおよそのことが分かり、求めるものかどうかを判断できるよう、概要を記述する。
- 発見：多数のもの（情報）の中から、特定のもの、あるいは条件に合致するものを見つける手段を提供する。
- 入手：求めるものがどこにあるか、どうやったら入手できるかの手段を示す。

＊ ISO 11179

＊ TimBL97

ものを識別して記述した「存在目録」は古くから存在し、古代バビロニアにまでさかのぼると言われます*。紀元前7世紀のアッシリアの宮殿跡から発掘された約2万5千枚の粘土板の文物は、束ねて内容を示すラベルが付けられ、表題やそれを構成する粘土板の数を記録した目録まであったようです*。主題によって書物を見つける手がかりを提供したアレクサンドリアの『ピナケス』はすでに検索目録の役割を果たしていたと考えられますが、目録に本格的に検索機能が導入されるのは、グーテンベルクの活版印刷の発明によって書籍が多量に作成されるようになってからのことでした。さらに産業革命、市民革命を経て書籍もモノも量が飛躍的に増大すると、発見のためのカタログの必要性がますます高まっていきます。

1.1.3　カタログとメタデータ

図書館の目録は、書籍という情報の塊に対して、タイトル、日付などその特徴を示す情報を加えるもの、すなわち「情報についての情報」です。一段階抽象化のレベルが上がることを表す「メタ」ということばを用いて、この「情報についての情報」を**メタ情報**と呼びます。

書籍のような形あるものばかりでなく、デジタルファイルなどのデータ管理にもファイル名、日付、サイズといったメタ情報が使われますが、こうした情報それ自身がまた、ファイルシステムなどに格納されるデータです。この場合のメタ情報は、「データについてのデータ」でもあるので、特に**メタデータ**と呼ばれます[注1]。

もっとも、メタ情報とメタデータということばは厳密に使い分けられるわけではなく、人の氏名や生年月日といったプロフィール情報をメタデータと称したりもします。本書でも、これから扱うさまざまなカタログ的記述情報全般を、より一般的な用語であるメタデータと呼んでいくことにします。

注1　メタデータ標準に関する国際規格の1つであるISO/IEC 11179*は、メタデータを「ある文脈において別のデータを定義し記述するデータ」としています。バーナーズ＝リーは、メタデータを「情報についての情報」とした上で、ウェブにおいては特に「機械可読な形で記された、ウェブリソースなどに関する情報」を指すとしています*。

1.2　文書のメタデータ

メタデータは情報を記述し、見つけやすくします。古くからカタログ作成に力を入れてきた図書館のシステムでそのメタデータの働きを振り返り、それとの対比でウェブ文書のメタデータを考えます。

1.2.1　図書館のカタログ

図書館で本を探すときは、コンピュータの端末を用いて、いくつかのフィールドにキーワードを入れて検索するでしょう。この検索システムOPAC（Online Public Access Catalog）は、その名の通り図書カタログのデータベースを公開し、オンラインで利用できるようにしたものです。

次に示すのは、国立国会図書館のOPAC書誌検索画面と、検索結果一覧、および個別の書誌情報です。図書目録の記述法はカードの時代から標準化が進められてきましたから、OPACの検索項目はどこの図書館でも基本的に共通しています。

図1.2

左上は国立国会図書館のOPAC検索項目、左下は検索結果一覧、右は一覧から選択した結果表示される書誌情報。

左上に示されている検索画面は、カタログの“発見”の側面にあたります。検索項目は書誌情報のサブセットになっていますが、情報はデータベースにあるわけですから、どの項目も検索対象とすることは可能です（実際、図1.5右に示す詳細検索画面を選ぶと、大部分の項目が検索条件として利用できるようになっています[注2]）。

左下の結果一覧は、検索条件にマッチする書籍を知るために、タイトル、著者、出版者、出版年の組み合わせで本を区別して表示しています。どんなもの（情報）があり、それらが具体的にどの本なのかを区別して示すという“識別”の機能です。

右の書誌情報画面は、直接本を手に取らなくても概要を確認し、必要なものを選択できるようにする“記述”の役割です。この本を借りたければ、請求記号をメモしてカウンターに依頼しますから“入手”のための情報も示します。さらに、結果画面をコピーしたり印刷して使う場合にどの本についての記述か分からなければ意味がありませんから、記述の中には本を“識別”するための情報も含まれています[注3]。

この書誌情報項目のうち、詳細検索（拡張検索）の対象となっているものを列挙してみましょう。検索項目と書誌情報で呼び名が異なるものは、括弧に入れて示します。

表1.1

項目	簡易検索	一般検索	書誌区分
タイトル	○	○	書誌記述
著者・編集者（責任表示、個人著者標目）	○	○	
出版者		○	
出版年		○	
出版地			
形態（物理的属性）			
注記			
識別子（ISBN/ISSN、書誌番号、請求記号、書誌ID、ほかにUS MARC番号など）		○	
本文の言語（本文の言語コード）			
件名（普通件名）		○	主題
分類（NDC、NDLC）		○	

▼13.2.4項参照

責任表示は、その著作物の内容に責任がある人や団体を示すもので、基本的には本の奥付に記された著者名です。個人著者標目[注4]は人物名と生没年や職業をセットにして、同姓同名の著者を区別し、また同一著者の異表記を集約できるようにした識別名です。識別子のISBNはおなじみですが、ISSNは同様にして雑誌などの逐次刊行物に与えられる番号（International Standard Serial Number）、また全国書誌番号は国立国会図書館が付与する出版物の識別番号です。

これらのうち、タイトル～出版年は検索結果一覧表での書籍区別にも用いられたように、その書籍を“識別”するための基礎項目であり、中でもタイトルと著者は簡易検索でも使

注2　検索の対象となる項目を「アクセスポイント」と呼びます。図書カードのように検索手段がタイトル、著者などの順に並べられたカードボックスしかなかったときは、こうした項目のみがアクセスポイントであったわけですが、OPACにおいては基本的にどのフィールドも検索対象になり得ます。ただし項目が多ければ使いやすいというわけでもなく、頻繁に検索に用いるフィールドはインデックスを作成して高速化を図りますから、何をアクセスポイントとして提供するかの検討は、データベースであっても必要です。

注3　ウェブ検索の場合、検索結果の一覧には識別情報と同時にスニペットやファイルサイズなどの“記述”も加えて、その場で概要の確認と選択ができるようにしています。さらにリンク先は資料自身なので、そこに“入手”の情報もあることになります。

注4　標目（Heading）ということばは、図書カード時代にカードの頭に見出し語として書かれていたことの名残です。OPAC以前は、タイトル、著者名、件名、分類を見出しとしたカードがそれぞれのボックスに並べられ、カードの頭を素早く見ながら必要な資料を探していました。この場合は、標目がすなわち「アクセスポイント」であったわけです。

われる最も代表的なメタデータです。同時に、タイトルから言語までの項目は、その書籍固有の客観的情報を“記述”するメタデータです。これらが基本的な**書誌記述**（Bibliographic Description）を構成します[注5]。

最後の件名と分類は、基本的な書誌記述とは毛色が異なり、書籍の内容を、あらかじめ用意した値セット（キーワードなど）から選んだデータで組織的に表現して“発見”の手がかりにしようというものです。これらのメタデータは**主題**（subject）と呼ばれ、客観的な記述とは異なる方法論による体系化が行なわれています。次項でこの主題によるメタデータを検討します。

1.2.2　主題あるいはキーワード

書誌情報の中で、件名、分類という名が与えられている項目は、目録作成者が資料（書籍）の内容を読み取って、その主要なテーマ（主題）をメタデータとして記述するものです。

書誌記述が特定の書籍を見出すための情報であったのに対し、主題は具体的な書籍が念頭にない場合に、何らかのテーマに関する本を探すといった目的のために提供されます。たとえば、プログラミング言語Pythonに関する本を見つけたいときに、図書館や書店でコンピュータ書籍が並べられた棚に向かうのは、「コンピュータという主題による分類」を手がかりとして候補（棚）を絞り込んでから、個別の書籍のタイトルや著者という「書誌記述」で内容を推定し、これと思ったものを手にとって吟味するわけです。

図書館の書誌情報の場合、主題を表すメタデータの値はあらかじめ定められた値のセットから取られるので、ことばの揺れなどによる検索漏れがなく、同じ主題を持つ資料を網羅的に検索できます。また通常この値は上位、下位などの関係で別の値と結びつけられているので、その関係をたどって関連する資料を取り出していくことも可能になっています。

件名標目

件名[注6]とは、主題を表す統制された（値のセットが定められた）キーワードです。資料（書物）の内容を端的に表すキーワードをメタデータとして付与していくわけですが、そのキーワードを思いつきで付けるのではなく、件名標目表として体系化された語の中から適切なものを選択するところがポイントです。

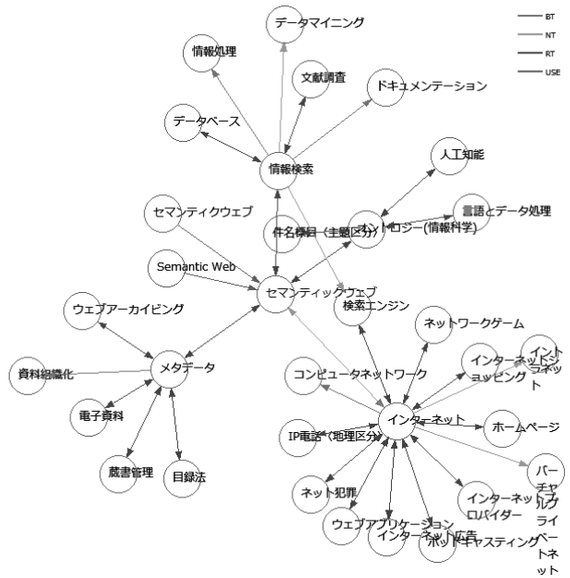
件名標目は通常、シソーラスとして広義語、狭義語、関連語などを定義してキーワードのネットワークを構成しているので、与えられた件名から範囲を広げた検索、あるいは絞り込んだ検索を行なうことが可能です。

注5　目録記述の国際的な標準規則である国際標準書誌記述（ISBD）は、書誌記述のエリアとして、1. タイトル、責任表示、2. 版、3. 資料あるいは刊行方式の特性（書籍ではあまり使用しない。電子資料の場合、画像、テキストなどタイプや、目的）、4. 出版、印刷、頒布（出版者、出版年、出版地など）、5. 資料の形態（サイズ、分冊など）、6. シリーズ関連、7. 注記、8. 識別子の8つを規定しています。NII（国立情報学研究所）の『目録情報の基準』では、識別子に相当する部分については「ID & コードブロック」という枠を設けて、「記述ブロック」とは区別しています。

注6　件名はsubjectの訳です。主題もsubjectの訳なのでややこしいですが、図書館ではキーワードとしてのsubjectを伝統的に件名と呼んでいます（電子メールのSubjectも「件名」です。このフィールドがTitleではないのが面白いところです）。

たとえば、図1.2の書誌情報を見ると、この本には「セマンティックウェブ」と「プログラミング（コンピュータ）」という2つの件名が与えられており、それぞれの件名の上位語や下位語、関連語も参考として示されています。これらの件名は検索リンクとなっており、同じ件名を持つ書籍を調べることができるわけです。

図1.3



件名表目標は、広義語、狭義語、関連語という関係で件名がリンクしている。図は国立国会図書館件名標目表の「セマンティックウェブ」が持つ関連を、筑波大学のHANAVIシステム（http://raus.slis.tsukuba.ac.jp/）で表示したもの。

図書館で用いられる件名標目には、この例の国立国会図書館件名標目表（NDLSH）のほかに、基本件名標目表（BSH）、米議会図書館件名表目標（LCSH）などがあります。1つの資料に対して複数の件名（つまりキーワード）を与えることができますが、どの件名表目標のキーワードを用いているかを明示しておくことが重要です。

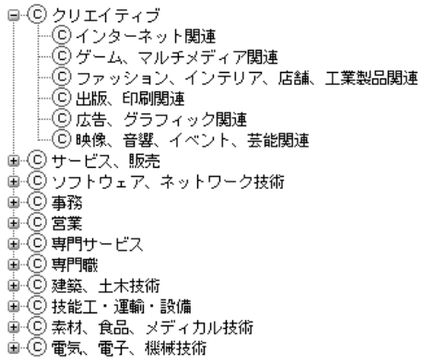
分類標目

分類とは文字通りものごとを何らかの「類」に「仕分け」していくことです。ここでいう「類」は一般にクラス、あるいはカテゴリと呼ばれ、対象領域を一定の基準に基づいて区分したものを指します（この基準を区分原理と呼びます）。「分かる」とは「分ける」ことだとよく言われますが、たくさんの資料をカテゴリによって分類することで、それぞれの資料を把握しやすくなり、また発見も容易になります。

区分原理は、分類の目的に応じてさまざまなものが用いられます。たとえば、人という対象を「職業もしくは社会的立場」という区分原理で分類するなら、学生、会社員、主婦、自営業…といったおなじみのアンケート項目ができあがりますし、同じ対象を血液型という区分原理で分類することもあり得ます[注7]。

また分類の区分は、階層的にすることもできます。人の職業を、まずおおまかに事務、営業、クリエイティブなどと区分しておき、クリエイティブをさらに出版、ゲーム、ファッションなどと順次細分化していくことで、体系的な分類を行なうことができます。

図1.4



あるサービスでは、職業を2レベルの階層で分類している。分類体系は木構造で、1つの下位区分が2つの上位区分に属することはない。

通常、1つの分類体系の中では、対象はどれか1つの区分に分類され、複数の分類に属することはありません。もちろん階層構造を持つ分類体系では、下位クラスに分類されたものはその上位クラスにも属します。図1.4の職業分類で「出版、印刷関連」を選択したら、「クリエイティブ」でもあることになります。しかし、血液型がA型であると同時にO型でもあるということはないように、同じレベルにおいて2つの区分に同時に属することは、基本的にはありません[注8]。書誌での分類は、配架（どの棚のどこに本を並べるか）と連動するので、1つに定めることが求められます。

書誌における分類体系は、図1.2に示されているNDC（日本十進分類表）、NDLC（国立国会図書館分類表）のほか、LCC（米議会図書館分類表）、DDC（デューイ十進分類）などがあります。また、複数の区分原理を組み合わせて1つの分類を作るファセット分類のような方法が導入されることもあります。いずれもキーワードではなく、数字やアルファベットの組み合わせが用いられるのが特徴です。

注7　区分原理は、包括的でかつ排他的なものが必要とされます。包括的とは対象領域全体をきちんと区分できる原理で、一部だけを区分して残りは「その他」では、あまり有益な分類はできません。排他的とは複数の区分に当てはまる要素ができない原理です。“使用しているパソコンの機種”という区分原理では、2種類のコンピュータを使っている人を適切に分類できないわけです。

1.2.3　ウェブ文書とメタデータ

図書館の目録は、資料を探したり記述したりするために必要な要素は何かを研究してきた長い蓄積が反映されており、文書のメタデータを考える上での1つの指標といえるものです。ウェブ文書の場合、このメタデータはどの程度提供されているでしょうか。表1.1で列挙した項目を、ウェブ文書での表現方法と対比させてみましょう。

表1.2

メタデータ項目	ウェブ文書での表現法	標準化
タイトル	title要素	○
著者・編集者	address要素あるいはmeta要素（author）	△
出版者	ホスト名？	？
出版年	HTTPヘッダ（Last-Modifiedフィールド）？	？
出版地	ホスト名（IPアドレス）？	？
形態	HTTPヘッダ（Content-Typeフィールド）	○
注記	meta要素（description）？	？
識別子	URI	○
本文の言語	langあるいはxml:lang属性	○
件名	meta要素（keywords）	△
分類	meta要素（keywords）	△

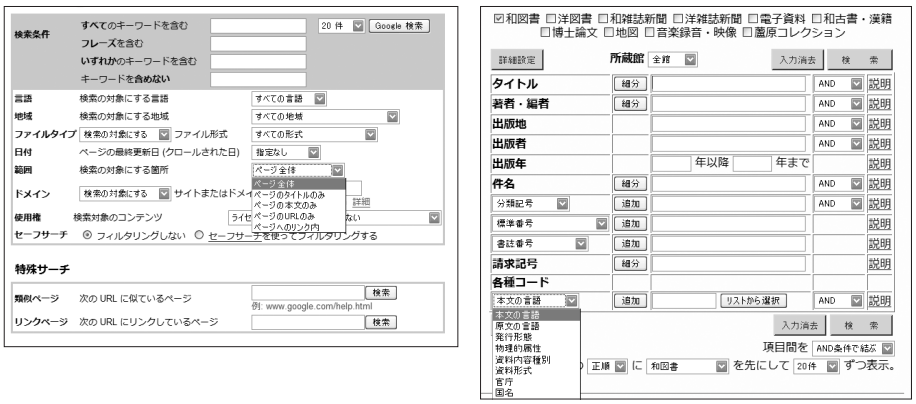
書籍特有でウェブにはあまり意味がない項目もあるとはいえ、書籍メタデータに対応する表現手段は、一通り揃っているようにも見えます。しかしこれらのうち常に得られるのはtitle要素とURI、Content-Type程度で、他はみなオプションだったり、標準的な使い方が定まっていなかったりで確実とはいえません（出版者と出版地は、少し意味合いは異なりますが、URIのホスト部と考えれば確実に得られる情報の部類とはいえます）。

作者情報は、書籍の場合は最も基本的なメタデータであり、重要度はウェブ文書でも同じはずですが、address要素に書く場合もあれば、meta要素にname="author"として記述するケースもありますし、書式も定まっていません。Last-Modifiedフィールドはサーバ側で設定しなければ得られませんし、ファイルを少しでも修正したらその日時となるので、作成日（あるいは発行日）である出版年情報は、標準的には提供されていないのです。

注8　一般に、クラス階層のツリーにおいて、メンバーが異なる枝に属するような設計は行ないません。複数の枝に属するメンバーを、多重継承という形で扱えるシステムもありますが、ここでは分類は排他性の原則に従うものとします。

ウェブ文書と書籍のメタデータの違いは、ウェブ検索エンジンの“検索オプション”で指定できる項目を見ると、よりはっきりします。

図1.5



ウェブ検索エンジンの設定項目は、書籍の場合とかなり異なる。左はGoogleの「検索オプション」画面、右は国立国会図書館OPACの「書誌 拡張検索」画面。

ウェブ検索の基本が全文を対象にした自由キーワードであることは別にしても、オプションのメタデータ項目もかなり異なります。ウェブ検索で最初に示されている選択肢の言語、地域、ファイルタイプは、OPACでは最後にまとめてプルダウンメニュー（本文の言語、国名、発行形態、物理属性など）で扱われているに過ぎません。逆にOPACでは重要な著者名や主題（件名、分類）[注9]は、ウェブ検索では項目自体がないという状況です[注10]。

タイトル検索は、独立した項目ではなく、キーワード検索の「範囲」を「ページのタイトルのみ」に絞り込むようになっています。日付項目はありますが、出版年ではなく更新日が用いられています。使用権（ライセンス）での検索ができたり、フィルタリングのオプションがあるのはウェブならではです。

この差異は、ウェブ文書と書籍の性質の違いによるところもありますが、図書目録が人手によって作られるのに対し、ウェブ検索のインデックスがコンピュータのプログラムによって生成されるという点も大きな要因でしょう。作者や作成日を収録して検索項目にしようとしても、記述方法が決まっていなくてはプログラムには見つけようがありませんし、判別できたとしても書式がまちまちで正規化するのが困難です。ウェブ検索オプションの選択項目の大半は、HTTPヘッダやURIのように、記述場所や書式が明確なものを利用していることが分かります。

これは逆に言えば、メタデータをプログラムで処理可能な一貫した形で記述する方法があれば、ウェブ検索でも、より精度の高いメタデータを利用する可能性があるということです。実際、「使用権」項目のライセンス検索は、7.4.2項で取り上げるライセンス記述の普及によってウェブ検索に加わってきました[注11]。

ウェブ文書にメタデータを埋め込む方法は、2008年ごろになってかなり規格が出揃い、実際の利用を考えるタイミングに差し掛かっています。次章以降で、こうした手法の扱い方を、順を追って検討していきます。