Linked Open Data による多様なミュージアム情報の統合

嘉村 哲郎[1,5] 加藤 文彦[2] 大向 一輝[2,1] 武田 英明[2,1] 高橋 徹[3] 上田 洋[4]

[1]総合研究大学院大学 複合科学研究科, [2]国立情報学研究所,

[3]ATR メディア情報科学研究所, [4]株式会社 ATR-Promotions ,[5]東京藝術大学

本論文は「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(LODAC)」プロジェクトの一環として取り組む，学術に関するデータを Linked Open Data(LOD)として公開・共有することで，学術コミュニティおよび社会で情報を活用できる，柔軟な情報流通の仕組みを構築することを目標としている．その最初の研究対象に，未統合かつ複雑な構造，語彙を有する芸術・文化情報を LOD として扱うプロトタイプシステム「LODAC Museum」を構築して，多様な情報と統合することで新たな発見や知の獲得の可能性を探る．

Integration of distributed of museum information with linked open data

Tetsuro KAMURA[1,5], Fumihiro KATO[2], Ikki OHMUKAI[2,1], Hideaki TAKEDA[2,1], ,
Toru TAKAHASHI[3], Hiroshi UEDA[4]

[1]The Graduate University of Advanced Studies (SOKENDAI), [2]The National Institute of Informatics,
[3]ATR Media Information Science Laboratories, [4]ATR-Promotions Inc. , [5]Tokyo University of the Arts

The museums in Japan maintain and publish museum information with the original metadata schemata. This leads to difficulty in crossover searching for museum information. This paper describes the method with linked open data (LOD) to integrate vast collections of museum information and relevant resources through the web. We introduce the prototype system (LODAC Museum) and attempt to aggregate information across multiple resources. We identify and associate artists and works from different museum collection to provide integrated views for them.

1． はじめに

本研究は，情報・システム研究機構 新領域融合研究センターのプロジェクト[1]，「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(LODAC)」の一環として取り組んでいるものである．本プロジェクトは，国内の学術に関するデータを Linking Open Data(LOD)方式で公開・共有することで，学術情報を学術コミュニティおよび社会で共有できる柔軟かつ広範な情報流通の仕組みを構築することを目標としている．本稿では，その最初の研究対象として未統合の情報かつ複雑な構造や語彙を有するミュージアム資料情報を多様な情報とゆるやかに統合する試み，「LODAC Museum」を紹介する．

本稿の構成を以下に示す．本章では国内の博物館資料基盤背景を取り上げ，2 章では LOD 化の必要性について論じる．3 章では LOD の概説と，これに関する関連研究を 4 章で紹介する．5 章前半では LODAC Museum 構築目的と目標，並びにプロトタイプシステムで使用する情報源を説明し，後半ではデータ統合に関する諸問題と解決方法を述べる．6 章ではサンプルデータによる動作検証結果，最後の 7 章では本研究の今後の展開について論じる．

1.1 国内ミュージアムの資料基盤背景

国内のミュージアムでは，1996 年のデジタルアーカイブ推進協議会の設立や(2005 年解散)政府の IT 戦略本部の「e-Japan 重点計画-2002」，総務省，経済産業省，文化庁等各省庁の情報基盤強化施策の後押しにより，90 年代後半からコレクションのデジタル化や DB 構築が活発に行われるようになった．当時は参考になるミュージアムに関するメタデータや仕組みとしてのデータモデルはほとんど知られていなく，各館独自のシステムとして構築されていた．この場合，複数館に対する横断的検索は難しいことから，当時の文化庁ではこれを解決するために「共通索引検索システム」の構築が進められていた．しかし，結果的にシステムをうまく機能させることなく，その構想は「文化遺産オンライン構想」へと引き継がれた．十数年経過した現在，日本には大小併せて 5773 館を超えるミュージアム[1]が存在し，現在も活発にデジタル化が行われている．一方で，国際博物館会議(CIDOC)による博物館資料構造化モデル(CIDOC CRM,ISO 21127:2006)や東京国立博物館 「ミュージアム資料情報構造化モデル」[2]が登場する等，資料基

[1]文部科学省「社会教育調査・博物館調査票」平成 17 年度による登録博物館，博物館相当施設ならびに博物館類似施設の数.

盤に関する部分で発展が見られた．しかしながら，上述するモデルや標準化規則が登場するものの，現在も多くの館が独自システムを使用している現状がある．さらに，市町村レベルの中小規模館では，未整理の資料が多く，とりわけDB化に着手出来ていない館が多数見られる．これは，一般に知られていない文化資源が多数存在することを意味し，これらを整理・公開・共有することでさらなる芸術・文化の価値創出が可能と考える．

2. 芸術・文化情報のオープン化

近年，幾つかの組織では複数ミュージアム間の資料を横断的に検索出来るシステムが構築されている．例えば，島根県の「しまねバーチャルミュージアム」や「独立行政法人国立美術館所蔵作品総合目録検索システム」は，システム上に登録された複数館の資料情報を同時に検索することができる．ところが，現在の横断検索システムは，複数館の情報を検索することが可能でも，他のシステムから資料情報を参照・連携などといった，情報の再利用については適さない仕組みである．例えば，東京にある博物館A の資料 a は奈良県にある博物館 B が持つ資料 b の一部であるとき，横断検索の結果はそれぞれが独立した情報として表示される．このとき，2 つの情報は同一資料であることから，資料情報の統合を考える．しかし，現在の博物館システムでは，異なる情報源を統合して 1 つの情報として扱う仕組みを持たないため，同一資料の情報が独立して複数箇所に存在する問題がある．LODAC プロジェクトでは，このような情報が国内に多数存在すると仮定し，ミュージアムが保有する資料情報や，それに関連する異なる情報源を Web 上でリンクさせて 1 つの情報として統合・共有することを提案する．しかし，これを実現するためには情報のオープン化が必要不可欠であり，現在の Web における情報流通が「利用・創造→公開→発信」のサイクルであるのに対し，柔軟に情報を扱うためには「利用・創造→公開→収集・蓄積→共有→利用・創造･･･」の循環型へと移行する必要がある[3].

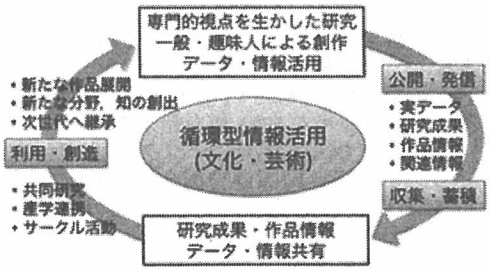

図 1.循環型情報活用

循環型 Web で情報を扱う場合，「共有する」という観点から，公開される情報には一元的な制限を受けることなく自由に共有・利用されなくてはならない．近い将来，ミュージアムや研究機関が保有する芸術・文化的価値のある情報を自由に共有・利用できる仕組みを設けることで，より柔軟に情報が流通し，それらが研究や創造に再利用されることで新たな知見を生み出す，文化・芸術分野における情報循環サイクルの形成が可能と考える[図 1].

3. Linked Data と Linked Open Data

3.1 セマンティック Web

セマンティック Web とは，Web の創始者であるティム・バーナーズ=リー[1]によって提唱された現在の Web を拡張した次世代 Web である．現在の Web は，主に HTML によって記述されるが，この構造は人間が内容を理解できるように作られており，コンピュータには HTML の構造やそこに記述された情報にどのような意味があるのか機械的に処理出来ない問題がある．

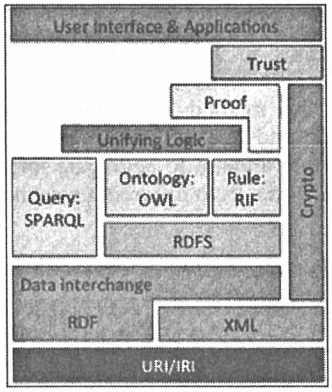

図 2. Semantic Web layer cake (2007)

この問題を解決するために，セマンティック Web では，Web 上のあらゆる情報に対して意味づけ(メタデータ付与)をすることで，人間とコンピュータの双方が効率よく情報を扱える仕組みを提供している．図 2 はセマンティック Web を表す階層図である．このうち，中間層に位置する RDFS や Ontology，Rule では，セマンティック Web を構成する言語として OWL や RDFS，検索クエリとして SPARQL を提供している．

セマンティック Web の世界では，階層図の下位層から上位層にかけて様々な研究が進められているが，普及をすすめるにあたって問題も数多く抱えている．とりわけ，概念レベルのオントロジー(ドメインオントロジー)構築がボトルネックになっている．ドメインオントロジーとは，特定分野・領域における概念を記述したものであり，ドメインオントロジーを構築する際は，その領域の専門家を交えて議論を積み重ねた上で構築される．例えば博物館領域の場合，コレ

クション記述規則から管理・運営まで幅広く概念レベルで定義する必要がある．オントロジー構築は多くのコストを要することから，セマンティック Web が広く普及しない要因の 1 つであった．

3.2 Linked Data と Linked Open Data

セマンティック Web でオントロジーを利用することは，共通の概念で情報を共有できることから，広く普及できれば恩恵も大きいが，構築までに多大な労力を要することから普及が遅れる懸念があった．一方，Linked Data とは，オントロジーの構築については別で検討し，まずは実体として存在するデータの共有と利用を広く進めてく点に特徴がある．Linked Data でデータを公開するにあたって，ティム・バーナーズ・リーは次のような原則を定めている[4].

1.あらゆる事象に対して URI を付与すること.
2.HTTP 経由で URI を参照できること.
3.URI を参照した際は情報が閲覧できること．また，データは RDF や SPRQL 等標準化技術で利用出来ること.
4.他の URI へのリンクを含めること.

上述する原則に則ってつくられるデータは，あらゆる事象に URI が付けられ，それぞれの事象が共有されると共に，リンクとして相互参照可能な仕組みを持つ．このような性質をもつ情報のことを Linked Data と言い，Linked Data による情報共有や利用を推進する活動を Linked Open Data(LOD)という．LOD によってもたらされる恩恵は様々に考えられるが，必要な情報を必要なだけ参照できることが Linked Data の特徴であると言えよう．つまり，従来の情報公開方法は，HTML や PDF 等であることから，例えば PDF 内に記述されたある部分の情報をデータとして利用したい場合には，データを抜き出すための二次加工をする必要があった．これに対し，LOD ではすべての事象に対して URI の付与，および RDF 化形式で情報記述・公開することから，PDF のようにデータを抽出するための加工処理等が不要である点が大きい．欧米では LOD として公開された大規模データ群を LOD Cloud 図として公表している[5]．LODAC では，Linked Data の特徴に注目し，国内に個別に分散する芸術・文化情報を収集・共有，さらにはミュージアムの類似文化施設である図書館や文書館，その他芸術や文化の担い手等，異なる流れの情報を統合することで，新たな発見や知の獲得の可能性を探る．本プロジェクトにとって Linked Open Data とは，それらを実現するための手段であり，活動である．

4．関連研究

4.1 英国 BBC

英国放送協会(BBC)では早くから Web サイトに Linked Data の仕組みを採用している．例えば，音楽関係のページでは音楽に関する独自オントロジーの構築やアーティスト情報の RDF 化がされており，これらを Linked Data としてテレビ・ラジオ番組等のページにコンテンツとして利用している．さらに，RDF はリソースとして BBC サイト内の他の RDF を参照している他，次のような記述を用いて外部で公開されているリソースを利用している点にも着目できる．

“<mo:wikipedia rdf:resource="http://en.wikipedia.org/wiki/Bon_Jovi"/>”.

この例は，米国のアーティストである Bon Jovi の概要を Wikipedia から自動的に参照してサイトに掲載していることがわかる．BBC では，その他ニュースを含めた様々な情報に対して URI を付与し，LOD として情報を利用・公開している．

4.2 Europeana

Europeana は，2005 年 9 月に欧州委員会によって計画された「i2010 欧州情報社会戦略」プロジェクトの 1 つである．図書館領域から開始したこのプロジェクトは，現在はオランダ国立図書館のチームをコアに，European Digital library network(EDLnet)と称する企業や研究機関を含めた活動を展開し，EU の図書館，博物館，文書館等のデジタルデータを収集し，600 万件を超える資料情報を公開している[6]．Europeana では EU 各国の文化施設から提供されたデータを利用していることから，これらを横断検索可能にするために，資料に対して共通に適用できる語彙として Dublin Core と 12 種類の独自語彙を使用している．膨大な情報を扱う Europeana では，Linked Data 化も進められており，SWI-Prolog をベースとするセマンティック Web サーバ “ClioPatria”[7]を用いている．セマンティック Web の対応では，既存語彙として米国 Getty 財団[2]が作成する人名典拠情報の ULAN や美術・建築シソーラスの AAT，各国の博物館系語彙等を RDF 化し，約 1600 万トリプルが実験的に格納されている．様々な試みが行われている Europeana であるが，膨大なデータと数十からなる参加国数ゆえに，検索時の言語表記問題や電子情報と実体としての資料取扱い関係，隣接領域の博物館・文書館の語彙やオントロジー利用等の課題がある．

4.3 国立国会図書館(NDLSH)

一方，国内では 2010 年 6 月に国立国会図書館が国会図書館件名標目表(NDLSH)を Linked Data 化し，SPARQL Endpoint とともに公開を開始した[8]．Linked Data として公開された NDLSH は，2006 年に公開されたテキストデータ版 NDLSH に SKOS(Simple Knowledge Organization System)を用いて表現している．SKOS は，件名標目やシソーラスの表現に適した語彙であることから，図書館系の情報を扱うには相性がよい．しかし，

---

[2] http://www.getty.edu/

SKOS のみですべての NDLSH 項目と SKOS の語彙への対応付けはできないため，Dublin Core(DCMI Metadata Terms)と独自に定義した語彙を DCNDL として使用している．日本語特有の問題である「読み」については，参考文献[9]の時点では検討課題としていたが，2010 年 6 月に公開されたスキーマの内容には“ndl:transcription”として読みに関する独自語彙を定義し，件名標目の下位に位置する構造として表現している．

5．LODAC Museum の構築

5.1 LODAC Museum の目的・目標

LODAC Museum の構築は，セマンティック Web と Linked Data の技術を用いて芸術・文化おける次の課題解決と検証を試みる．

1. 国内のミュージアム資料情報は，各館が独自に管理していることから，全国各地に分散している状態にある．すなわち，情報の統合化が行われていないため，情報検索したときに，断片化した情報しか入手できない問題がある．これらを Linked Data 化することにより，複数情報源からの情報統合化が可能であると考え，プロトタイプシステムを構築して検証する．
2. ミュージアム資料情報だけでなく，図書館や専門的なシソーラス等，外部の情報を組み合わせることで，通常の Web 検索や単一の DB 検索では得られない発見や知見の獲得が可能であると仮定し，プロトタイプシステムを構築して検証する．
3. 多種多様な情報を LOD 化することで，特定分野のみの情報利用にするのではなく，多岐にわたる分野で共有・利用可能にすることで，国内の情報流通における流動性と柔軟性を向上させる．そして，最終的には情報発信主体に還元できる仕組みの構築を目指す．

本稿では，プロトタイプシステムを用いて 1，2 を明らかにし，3 は芸術・文化における LOD の可能性を考察する．

5.2 LODAC Museum の情報源

プロトタイプシステムでは，実在するミュージアムが保有する資料情報とそれらの関連情報，性質が異なるその他関連情報を用いる(表 1)．表 1 の (1)〜(14)は，ミュージアム資料の情報源として，実在するミュージアムの Web サイト上に DB，或いはコレクション情報を公開している都道府県の国公立館を選んだ．ただし，今回は資料自体が非常に複雑な性質を持つ生物，考古系等は対象から外し，美術品等文化財に限定した．関連資料のうち，(15)日本美術シソーラス DB 絵画編[3]は，筑波大学日本美術シソーラスデータベース作成委員会によって構築されたものである[10]．日本美術に関する多数の項目が構造化されており，特に作品，作者，主題，時代，名号，所蔵，地域に関する情報が個別に格納されていることから，これらを Linked Data 化することで(1)〜(14)で取得した情報と統合出来る可能性が高いと考えた．今回，プロトタイプシステムに利用するに当たり，福田氏，五十殿氏の了承を得てソースデータを拝借させて頂いている．(16)文化遺産オンラインは，館に関する施設情報(開館時間や連絡先等)があることから，これを Linked Data 化した．作品情報については，文化遺産オンラインの性質上，情報源から除外した．(17)国指定文化財データベースは，美術品カテゴリのうち，国宝・重要文化財のデータを利用した．都道府県別に分類されていることや，情報によっては解説文や独自の情報項目があることから，ミュージアム資料情報とこの情報を統合することで，より多くの資料情報を提示できる可能性があると考えた．

表 1. LODAC Museum の情報源

A.ミュージアム資料
(1) 東京国立近代美術館
(2) 国立西洋美術館
(3) 京都国立近代美術館
(4) 国立国際美術館
(5) 京都国立博物館
(6) 奈良国立博物館
(7) 福島県立美術館
(8) 栃木県立美術館
(9) 秋田県立近代美術館
(10)岩手県立美術館
(11)徳島県立近代美術館
(12)山梨県立美術館
(13)東京都現代美術館
(14)香川県立東山魁夷せとうち美術館
B.関係資料
(15)日本美術シソーラス DB
(16)文化遺産オンライン
(17)国指定文化財データベース
C.その他関連情報
(18)国土交通省国土計画局 GIS
(19)日本語版 DBpedia

情報源 A,B の (15)を除くデータは Web サイト上の情報であることから，HTML や DB 検索結果から必要なデータのみをプログラミングによる自動処理で取り出すスクレイピングと呼ばれる方法でデータを収集した．

C のその他関連情報では，美術品や文化財に直接関連しない，性質の異なる情報を使用した．(18)国土交通省国土計画局 GIS からは，一般公開されている国土数値情報のうち，全国博物館総覧(2005 年度版)の情報が含まれている公共施設データを LOD 化した．これにより，ミュージアムの位置情報と Google Maps を連携することで館の所在地を地図として提供可能になった．但し，ミュージアム以外の公共施設情報も多数

[3] http://www.tulips.tsukuba.ac.jp/jart/mokuji/index.html

含まれ，情報検索の際に地名や施設名がノイズデータとして抽出されることから，現在はデータベース上から取り除いている．(19)日本語版DBpediaは，日本語Wikipediaに情報を参照するためのLinked DataのハブとしてLODACプロジェクトが試験的に運用しているLODサービスである．LODAC Museumでは資料情報に対応する関連情報がWikipedia上にあった場合，これらの情報も参照出来るようにする．

5.3 Linked Data化とデータ統合

プロトタイプシステム構築に際してLODACプロジェクトでは，ミュージアム，日本美術シソーラスDB絵画編，国指定文化財データベース等，複数情報源から収集した情報統合するため，データのLinked Data化と統合方針を決めておく必要がある．本節ではデータ統合化を進めるにあたり各問題とその対応，統合ポリシーについて述べる．

5.3.1 実体と典拠の記述方針

ミュージアムの隣接領域である図書館では，実体として存在する書誌，並びに情報として存在する典拠に関する取扱い手法が整理されている．例えば，国立国会図書館ではそれぞれを書誌情報と典拠情報を分離して扱い，各関係を記述出来るように設計されている．しかし，ミュージアムでは実体としての資料はあるが，それが典拠として何を示すのかは必ずしも確立されているとは言えない．すなわち，実体資料に関する情報と典拠情報の取り扱いに関して確立された方法論がないため，本プロジェクトでは資料情報として得られた情報はまとめて1つの情報として扱うことにした．

5.3.2 データのLinked Data化

LODにおいて，公開されているデータをLinked Dataとして使用する場合，通常はデータを保有する主体がLODとして公開しているものを使う．しかし，今回の場合はLODACプロジェクトが既存ミュージアムのWebサイトから収集したデータをLinked Dataとして新規に構築することから，サイト上のデータの中身に対する変更や根拠，権限を持たない第三者の立場で扱う．そのため，LODAC Museumで扱う際の基本方針を設定した．具体的には，作品や資料説明等が記述されている箇所はそのままデータとして使用し，データに対する付帯情報(メタデータ)に対してのみ加工する．したがって，資料の解説文や名称等に明らかな不具合があった場合でも，LODAC Museumでは各ミュージアムから収集したデータをそのままの状態で扱う．

5.3.3 ソースデータの更新

Linked Dataとして処理されたデータはLODACプロジェクトのサーバに格納される．しかし，取得したソースデータに変更或いは更新があった場合に，どのようにしてLinked Data化したデータに反映させるのか，ソースデータ更新の問題がある．資料情報の一部が追記等，その情報に対する一意性が失われていない更新であれば，再スクレイピングをすれば良いが，一意の識別子であるURL自体が変更された場合は，Linked Data化したデータと元データが同じ情報であると判断する事が難しくなる．このような問題を解決するために，LODACプロジェクトでは一意の識別子であるURLの他，資料情報にミュージアム固有の資料管理IDや作者名などユニークと判断できる複数の情報を用いた文字列マッチなど対処している．ただし，いずれの場合も確実性が保証された手法ではないため，データの更新問題はLODを扱う分野全体の課題になっている．

5.3.4 REFリソースとIDリソース

LODAC Museumでは，REFリソースとIDリソースという2種類のリソースを管理する．本プロジェクトでは，各情報源から収集した情報に一意の識別子を付与したものをREFリソースとして定義した(図3)．一方のIDリソースに対しても一意の識別子を付与し，外部から参照されるためのポインタとしての役割とREFリソースを間接的に参照するための関係を記述するリソースとして定義した(図4)．

```
<http://lod.ac/ref/18731>
  <http://lod.ac/ns/lodac#exhibitionHistory> "個展（東京、東京藝廊 1969)" ;
  <http://lod.ac/ns/lodac#genre> "Prints:", "版画 :" ;
  <http://purl.org/NET/cidoc-crm/core#P62i_is_depicted_by> "右下に署名（絹）" ;
  <http://purl.org/dc/elements/1.1/creator> "横尾忠則 " ;
  <http://purl.org/dc/terms/created> "1969", "昭和 44" ;
  <http://purl.org/dc/terms/extent> "90.0×90.0", "on paper, acrylic films and acrylic sheet90.0×90.0" ;
  <http://purl.org/dc/terms/identifier> "P01847" ;
  <http://purl.org/dc/terms/isReferencedBy> <http://lod.ac/id/18731> ;
  <http://purl.org/dc/terms/medium> "silkscreen", "シルクスクリーン・紙、アクリルフィルム、アクリル板・1" ;
  <http://purl.org/dc/terms/provenance> "平成 17 年度購入 P01847" ;
  <http://purl.org/dc/terms/source> <http://search.artmuseums.go.jp> ;
  <http://purl.org/dc/terms/title> "Landscape No.1 Girl", "風景 No.1 女の子" ;
  a <http://lod.ac/ns/lodac#WorkReference> ;
  <http://www.w3.org/2004/02/skos/core#prefLabel> "Landscape No.1 Girl", "風景 No.1 女の子" .
```

図3. REFリソース(ttl形式)

```
<http://lod.ac/id/18731>
  <http://purl.org/NET/cidoc-crm/core#P55_has_current_location> <http://lod.ac/id/912> ;
  <http://purl.org/dc/terms/creator> <http://lod.ac/id/874> ;
  <http://purl.org/dc/terms/references> <http://lod.ac/authority/18731> ;
  <http://purl.org/dc/terms/title> "Landscape No.1 Girl", "風景 No.1 女の子" ;
  a <http://lod.ac/ns/lodac#Work> ;
  <http://www.w3.org/2004/02/skos/core#prefLabel> "Landscape No.1 Girl", "風景 No.1 女の子" .
```

図4.IDリソース(ttl形式)

REFリソースに記述される内容は情報源からの情報を忠実に記述するのみにし，責任・権限については情報源に委ねることにした．一方，IDリソースは，情報源から作成された多数の作品や作者などのリソース情報を統合して記述することから，IDリソースの統合・編集内容に関する部分についてはLODACプロジェクトが内容の責任を持つ．

5.3.5 リソースの恣意性

各リソースをLinked Dataとして使用する場合，どのデータを基準，基点として関連づけをするか決める必要がある．LODAC Museumでは，基準となるデータにはなるべく情報量が多く格納されている日本美術シソーラスDBを用いた．これには他の情報源に含まれる項目がほぼ網羅されていることに加え，各項目がある程度構造

化されていたことから，他の情報源との関係を発見する際の基準になりえると考えた．データの基点については，作者名・作品名・所蔵館それぞれを ID リソースとして作成し，これらには必要最小限の情報が記述される．一方，これらに関連する情報は別途 REF リソースとして作成される．例えば，ある作品に関する ID リソースは，2 個以上の dc:references で接続される REF リソースの和として表現される(図 5).

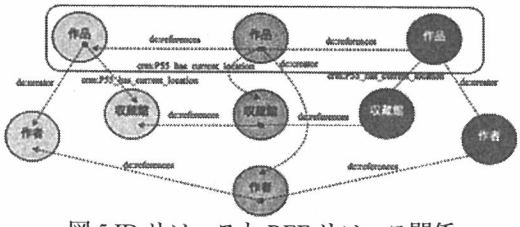

図 5.ID リソースと REF リソース関係

5.3.6 同一内容リソースの同定問題

ある ID リソースを LOD として公開したとき，全く同じ内容の ID リソースが複数存在すると，その情報を使用したいユーザにとってはどちらの情報を使えばよいのか，判断が困難になる場合がある．このように，複数の ID リソースが同一の内容である判断できたときは，次のような手順で対応する．

1. ID100 と ID700 が同一と判明したとき，ID100 に記述されている REF700 の内容を削除する．
2. ID700 にリンクしている他の ID の記述を ID100 にリンク先を変更する．
3. ID100 側には ID700 としてアクセスされるという情報を持っておく．
4. ID700 にアクセスがあった場合は ID100 へリダイレクトさせる．
5. 結果，実体は 1 つのリソースになるが，2 つの ID からアクセス可能になる．

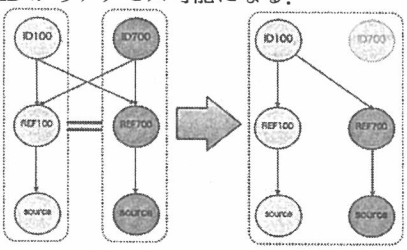

図 6.リソースの再リンク

同様の問題は DBpedia にもあるが，DBpedia の場合はリソースを削除せず，重複したリソースに対して dbpprop:redirect のプロパティを使用し，参照先へリダイレクトさせて対応している．

5.4 語彙のマッピング

5.4.1 メタデータの扱い

この 10 数年の間，ミュージアムのメタデータや関係語彙について，スキーマ共通化の必要性等，膨大時間を費やして議論されてきたが，具体的な結論には至っていない．その間，国際標準規格化された CIDOC CRM や東京国立博物館モデル等が登場したが，国内においてはいずれもスタンダードなものとして普及していない．とくに，CIDOC CRM はセマンティック Web を意識し，OWL や RDF スキーマ，RDF データによる情報の表現を採用しているため，定義すべき項目，項目間(関係記述規則)の制約がある．海外では研究事例として幾つか報告されているが[11][12]，国内では複雑さと扱いにくさが影響し，利用に対する関心は薄れつつある．項目間に厳格な記述規則を必要とするオントロジーアプローチに対し，Linked Data によるアプローチは，基礎的な情報に対しては一般的な語彙を用いてデータ構造を記述しつつ，URI で表現されるリソース間に新たな関係性を自由に定義・記述できる余地を与えることで，情報に柔軟性と流動性を与えられるメリットがある．本研究では，芸術・文化情報のオープン化を目指すことと，国内ミュージアムの性質を鑑み，ある程度の自由度があり，リソース同士の関係性を記述できる，Linked Data を採用した．具体的には，収集した資料情報に記述されている作者名や法量，収蔵館等の項目とそれに記述される情報を参考にして LODAC プロジェクト側が新たに語彙をマッピングしていく．このとき，作品や資料に対して詳細な語彙を記述するのではなく，共通性のある項目に対して語彙を割り当てていく．LODAC Museum で使用している語彙は DC, DC Terms, SKOS, iCal, FOAF, NDLSH, RDA, CIDOC CRM, OWL, RDFS から必要な項目のみを参照した．現在の語彙数は作品(45/16)，人物(23/12)，施設(13/10)，書誌(12/1)の 4 種に分けられ，計 93 個使用している(カッコ内は合計語彙数/独自定義語彙数).

5.4.2 作者名と日本語読みの問題

芸術関係の情報を Linked Data 化する場合，特に名前に関する取扱いは重要になる．例えば，一人の芸術家の名前に関係する情報には，一般的な作者名の他に複数の作者名称や名号等，作者名義の情報を数多く持つ．LODAC Museum ではこれに対応するため，作品情報に対して複数の作者情報を別の語彙として記述している(表 2).

表 2.人物名に関する語彙

Person Reference	property
作者名(一般)	foaf:name / skos:prefLabel
作者名読み	foaf:name @ja-hrkt / skos:altLabel
名号	foaf:nick
名号読み	foaf:nick @ja-hrkt
作者英名	foaf:name @en / skos:altLabel

日本語特有の読みの問題については，表 2 にあるとおり，言語タグを用いて表現している．言語タグは作者名の他，作品名にも同様の処理を与え，同一語彙に複数の値を定義して表現している(表 3)．これは，同一人物に対して複数の作者名があった場合，読みも複数定義されるこ

とから，読みと作者名を対応づけるために構造化している．

表 3. 人物名と言語タグ表現

```
foaf:nick [
   a lodac:Name;
   lodac:label "武田"@ja;
   lodac:label "たけだ"@ja-hrkt;
   lodac:label "Takeda"@en;
   ].
```

6．プロトタイプシステムの検証

6.1 格納資料情報

以上のような情報統合に関する基本方針を踏まえ，本節では Linked Data 化された資料状況を説明する．表 4 は，表 1 の情報源から作成した作品，所蔵館情報，人物，グループに関するデータ数である．

表 4.情報源と使用データ数

情報源	情報種別	データ数
国立美術館(西美を除く 3 館)	作品	25180
国立西洋美術館	作品	4373
京都国立博物館	作品	5819
奈良国立博物館	作品	431
福島県立美術館	作品	20
栃木県立美術館	作品	32
秋田県立近代美術館	作品	22
岩手県立美術館	作品	1558
徳島県立近代美術館	作品	18482
山梨県立美術館	作品	262
東京都現代美術館	作品	5416
香川県立東山魁夷せとうち美術館	作品	266
日本美術シソーラス DB	作品	3800
日本美術シソーラス DB	人物	1332
日本美術シソーラス DB	グループ	289
日本美術シソーラス DB	所蔵館情報	648
文化遺産オンライン	所蔵館情報	915
国指定文化財データベース	作品	10115
合計		103096

これらのデータから生成した総リソース数は 529,449 件になり．Linked Data として作成された総トリプル数は 1,915,586 になった(トリプル数にはブランクノードや情報種別が付けられていないノードも含む)．これらのデータは RDF ストアである 4Store 上に格納されている．

6.2 同一内容リソースの統合

同一内容を示すリソースの同定処理について検証する．まず，リソースを同一のものとして認識するために，単純にタイトルの文字列マッチによる抽出を行った．対象のデータは A.日本美術シソーラス DB 所蔵館情報(648 件)と B.文化遺産オンライン(915 件)の所蔵館情報のタイトルをキーにして文字列の完全マッチを行い，重複するリソースがどの程度抽出されるか検証を行った．その結果，77 件の所蔵館情報が抽出された．例えば，京都文化博物館の場合，異なる ID リソースとして A=/id/3341，B=/id/8057 が抽出された．各 ID に対応する REF リソース A には，作品情報，住所，電話番号のデータが格納され，対する B には休館日や開館時間，アクセス方法が記述されている．双方のリソースが指す実体は京都文化博物館で，同じものを示しているが，A と B には異なる情報が記述されている事がわかる．このように，それぞれが異なる情報を持ち，同じ実体を指すリソースを統合することで 1 つの情報源からは得られない情報を抽出することが可能と判明した．次に，タイトルに含まれる一部の文字列によるリソースマッチをした．例えば，昭和女子大学光葉博物館ならば昭和女子大，武蔵野市立吉祥寺美術館ならば吉祥寺として抽出した．結果，前者は A=/id/3534，B=/id/7672 として，双方のリソースが示す情報を同定することができたが，後者の例では A=/id/3544 の吉祥寺という寺のリソースを抽出し，マッチすることが出来なかった．この他，一般的に使われている国分寺や地名に関する部分マッチはいずれも失敗していることから，汎用性高い名称については別の手段でリソースの同一化・統合を検討する必要がある．

6.3 Linked Data による情報統合例

表 4 のデータを Linked Data したことで，複数館にある作品情報が統合化された例を示す．次の例は，作者名をキーに日本画家である「下村観山」の ID と REF リソース表示させた例である．各ページ上段には作者名が表示され，ID リソースには lodac:creates として 14 件の下村観山の作品と dc:references として 1 件の作者情報にリンクが作成されている(図 7 左)．

図 7. 作者 ID と REF リソース

dc:reference のリンク先は REF リソースになる(図 7 右)．この REF リソースには，日本美術シソーラス DB を情報源とする作者情報が表示され，下村観山に関する作品 2 件がリンクとして記述されている．つまり，基準となる情報源には 2 件の作品情報しか情報はないが，データを Linked Data 化し，統合化することで 12 件の追加の作品情報を得られたことになる．今回の下村観山のデータについては，徳島県立美術館，日本美術シソーラス DB(作品)，日本美術シソー

ラス DB(書誌情報)，国指定文化財データベース，国立美術館総合目録データベース，福井県立美術館の計 6 件の異なる情報源からの情報統合を確認することが出来た．

7. 考察と今後の展開

本論文は，未統合かつ複雑な構造・語彙を使用する芸術・文化情報を LOD として扱うために，LODAC Museum のプロトタイプシステムを構築した．これにより，分散するミュージアム資料情報と異なる種類の情報を統合することで次のような結果を得られることがわかった．1.ミュージアム資料情報を Linked Data として共通の形式で記述することで，複数の情報源に横断検索できること．2.複数の異なる情報源から資料情報を統合することで，単体では得られない情報が獲得できること．3.特定の情報項目を統合することで，元のデータベースにはない情報が統合されて表示されたこと．

以上のことから，より多くの情報を用いることで，さらなる発見や知見の獲得が可能と考える．今回のプロトタイプシステムでは，構築上の問題も多数表面化したことから，これらをついて解決していく必要がある．1.情報統合の失敗では，失敗したリソースに誤りの記述があったため，うまく統合できない場合があった．このような誤った記述を発見した場合は，情報源に情報提供し，校正されることで精度の高い情報公開が可能になる．2.作者名の別名問題では，作者が複数名前を持つ場合，参照情報としてアーティストシソーラス等が必要になることから，これの情報源を検討する必要がある．3.作品名のゆらぎ問題では，ある DB には正式名称で記述され，もう一方には省略された名称で記述されるために作品の同一化が難しい問題がある．このような場合の同一化手段を検討する必要がある．4.リソースの更新問題では，スクレイピングによる元データとの同期方法の検討が必要である．5.外部情報へのリンクでは，多様な情報とリンクするために，国会図書館の書誌情報や Europeana 等，LOD で公開している情報へのリンクを視野に入れたデータ構造を検討する．6.LOD へ参加問題では，保有するデータを CSV でインポートできる等，容易に Linked Data として扱える仕組みを検討する必要がある．以上のような問題解決をすすめるとともに，LOD の活用について考えていく必要がある．例えば，芸術・文化情報が LOD として公開された場合，GIS 情報と日本各地の資料情報を組み合わせることで，ある資料を追跡する日本縦断ミュージアムツアーや，仮想的に資料を体験するバーチャルミュージアム構築等が可能になる．また，ユーザ参加型の例では，これまでは権利関係上所蔵館が資料に対してキャプションし，公開していたものを，一般ユーザによるキャプションやコメント等の作品解説が可能になることや，公開資料をベースにユーザが創作した作品とそれらの派生情報が LOD でリンクすることなど，インターネットを利用したサービスが考えられる．その他，教育分野におけるワークショップへの利用やミュージアム・リテラシー向上[13]など，博物館教育への応用等が考えられる．このように，LOD として情報を公開・共有・利用することで，その可能性は限りなく広がり，芸術・文化だけでなく他の分野への応用など，幅広く社会に対して貢献できるものと考える．

8. 謝辞

LODAC Museum のプロトタイプシステム構築にあたり，日本美術シソーラス DB データ利用にご協力下さいました，跡見女子大学の福田博同氏，並びに筑波大学の五十殿利治氏に謝意を表する．また，本論文執筆にはプロジェクトミーティングに参加頂き，LOD に関する積極的な議論展開を頂いた産総研の濱崎雅弘氏をはじめとするプロジェクトメンバーに感謝致します．

参考文献

[1] http://lod.ac/
[2] http://webarchives.tnm.jp/docs/informatics/smmoi
[3] 武田英明:日本における Linked Data の現状と普及に向けた課題,情報処理，Vol.52，No.2，2011.(予定) (http://tinyurl.com/lod-japan 参照)
[4] http://www.w3.org/DesignIssues/LinkedData.html
[5] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[6] http://www.europeana.eu/portal/
[7] http://e-culture.multimedian.nl/
[8] http://id.ndl.go.jp/auth/ndlsh/
[9] 永森光晴,杉本重雄:国会図書館件名標目標(NDLSH)の SKOS 化とそのグラフィカルブラウザの作成，情報処理学会研究報告．情報学基礎研究会報告,Vol.118, pp.11-19, 2006.
[10] 福田 博同, 五十殿 利治:美術シソーラスデータベース形成の諸問題,情報管理，Vol. 40, No. 9, pp.790-809, 1997.
[11] Chryssoula Bekiari, Leda Charami, Martin Doerr, Christos Georgis, Athina Kritsotaki: DOCUMENTING CULTURAL HERITAGE IN SMALLMUSEUMS, 2008 Annual Conference of CIDOC, 2008.
[12] Ceri Binding, Keith May, Douglas Tudhope: Semantic Interoperability in Archaeological Datasets:Data Mapping and Extraction via the CIDOC CRM,research and Advanced Technology for Digital Libraries, pp.280-290, 2008.
[13] 嘉村 哲郎，加藤 舞，北岡 タマ子: ミュージアム・リテラシーに関するワークショップ実践報告．日本ミュージアム・マネージメント学会会報，Vol.14,No.2, pp.23-25, 2009.