

メタデータを巡って

いりえ しん
入江 伸

(メディアセンター本部課長)

1. はじめに

最近、メタデータという言葉をよく耳にする。ところが、このメタデータという言葉は、とても厄介で話す人や脈絡によっていろいろな意味で使われ、しばしば混乱さえ起こしている。メタデータという用語は、W3C (World Wide Web Consortium) のようなインターネットに関する研究者や、博物館・文書館や図書館の人達、e-Learningシステムなどの業務系システム、GIS (Geographic Information System。以下、GIS) システムや動画システム等の基礎的な研究分野でもキーテクノロジーを示す言葉として議論されている。この背景とその多様な使われ方について、個人的な見解を含めて説明し、図書館の今後の役割と展望を明確にしたい。

2. メタデータの定義

インターネットでは、国や言語・分野を越え、あらゆるコミュニティから発信される情報が同一のネットワークで接続され、テキスト・画像・動画など多くのメディアが使用されている上に、その情報が刻一刻と変化している。この成長を続けるインターネット世界において効果的に情報を取得し、データを運用していくためのキーテクノロジーを示す言葉としてメタデータという用語が使われている。

辞書ではメタデータの広義な定義として「データについてのデータ」と説明してある。またJISX0017「情報処理用語(データベース)」では「データ記述を含むデータ要素に関するデータ、並びにデータの所有者、アクセス経路、アクセス権及びデータの変更度に関するデータ」となっている。しかし、この説明では図書館や情報管理のコミュニティで古くから使っている二次情報という用語と区別ができない。「メタデータ」はこれまでの二次情報という概念とは異なり、インターネット環境における新しい二次情報の考え方であるといえる。

3. コミュニティとメタデータ

W3Cは、HTMLという表示形式を拡張し、XMLとSemantic Webというキーワードで示されるように、インターネットの世界をグローバルな知識ベースとしていくために取り組んでいる。これまでのHTMLでは、表現形式しか表わせないためそれぞれのページの持つ内容は表示された画面を人間が見なければ理解することができなかった。それでは、インターネット上の膨大な情報を選別するのに、とても大きな労力が必要となってしまう。そのため、一定の選別をプログラムで行う仕組みをページ内に組み込むことによって、あるレベルまでの選別をプログラムが自動的に行う試みが進められている。XMLとはページ内に内容を表現するタグを記述するための形式である。更に、プログラムが内容を理解できるページを知識データベース(以下、DB)としてとらえ、インターネット上のページをコンピュータによって評価する知識構造の構築を目指している。「メタデータ」という言葉は、この夢のような取り組みの中で、知識表現のためのデータ定義という意味でも使われている。

W3Cの動きに影響されて、業務アプリケーションの世界でもメタデータが議論されている。内容を示すタグを国際標準とし、特定の分野のデータを国際的なレベルで相互交換や相互利用を可能にしようとするものである。業務アプリケーションで使われる「メタデータ」とは、XMLを使った国際的な相互交換可能なデータフォーマットという意味である。

図書館界の「メタデータ」への取り組みは、W3C等の研究に大きく影響されながら進んでいる。図書館は将来の地位を確保するためにも、インターネット上の情報資源の目録を作成し利用者へ提供するという使命がある。そのためには従来目録規則とMARCフォーマットを見直す必要がある。一方で、これまで作ってきた巨大な目録DBとの整合性を確保することも大きな課題である。そのような中

で議論されたフレームワークがDublin Coreという考え方であり、実証プロジェクトの一つがOCLCのCORC (Cooperative Online Resource Catalog)である。

4. メタデータ間の対応

このように、多くの分野からいろいろな目的を持ったメタデータが提案されている。しかし、多種類のメタデータが提案され運用されると、インターネット上の情報を統一して検索することが難しくなる。例えば、イメージデータの記述に特化したメタデータには、データ形式等のイメージデータの特性を記録する項目に加え、作成日・時間・撮影者・内容等が必要である。このイメージデータ用の記述要素と書誌データの記述要素は当然に異なってしまう。記述要素を統一するために、Dublin Coreのような汎用的な記述要素を使ってデータを作成すると、固有の記述要素をコンピュータでは識別できなくなってしまう。このようにデータを単純化することは、メタデータの目的とは相容れない。しかし、インターネット上の全ての情報資源から、媒体を問わずあらゆるデータを同一方式で検索したいとすれば、それに適した汎用的なデータ定義が必要になってくる。この相反する事態を解決するため、メタデータではそれぞれの記述要素の関係を定義することで、自動的に対応させて処理する機構が用意されている。このメタデータ記述要素間の相互関係は、資料や媒体毎に詳細に定義されているメタデータと、汎用的で比較的単純なメタデータとの階層関係として理解することもできる。

5. 資料・媒体単位の詳細なメタデータ定義

資料・媒体毎にその個別の特性を表すための詳細なメタデータ定義がある。このような固有の特性を表わすためのメタデータの標準化を進めているグループにGISシステムや、音声・映像情報システム等がある。どちらのグループも、特定のソフトウェアにデータを通して初めて内容を確認できるため、異なったシステム間の相互利用やデータ管理のためにもメタデータの標準化は重要な課題となっている。

業務モデルを丸ごと標準化することによって、異なったシステム間でのデータ共有や交換を可能にす

る試みも行われている。e-Learningシステムでは、番組提供データや受講者データまでメタデータとして標準化しようとしている。これは、e-LearningシステムのDBを標準化して、どのメーカーのシステムを採用しようとも、過去のデータをそのまま利用できることを目的にしている。この考え方は、図書館システムにも十分に適用できるものである。

すべての情報を平坦に表すメタデータ定義として代表的なスキーマがDublin Coreであり、これは図書館で行なわれているメタデータについての議論の中心である。Dublin Coreは、インターネット上の情報資源を資料形態やデータ形式に関係なく、国際標準フォーマットで効率的に表すための記述規則である。その他の多くの種類のメタデータ定義から、この記述へのリンクを維持することで、いろいろなデータを同一フォーマットで扱うことを可能にしている。この記述は、資料形態を問わず使用可能なため単純で作成が容易に思えるが、要素定義が曖昧なため厳密な記述には適さない。

このように、メタデータと一言で言っても、フォーマットはいくつもあり、利用目的や運用方法により、どのフォーマットを採用するかということや他フォーマットとの対応の検討が重要となっている。

6. Dublin Core について

紙の出版物を整理するためには、これまでの目録規則やMARCフォーマットは極めて優れたものである。詳細な記述をDublin Coreで記述することによって、これまで識別できた多くの項目が識別できなくなるようでは、コンピュータで意味を解釈しようという目的から外れてしまう。また、詳細な項目定義を求めてDublin Coreの項目が細分化されていたのでは、Dublin Coreの単純性という大きな意味が失われてしまう。といっても、MARCフォーマットは、インターネット環境で利用するには、あまりにも図書館的(目録カード的)であり、一般性がない。最近、この矛盾を整理した新しい記述方法としてMODS (Metadata Object Description Schema) が提案されている。

(<http://www.loc.gov/standards/mods/>)

インターネットへ向けての情報発信には、目的にあった最適なデータ設計やDB設計が必要である。

そのためには、いろいろなメタデータや標準化についての知識とデータを作成してきた豊富な経験が必要となる。カタログラーは、これまでの経験を踏まえ、インターネットへの情報発信を行うためのメタデータ設計やデータ作成をリードできるスキルを磨いていくことが重要となっていく。

7. RLG CMIのメタデータ

メディアセンター(以下、MC)では、写真DBをRLGのCultural Materials Initiative(以下、CMI)へ登録する事業に取り組んでいる。慶應義塾がRLGの正式なメンバーとして進めている国際協力事業である。CMIは各大学の持つ貴重な資料をメタデータと画像データで登録して提供しようというものである。RLGは、Dublin Coreという既存のフォーマットよりも、CMIの持つ検索機能との適合性に注意を払っている。登録においては、タイトルの英訳やLCSH件名付与といった国際標準への対応が必要になっている。漢字・ローマ字・英語の3形式を用意するが、漢字での検索や表示の処理方法については不明な点が多い。これからは、国際的な共同事業の中で、漢字の処理方法についても合意を取っていくことが重要な課題となっていく。

8. メタデータデータベース小委員会

昨年、国立情報学研究所(以下、NII)のメタデータデータベース小委員会に委員として参加している。この小委員会は、NIIが進めるメタデータDB共同事業の意見調整と推進のために全国の大学図書館から委員を招集し設置されたものである。メタデータDB共同事業とは、参加大学の共同事業によってWeb上の学術情報のメタデータを作成し、国内における学術情報のポータルサイトを作成しようというものである。この小委員会は、学術情報資源を、大学が生産する情報の発信(SetA)と、大学が利用する情報(SetB)に分け、これを大学共同事業として取り組むための推進機関となっている。このDBは、基本フォーマットをDublin Core準拠、交換形式をXMLと規定し、世界への情報発信と相互交換を目標にしている。今後この事業は、以下のような課題を解決しなければならない。

- ・大学から世界へ発信している情報は少ない。
- ・商用ポータルサイトと区別して利用される大学ポータルサイトに育てていくためには、明確な目的意識と組織的な対応が必要となる。

9. 慶應義塾ポータルサイトの構築に向けて

慶應義塾ポータルサイトを立ち上げ、慶應で生産される教育・研究情報を積極的に発信していくことが重要になっている。そのためには、その事業の全体的承認と、定常的に情報収集・資料のデジタル化・メタデータの作成・ポータルサイトの運用を担える組織を立ち上げる必要がある。この組織は、独立した組織というよりも、組織的な妥当性と経験から得意な部分を補完し合う、複数の組織の共同事業体となっていく可能性がある。MCとしては、デジタル化やメタデータ作成等の業務において、これまでの経験を活かして取り組む必要があると考える。ポータルサイトの構築には多くの課題があるが、有用なものにするための留意点を以下に示す。

- ・網羅性 情報が網羅的に収集されていること。
- ・信頼性 掲載される情報が正確であること。特に人・組織に関する情報は正確であること。
- ・利便性 体系的で利用しやすいこと。検索機能が充実していること。
- ・学術情報DBを維持し、メタデータスキーマが標準的であること。使いやすいキーワードや統制された件名で検索可能なこと。
- ・メタデータとして相互交換が可能なこと。
- ・デジタルアーカイブを維持し、一次情報へのアクセスが可能なこと。

10. まとめ

メタデータに関わる業務は、今後ますます意味を持つようになっていくであろう。その中で大学が教育・研究情報をどのように扱い、情報を発信していくかということは大きな課題である。適切な情報を公開するために目的を明確にし、その目的に適合するメタデータを採択してデータを作成することが重要となる。MCとして、この課題に取り組み、寄与していきたいと思う。