社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DD－57 (5)
2006／9／29

デジタルドキュメント流通のためのメタデータ設計に関する比較報告

嶋津恵子† 齋藤功† 有澤達也† 吉永早織‡ 古川康一‡
†慶應義塾大学デジタルメディア・コンテンツ統合研究機構
‡慶應義塾大学大学院政策・メディア研究科
shimazu@dmc.keio.ac.jp

本書は，メタデータの情報構造を調査した報告である．我々は，デジタルコンテンツの流通を目指した代表的なメタデータ，Dublin Core と LOM，およびドイツ国立美術館が考案したメタデータの 3 つを対象として情報構造を調査した．その結果，デジタルコンテンツの流通に必要不可欠な属性情報の把握と，専門性の高いコンテンツと一般利用を目指したコンテンツのメタデータの違いを確認した．さらにこの結果を参考に，我々が開発するデジタルコンテンツ流通システムにおける標準メタデータの構造を考案した．

A survey report on information structure of digital contents metadata

Keiko Shimazu†, Isao Saitoh†, Tatsuya Arisawa†, Saori Yoshinaga‡, and Koichi Furukawa‡
† Research Institute for Digital Media and Content, Keio University
‡Graduate School of Media and Governance, Keio University

This paper is a report of survey about information structure of metadata on digital contents. We studied not only well-known metadata, Dublin Core and LOM, but also Dresden's metadata. The third one is remarkable output of a project of The Dresden State Art Collections. Their aim was creating digital image contents from actual historical art objects, and developing metadata of the digital contents to be shared and utilized by nonspecialist. We recognized essential information as metadata, and realized differences between contents used by specified users and contents used by the general public, through this survey.

1. はじめに

人は経験から学んだ知識を，自身の記憶の再起を助けるために，もしくは後世に伝えるために記録した．15 世紀の印刷革命後，ドキュメントを介した知識の共有が爆発的に広がり，1900 年代前半の乾式電子写真(ゼログラフィ)の発明により，より手軽にかつ身近にこれをおこなえるようになった．さらに，1900 年代後半のインタネット技術の台頭により，ドキュメントの共有は地理的条件に依存しなくなっただけでなく，(特にハイパーテキスト技術により)知識の関連を記述することとそれを後からたどることが可能かつ容易になった．

その後，より高度にかつ効果的に Web 上の情報を活用することを目指し SemanticWeb が提唱された．提唱者の Sir. Tim Benrers-Lee によると SemanticWeb の目的は，Web 上にあるデジタルドキュメントの意味を取り扱い，情報検索等の問題解決に貢献させようとするものである．SemanticWeb は，メタデータの空間を取り扱う．人は HTML 文書を読み，ソフトウエアエージェントは HTML 文書を説明したメタデータを読み，人と同様の処理を機械的におこなう[1]．原理的には単純であるが，曖昧

性なく記述し誰もが正しく操作するために統一した定義が必要であり，一般には共有するサイトを特定し name space を使ってオントロジーを交換する方法が採用される．
現在筆者らが所属する慶應義塾大学デジタルメディア・コンテンツ統合研究機構（以降 DMC と表記）では，専門分野を横断して研究成果であるデジタルコンテンツとドキュメントを流通させるシステムを構築中である．これを実現するための施策のひとつとして，効率的・効果的に流通させるための標準メタデータフォーマットを検討中である．本書は，デジタルドキュメントの共有を目的とした代表的なメタデータを対象に，それらの構造を比較調査した報告である．
本書は次の構成を取る．2 章にメタデータの構造の調査対象としたデジタルドキュメントの特徴を述べる．3 章に，メタデータの構造を比較するために用いた ER モデモデリングを概説する．4 章に比較結果とそれに対する考察を述べ，5 章に我々が提案する専門分野を横断したデジタルドキュメント流通用のメタデータの構造を示し，6 章にまとめを述べる．

2. メタデータを活用したデジタルドキュメントとコンテンツの代表事例

2.1 デジタルライブラリー

昨今，図書館を取り巻く環境が変貌している．インタネットの台頭により物理的な制約を越えて，情報資源にアクセスしたいという要求は世界的に高まっている[2].これに応えるために情報資源の作者，タイトル，作成日といった書誌情報の特徴を統一的に記述するメタデータの要素集合（ボキャブラリ）が定められ Dublin Core と命名された[3]．このメタデータ・ボキャブラリの定義を策定する中心的な役割を担っていたのが米国の図書館コンソーシアムであったことも追い風となり，デジタル書籍のほとんどがこれに従ったメタデータを持つようになった．今では書籍以外の多くの Web 上の情報資源のメタデータの記述に Dublin Core が用いられている．当初 13 個のメタデータ項目で定義されていたが，その後 1996 年には 15 に拡張された．さらにより詳細な意味をもたせるために意味を深化させる修飾子が定義された．Dublin Core は，図書館司書や記載内容に精通している専門家でなく，一般的なデジタル文書の作成者が記述可能であり，また利用可能であることを目標に設計されている．

2.2 e ラーニング教材

e ラーニングの” e”は electronic の頭文字であり，コンピュータやネットワークなどを利用して学習することをいう．広い意味ではパソコン を使った学習全般のことを指すが，一般的にはネットワークに接続し，ブラウザ上で学習することをいう場合が多い．従って多くの e ラーニング教材はネットワークを流通するデジタルコンテンツとして作成される．LOM（Learning Object Metadata）は，IEEE で標準化作業が進められている e ラーニング教材用のメタデータである．教育を目的にするコンテンツを対象にしているため Dublin Core で定義する書誌情報だけでなく，対象年齢や学習指導要領項目等学習教材としての属性情報が設定される．現在マサチューセッツ工科大学を中心した OCW(Open Course Ware)の活動が世界中に広まりつつある．これは正規の科目として実際に行われる講義のオリジナル資料をホームページで公開し，その情報を利用して誰もが「知」を高めることのできる機会を提供する仕組みである[1]．この OCW 上での教材コンテンツに LOM が利用され，広く知られるメタデータの一つになった．

---

[1] 特色ある教育支援プログラム「大学連携による新しい教養教育の創造」公開シンポジウム「拡大するオープンコースウエア」実施報告

2.3 ドレスデンのデジタル美術画像

ドレスデン美術館（Staatliche Kunstsammlungen Dresden）は，現在世界で最も価値のある美術品を所蔵する美術館のひとつである．ラファエッロ，ジョルジョーネをはじめとするヨーロッパの古典絵画を展示する古典絵画館（Gemäldegalerie Alte Meister）と，19 世紀から 20 世紀の絵画を集めた近代絵画館（Galerie Neue Meister）のほか，陶磁器コレクション，古代彫刻コレクション，「緑の丸天井」(ザクセン王家の宝物館)など，ドレスデン市内にある計 12 の博物館から構成されている．この美術館が 1990 年代の後半から，デジタル画像による所蔵美術品の一般公開をテーマとするプロジェクトに取り組んだ[4]．世界の多くの美術館が貴重品の所蔵に興味を持っているのに対し，(デジタル画像化されたコンテンツとは言え)所蔵品の一般公開に注力したことがこのプロジェクトの特徴である．ドイツ国内の美術館のデータベースの調査から，専門性や詳細な情報の整備の充実度と，利用者によるそれらに対する満足度が比例しないことが発見され，さまざまな異なる目的を持つ利用者や，専門家でない利用者から満足されるためのメタデータ構造を検討し，実装した．

3. データモデル

情報システム構築の工程で，モデリングは，情報システム化する対象の本質を統一的な方法で記述することを指し，データモデルは対象の中でデータに関するものである．したがって一般には情報システム構築の上流工程の作業としてデータモデリングを行う．データに関するモデリングの手法の中で代表的なものの一つが，ER(Entity Relationship)モデルを使うものであり，実体と実体間の関係情報を用いて，対象データの意味的な構造を統一的に表現する．このモデリング手法は，結果を一意に関係データベースのテーブルへ変換可能なことから専用のツールも多く開発され，実システム構築上多く用いられている．

一方，Web コンテンツのメタデータを設計する際に用いられるのが，RDF（Resource Description Framework)によるモデルである．一般的にメタデータは，HTML を拡張した XML で記述される．この XML の記述内容を RDF に変換し，コンピュータが情報の分類や検索を効率的に処理する．RDF は，RDF そのものの構造や構文を定める Model and Syntax と，メタデータのボキャブラリを規定する Schema の 2 つの仕様から構成される．RDF (Resource Description Framework)によるモデルは，前者を指す．従って RDF モデルは有効グラフとして表現され，メタデータを対象にリンクのたどり方を構造的に表現する．

今回我々は，デジタルコンテンツのメタデータの情報構造の特徴を調査することが目的である．そこで 2 章で述べたデジタルコンテンツのメタデータを対象に，ER モデルに展開し比較をおこなった．

4 メタデータの構造調査

4. 1 ER モデル図によるメタデータの構造

デジタルライブラリで広く用いられているメタデータ(Dublin Core)，e ラーニング教材で用いられているメタデータ(LOM)，さらにドレスデンのデジタル美術画像用に開発されたメタデータ(本書では“ドレスデンメタデータ”と表記)を，それぞれ ER モデル図の構造に展開したものが，図 1，図 2 そして図 3 である．

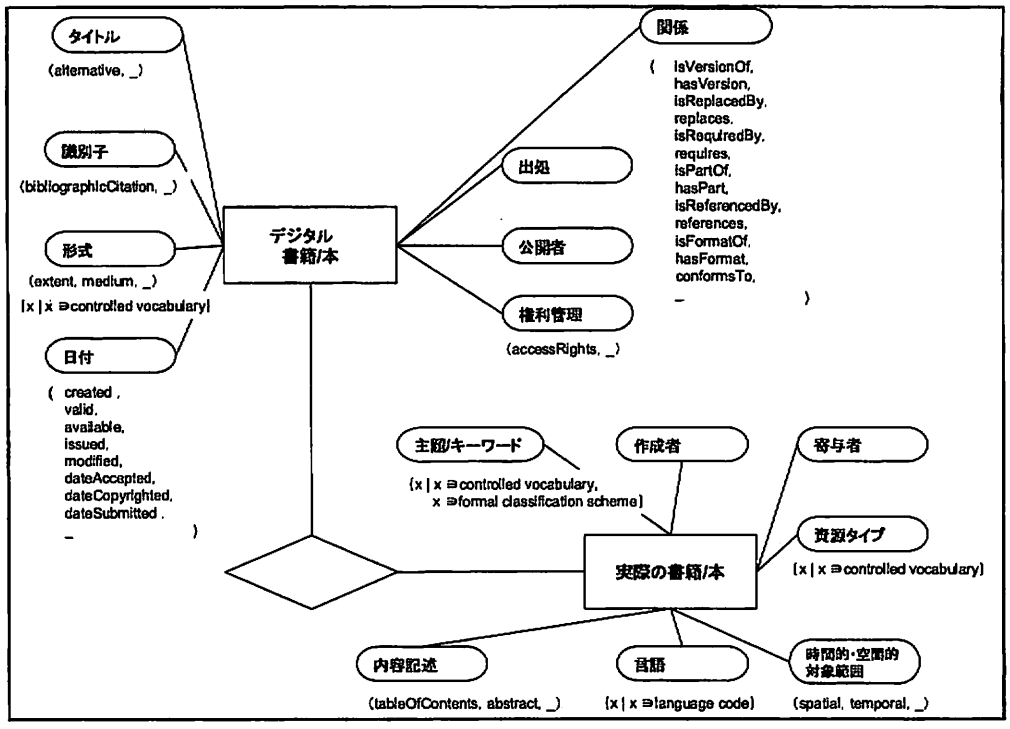

図 1　Dublin Core の構造

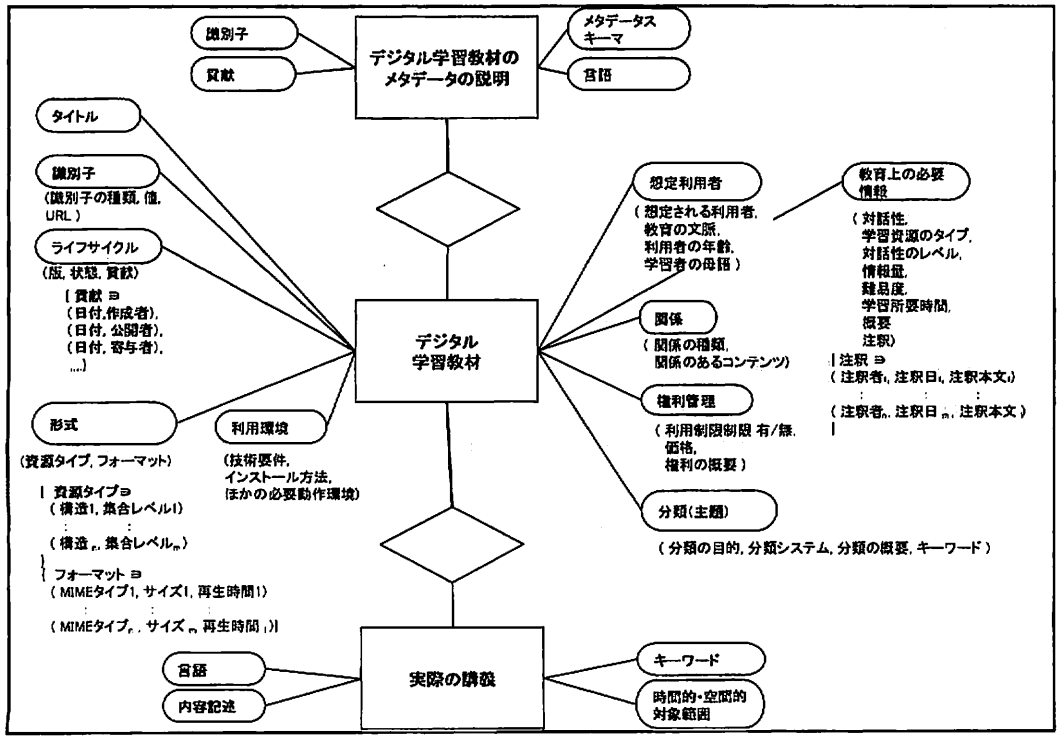

図 2　LOM の構造

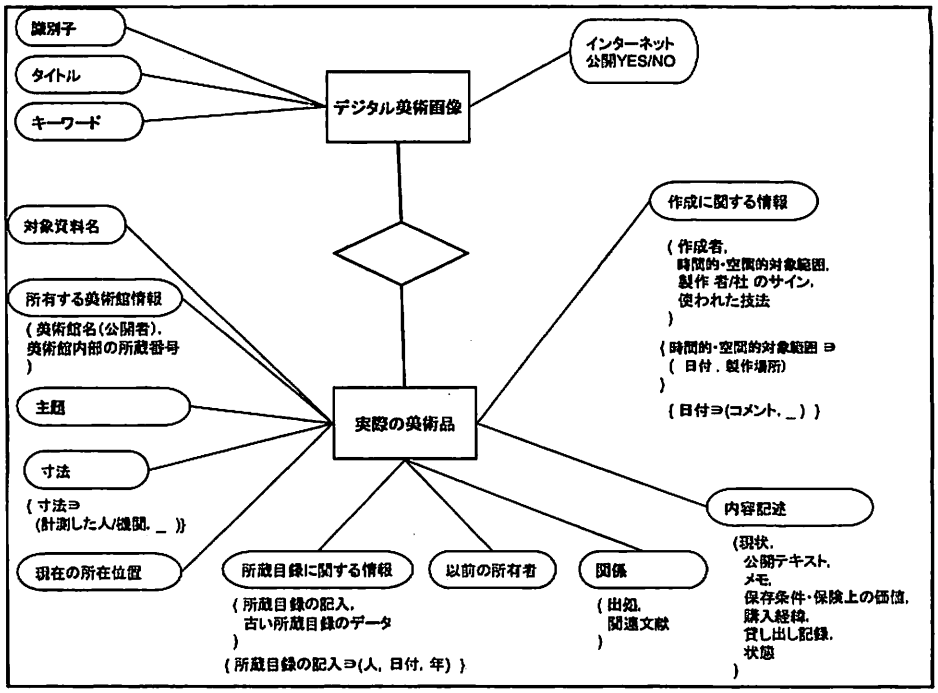

図 3　ドレスデンメタデータの構造

4. 2　メタデータの構造

3 つのメタデータの情報構造の ER 表現上の特徴をまとめたものが表 1 である. メタデータとして扱う情報の体系を見ると, Dublin Core とドレスデンメタデータは, 2つの実体とそれらを関係付ける比較的単純な構造を取る. 実体の特徴を表す属性の数も 16,14 と類似している. 一方 LOM は, 3 つの実体と 2 つの関係から成り, 属性数も他の 2 つの倍以上の 37 ある. LOM のメタデータの情報構造は, 他の 2 つに無い情報管理用の実体を持っており, また動作環境に関する属性を備えている. 実体が持つ属性の数を見ると, デジタルライブラリでは実体それぞれ(デジタルコンテンツとその元となったオブジェクト)に大きな差が無く, 9 件と 7 件である. ドレスデンメタデータは, デジタルコンテンツが持つ属性の数は4, デジタル化の対象物の属性の数は10であり, 差を認められる. 一方, LOM はそれぞれ 4 件, 10 件であり Dublin Core のそれらと数が逆転している. また属性が取る値に注目すると, Dublin Core とドレスデンメタデータはその多くが任意の一つの値をとるか, 2および3の要素の組からなっている. 値がさらに階層的な構造は持つものは無い. 一方 LOM は, 4 つ以上の要素からなる値を持つ属性が3つあり, さらその中の要素が階層構造を持つものがある. 具体的には, 属性"教育上の必要な情報"の要素"注釈"は, さらに 3 つの要素の組の集合からなる.

このような属性の詳細化は, 情報システムの入力フィールドに展開したときに影響する. 表 2 に示すとおりメタデータの属性の数より多い入力フィールドが用意される. 例えば, 図 2 の LOM の実体であるデジタル学習教材の属性"形式"は, Dublin Core の実体"デジタル版書式/本"の属性"形式"に意味的に同一である. 一方 Dublin Core のその値が 3 つの要素の組からなるのに対し, LOM では2つの要素から成り, さらにそれぞれが2つおよび3つの要素から成る組の集合で構成されている. これらの構成要素すべてを入力値として必要とすることから, 情報システム上のフィールドとして展

開したときに，その数が増加する．
また3つのメタデータともに共通していることは，2つの実体(デジタルコンテンツと，その元になったオブジェクト)を持つことと，それらが次の属性を持つことである．作成者および作成に関係した人物に関する情報，作成日に関する情報，コンテンツの表題，デジタルコンテンツ化された状態(形式や大きさ)，コンテンツのアクセス先(URL)，アクセス権利情報である．

表 1　ER モデル表現によるメタデータの構造

D.C.(Digital Contents)；デジタルコンテンツ A.C.(Analog Contents)：元となったオブジェクト M.D.(Metadate Discription)：デジタル学習教材のメタデータの説明	実体の数	関係の数	属性の数			
			合計	D.C.	M.D.	A.C.
Dublin Core	2	1	16	9	−	7
LOM	3	2	18	10	4	4
ドレスデンメタデータ	2	1	14	4	−	10

表 2　情報システム上で入力が必要となる項目数

D.C.(Digital Contents) ；デジタルコンテンツ A.C.(Analog Contents) ；元となったオブジェクト	値の入力が必須な項目の数			値の入力が推奨される項目の数			それ以外の項目の数			全合計数
	計	D.C.	A.C.	計	D.C.	A.C.	計	D.C.	A.C.	
Dublin Core	システムに依存			15	8	7	31	27	4	46
LOM	8	7	1	50	47	3	0	0	0	58
ドレスデンメタデータ	27	4	23	5	0	5	5	0	5	37

4．3 考察

3種類のメタデータの構造の比較からLOMが最も詳細化・精密化され深い階層構造をとっていることがわかる．これは，メタデータが対象としているコンテンツの特徴によると考えられる．デジタルライブラリーとドレスデンのデジタル美術画像が広く専門領域を超えた一般の利用者に共有されることを目指しているのに対し，eラーニングのコンテンツは，特定もしくは限定された利用者を対象にしている．実体“デジタル学習教材”の属性“想定利用者”や“教育の必要な情報”が用意され，さらにそれらの値が詳細化された要素から成ることもこの事実を裏付けている．同様の解釈は，ドレスデンメタデータにも言える．広く一般の利用者を対象にしている実体“デジタル美術画像”の属性数に比べ，専門家が調査研究の対象とする実体“実際の美術品”の属性数が多くなっている．
さらにeラーニングコンテンツは，一般に(実際の授業の録画を元に)動画等のリッチコンテンツを利用していることが多く，コンテンツの利用環境を指定・推奨する属性も存在する．これもまた，利用者を特定する要素となっている．これらのことは，我々の事前の仮説「専門家もしくは利用者を特定したいコンテンツではメタデータの詳細化・精密化が必要であり，一方専門領域を横断もしくは非専門家を対象とした利用を目的とするコンテンツは，大略的な表現や区分によるメタデータが適している」を裏付ける結果となった．

5.　DMC context accumulation system

2004 年，慶應義塾大学が文部科学省の科学技術振興調整費の「戦略的研究拠点育成プログラ

ム」に 提案し採択された. これを受け, 同年 DMC が設立された. この機構では, これまで紙媒体に文字情報を中心に記録され伝達されてきた知的資産を, デジタルコンテキストとして創造することを促進するための次世代型情報システムやメディア技術を研究開発する. 開発したシステムやメディア技術を実際に利用し, 知の提供と知的インタラクションのあり方を一新するという, わが国が「知識創造立国」として世界の知識社会に貢献するための先導的役割を果たそうとするものである. ここでの研究成果の一つとして現在構築が進められているものの一つに DMC context accumulation system がある. このシステムは, 慶應義塾の研究者が発信するデジタコンテンツを, 専門領域を横断し, また学外の利用者にも共有と活用を実現することを目指している. 今回我々は, これを実現するためのデジタルコンテンツ用メタデータの基本形を考案した.

5. 1 DMC デジタルコンテンツメタデータ構造

我々は, 3つの代表的もしくは注目すべきデジタルコンテンツのメタデータの構造を調査し, DMC context accumulation system の目的に最適なメタデータの構造を設計した(図 4). 特徴は, (1)前章で述べたメタデータの構造の共通特徴を持つこと, (2)専門的利用よりも一般的な利用や専門領域を超えた利用を想定したものであること, (3)専門的利用要求に対応可能なように拡張性をもたせたこと, (4)情報システムに登載したときにシステムが自動獲得できるものを活用し利用者による手作業入力作業を最小限にとどめたこと, そして(5)独自の要求である作成途中(作成過程)コンテンツの情報も入手できるようにしたことの 5 点である. このメタデータの構造と前述の 3 つを比較した結果を表 3, 表 4 に示した.

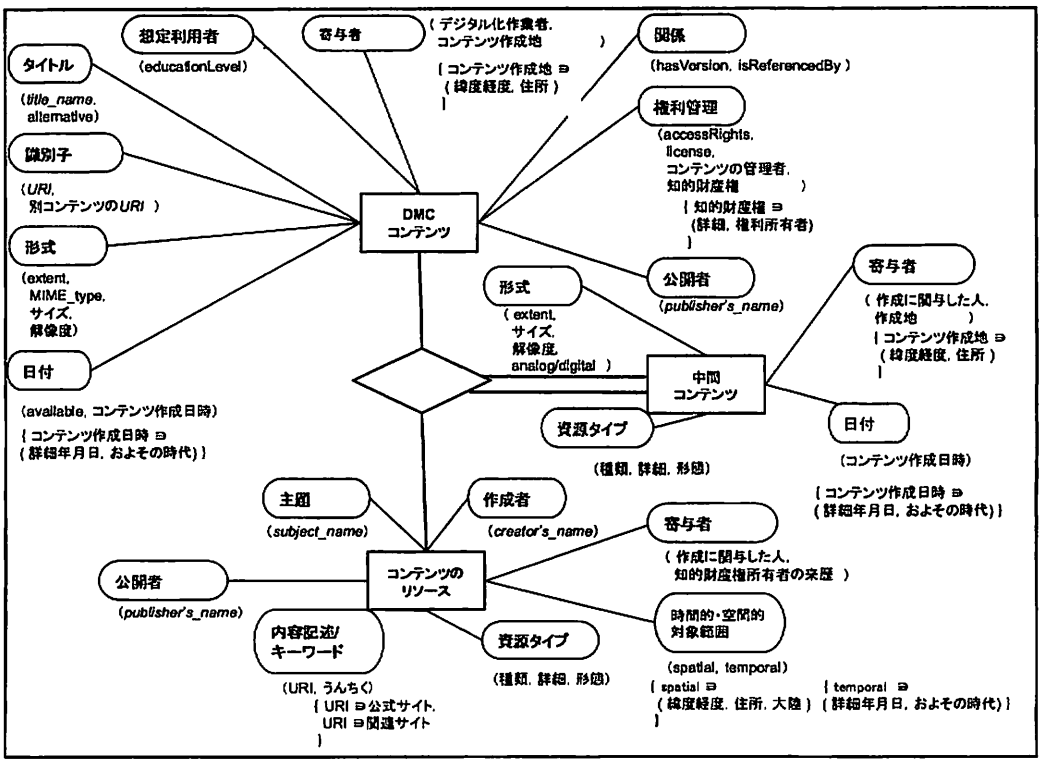

図 4 DMC デジタルコンテンツメタデータの構造

表 3　DMC 考案メタデータの構造

D.C. (Digital Contents) : デジタルコンテンツ A.C. (Analog Contents) : 元となったオブジェクト I.C. Intermediate Contents : 中間コンテンツ	実体の数	関係の数	属性の数			
			合計	D.C.	I.C.	A.C.
DMC 考案メタデータ	3	1	20	9	4	7

表 4　DMC 考案メタデータの情報システム上で入力が必要となる項目数

D.C. (Digital Contents) : デジタルコンテンツ A.C. (Analog Contents) : 元となったオブジェクト I.C. Intermediate Contents : 中間コンテンツ	値の入力が必須な項目の数				値の入力が推奨される項目の数				それ以外の項目の数				全合計数
	計	D.C.	I.C	A.C.	計	D.C.	I.C.	A.C.	計	D.C.	I.C.	A.C.	
DMC 考案メタデータ	9	7	0	2	0	0	0	0	47	18	13	16	56
※1	(6)	(6)		(0)					(25)	(9)	(6)	(10)	(31)

※1 括弧内は、システムが自動/半自動で取得する項目数

6　まとめ

我々は，デジタルコンテンツの流通を目指した代表的なメタデータ，Dublin Core と LOM，およびそれまで専門家間で研究材料とされてきた美術品をデジタル画像化することで広く一般の利用者へ共有することを目指したドレスデンの活動から生まれたメタデータの 3 つを対象として情報構造を調査した．その結果，必要不可欠な属性情報の把握と，専門性の高いコンテンツと一般利用を目指したコンテンツのメタデータの違いを確認した．この結果を参考に，我々が開発するデジタルコンテンツ流通システムにおける標準メタデータの構造を提示した．今後は実際にこのメタデータを搭載したコンテンツを流通させその有効性を確認する実証実験を開始する予定である．

謝辞

本研究は，文部科学省科学技術振興調整費の支援によるものである．

参考文献

[1] 萩野達也，神原顕文，清水昇，豊内順一，細見格，津田宏，白石展久，韋慶傑:"センマンティック Web とは",情報処理,Vol.43,No.7,pp.709-717(2002)

[2] 鹿島みづき，山口純代，小嶋智美:"パスファインダー・LCSH・メタデータの理解と実践 : 図書館員のための主題検索ツール作成ガイド",愛知淑徳大学図書館(2005)

[3] 松井くにお，津田宏，上田健次，小泉雄介，豊内順一，布目光生:"セマンティック Web におけるメタデータとその活用",情報処理,Vol.53,No.7,pp.718-726(2002)

[4] イーゴル・イェンツェン:"文化遺産の継承と新たなる挑戦に向けて",世界遺産 古都ドレスデン デジタル技術が支える人類の記憶,慶應義塾大学デジタルメディア・コンテンツ統合研究機構編,pp.67-91,(2006)