Web 上で公開された博物館資料メタデータの評価：値の記述率，値の形式的妥当性，値の表記一貫性の観点から

矢代 寿寛　　宮澤 彰
総合研究大学院大学／日本学術振興会　　総合研究大学院大学／国立情報学研究所

抄録　Linked Open Data 化を企図して収集された Web サイト上の博物館資料メタデータ約37万件について，分析と評価を試みた．記述項目の定義率，値の記述率，項目と値の対応における形式的な妥当性，値の表記の一貫性などを分析した．分析結果から，二次的利用を行う立場からの評価を行った．「作品名」などの主要な項目は，定義率や値の記述率が高く，有用な情報源となりうるものの，半数以上で項目名が明示されていないなど，機械処理に適さない面がみられた．

A metadata evaluation approach of Japanese museum's collections on the Web: blank value, coherence, consistency

Kazunori Yashiro
Department of Informatics
SOKENDAI

Akira Miyazawa
Department of Informatics
SOKENDAI

Abstract　This research, aiming at the support for create "Linked Open Data" of LAM(Library, Archives, Museum), tries to analyze and evaluate the scraped dump data of museum collections that has 370,000 object records. The result, core metadata keys (e.g., "Title") is popular, has less blank value, and has potential of become the useful resource. However, that was not explicitly named for human-readable in more than 50%, that unsuitable for machine processing.

１．はじめに

学術情報の共有をより促進するため，情報・システム研究機構新領域融合研究センター「学術リソースのためのオープン・ソーシャル・セマンティック Web 基盤の構築」プロジェクトの一部として，国立情報学研究所らにより，LODAC プロジェクト（Linked Open Data for Academia. 以降，LODAC）が進められている[1]．LODAC のプロトタイプシステムの一つに，我が国博物館分野の情報を主たる対象とした LODAC Museum がある．LODAC Museum では，Web 上で公式に公開されている博物館資料のメタデータを収集し，それらを機械的に Linked （Open） Data 化している．

Linked （Open） Data（以降, LOD）は，データ共有技術の一種で，機械可読形式で構造的に記述されたデータまたはメタデータである[2]．LOD の要件として，a）URI（URL）によるリソースの識別，b）HTTP による参照，c）標準技術による有用なデータへのアクセス提供，d）他の情報資源へのリンク，の四点が提唱されている[3]．また，ライセンスや公開の状態によって，1）オープンライセンスでのデータ公開，2）機械可読形式での公開，3）非独占の機械可読フォーマットでの公開，4）Web 標準技術での参照・識別が可能，5）他の情報源へのリンク，の五段階格付け評価を行うことができる[4]．実装として，RDF 形式のメタデータを RDF ストアに蓄積し，SPARQL Endpoint などを通じて提供することが一般的である．RDFa や Microformats のメタデータ埋め込みによる XHTML の更なる構造化という簡易な方法も可能である．2011年には，LOD-LAM という博物館・図書館・文書館分野における国際会議が開催され，また W3C の Library Linked Data Incubator Group が最終報告書を公開するなど，今後の人文科学分野への浸透が期待される[5][6].

2011年11月現在，LODAC Museum では，53機関の約82,000件が利用可能になっている．これは，

2010年度に Web 上で収集した14機関の博物館資料メタデータ61,861件（以降，第一次収集メタデータ）と，2011年度に収集した70機関の421,446件（以降，第ニ次収集メタデータ）,合計84機関483,307件のうち，処理済みの分である．また，一部に日本美術シソーラスが収録している作品情報を含んでいる[7]．これらのデータは，Wikipedia（DBpedia）などとリンクした状態で統合的に検索可能になっており，博物館資料や作家情報の擬似的な総合目録システムとして機能している．

本研究では，LODAC が収集し，LOD 化処理をする前の，Web 上で公開されていた博物館資料メタデータのダンプと，LODAC とは別に筆者が収集したメタデータのダンプの分析・評価を行っている．公開されているメタデータの状況を概括することにより，二次的利用者による複合的な活用，いわゆるマッシュアップを促すことを目的としている．これまでに，第一次収集データについて分析し，二次的利用を行う立場からの評価を試みた[8]．本稿では，第二次収集メタデータに独自収集データ追加し，さらに分析困難なデータを除外した上で，分析した結果について述べる．分析対象の変更と量の増加に加え，方法として，第一次収集メタデータの評価試論に関する議論を踏まえ，メタデータの値の表記一貫性をより高い精度で分析することを試みる．分析結果から，二次的利用を行う上での機械的な処理における利用しやすさの評価を行う．

2．博物館資料メタデータの種類

分析・評価を行う前に，本稿で対象とする Web 上で公開された博物館資料メタデータの位置づけを整理する．博物館資料のメタデータには，作成主体，利用範囲，媒体などによって異なる記述の形式や項目，値を持つものが存在する．作成主体は，資料の所蔵機関である博物館・美術館とそれ以外に大別できる．利用範囲は所蔵機関内に限定されるものとされないものがある．所蔵機関内に限定される場合は，さらに利用対象が，内部向けと外部向けに分けられる．媒体としては，電子的な媒体とそれ以外の紙を主とする媒体に分けられるが，近年では区別する意義を見出し難い．それぞれの分類に該当するメタデータ（を収録するリソース）について，代表的な例を表１に示す．

表１　代表的な博物館資料メタデータの種類

<table>
<tr><th rowspan="3">利用範囲</th><th rowspan="3">利用対象</th><th colspan="4">作成主体</th></tr>
<tr><th colspan="2">所蔵機関（博物館・美術館）</th><th colspan="2">その他（作者・出版者・研究者等）</th></tr>
<tr><th>電子媒体</th><th>その他媒体</th><th>電子媒体</th><th>その他媒体</th></tr>
<tr><td rowspan="3">機関内限定</td><td rowspan="2">内部</td><td colspan="2">資料台帳・目録</td><td></td><td></td></tr>
<tr><td colspan="2">カルテ</td><td></td><td></td></tr>
<tr><td>外部</td><td>閲覧システム</td><td>展示解説</td><td></td><td></td></tr>
<tr><td rowspan="9">限定なし</td><td rowspan="9"></td><td colspan="2">展覧会カタログ</td><td colspan="2">参考図書</td></tr>
<tr><td colspan="2">出品リスト</td><td colspan="2">ポートフォリオ</td></tr>
<tr><td>コレクション紹介</td><td>所蔵資料目録</td><td colspan="2">競売目録</td></tr>
<tr><td>コレクション DB</td><td>展示ガイド</td><td>展覧会特設サイト</td><td>ガイドブック</td></tr>
<tr><td>仮想展示・展示ガイド</td><td>鑑賞教材</td><td>シソーラス</td><td>学術書</td></tr>
<tr><td>モバイルアプリ</td><td></td><td></td><td>研究報告</td></tr>
<tr><td>統合 DB</td><td></td><td></td><td>画集・レゾネ</td></tr>
</table>

所蔵機関が作成主体となるメタデータのうち，最も信頼性があり，全ての基本になると考えられるのは，所蔵品管理に用いられる資料台帳・目録である．『平成 20 年度博物館総合調査』によると，およそ 90%（n=2257）の機関で何らかの資料台帳が作成されている[9]．所蔵機関内限定かつ内部向けの業務用メタデータであり，同種のものにはカルテがある．カルテは資料の貸借や保存における状態および作業内容の記録として作成される．これら他に，一時的な作業に伴うデータやカードが個人単位で作成される場合もありうる．これらのメタデータは，収蔵に伴う経緯のような個人情報などを含む場合があり，これらは外部に公開されないため，情報量も他のリソースより多くなると考えられる．

利用範囲が機関内に限定され，かつ外部利用者でも利用可能なメタデータには，展示室内に設置されている専用の閲覧システムや解説が挙げられる．展示解説は，作成対象が展示資料に限られており，全資料に作成されるものではないという点で，規模に劣ると考えられる．なお，鑑賞・学習の補助として，展示解説をカード等の形式で頒布している場合があり，これは機関内に限定されない[10].

利用範囲が機関内に限定されない資料メタデータは，企画・常設の展覧会に伴うものと伴わないものに分けられる．展覧会に伴うものとしては，展覧会カタログや出品リストが代表的である．展覧会カタログは，展覧会に伴う出版物であり，学芸員らにとっての研究成果物に相当する．展覧会カタログが内包する博物館資料メタデータは，図版解説が中心である．出品リストが収録され，二種類の記述が混在する場合もある．展覧会会期後に，情報の追加された豪華版や作品集が出版されることがある[11]．単独の出品リストは，展覧会カタログの出品リストを抜粋したものとそうでないものがあり，両者はメタデータの記述が異なる．主に紙媒体で展覧会会場において無償配布されるが，東京国立近代美術館などの一部の機関では，単独の出品リストを Web サイト上で公開している[12]．この他，近年では展覧会の鑑賞補助として，モバイル機器向けに専用のアプリケーションを提供する例が見られる[13]．国立文化財機構の「e 国宝」のように，展覧会に伴わないアプリケーションも存在する[14].

外部向けの資料台帳・目録に相当するのが，Web サイト上のコレクション DB（スプレッドシート）と，主に冊子体で出版される所蔵資料目録である．機関の Web サイトには，DB ではないコレクション紹介のページが設けられている場合があり，主に資料的価値の高いものなどに限った公開が行われている．所蔵資料目録は，「美術工芸編」といった特定コレクションのみに収録範囲が限定される場合がある．この他，実際の展示室を模して，学習・鑑賞の補助とするページや，デジタル化された資料を公開するバーチャルミュージアムなども設置されている場合がある．展示施設を持たず，バーチャルミュージアムの形式でのみ博物館資料の情報を公開している例も見られる[15]．これらに加えて，個々の機関の DB を統合した DB が，国や自治体により作成されている．統合 DB では，個々の機関で用いられている記述の差異をマッピング等により埋めることで，ほぼ共通のフォーマットでデータが公開されるため，ローカルの Web サイトの記述と異なる場合がみられる．Web サイトではメタデータをほぼ公開せず，統合 DB にのみデータを提供する例もある[16].

本稿で分析・評価の対象とするメタデータは，所蔵機関の Web サイト上で公開されているコレクション紹介および DB のデータの一部であり，詳細なレコードページまたは表ページに存在する．博物館資料メタデータ全体からみると，利用範囲・利用対象に制限がなく，電子的な媒体であり，展覧会に伴うものではなく，鑑賞・学習補助に特化していない，という性質を持つものといえる．資料台帳・目録より信頼性は劣るものの，電子化を必要しないコンピュータ上で利用できるメタデータとしては，最も汎用性が高いと考えられる．また，既に公開されたデータを活用するため，所蔵機関に更なる負担を求めないという利点がある．

しかし，実際に分析することができるのは，スクレイピング・スパイダリングにより収集したダンプデータであるため，厳密には Web サイト上のデータそのものであるとはいえず，留意が必要である．もし収集元と対照させたときにメタデータが異なる状態であった場合，それがスクレイピングの失敗によるものか，収集後の変更によるものかは，個別に確認しなければ明確にいえない．LODAC が収集した数十万単位のメタデータに関して，このような同一性の確認を行うのは現実的ではない．よって，本研究での分析や評価は，Web サイト上で公開されるメタデータそのものの改善に直接資するものではない.

3．先行研究と本研究における分析・評価方法

我が国の博物館資料を対象としたメタデータ分析・評価に関して，業務ではない研究としては，これまでほとんど行われてこなかった．資料台帳・目録へのアクセスが限られていたことと，大規模館

以外でのメタデータ自体の整備が遅れていたことが遠因と考えられる．本研究では，我が国における列挙書誌および文献目録の評価，先行する海外での図書館資料の書誌レコード評価と博物館資料メタデータ分析事例から，分析・評価方法の我が国博物館資料への応用を試みる．

我が国での大規模な試みとしては，1980年代の小田ら日本索引家協会のグループによる冊子体列挙書誌の評価項目調査が挙げられる[17]．この中では，当時の既存評価研究から，主たる12点が評価基準として抽出された．それらの内，分野の違いを考慮すると，博物館資料メタデータ（を収録するリソース）に対して応用可能なのは，収録範囲，編纂方法，構成および配列，記入，適時性，正確さ，継続性，と考えられる．海外では，2000年代以降の図書館において，Bruce らにより提案された，1）Completeness，2）Accuracy，3）Conformance to expectations，4）Logical consistency and coherence，5）Accessibility，6）Timeliness，7）Provenance，の7軸のフレームワークと，それを拡張した方法論が中心となっている[18]．これらの研究は小田らと時期が大きく異なるため，OPAC等の電子化された書誌レコードが対象である．7軸のフレームワークは，読み替えによって全て博物館資料メタデータへの適用が可能と考えられる．博物館資料メタデータを対象とした海外での分析では，比較的新しいものとして，Hooland らによる LOD 化プロジェクトが挙げられる[19]．この中で，Powerhouse Museum の博物館資料メタデータを LOD 化する下処理として，空白項目率の調査や Google Refine というデータクレンジングツールを用いた表記揺れの分析などが行われている．

本研究では，これら既存研究と，第一次収集データの分析・評価の試行結果から，資料の非専門家であっても可能な分析・評価を改めて試みる．方法としてまず，第二次収集データ全体に関して，定義されている記述項目の名称と数を整理し，概要を明らかにする．次に，一度に複数の機関の博物館資料メタデータを用いるため，個々のメタデータレコードを LODAC 共通の形式へと便宜的にマッピングし，同一の枠組みで分析可能にする．そして，各項目の値について，1）値の記述率，2）値の形式的妥当性，3）値の表記の一貫性，を分析する．1）値の記述率，は用意された項目に対する値の記述割合である．Hooland らの研究における空白項目率に相当する．例えば，値が分からない場合に「不詳」等の記述がなく空値であれば，記述率は低くなる．2）値の形式的妥当性，は用意された項目に対して記述された値が，形式上妥当であるかである．ここでの形式とは，例えば，「寸法・法量」の項目には縦横径奥行など部位名やセンチメートル単位の数値が記入される，といった項目と値の基本的な対応関係を指す．仮に，「寸法・法量」に和暦が記述されていれば，その値が妥当ではないことは非専門家でも比較的容易に判断可能である．専門家による正確さの評価を代替することを意図している．対応関係をみるため，記述項目が明示されていないものはここで分析対象から除外する．3）値の表記の一貫性，は記述された値の表記が，記述規則の中でどれだけ一貫しているかである．例えば，「寸法・法量」の項目で「縦10 横10」と「10×10」が混在していれば，一貫性は低下する．

以上の観点について， Hooland らが利用した Google Refine を用いて分析を試みる[20]．

４．分析・評価対象の概要 とメタデータの項目

本研究では当初，第二次収集データ 421,446 件に，独自収集したメタデータ 1,041 件を加えた 422,487 件を分析対象としていた．しかし，このデータ全体から図書資料，遺跡等の非モノ資料のメタデータ，明らかに収集に失敗した分を除いて分析を試みたところ，自然科学資料が人文科学資料（美術工芸資料や考古歴史資料など）と大きく異なる傾向を示したため，本稿では人文科学資料のメタデータ 61 機関（87.1%，n=70）71 規則（62.3%，n=114）367,375 件（87.0%，n=422,487）のみについて述べる．2008 年度の『社会教育調査』によると，我が国博物館の人文科学資料（模型および図書資料を除く）の総数は 5775 機関 97,123,231 件であるため，重複などを考慮しない場合，これは約 0.4%に相当する．このうち 245,337 件（66.8%）は唯一の国立機関である国立民族学博物館のデータであり，全体に対する影響が過大であると考えられる．そこで，第一次収集データで行ったような全

体的な値の記述傾向に関する分析よりも，個々の機関単位での分析をまとめる方向とした．概要として，機関の内訳，メタデータの記述規則数，定義項目数，記述形式，項目名の明示状況，主要定義項目，を表 2 に示す．

表 2　第ニ次収集データ概要

メタデータ収集元機関数（割合，n=62）				
設置	国立	都道府県立	市区町村立	私立
	1(1.6%)	29(47.5%)	27(44.3%)	4(6.6%)

館種*	総合	美術	歴史	不明	VR
	6（9.7%）	38（61.3%）	6（9.7%）	7（11.2%）	5（8.1%）

記述規則（記述項目のセット）			
記述規則数	単一記述規則 利用機関数(n=61)	複数記述規則 利用機関数(n=61）	複数利用機関における 平均記述規則数
71	54（88.5%）	7(11.5%)	2.3

一記述規則における定義項目数の代表値（資料画像，関連リンク等含む）						
最大値	最小値	平均値	中央値	最頻値	標準偏差	分散
24(1 機関)	3(4 機関)	10.5	10	9	4.69	21.97

詳細レコードの記述形式件数（割合，n=71）			
単票	帳票**	リスト・表	その他
17(23.9%)	50(70.4%)	4(5.6%)	0

一部または全部の項目名の明示件数（割合，n=71）		
表の項目	凡例	明示なし
43(60.6%)	3(4.2%)	25(35.2%)

主要記述項目の定義率（n=71）（カッコ内は項目明示割合，n=当該項目定義規則数）						
作品名	制作者	制作年	材質・技法	員数・形状	寸法・法量	概要・解説
100% （42.3%）	64.8% （60.9%）	83.1% （64.4%）	67.6% （58.3%）	25.4% （61.1%）	77.5% （63.6%）	70.4% （46.0%）

*）VR 以外は『全国博物館総覧』加除式に掲載されていた分類を用いた．

**）一部項目が表などの構造もち，他の項目は視覚的な配置によって人間に解釈させる形式．例えば，本稿初頁の標題，著者，所属，抄録．

機関の種類としては，国都道府県立のみであった第一次収集データに比べ，市区町村立と私立を含んでおり，やや規模の幅が広いことがいえる．公立機関が中心となったのは，特例市以上規模を持つ主要自治体の機関の Web サイトを網羅的に調査した上で収集したためである．また，収集の前提として，Web サイト上で博物館資料メタデータの公開が行われている必要があり，準備段階で調査した範囲では，ほとんどの私立機関で公開が行われていなかったためである．館種として，美術系が多い点は第一次収集データと共通している．

定義項目の組み合わせを記述規則と見做し，各機関における使い分けを調査したところ，7 機関で資料種別ごとの使い分けが行われていた．種別にかかわらず単一の記述規則を用いていたのは，54 機関であった．複数利用の 7 機関には，平均 2.3 種類の記述規則があった．詳細なレコードの記述形式として最も多かったのは，本研究で便宜的に「帳票」と呼称している形式であった．

定義されていた項目は，明示・非明示問わずの最多が「作品名」に相当する項目であった．定義されていなかったのは，本稿の分析対象外である 1 機関 3 記述規則の，不定形な記述が行われている人間可読も困難なメタデータのみであり，例外といえる．他には，「資料画像」（94.4 %），「制作年」，「概要・解説」，「作品名よみ」（45.1%）の定義率が比較的高かった．これら単独での定義率が高い項目を，最少定義項目数である 3 組にした場合，最も定義率が高い組み合わせは，「作品名」「制作年」「資料画像」の 51 機関（83.6%，n=61）56 規則（78.9%，n=71）であった．規則よりも機関の方で割合が高いため，特定の機関が多くの規則で同じ項目のセットを定義することにより割合が

高いのではなく，多くの機関で定義されているから率が高い，すなわち普遍性のある組み合わせと考えられる．しかし，組み合わせる項目数を増やすと定義率は大幅に低下し，上位 4 項目の組み合わせではおよそ5割，上位5項目では3割程度となった．

「制作者」は，第一次収集データの分析・評価において，定義率・記述率・表記一貫性の高さから，機械処理しやすい項目であると結論づけられた．しかし，今回の第ニ次収集データでは，全体で 6 割，平均以上の項目定義数がある 28 規則でも約 3 割で定義されていなかった．精査した結果，資料種別に依存する性質があるのではないかと考えられるが，定量的にはいえる状態になかった．また，「制作者」ごとに Web ページのディレクトリ分けをサイト上ですることにより，レコードでの項目としての定義率が低くなるのではないかとも考えられる．

定義項目の明示状況に関して，全項目の名称を明示していたのは6機関のみであった．項目別では定義率の高い「画像」の項目名明示が 7.5%で最低であった．視覚的に何を意味するか解釈しやすい項目であるためと考えられる．「作品名」に関しては，定義率は高いものの，明示は半数でしかされていなかった．また，「作品名よみ」も 28.1%とやや低い値を示した．収集元の Web サイトで調査したところ，収集時期と異なるため厳密にはいえないものの，画像と同様に視覚的な配置によって「作品名」とその「作品名よみ」であると人間に解釈させる傾向がみられた．この他に特徴的であったのは，「員数・形状」「管理番号」の項目で，定義率は低いものの，定義されていた場合の項目明示率は 6 割以上と高かった．これらは人間であっても非明示では値の解釈が困難なために明示されると考えられる．「制作者生没年」は，定義率（23.9%）・明示率（17.6%）ともに低く，対応関係の解釈は「制作者」との位置関係に依存していることがいえる．美術史等の研究において重要と考えられるものの，明示率・定義率が低いため「制作年」等の値と混同しやすいため，留意が必要と考えられる．この他，特定種別の資料に関しては，美術工芸資料における「展覧会歴」，考古資料における「出土地」「地名」の定義が特徴的にみられた．

５．値の記述率

まず，項目が定義されていた場合の値の記述率を分析した．主要な項目について，規則ごとでの最高・最低の記述率を，表3に示す．参考として，国立民族学博物館を除いた全体での項目ごとの記述率と，記述率の中央値，平均値を併せて示す．

表3　値の記述率

項目	規則ごとの記述率と該当規則数		全体**	記述率の中央値*	記述率の平均値*
	最高	最低			
作品名	100%（43）	19%（1）	95.3%	100%	96%
制作者	100%（18）	7%（1）	90.6%	100%	80%
制作年	100%（14）	6%（1）	68.4%	90.9%	76%
材質・技法	100%（11）	10%（1）	95.0%	99%	84%
員数・形状	100%（4）	12%（1）	89.0%	93%	75%
寸法・法量	100%（13）	3%（1）	61.3%	94%	73%
概要・解説	100%（13）	1%（3）	32.4%	88%	67%

*）小数点以下切上げ.
**）項目が定義されており，かつ記述率が0ではない記述規則のデータ件数合計が母数.

取り上げた主要項目すべての記述率が100%であったのは1機関であった．いずれかの規則で定義された全ての項目のうち，最高記述率が100%ではないものはみられなかった．定義率が最も高かった「作品名」は，4規則を除いて90%以上の記述率を示した．「制作者」は定義割合が低いものの，定義された場合は何らかの記述が行われる傾向にあったといえる．一方で，「制作年」は定義割合が高いものの，値が不明な場合などで，「不詳」等の記述が行われず，空値になる傾向があった．また，記述されていても「＊」「-」などの代替文字がいくつかの規則で存在していた．「材質・技法」に関しても，代替文字が記述されているレコードが半数以上を占める例が，公開件数1000件以上の機関

で複数みられた．「員数・形状」に関しては，「１点」など何らかの値によって記述がなされる傾向にあった．また，一部の機関では，単一の資料に対して「１点」と「六曲一隻」のような異なる値が複数「員数・形状」に記述されている例があった．これは，予め「１点」などの値を入力していることが原因と考えられる．「寸法・法量」は，ある程度公開件数が多い機関での記述率が低く，また設置率が高いために全体としては割合が低くなっていた．「概要・解説」は，記述率が高いか極端に低いかの二極化傾向にあった．

６．値の形式的妥当性と表記の一貫性

次に，記述された値が項目と適切に対応しているか否か，形式的な妥当性について分析した．全体的に，妥当でない値が記述される例はほとんどみられなかった．一部記述規則の特定の項目にのみ多数存在していた．このうち，全レコードの「概要・解説」の項目に「解説」という値のみが記述されていた一例以外は，データ収集の際の誤りによるものである可能性が高いと考えられる．誤りによる場合はいずれも，項目名が明示されていない機関のレコードであり，位置的に近い項目の値が空値であるためにずれが生じたと考えられる．

そして，値の表記の一貫性について，第一次収集データの分析・評価の試行結果から表記揺れが起こりやすいと考えられる項目ごとのパターンの数と代表的な表記例を表4に示す．パターンとしてまとめる際には，Google Refine の出現頻度でのソートや N-Gram ベースのクラスタリング機能による分析を基本に，カッコやカンマ，ハイフン等の記号による値の区切り方や，「初期」や「頃」といった日付に付与する語の付与位置などから，判断した．

表4 第ニ次収集データにおける表記揺れ

項目名	一規則の最大揺れ数	一規則の最小揺れ数	主な表記の例
制作年（西暦）	16（1）	1（2）	"YYYY"，"YYYY 年"， "YYYY（和暦年号 YY）"
材質・技法	10（2）	3（2）	"材質・技法"，"材質／技"， "技法・材質"，"技法 / 材質"
員数・形状	6（1）	1（1）	"N 点", "N 幅", "六曲一双"
寸法・法量	20（1）	1（4）	"X × Y", "X×Ycm"，"X",

「制作年」は，年代や日付が不確かであれば本質的に表記揺れを避けることが困難なメタデータであると考えられる．しかし，第二次収集データでは，全角や半角の違いによって生じた表記揺れが最多であった．「材質・技法」は，機関ごとに表記の順や区切り記号が異なるものの，一規則単位ではさほど表記揺れが生じておらず，主に材質等の量により変化するのではないかと考えられる．「員数・形状」については，項目を定義している規則自体が少なかったものの，いずれも全角半角の違いや助数詞の有無により表記揺れが生じていた．「寸法・法量」に関しては，記述項目名が「縦」「横」などの部位指定になっている規則が複数存在しており，そこでは数値のみが記述されるため，表記が揺れない例が少なからずみられた．資料の種別によって分けた記述規則ごとに分析を行ったため，第一次収集データに比べ，揺れの幅は小さくなったのではないかと考える．しかし，第一次収集データにおいて表記揺れを調査した「制作者」に関しては，一記述規則内で50以上のパターンがみられたため，本稿では分析対象から外した．

７．おわりに

本稿での分析結果から Web 上の博物館資料メタデータの評価を試みたとき，項目の明示が全くない規則が約3割も存在することが，機械処理における最大の問題点として挙げられる．記述されている値が何を意味するか明らかでなく，二次利用者側で対応関係を推論する必要が生じる．また，対応関係が一貫しているかの判別も困難である．テキストのクラスタリングによる判別という方法もあり

うるが，「作品名」のように，ある種のパターンがほとんどない値を持つ項目で明示されない割合が高いため，容易ではないと考えられる．記述率に関しては，全体的に高い傾向を示したものの，空白の代替として用いられた値によって過大になっているといえる．形式的妥当性については，一部例外を除いて問題はないと考えられる． また，表記揺れは，ツールの存在により二次的利用者側がクレンジングを行えるようになってきたため，機械的処理のしやすさという評価の観点からは，不要になりつつあるのではないかと考える．

今後は，第一次収集データ，第ニ次収集データ，独自収集データを一括して分析する予定である．これまで試行した形式的な分析・評価の他に，適時性や正確さなどの質的な評価についても可能であるか検討している．

付記　本研究で使用しているデータは，LODAC プロジェクトにご提供いただいたものです．国立情報学研究所の武田英明教授をはじめ，メンバーの皆様にお礼申し上げます．研究の一部は，日本学術振興会特別研究員奨励費によるものです．

参考文献

[1] "LODAC Museum"．http://lod.ac/
[2] Bizer, C., et al.: Linked Data の仕組み Linked Data-The Story So Far，情報処理，Vol.52, No.3, pp.284-292 (2011).
[3] Berners-Lee ,T.: Linked Data - Design Issues，
http://www.w3.org/DesignIssues/LinkedData.html
[4] "Linked Data star scheme by example"．http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/
[5] "LODLAM"．http://lod-lam.net/summit/
[6] "Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets"．http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/
[7] "日本美術シソーラス DB 絵画編索引"．http://www.tulips.tsukuba.ac.jp/jart/mokuji/
[8] 矢代寿寛, 宮澤彰: Web 上で公開された博物館資料メタデータの評価の試み. 情報処理学会研究報告 CH，Vol.2011, No.7，(2011).
[9] "平成 20 年度　日本の博物館総合調査研究報告書：文部科学省".
http://www.mext.go.jp/a_menu/01_l/08052911/1282292.htm
[10] "アートカード・セット | 学習 | イベント・学習 | NMAO：国立国際美術館".
http://www.nmao.go.jp/study/art_card.html
[11] "展覧会カタログに関する取扱い及び解説"．http://www.nii.ac.jp/CAT-ILL/PUB/nl2/No18/f6_p1.html
[12] "生誕 100 年 岡本太郎展　出品リスト"．http://www.momat.go.jp/Honkan/okamoto_taro/list.html
[13] "「フェルメールからのラブレター展」チャリティアプリ - Android マーケット".
https://market.android.com/details?id=com.hatais.vermeer
[14] "App Store - e 国宝"．http://itunes.apple.com/jp/app/id413457009
[15] "台東区ヴァーチャル美術館"． http://www.city.taito.lg.jp/bunkasinko/virtualmuseum/
[16] "文化遺産オンライン"．http://bunka.nii.ac.jp/
[17] 小田光宏, 樋口恵子, 河島正光, 長澤雅男: 書誌の評価基準, 書誌索引展望, Vol.10, No.1, pp.1-23, 1986.
[18] Bruce, T.R.,et al.: The Continuum of Metadata Quality: Defining, Expressing, Exploiting, Metadata in Practice, ALA, (2004).
[19] Hooland, S.V.,et al.: Free your metadata: Integrating cultural heritage collections through Google Refine reconciliation, (2011).
[20] "google-refine - Google Refine, a power tool for working with messy data (formerly Freebase Gridworks) - Google Project Hosting"． http://code.google.com/p/google-refine/