平成 29 年度「ジャパンサーチ（仮称）」
利活用フォーマット検討成果物

平　成　３０　年　３　月

国立国会図書館
National Diet Library, Japan

国立国会図書館
平成 30 年 3 月

国の分野横断統合ポータル「ジャパンサーチ（仮称)」における利活用のためのメタデータフォーマットの検討結果について

国立国会図書館は平成 29 年度に国の分野横断統合ポータル「ジャパンサーチ（仮称)」における利活用のためのメタデータフォーマットの検討を行い、次の資料を作成しました。

- 概念モデルと語彙検討報告書
- 利活用メタデータフォーマット仕様案
- 利活用メタデータフォーマット仕様案を用いたデータ例

検討の支援作業を、ゼノン・リミテッド・パートナーズに委託しました。

検討にあたっては分野横断統合ポータルの連携候補等機関からメタデータの提供を受け、上記資料作成の参考としました。

また、当館職員、外部有識者、及び受託者からなる検討会を開催しました。そこでの議論から得られた知見は、上記資料に反映されています。

ご協力いただいたみなさまに御礼申し上げます。

「概念モデルと語彙検討報告書」執筆者

- 神崎 正英 （ゼノン・リミテッド・パートナーズ)

検討会開催記録

- 第 1 回　　平成 29 年 9 月 7 日（木）
- 第 2 回　　平成 29 年 11 月 16 日（木）
- 第 3 回　　平成 30 年 2 月 9 日（金）

検討会に参加いただいた外部有識者（五十音順、敬称略）

- 嘉村 哲郎 （東京藝術大学芸術情報センター）
- 寺澤 正直 （内閣府大臣官房公文書管理課）
- 村田 良二 （東京国立博物館学芸企画部）

※ 所属は平成 30 年 3 月現在

ジャパンサーチ利活用フォーマット
概念モデルと語彙検討報告書

Xenon Limited Partners
2018 年 3 月 26 日

目 次

1　検討作業とモデル設計の考え方

分野横断統合プラットフォームのの実現へ向けて、メタデータ要件に関する資料をふまえ、先進事例の調査などに基づく概念モデルを作成し、有識者を交えた検討会を経て、メタデータフォーマット仕様案を作成した。ここではその概念モデル設計について報告する。

1.1　統合プラットフォームの目的確認

「知的財産推進計画 2017」に従えば、デジタルアーカイブを活用するための「つなぎ役」として、分野横断の検索機能のほか、各アーカイブ機関やつなぎ役が整理したメタデータを集約・共有化し、活用者による様々な形での利活用に資する統合ポータルを構築する。

- デジタルアーカイブ化によって分野・地域を超えた知を集約し、学術研究・教育・防災・ビジネスへの利活用が期待できることに加え、海外発信機能を付加・強化することにより、インバウンドの促進や海外における日本研究の活性化にもつながりうる。
- デジタルアーカイブが国内外において日常的に活用され、新たなコンテンツやイノベーションを生み出すための基盤となる社会を実現するため、今後、各アーカイブ機関を結ぶ「つなぎ役」と国等が一体となった取組を加速することが必要。
- 「つなぎ役」は、分野内のメタデータ項目を標準化するために分野ごとに標準メタデータ項目を作成していくこと、さらに、その分野において、長期に渡ってデジタルアーカイブ基盤を維持できるよう、アーカイブ機関の技術、法務上の課題等に対応できる人材の育成をサポートしていく役割が求められる。

（知的財産推進計画 2017 より）

ただし統合プラットフォームは、書物の書誌データのように必ずしもデジタルデータではないリソースのメタデータも対象とする。上記のデジタルアーカイブをより広い知的文化財資源（CHO）と読み替え、「各分野の CHO メタデータを集約・共有化し、横断検索をはじめとする様々な活用の促進に資する」ことと考える。

1.2　メタデータの役割と基本的な要件

各機関が保有する資料を、**検索・表示・提供**するための情報。

- つなぎ役（アグリゲータ）との連携が進むと想定されるため、つなぎ役の参加機関の名称等を検索・表示・提供時に出力できるようにする（参加機関のインセンティブ向上）。
- デジタルコンテンツの有無、ライセンス情報の区分等でも絞り込み等ができるようにする。
- 従来の図書館資料で扱ってこなかった各種文化財についても、適切な形で検索・表示・提供できるよう、必要な情報を記録化できることが望ましい。例えば、材質・構造・技法等のモノに関する情報、その由来情報、地理的な情報、いわゆる著者以外の関係者に係る情報、などが考えられる。

2　基本データモデル

提供機関から収集したメタデータはそれぞれ異なる目的のために設計されたもので、その意図を十分生かしながら共通のメタデータにマッピングすることは難しい。そこで統合プラットフォームにおいては、収集した**ソースデータ**を、(1) 利用者タスクに対応して整理した分野横断モデルのデータ（**共通アーカイブ情報**）に変換する；(2) 併せて元のデータを収集元などのメタデータとともに**ソース情報**として保持する；ことで共通性と個別性を両立させる。

- 項目を指定した検索、結果の識別・選択、資料アクセスのための情報提示には、共通アーカイブ情報を利用する。
- 自由キーワード検索はソースデータを対象とし、結果表示に共通アーカイブ情報を利用する[1]。
- 検索結果の個別詳細において、共通アーカイブ情報に加えてソースデータの確認を可能にする[2]。資料へのアクセス提供情報の表示を工夫する

集約データ（共通アーカイブ情報）とソースデータを分離したうえで連動させるために、Europeana 型の EDM モデルを念頭に置きつつ、扱いやすいモデルを設計する。

2.1　共通アーカイブ情報と利用者タスク

基本要件（§1.2）における「検索・表示・提供」機能を、FRBR §2.2 Scope での利用者目的分析にもとづいて表 1 の 4 つのタスクで捉え、それぞれのタスクの観点から共通アーカイブ情報に求められるメタデータ項目（プロパティ）を分類整理する。

表 1: 利用者の観点で統合プラットフォームに求められる機能

機能	内容
発見（検索）	キーワードなどで検索するための情報（領域の知識を前提とせずに）
識別（表示）	示されている対象が何なのか、求めている（既知の）資料かどうか判断できる情報
選択（表示）	示されている対象が求めている（未知の）資料かどうかを判断でき、比較検討が可能な情報
取得（提供）	個別リソースあるいは詳細メタデータにアクセスするための情報

このタスクに基づく具体的な項目は、具体的な語彙とは切り離した概念モデルとして §3 において詳細に検討し、その上で必要な語彙を §4 で検討する。

2.2　ソース情報の分離と EDM

EDM は ORE の集約モデルをベースにいくつかのパターンを提示しているが、Europeana が採用しているのは、集約データにもとづいて Europeana が付与したメタデータと、集約元のメタデータを

[1] 当初案で検討した全文テキストの RDF グラフは用意せず、全文検索はソース情報側に委ねる。SPARQL クエリはグラフ構造を利用した検索が利点であり、全文テキスト検索のニーズは高くないと思われる。

[2] 提供元ページへのリンク、およびソースデータの表示・ダウンロードも提供する（ソースデータはオープンライセンスで公開とする）。ソースデータは、項目名や構造は提供されたままとするが、処理を容易にするため JSON などのフォーマットに変換して保持する。

分離して扱うものである。統合プラットフォームで言えば、前者が共通アーカイブ情報、後者がソース情報に相当する。モデルは 3 つのリソース型を核に構成される。

- 作品（資料）の実オブジェクトを `ProvidedCHO` とする
- 関連リソース（作品のデジタル化表現や情報ページなど）と作品オブジェクトを `Aggregation` として集約（グループ化）する
- 複数の視点（`Aggregation`）からの実オブジェクトに関するメタデータを、それぞれ代理オブジェクト `Proxy` によって記述する

統合プラットフォームに対してこれをストレートに適用した場合、おおむね図 1 のようなグラフを考えることができる。

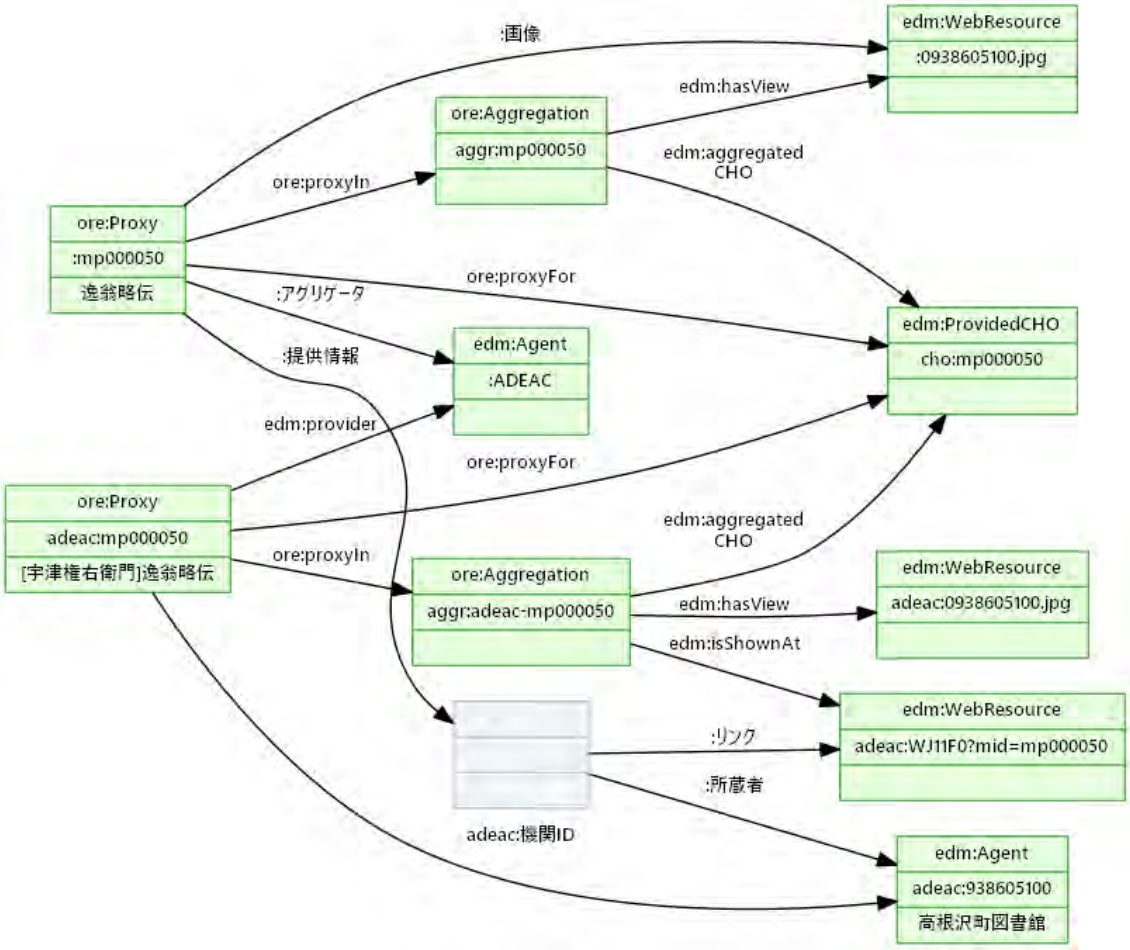

図 1: Europeana の EDM に近い形のモデル。この図はノードの上中下段にリソースの型、URI、ラベルをまとめることでグラフをコンパクトにしている。図右上の画像は、右中ほどの画像からサムネイルを生成したと想定。

2.3 統合プラットフォームのソース分離モデル

Europeana との互換性を重視する場合は図 1 を検討する価値があるが、このモデルは利用者にとってかなり複雑で理解し難い。関連リソースの `Aggregation` は直接収集したもの以外は提供側にあるの

で、統合プラットフォームのメタデータとして扱うのは複雑である。

そこで、EDM のリソース型概念を前面に出さず、収集元の視点によるメタデータを「ソース情報」という枠組みで分離し、構造を簡潔にしたモデルが図 2 である[3]。

統合プラットフォーム側の `Aggregation` に相当するノードは「提供情報」とし、収集するオブジェクトに限らず資料アクセス関連の情報を集約する[4]。ソース情報側は、メタデータとリソース集合を分離していないので ORE/EDM のモデルには直接対応せず[5]、全体で提供元の視点によるメタデータと関連リソースを表す[6]。

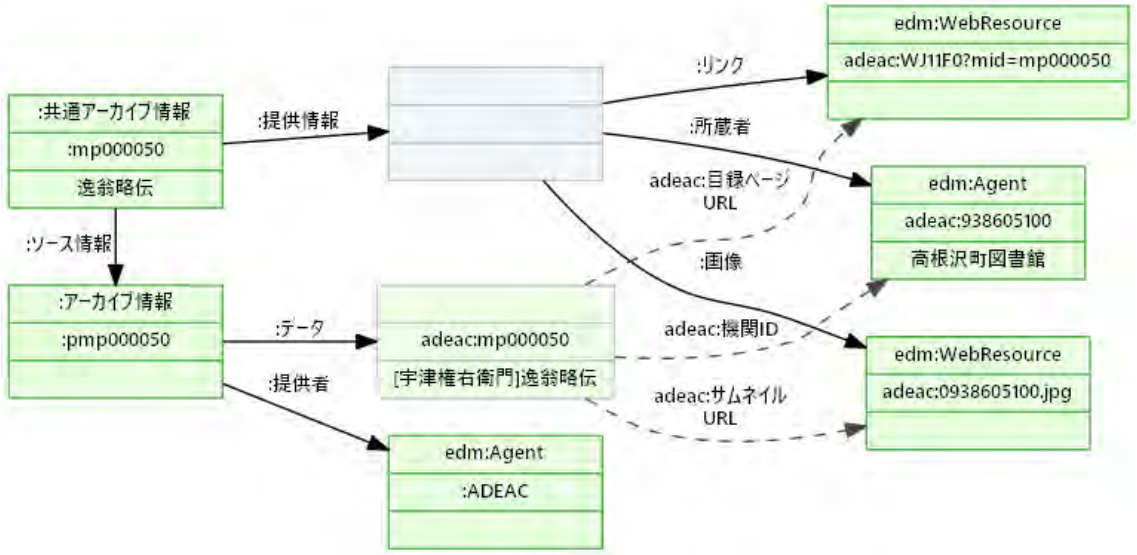

図 2: 共通アーカイブ情報から「ソース情報」プロパティで収集元のプロキシに結びつけるモデル。

2.4 データの収集と分離モデル

データ提供者（アグリゲータ）から統合プラットフォームが収集するデータ（**ソースデータ**）は、原則としてそのままソース情報の中に収める。すなわち収集対象は、データ提供者の項目名、構造による元データであり、そのデータをソース情報内に格納すると同時に、必要な情報を抽出して共通アーカイブ情報に変換する。

2.4.1 NDL サーチとソースデータ

NDL 書誌情報およびすでに収集対象になっている外部ソースは、NDL サーチをアグリゲータと位置付け、そのデータから共通アーカイブ情報を抽出（もしくは変換）する。すなわち、ソース情報（アーカイブ情報リソース）を生成した上で、そこからソースデータとしてリンクする[7]。

[3]当面 ProvidedCHO に相当するリソースは独立させず、共通情報側にそれぞれの内容に対応した基本型（絵画など）を併記している。

[4]リソースの説明も含むので、ORE では Aggregation の説明記述である ResourceMap に近い位置づけになる。

[5]EDM との対比のためには、提供元の視点を「代理する」という意味で「ソース情報リソース」を Proxy、画像などのリンクを含むという意味で「ソースデータ」を Aggregation に当たるものと考えると、対応関係を描くことはできる。

[6]DPLA では、収集側の Aggregation を起点にして収集側メタデータを sourceResource で関連付け、元データは RDF 化されないそのままのデータを originalRecord として関連付けている。

[7]Proxy に相当する「アーカイブ情報リソース」のみ独立したものとして生成する。ソースデータは別途保存するのではなく、NDL サーチのレコードにリンクする。§ 4.3 の Linked Data の場合を参照。

2.4.2　ソースデータの提供とマッピング

ソースデータは提供元の項目名、構造によるデータをそのまま受け取る[8]。中間的な交換フォーマットは定めないが、受取可能なデータ形式（JSON、XML、TSV など）およびプロトコルは、別途規定する。

ソースデータから共通アーカイブ情報への変換は、項目名や変換規則などに基づいてマッピングを自動生成するツール（メタデータアナライザー）で仮マッピングを準備したうえで、個別調整する。仮マッピングは提供元も確認できるようにし、これをもとに協議の上、提供元がソースデータを加工するなどして「共通アーカイブ情報変換用の追加データ項目」を提供可能にする[9]（ソースからの変換については、§5 での実装レベルの検討も参照）。

2.4.3　分離モデルと検索

共通アーカイブ情報は、REST API および SPARQL により、項目を指定した検索を提供する。

ソースデータは（必要に応じて JSON などの形に統一した上で）全文検索の対象とする[10]。この全文検索の結果表示に、共通アーカイブ情報で整理した共通項目を利用する。

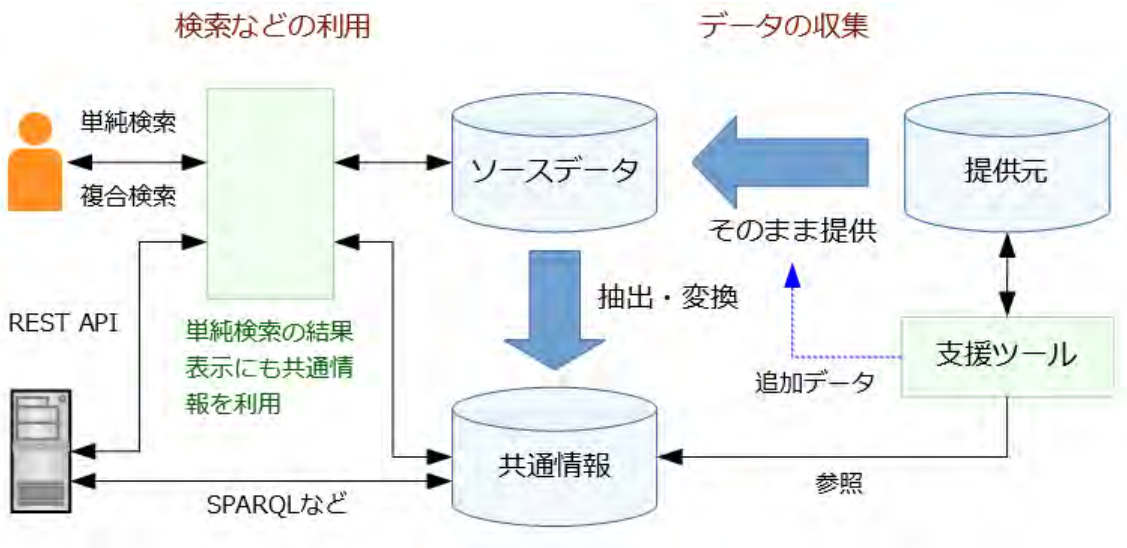

図 3: 提供元からのデータを共通アーカイブ情報に変換して検索などに利用する流れ

資料 URI へのアクセスに対しては、リクエストに応じて HTML あるいは RDF で共通アーカイブ情報のデータを返戻する。

[8] ソースデータは全てオープンライセンスで公開とするため、公開不可データは提供側があらかじめ除去することを原則とする。場合によっては、マッピング時に非公開項目をソースから除去する。

[9] 例えば、和暦年しかないデータに西暦年を加える、「染織 (I)」といった独自分類に「染織」という一般分類項目も追加する、など。追加データ項目は、項目名にアンダーバーを前置するなどして区別の上、ソースデータに一体化する。

[10] 元データ項目名を利用した検索機能を提供してもよい。

3　メタデータのプロパティと構造

3.1　メタデータ項目と優先度

利用者タスクを踏まえ、共通情報を表現するための概念モデルとそのプロパティ項目を検討する。検討会や質問・意見を経て整理したメタデータ項目を表 2 に示す。

表 2: 統合プラットフォームが持つメタデータ項目

基本項目	内容	F	I	S	O
タイプ	資料の基本区分+共通情報を表す型	○	-	○	-
ラベル	一覧などに表示し資料を識別する名前	-	◎	○	-
名称	タイトル、別名、読みなど検索対象とする名前	◎	○	○	-
寄与者／関係	資料に寄与した人／組織。出演者等も含む	△	○	○	-
	寄与者一般としてのショートカット	○	○	○	-
作者	寄与関係 [制作] へのショートカット	◎	◎	○	-
発行者	寄与関係 [出版] 相当へのショートカット	○	○	○	-
場所／関係	場所に関する構造化情報。制作／内容は関係で区別する	△	○	◎	-
	場所に関するショートカット	○	△	○	-
時間／関係	時間に関する構造化情報。制作／内容は関係で区別する	△	◎	◎	-
	時間に関するショートカット	○	○	○	-
	時間のリテラル値（標準プロパティがある場合）	△	○	○	-
主題・区分	主題、分類、区分（キーワード含む）	○	△	○	-
識別子	ISBN など体現形レベルの識別子	○	○	-	△
言語	資料の記述言語	△	◎	◎	-
画像	資料の特徴を確認するための画像	-	◎	◎	-
記述	個別項目に収録できない情報	△	○	◎	-
上位資料	現在の資料がその一部である上位資料	○	○	○	-
提供情報	資料の提供・アクセスに関する情報	-	-	-	-
提供者	資料（に関する情報）の提供者。保管者は別プロパティで定義	-	○	△	◎
リンク	資料の紹介ページやアクセス情報が記載されたページ	-	-	-	◎
オブジェクト	資料のデジタル画像、音声動画など	-	-	○	◎
権利情報	資料利用のライセンスおよび権利	-	-	○	○
個別識別子	提供元が付与する識別子	△	△	-	◎
ソース情報	ソース情報およびその提供者に関する情報	-	-	-	-
提供者	ソース情報の提供者（アグリゲータ）	-	-	-	○
データ	プラットフォームが保持・提供するソースデータ	-	-	-	○
リンク	アグリゲータの情報ページ	-	-	-	○
更新日	収集元データの更新日	-	-	○	○

3.1.1 単純プロパティと構造化プロパティの併用

利用者の**発見**タスクのためには、プロパティ項目は単純であることが望ましい。一方で**識別**や**選択**タスクのためには、プロパティは適切な詳細度が必要である。

たとえば映画作品の情報には監督、脚本、撮影、音楽などさまざまな役割がある。制作に関与した人を調べるためにはこれらを同一の単純プロパティ（寄与者）で検索できると便利である。しかし結果を適切に識別し選択するためには、その人がどのような役割で関与したのかも知る必要がある。

時間・場所についても同様に、制作（creation）だけでなく採集、撮影、主題などさまざまな種類がある一方、利用者はそうした違いを意識せずに時間・場所を検索したい場合が多いと思われる。

そこで、これらの「いつ」「どこで」「だれが」に関しては、まず単純なプロパティで記述し検索などの利用を容易にすると同時に、提供データの詳細を構造化して記述するプロパティを併用する（表2においては単純プロパティを「ショートカット」としている）。また資料の提供情報とソースデータ情報の関係を整理し、それぞれを構造化した。

以下においてこのメタデータ項目を、資料自身の記述に関するもの（§3.2）と、そのアクセスに関するもの（§3.3）に大別して検討する。なおソースデータについての記述である「ソース情報」の内容も、便宜上§3.3の一部として扱う。

3.2 資料を記述するメタデータ各項目

3.2.1 タイプ

資料の基本区分を表す型を付与する。資料の基本区分は、複数種類の資料情報が混在するアーカイブにおいて、情報を大きく区分するために重要である。

また共通アーカイブ情報自身を表す型も（EDMでのProxyのように）与える。ただし、「文化財」でもあり「アーカイブ情報」でもあるというリソースは混乱を招く恐れがある一方、利用する立場からは資料の基本区分も`rdf:type`として確認したい。そこで、資料の基本区分を表す名称を用いつつ、それらを共通アーカイブ情報型のサブクラスとして定義することで、両者をともに`rdf:type`で示すものとする（§4.4.1も参照）。

なおこれらのタイプは統合プラットフォームにおいて定義し、データ提供者やつなぎ役が共通のタイプを記述できるようにする。

3.2.2 ラベルと名称

資料を識別するための名前を**ラベル**とし、汎用的なプロパティで付与する。ソース情報にラベル相当の名前がない場合は、アーカイブで連番などに基づいて生成する。言語タグは与えない。

ラベルは原則としてすべての資料が持つべき必須項目であり、アーカイブ外から利用するときにも取得した情報を確認する基本項目として重要である。

加えて、作品などに与えられた名前を**名称**とする。原則として全て言語タグを付与し、読みも`@ja-Kana`などの言語タグによって区別する。サブタイトル、別名も名称として並列に扱う（主タイトルはラベルと一致するものを調べることによって区別する）。

必須のラベルとは逆に、ソース情報に名称相当の項目がない場合は、名称は設けない。

タイトルと別名の区別　タイトルと別名などを同じ名称として扱うのは、異なる名前を単一プロパティで検索できるようにするためである。標本の正式名称が「はえ追い」で別名が「払子（蠅払い）」となっているとき、どちらの名称でも区別なく検索できる方が利用者にとっては便利だといえる[11]。

読みを構造化せず、別プロパティにしないのも同じ理由からである。またラベル（よみ）のように読み項目を設けずラベルや名称に付記する形で読みを加えているソース情報もあり得る。

ただし、英文でも主タイトルと別名が提供されている場合、両者を等しく名称として扱うと主タイトルが区別できなくなる。サンプルデータでは見当たらないが、もしこの区別が必要であれば、ラベルにも言語タグを付与する必要がある[12]。

3.2.3　作者、寄与者および発行者

作者、発行者などはすべてリテラルではなく実体（Agent）として記述する。NDLA の名称典拠とマッチングさせて典拠 URI で記述することを目指す。典拠 URI が得られない場合でも、ソースデータでの ID などに基づき、仮 URI を付与する[13]。

寄与者は役割を記述するために図 4 のようなロール・モデル[14]を用いて寄与関係として記述する。

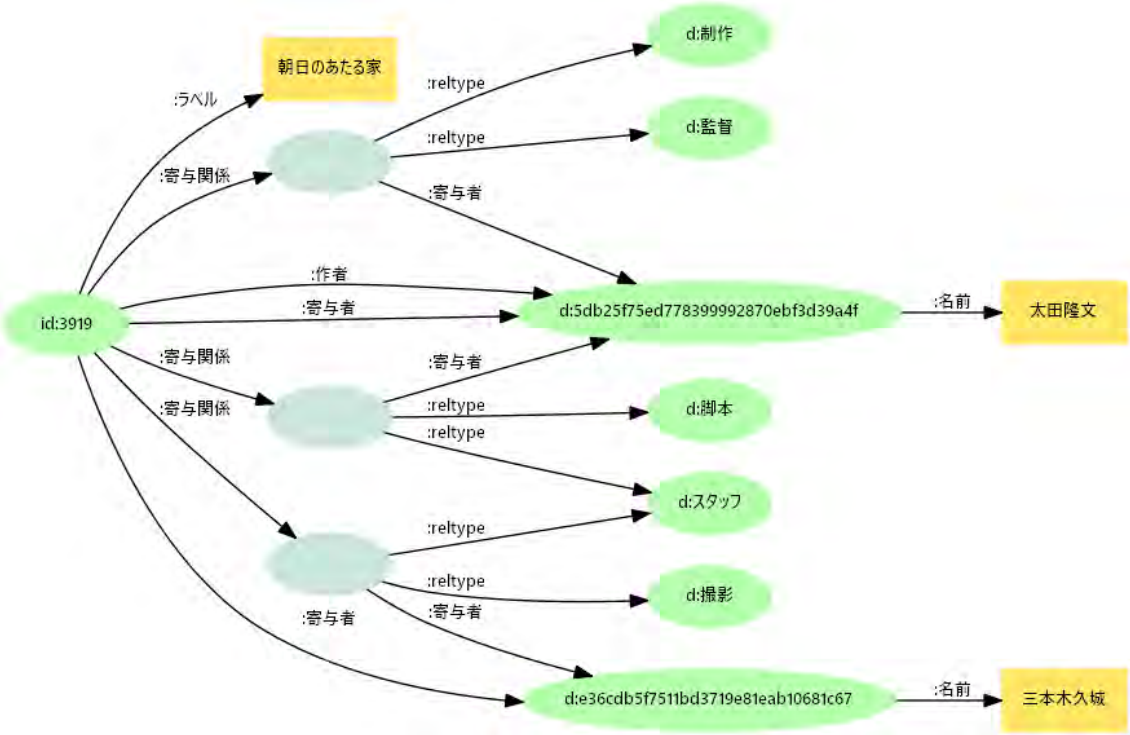

図 4: :役割を作品と寄与者実体を直接結ぶプロパティとして用いる

[11] もちろん 2 つ以上の項目を対象に検索することはできるが、シンプルな検索で見つかるほうが利用しやすい。

[12] なお言語タグの与え方としては、(1) 他言語もしくは読みがある場合は、区別のために各ラベルに言語タグを加え、(2) 単一ラベルの場合には言語タグは付与しない、という方法も考えられる。本来は、言語タグを用いるなら全てのラベルにタグを付与するべきだが、ソース情報の同一ラベル項目に複数言語が混在している可能性があり、これに一律の言語タグを付与するとかえって不正確になる。なお、同じリソースに対するラベルに言語タグ付きと無しが混在するのは、アプリケーションにとって扱いが難しいため、避けるべきである。

[13] ID がない場合でも、機関名＋氏名のハッシュなどで URI を生成する。最小限でも同一ソース内で同一実体を集約できる他、後日典拠 URI との一致が得られれば、典拠 URI に置き換えた上で旧 URI は転送するなり sameAs を用意することで恒久性を確保できる

[14] Schema.org のプロパティ反復形を参考にしている。役割を階層語彙として定義するのも検討に値する。

ここで、寄与者は作者などより一段深い位置におかれて利用しにくい可能性や、役割が明示されなければ単に冗長であるという問題にも対応するため、単純プロパティ（ショートカット）も併用する[15]。

- 作者も常に寄与関係の一つとして併記し、役割を「制作」などにする。作者も含め関連する人（実体）を調べたい場合は寄与関係として検索できる。
- さらに作品と寄与者実体を直接結びつける寄与者プロパティを追加する。プロパティが不明でも作品に直接つながる実体を検索すればよい（図 4）。

発行者も、「出版」もしくは「発行」という関係において寄与する実体の一つとして捉えることができる。併用する単純プロパティを、一般的な「発行者」とする。

出版に関するイベントモデルは採用しないが、以下に述べるように場所、時間情報についても関係型モデルを用い、その関係を「出版」とすることで、出版に関わる実体やデータを集中させられる。

3.2.4 時間関係

作成、公開、発見、内容[16]など対象に関する時間範囲を実体（時間実体）として記述する。またどのような関係でその時間とつながるかを示すための時間関係も併記する（図 5）。

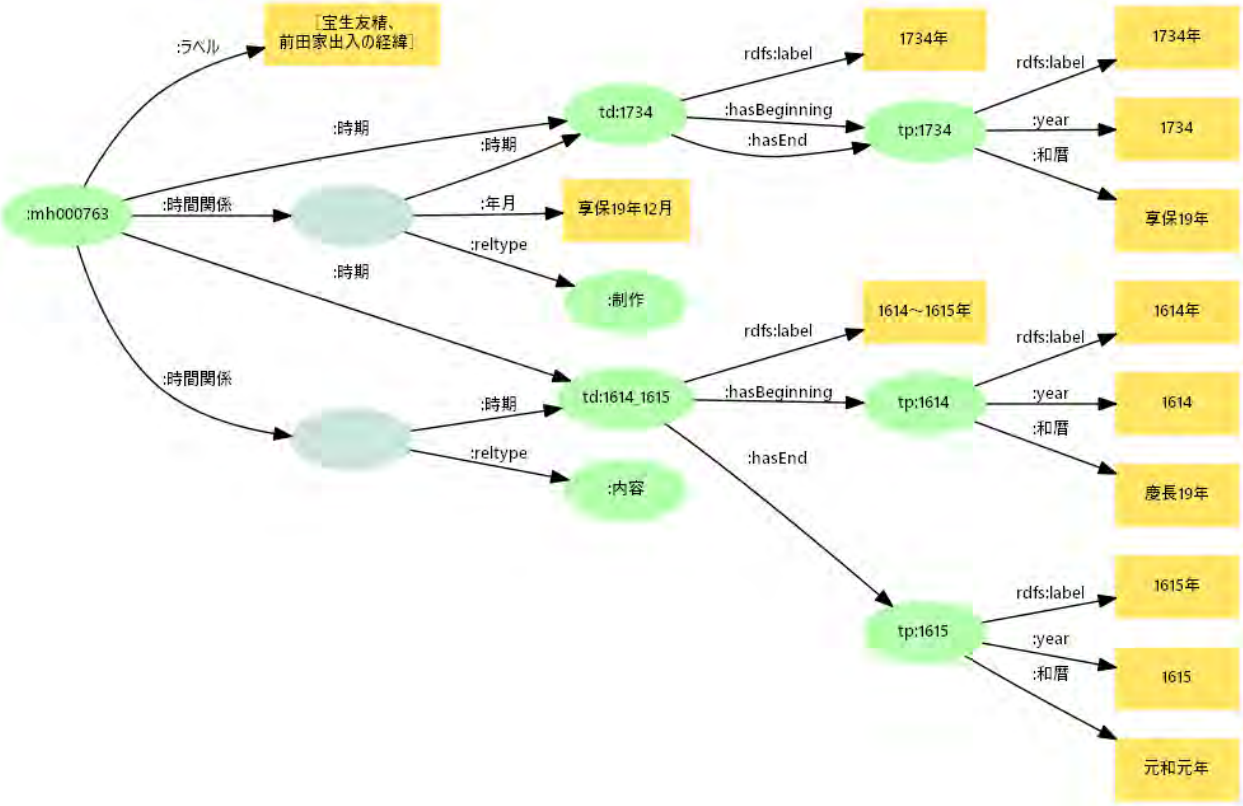

図 5: 時間範囲を実体として扱う。時間関係ノードに「作成」「内容」などの型を与えることで区別する。また個別レコードに記述するのは接頭辞 td:のノードまでで、それ以降は共通の情報となる。

[15] メタデータフォーマット仕様では、分かりやすさのためにショートカットを前面に出している。発見タスクの観点では、ショートカット、構造化プロパティは未知の度合いがそれぞれ高い場合、低い場合に利用できるものと言える。

[16] 時間関係による「内容」とは、作品の記述対象である時代など、dct:temporal として記述されるものを念頭に置いている。

- 時間実体は、開始年、終了年（いずれも時間実体）を持ち、範囲検索を可能とする。また 1901〜2000 年の別名実体（sameAs）として「20 世紀」、1603〜1868 年の別名実体として「江戸時代」を定義することで、時代区分も同様に表現可能とする[17]
- 時間関係は、寄与関係と同様に、リソースと時間実体の関係を中間ノードを用いて記述する。また時間実体は年を単位とするので、月日などはこのノードにリテラルで追記する。
- 制作、出版など標準的に用いられるプロパティに対応する場合、時間リテラルをリソースの直接プロパティとして付与する。

時間実体に関する情報（開始/終了年、別名実体など）は、レコードごとに作成するのではなく**時間オントロジー**として共有し、煩雑さを回避するとともに、多様な利用を可能とする（図 6）。

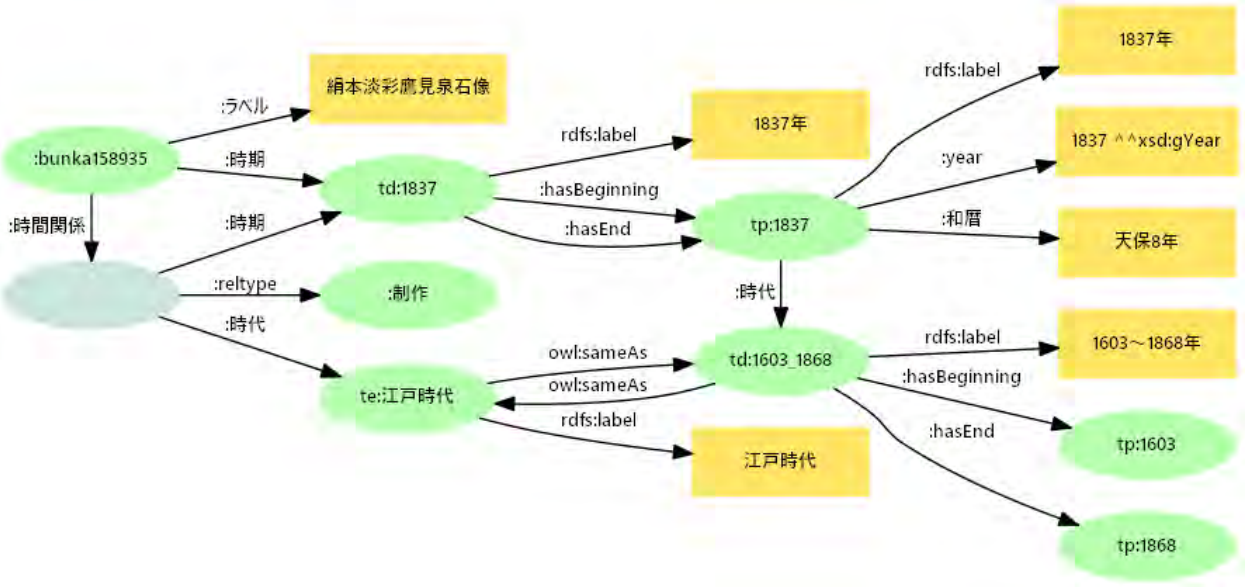

図 6: 時代、和暦、開始/終了年などを時間実体の情報として共有する。

3.2.5 場所関係

時間と同様に、作成、公開、発見、内容[18]など対象に関する場所の情報を一括して「場所関係」とする。また場所を示す実体をショートカット結びつける。時間以上に場所を表す項目はさまざまなものが見られるので、プロパティを統一した上で関係を示すモデルは共通利用に有益と思われる（図 7）。

ここで、場所も時間と同様に、国内は都道府県、海外は国単位の統一実体を定義することが、地域情報との連動など利活用の観点では非常に重要になる。ただしサンプルデータの段階でも場所は「日本」から「白金台」まで粒度がさまざまであるため、正規化には一段の工夫が必要。

[17]範囲ではない特定時期も、開始年と終了年が同一である範囲として扱う。DPLA では dc:date の値を常に構造化し、edm:begin、edm:end によって範囲を示している。
[18]場所関係による「内容」とは、作品の内容となる対象地域、舞台など、dct:spatial として記述されるもの。

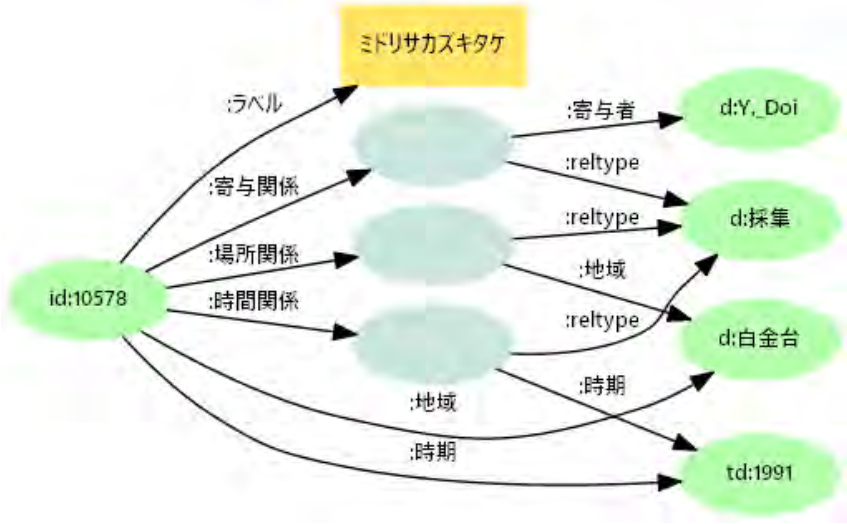

図 7: 場所情報の関係を示すことで、同じ「採集」関係にある寄与者、時間情報が集約できる。

3.2.6 主題と区分

資料の主な内容やテーマを表現する件名および分類を、主題とする。NDLSH など統制された URI を持つ値を利用する。キーワード文字列のみの値は、マッピング時の正規化を検討するが、困難であれば特別な名前空間でキーワードを直接 URI 化する[19]。

主題とは別に、分野における区分けもしくはジャンル（国宝、ニュース、ドキュメンタリーなど）を「区分」として別のプロパティを付与することを考えてきたが、使い分けが難しく利用者にとっても単純化したほうがメリットがあると判断し、統合することとした。したがってこの項目は、（キーワード的な）共通認識があり、ファセット分類や検索結果の絞込などに利用できる用語を記述するもの、という位置づけとなる。

3.2.7 識別子、言語、画像

ISBN 等の共有されている資料識別子を収める。導入句付き、NDL サーチのようなデータ型付き、あるいは URN など値の持ち方は別途検討する[20]。所蔵館がアイテムに与える個別識別子は、後述の提供情報の一部として扱う。

国際的な利用も踏まえ、内容の言語を URI として収める。複数言語で提供されるものはプロパティを反復する（アブストラクトのみ英語ありといった、部分的な要素は言語情報の対象としない）。

また資料を識別するために、特徴がわかるサムネイル～低解像度画像を用意する。これはアクセス情報として扱う（ある程度高解像度の）画像とは別に、統合プラットフォームで保持する。

3.2.8 記述

他のプロパティで表現できないテキスト情報を、元項目を導入句として加えた形で収める。なお、検討段階では次の両者を区別することを考えたが、利用者にとって単純化したほうがメリットがある

[19]たとえば CiNii は論文のキーワードをそのまま URI 化して foaf:topic の値としている。同音異義語の衝突可能性を了解した上で利用すれば、十分有益だと思われる。

[20]導入句を加えるとプロパティ値が識別子自身と違ってしまう。人間が確認するという意味での利用者タスクにとっては問題ないが、項目名ではなく値については機械利用という観点も考える必要がある。

と判断し、統合することとした。

資料体の記述　媒体（メディア）、形態、サイズ、数量など試料の物理的な特徴を、導入句付きの「記述」として扱う[21]。

内容記述及び注記　概要・要約、注記、備考など物理特徴以外のその他の情報を、同じく（必要に応じて導入句付きの）「記述」として扱う[22]。

項目（プロパティ）を細分化しないのは、その方が検索利用しやすいこと、識別・選択のためには導入句があれば用が足りること、分野／提供者ごとに異なる多様な項目をアーカイブで反映するのは困難なこと、詳しくはソース情報によって確認できること、による。

なお情報の表示にあたっては、導入句を値から切り離し、項目名に付加して「記述（注記）」のような表示項目として用いることで、より確認しやすくなると考えられる。

3.2.9　上位資料

ソースデータ内に上位資料を識別できる ID がある場合に設定する。このとき上位資料は：

1. ソースが上位資料を独立レコードとして含んでいれば、それに対応する共通情報リソース
2. 上位 ID はあるが、その ID に対応するレコードとしては記録されない場合は、（ラベルがあればそれのみを持つ）接続情報としてのノード
3. 上位の名称のみがある場合は、名前に基づく仮 URI を生成して接続ノードを作る[23]。

[21] サイズ、数量などは形式ばかりでなく単位のあり方もばらばらなので、元の項目名を保っても機械利用は難しい。現状では、人間利用者の識別、選択タスク用である。そうであるならば、当初検討した「資料対記述」プロパティを「記述」と区別するメリットもなくなる。また、値を URI で表現し、区分に準ずる扱いとすることも検討したが、現実的ではないので見送った。

[22] 概要・要約を独立させることも検討したが、提供される割合が低く、逆に検索漏れを生じる可能性が高くなることから、記述と一体化する。

[23] 名前からの URI 生成は、データの正確性に依存する。表記の揺れや句読点違いなどがあると成立しない。

例 1：国立公文書館のデータ　資料群、簿冊、件名/細目がそれぞれ別レコードとして提供され、かつ「親メタデータ ID」によって上位と関連付けられている（図 8）。

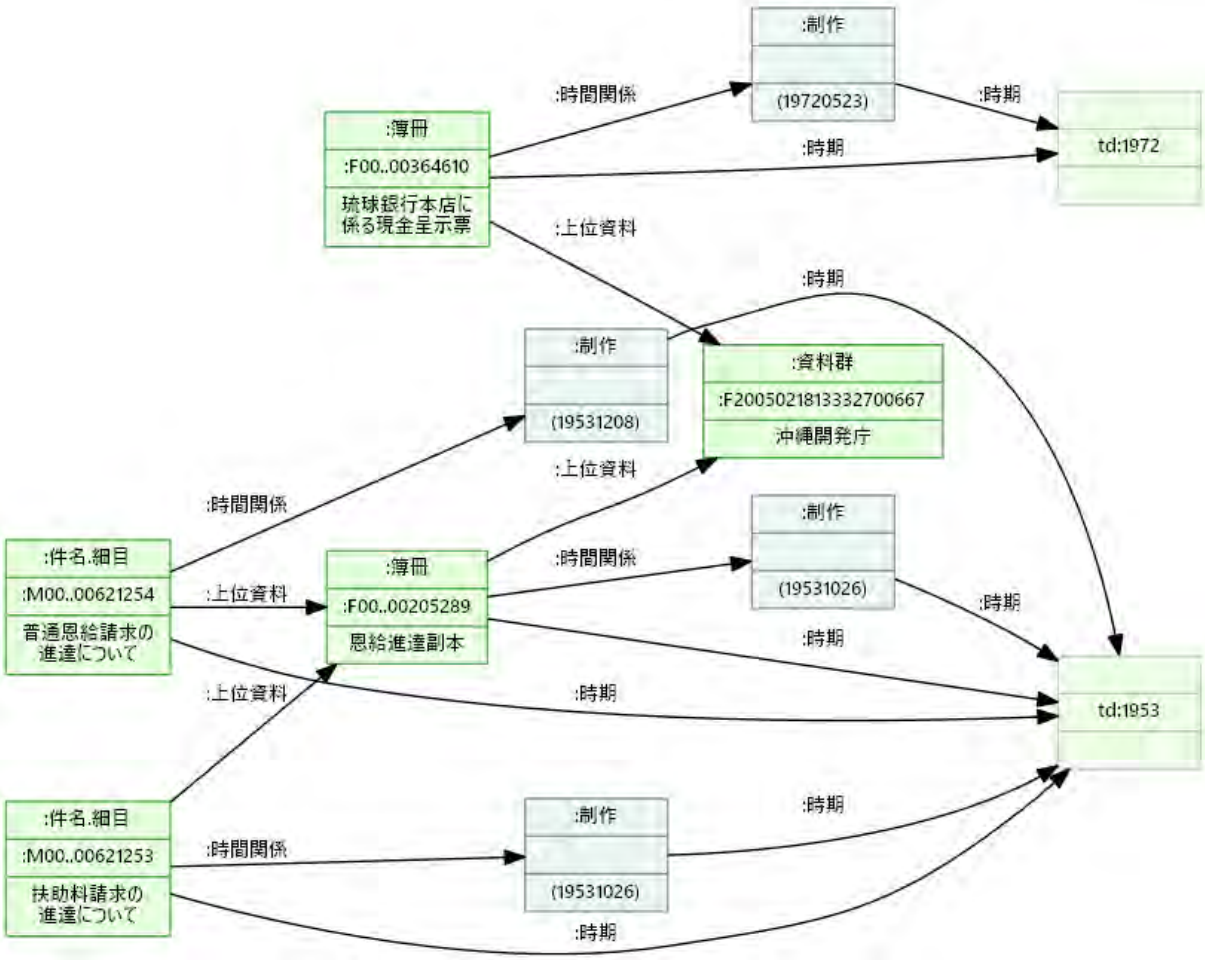

図 8: 上位 ID がレコードとして含まれる場合

この関係をたどり、公文書などで重要とされる「最上階層レコードとの結びつき」が確保できる。

例 2：ADEAC のデータ　各資料が含まれる「刊本 ID」「刊本名」を持つが、刊本 ID に対応する独立したレコードはソースデータには存在しない（図 9）。

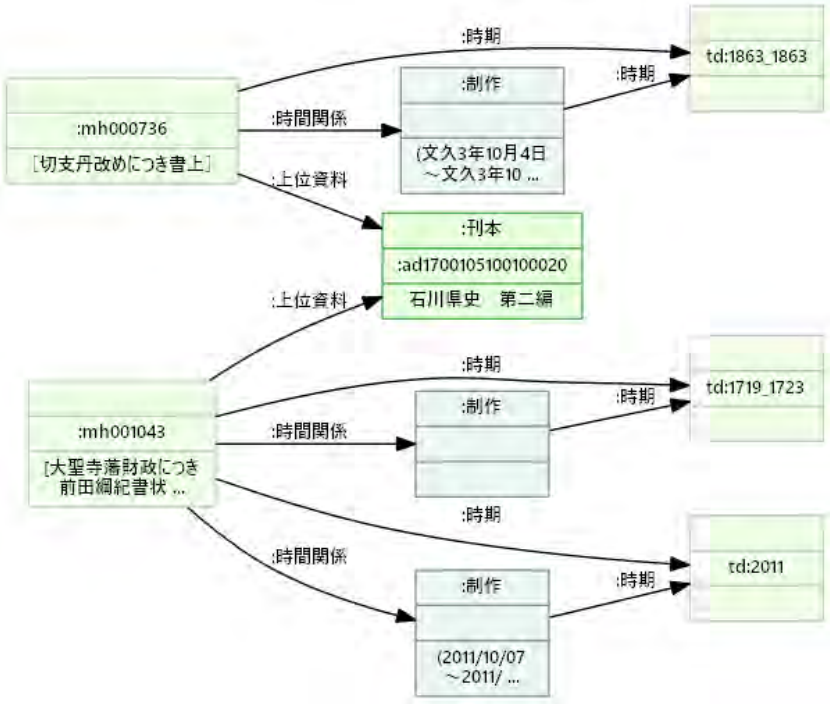

図 9: 上位 ID はあるがレコードに含まれない場合

この場合、上位資料についての情報は共通アーカイブからは得られないが、同じ上位資料を共有するレコードのつながりを見出すことができる。また上位資料の URI が参照解決可能（リンクしている）であれば、リンクを辿ってその情報を取得することが期待できる。

例 3：NDL サーチのデータ　書籍のシリーズ、雑誌の特集などは `dcndl:seriesTitle` にリテラル値があるのみなので、リテラル値のハッシュなどから仮 URI を生成して同じシリーズを結びつける[24]（図 10）。

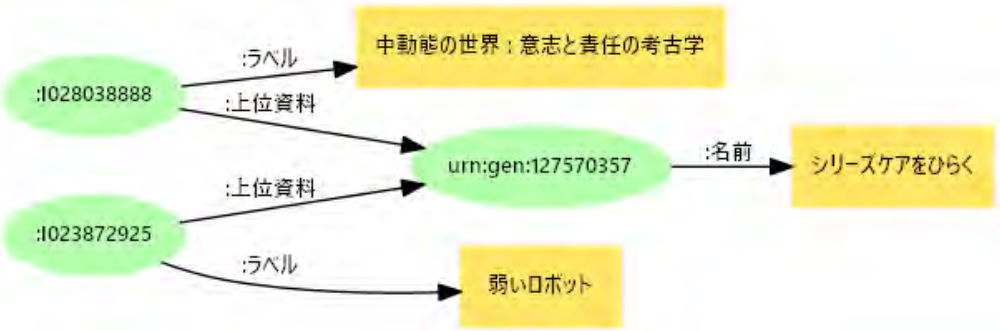

図 10: 上位については名称のみがある場合

[24]NDL サーチのウェブページ詳細画面では検索リンクが提供されているが、データとしては関連付けられていない。ハッシュに基づく URI であれば、同じタイトルは同一の URI として表現し、結びつけることができる。

掲載号と掲載誌　記事・論文等の掲載誌も、この上位資料として扱うことができる。この場合、雑誌の掲載号が記事の上位資料となり、巻・号などのリテラル属性を持つ。雑誌はさらにその上位資料として位置付けられる[25]。

掲載ページは、記事と掲載号の“ 関係（上位資料プロパティ）が持つ属性 ”であるため、寄与関係などと同様に関係ノードを設けて扱う[26]。これを区別することで、同じ号に掲載された記事同士を掲載号ノードによって結びつけることができる（図 11）。

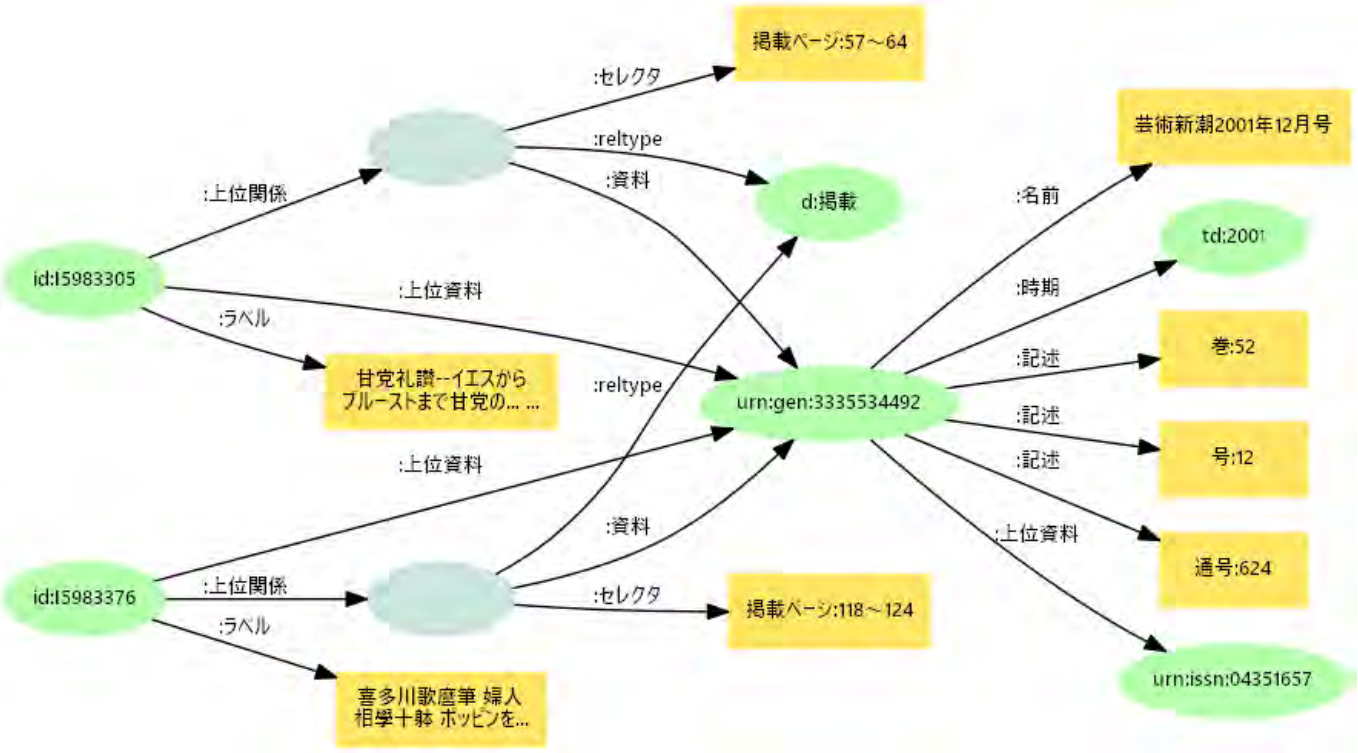

図 11: 記事、掲載号、掲載誌の関係グラフ。urn:gen... が掲載号、urn:issn... が掲載誌を表す。

この上位関係は、Web Annotation でリソースの部分を指定する特定リソース（Specific Resource）になぞらえることができる。特定リソースではリソース全体を `source`、その部分を示す方法を `selector` としており、ちょうどこのノードにおける「資料」と「セレクタ」に対応する[27]。この関係を利用して、CD のトラック、全集中の巻号なども表現できる。

[25]NDL サーチには掲載号という単位はないが、巻号情報を利用すれば一意の URI を付与することは難しくない。掲載記事の上位概念は、Web of Science では掲載号だが CiNii では掲載誌などデータベースによって異なっている。掲載号を考えるモデルは一段階複雑ではあるが、テレビ番組の階層などと対比しても、検討して良いモデルだと思われる。

[26]データモデル上。利用者への画面表示は巻号とページを同列に扱ってよい。

[27]Web Annotation ではセレクタも構造化してその種類を示すようになっているが、アーカイブにおいてはそこまでは必要ないと思われる。

3.3 資料のアクセスとソース情報に関するメタデータ項目

3.3.1 資料アクセスの提供情報

資料オブジェクト自身、そのデジタル化や複製物、および関連情報へのアクセスを提供するための情報を構造化して記述する。EDM の `Aggregation` に近いが、直接収集したものに限らず、URL によって提供元にアクセスできるもの[28]や、現地に行って閲覧できるものも含む。以下の項目を持つ。

- **提供者**：資料（の複製物）の提供者を識別する URI。**保管者**が別であればその URI も[29]。資料自身のほかに解説や複製物などが複数箇所から提供される場合、それらは次のリンクで示し、提供者は資料自身に最も近いものを提供する博物館などを示す。なお、ソース情報がアグリゲータ経由で提供される場合、その元となる一次データの作者がこの提供者となる[30]。
- **リンク**：資料の紹介ページやアクセス情報が記載されたページの URL。画像資料の IIIF マニフェストなど特別な URL は、クラスを付与して判別できるようにする。
- **オブジェクト**：資料のデジタル画像もしくは音声動画などの URI。共通項目の識別用画像とは別の、提供元が保持するもので、複数の場合もあり得る（画像閲覧システムの該当ページは「リンク」で）
- **権利情報**：提供者が権利情報を示していれば、そのテキスト。画像などオブジェクトのライセンスがある場合は、オブジェクト URI のプロパティとして**ライセンス**の URI 記述するか、提供オブジェクトのライセンスを各オブジェクトを主語として記述する（§ 3.3.4 も参照）。
- **識別子**：提供者／所有者が管理するアイテムとしての識別子（請求記号など）。
- **記述**：その他、資料へのアクセスに関連する情報があれば導入句付きの説明記述。

ここでの権利情報は、提供される資料（の複製物）に関するものを扱う。メタデータ（共通情報および次項の公開用ソースデータ）は原則として CC0 レベルのライセンスとし、全体に共通する情報として共通アーカイブの説明ページなどに明記する。

3.3.2 ソース情報

資料に関する元メタデータ（収集したソースデータ）についての情報を記述する。§ 2.3 のソース分離モデルにおけるソース情報に相当する。

ソースデータの**提供者**のほか、**更新日**などその管理メタデータを示し、アグリゲータに目録等の情報ページがあれば**リンク**としてその URL も加える[31]。さらにソースデータをそのまま[32]手を加えず**データ**として提供する[33]。

アグリゲータを介する場合は、アグリゲータがこのソースデータ提供者という位置づけになる。資料へのアクセスを提供する主体とソースデータ提供者が同一という場合もあり得る。

[28] EDM においては、これは提供元側の Aggregation として別のグループを構成する。

[29] 提供者、保管者の扱いは、ソース情報で公開されるものに準ずる

[30] EDM での dataProvider に相当する。

[31] デジタル画像などがアグリゲータ経由で提供されていても、これらはアクセス提供情報の“ オブジェクト ”として扱う。これは § 2.3 で提供情報を Aggregation と位置付けたとおり、資料に関連するリソースへのアクセスは提供情報に集約するため。アグリゲータの目録は、共通情報の元であるソースデータをウェブページで確認するという目的のためにソース情報の URL とする。こちらもアクセス提供情報の一部にしたほうが分かりやすいという議論はあるかもしれない。

[32] データ形式は JSON 等の共通フォーマットに変換するが、項目名（たとえば JSON キー）、値は元のままにしておく。

[33] ソースデータ提供時に公開不可項目はあらかじめ提供側が取り除き、共通情報と同じく CC0 で提供する。ライセンスは全体で共通して表示する。

3.3.3　アクセス提供情報とソース情報の関係

アクセス提供情報とソース情報の提供者が同一の場合　一次データを直接収集する場合は、提供情報とソース情報の提供者が同一となる。たとえば民博ビデオテーク・データベースはソース情報を提供するとともに資料へのアクセスも提供する（図 12）。

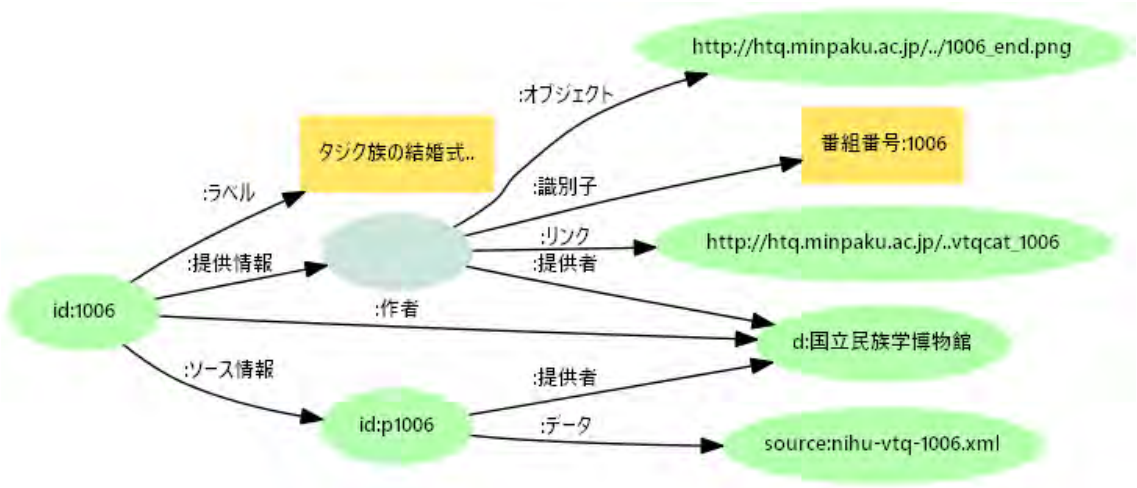

図 12: アクセス情報とソース情報の提供者が同一となるグラフ。ここでは提供者が作者でもある

アクセス提供情報とソース情報の提供者が異なる場合　アグリゲータとは別に資料の提供者（保管者）がある場合は、2 つの情報は基本的には分離される（図 13）[34]。

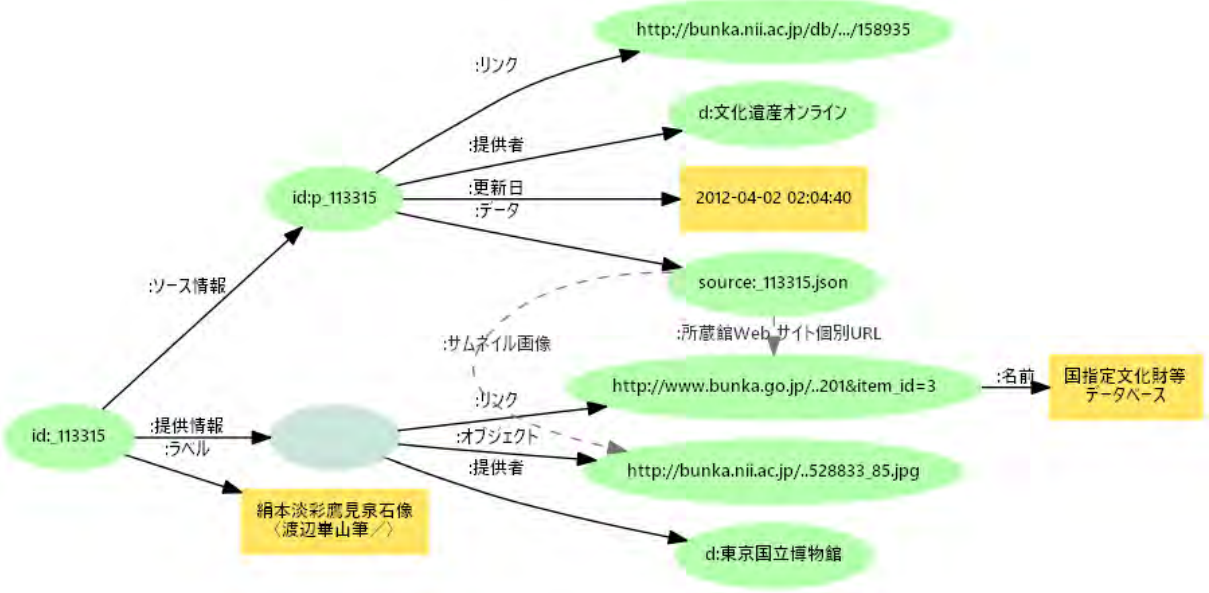

図 13: 提供情報とソース情報の提供者が異なるグラフ。画像はアグリゲータ経由であっても提供情報としている

[34]文化遺産オンラインの場合は、提供者リンクにさらに e 国宝があるが、ここでは省略した。ソース情報のデータ内には資料提供に関する情報が含まれるので、これが RDF であれば、図の破線のようにつながりができることになる。

3.3.4 アクセス情報リンク先リソースのプロパティ

権利の記述　アクセス情報として提供する画像・動画などのリソースについては、それぞれを主語として権利（ライセンス）を URI で記述する[35]。

検討段階では、プロパティの使い分けによる区別も考えた。

- 「オブジェクト」プロパティを用いた場合は自動的にライセンスが CC0 とみなせる、などの規約を設ける
- これに該当しないリソースは、アクセス提供情報での記述に別のプロパティを用いる

この方法はモデルとしてはシンプルになるが、ライセンスに関係なく画像を調べるといった検索時に複数プロパティを調べなければならないというデメリットも伴うため、最終検討会において不採用となった。

リソースタイプの記述　リンクあるいはオブジェクトでアクセス提供情報に関連付けるリソースには、型情報を与える。§ 3.3.1 でも示したように、IIIF マニフェストなど特別な情報を提供するリソースについては、そのためのプロパティを毎回定義するよりもリソースの型によって判別可能にするほうが利用しやすい。

このとき、特定のリソースだけが型を持つよりは、全てのリンク先リソースが型を持つほうが一貫性があり、利用者にとっても分かりやすい。またオブジェクトについては、画像以外のメディアの提供が今後増加する可能性がある。これらを考えると、最初から全てのリソースに型情報を付与しておくのが賢明と言える。

[35] ソースデータにライセンスのプロパティがない場合に統合プラットフォームで CC0 などのライセンスを記述するのはメタデータ管理上どうかという疑問もあったが、検討会において特に違和感はないとの意見であった。

4　語彙の検討

4.1　基本記述語彙の選定

共通アーカイブ情報およびソース情報を記述する語彙については、次の点を考慮して検討する。

1. モデルの狙いが的確に表現でき、各項目（トリプル）の意味に整合性があること
2. 少なくとも基本的な部分については、広く理解される既存語彙を再利用できることが望ましい
3. 多数の語彙を複雑に組み合わせるよりは、シンプルな語彙（名前空間）構成のほうが、データを扱う人間にとって把握しやすい

ユーザのタスク（§2.1）に照らしてみると、特に「発見」に用いられるプロパティには広く理解される語彙を用いることが望ましく、「識別」「選択」については的確な表現を重視する必要があるといえるだろう。

この観点から、Dublin Core、Schema.org、BIBFRAME を候補として、基本データモデルをどの程度記述できるかを比較してみる。そのうえで、3. を踏まえて基本記述語彙を一つ選び、不足する部分は原則として独自定義（もしくは DC-NDL を拡張）として、統合プラットフォーム情報の記述語彙を選定する。

なお独自語彙を定義する場合、ツールでの扱いやすさや他語彙との親和性、国際的な利用も考慮し、用語は英単語（の組合せ）での命名を原則とする。

4.1.1　基本プロパティの比較

3 つの語彙で基本項目に該当するプロパティを表 3 に示す（*印は定義とずれるもの／提案中）。

BIBFRAME　作者も含め寄与者のモデルのみで記述するため、作者に相当する短縮プロパティがない。また出版事項にイベントモデルを採用していることから、発行者の短縮プロパティがないほか、発行に関する時期、地域も異なるモデルとなるなど、基本情報については該当なしの項目がやや多い。

一方、基本モデル案では記述としてまとめている資料体情報は `extent`、`dimensions`、`baseMaterial`、`appliedMaterial`、`bookFormat` など詳細に記述するプロパティが用意されている。

Dublin Core　検索に用いる基本項目は、区分を除きカバーできている。画像を示すプロパティはない。

名称に対応するプロパティは `title` となるが、これは作品などの（作者が与えた）名称という受け止めが多いかも知れず、標本などの名称としては違和感が残る可能性がある。また `spatial`、`temporal` は `coverage` のサブプロパティであるため、制作に関する時間、場所の記述に用いるのは不正確（モデルの的確で整合性のある表現ができない）といえる。

Schema.org　基本項目は画像も含めカバーしている。対象とする分野も広い。

`name` は、リソース一般の名称としては DC の `title` より違和感が少ないと思われる。`spatial`、`temporal` は現在のところ Dataset の対象地域・時期を示すためのものとして定義されているが、CreativeWork 一般に適用できるよう提案中である[36]。

[36] また資料体記述を独立させることを検討していた段階では、W3C の Schema Architypes コミュニティ・グループと協力して physicalDescription も提案していた。

表 3: 基本プロパティの比較

基本項目	Dublin Core	schema	BIBFRAME	F	I	S	O
タイプ	rdf:type	rdf:type	rdf:type	○	-	○	-
ラベル	rdfs:label	rdfs:label	rdfs:label	-	◎	○	-
名称	title	name	title	◎	○	○	-
寄与者／関係			contribution (+role)	△	○	○	-
	contributor	contributor	↓	○	○	○	-
作者	creator	creator	↓	◎	◎	○	-
発行者	publisher	publisher	(provisionActivity + agent)	○	○	○	-
場所／関係				△	○	◎	-
	spatial*	location*	place	○	△	○	-
時間／関係				△	◎	◎	-
	temporal*	temporal*		○	○	○	-
	created	dateCreated	creationDate	△	○	○	-
主題・区分	subject	about	subject	○	△	○	-
識別子	identifier	identifier	identifiedBy	○	○	-	△
言語	language	inLanguage	language	△	◎	◎	-
画像		image		-	◎	◎	-
記述	description	description		△	○	◎	-
上位資料	isPartOf	isPartOf	partOf	○	○	○	-
提供情報				-	-	-	-
提供者		provider		-	○	△	◎
リンク	related	url	relatedTo	-	-	-	◎
オブジェクト				-	-	○	◎
権利情報	rights	license		-	-	○	○
個別識別子	identifier	identifier	identifiedBy	△	△	-	◎
ソース情報				-	-	-	-
提供者		provider		-	-	-	○
データ		distribution		-	-	-	○
リンク	related	url	relatedTo	-	-	-	○
更新日	modified	dateModified	changeDate	-	-	○	○

4.1.2 基本記述に用いる語彙

以上の比較をもとに、基本プロパティの記述には Schema.org を採用した。

Schema.org は W3C の Schema Bib Extend コミュニティ・グループにより書誌関連の拡張が導入され、VIAF や WorldCat の LD 記述に採用されている。さらに前述の Schema Architypes コミュニ

ティ・グループが、アーカイブ全般の記述についての拡張提案を準備している。また、商品、イベント、レシピなど多様な分野の記述が可能で広い範囲で利用されているため、領域を超えたデータの活用に大きく資することが期待できる。

4.2 関係モデルの構造化プロパティ

関係モデル（寄与者／関係、場所／関係、時間／関係）のうち、寄与者については既存語彙も記述方法を提供している。たとえばSchema.orgはプロパティを反復するRoleモデルを示している。BIBFRAMEでは`contribution`＋`role`、`agent`によってそのまま寄与者関係が記述できる。Schema.orgのRoleモデルは、時間／場所関係にも適用できなくもない。

ただしSchema.orgは基本語彙（ショートカット）として用いることにしており、BIBFRAMEでは時間／場所関係を記述できない。そのため、関係モデルには独自の語彙を用意し（接頭辞`v:`で示す）、可能な部分はSchema.orgの用語（接頭辞`schema:`で示す）で記述することにする（表4）。

表 4: 関係モデルに用いるプロパティ

項目	内容	プロパティ
寄与者関係	制作に関与した人／組織の関係	v:contribution
	寄与のタイプ。領域を超えて一般化できる制作、編集、翻訳など	v:relationType
	寄与した人物・団体	schema:agent
	領域固有の役割名、キャストの配役など補足記述	schema:description
	関連リンク	schema:url
時間関係	作成、公開、発見、主題など資料に関する時間	v:temporal
	時間関係のタイプ。制作、主題など	v:relationType
	時間関係における時間（範囲）を示す	v:temporalValue
	時間関係における時代を示す	v:era
	月日ほか補足情報	schema:description
場所関係	作成、公開、主題など資料に関係する場所・地域	v:spatial
	場所関係のタイプ。制作、主題など	v:relationType
	場所関係における場所を示す	v:spatialValue
	より詳細な地名・住所など補足記述	schema:description
上位資料関係	上位資料との関係（掲載誌のページ）などを記述する	v:partOf
	上位資料との関係のタイプ。掲載など	v:relationType
	上位資料関係における上位資料本体を示す	v:source
	上位資料内の特定部分を示す	v:selector
	補足記述	schema:description

寄与関係にはSchema.orgの`Role`のプロパティ利用を想定していたが、場所、時間関係にも独自プロパティを導入することから、独自の`relationType`を用いる。たとえば映画において監督の`relationType`を「制作」とすることで制作に関する日付や場所とも関連付ける。このとき`description`に「監督」を記述し、領域特有の役割表現を示す。

上位資料に関しては、§3.2.9 の例 3 に示した関係の記述のために部分指定のプロパティを導入するが、上位との関連を示すだけで十分な場合は構造記述を追加する必要はない。

4.3 アクセス提供情報とソース情報

アクセス提供情報及びソース情報に用いるプロパティは、表 5 のように定義する。

表 5: アクセス提供情報及びソース情報に用いるプロパティ

項目	内容	プロパティ
提供情報	資料の提供・アクセスに関する情報	v:accessInfo
	資料（に関する情報）の提供者	schema:provider
	資料の保管者（提供者と別の場合）	v:contentHolder
	資料の紹介ページやアクセス情報が記載されたページ	schema:url
	資料のデジタル画像、音声動画など	v:digitalObject
	資料のデジタル画像、音声動画などのライセンス URI（画像などに対するプロパティ）	schema:license
	資料自身の権利記述	v:contentRights
	提供元が付与する識別子	v:contentId
	補足記述	schema:description
ソース情報	収集したデータ及びその提供元に関する情報	v:sourceInfo
	ソース情報の提供者（アグリゲータ）	schema:provider
	集約元情報の公開ページ	schema:url
	プラットフォームが保持・提供するソースデータ	v:sourceData
	ソース情報が RDF である場合	rdfs:seeAlso
	補足記述	schema:description
	収集元データの更新日、もしくは収集日	schema:dateModified

アクセス提供情報は Schema.org の `Service` にやや近いとも考えられるが、サービスで提供するものの内容（画像など）も記述するため、その部分は独自のプロパティを用いる[37]。プロパティ名は、この点についての誤解を与えないよう `content` などを前置したものとする。

ソース情報は Schema.org の `Dataset` にほぼ準ずるものとして検討したが、定義範囲の違い[38]もあり、独自に定義した。なおソース情報が Linked Data として提供される場合は、元のデータを直接 `rdfs:seeAlso` で参照してリンクする[39]。

それぞれのリンク情報に `schema:url` を充てている。これは各情報リソース自身に対応する URL を意味することになる[40]が、誤解の可能性は低いと考えて採用した。

[37] ここで例えば schema:image を用いると、提供情報／サービスを表す画像を意味してしまうので、独自のプロパティを定義する。

[38] 提供データはデータセットではなくその中の 1 レコードであり、schema:Dataset そのものではない。また §2.3 で述べたように、ソース情報は EDM の Proxy に相当するものとして「アーカイブ情報」クラスを与える。

[39] 共通情報の抽出や基本検索のため、Linked Data であってもソースデータは収集し、内部的に保持する。

[40] 提供情報リソースの schema:url として、資料アクセスの方法が示されたページの URL は適切といえるが、資料提供元の目録ページなどは適当ではないかもしれない。最小名前空間の方針からは外れるが、オブジェクトも含め、EDM の isShownAt などの WebResource を示すプロパティを利用する方法もある。

4.4 クラスと概念体系

4.4.1 クラス定義

統合プラットフォームにおいては、収集者の視点によるメタデータレコードを表現するクラス（§2.3 参照）として**アーカイブ情報**を定義する。また共通アーカイブ情報の各レコードにはそのサブクラスである**共通アーカイブ情報**クラスを用いる。

また資料の基本区分として、各提供データの分類などに基づくクラス（**文化財**など）をそれぞれ定義する。これらは§3.2.1 でも述べたとおり、「文化財」でありかつ「アーカイブ情報」でもあるというリソースとして混乱を招かないよう、「共通アーカイブ情報」のサブクラスとし[41]、基本的なものを統合プラットフォームで定義する。

4.4.2 関係概念

関係モデルで用いる `relationType` の値は、クラスとして定義せずに `skos:Concept` として扱う。これらをクラスとして扱うと、たとえば寄与者関係ノードと時間関係ノードを表すクラスはそれぞれ別の基底クラス（寄与者関係クラス、時間関係クラス）のサブクラスとしなければならず、おなじ「制作」関係を通じてノードを結びつけることができなくなるからである。

時間に関しては、§3.2.4 で示した時間オントロジーを定義する。

4.4.3 主題とキーワード

主題は、可能であれば NDLSH を用いるが、各提供データが独自の主題体系（コード化されるなど URI 化可能なもの）を用いているときは、原則としてそれをそのまま共通情報にも用いる。

文字列のみのキーワードの場合は、§3.2.6 で示したように、特別な名前空間を用いて文字列を直接 URI 化する。

なお、場所関係における場所を示す URI については、国内は都道府県、海外は国（もしくは日本十進分類に示されるレベル）を念頭に集約・正規化して URI を付与することを目指すが、困難である場合はキーワードと同様に扱う。

4.5 Linked Data としての共通アーカイブ情報

統合プラットフォームの共通アーカイブ情報を広く活用できるようにするためには、Linked Data の標準に従ったデータモデル、語彙、URI を採用するだけでなく、LD クラウドのハブとなるようなデータセットとの結びつきが必要である。

出発点としては、主題や著者の記述に Web NDLA を可能な限り用いることで、ここを経由して VIAF、DBpedia などと間接的にリンクしていく。主題などに各機関のシソーラスなどを用いる場合も、この視点からできるだけ NDLA や DBpedia との関連付けができるよう、調整を進めることが望まれる。

一方、個々の作品（レコード）については、極めて有名なもの以外は外部にリンクできるリソースがない場合が大半と考えられる。これらはむしろ、日本（関連）の文化資源について言及する場合に、統合プラットフォームの URI が標準的に用いられるようにすることで、日本文化情報の**リンクのハ**

[41] クラス名としては「絵画」「書籍」など一般的な名称を用い、RDF を意識しない利用者にも違和感がないようにする。

ブになることを目指す。統合プラットフォームは、単独でメタデータを提供するだけでなく、こうした具体的記述情報と結びついてこそ、その価値を発揮することができる。

なお、RDF グラフにおいては、共通情報に対応する「コンテンツ..」は共通情報 URI に`#work` を加えて識別し、そこから共通情報と関連付ける（共通情報を複雑化させないため、「共通アーカイブ情報」からコンテンツ自身に関連付けるプロパティは用意しない）。

4.6 標準語彙との関係

独自定義したプロパティは、後日 RDF スキーマを整備する段階で、Dublin Core など標準的な語彙とのサブプロパティ関係を定義する。

5　マッピングと実装

5.1　マッピングの実装レベル

本メタデータモデルは、データのシンプルな利用と精緻な利用の両方に対応するために、単純プロパティと構造化プロパティの並列記述を用いる。このモデルを十分に生かすためには、プロパティ値として用いる時間／場所 URI や構造化の関係型プロパティ値（関係概念）の正規化が必要となるが、元データからのマッピングに際しては、十分な正規化が容易ではないことも考えられる。

そのためここではマッピングを 3 つの実装レベル（案）に分け、段階的な導入も可能とする。

5.1.1　レベル 1：最小マッピング

共通アーカイブ情報は、「いつ」「どこで」「だれが」「何を」を共通のプロパティによって調べられることを狙いとする。そのため、元データからこれらに対応する時間関係、場所関係、寄与者、ラベルへのマッピングを確実に行なうことが重要である。またアクセス提供情報、ソース情報は、出所の確認と資料の取得という目的を的確に果たせるようにする。

- 値の URI 化は、「いつ」については年レベル、「どこで」については都道府県レベルで行なう。
- 「だれが」については、提供元データに ID があればそれを用いた URI 化、なければ氏名（リテラル値）のハッシュを用いた URI 化を行なう。
- 「何を」については、ラベルは必ず生成する。名称は、項目名などから言語が明らかな場合のみ言語タグ付きとする。
- 構造化プロパティの関係概念については、「制作」「出版」に関しては最小限共通 URI を用いる（元項目名から判別可能と思われる）。それ以上の正規化が困難であれば、元データの項目名を URI 化するか、`schema:description` のようなリテラル値として項目名を保持する。
- 上位資料については、提供元データに対応する ID があるものは、確実にグラフがつながるようにする。
- 主題、キーワードについては、提供元データに ID があるものはそれによる典拠 URI 化、そうでなければキーワード URI 化（§ 3.2.6 参照）する。
- アクセス提供情報については、提供者、情報ページ URL、デジタル化オブジェクト URL、権利記述を最小限マッピングする。
- ソース情報については、提供者（アグリゲータ）、更新日（収集日）をマッピングする。

5.1.2　レベル 2：中程度マッピング

レベル 1 に加え、次の情報を正規化してマッピングする。

- 時間関係における時代、期間（範囲）および場所関係における国、地域を URI 化する。基本的な時間オントロジーを整備する。
- 構造化の関係について、実用になるレベルの標準関係概念をいくつか定義してマッピングする。標準概念で表現できないものはレベル 1 の形で保持する。

- アクセス提供情報のデジタル化オブジェクトに関するメタデータ（型、フォーマット、サイズ、ライセンスなど）を付与する。

5.1.3　レベル 3：高度マッピング

原則として共通アーカイブ情報の全要素をマッピングする。

- 寄与者（だれが）、主題 URI について、可能な範囲で NDLA と対応付ける。
- 時間オントロジーを構築する。
- 構造化関係の概念はすべて何らかの形で URI 化する。
- 上位資料は、ID がなくても一貫したラベルで示されている場合は、ハッシュURI を用いて関連付ける（§ 3.2.9 例 3）。

5.2　段階的実装と運用

導入にあたっては、利用者にとって実用的レベルであることと運用が現実的であることのバランスを考慮し、諸条件がゆるす範囲で実装を行なう。

5.2.1　段階的な導入とレベルの組合せ

必ずしも全てのソースデータが同じレベルでマッピングされなければならない訳ではなく、いくつかのコアデータをレベル 2 でマッピングしつつより多くのデータをレベル 1 で追加するなどの組合せも考えられる。

5.2.2　運用のレベル

提供元のソースデータ更新にすぐに対応できる随時更新に越したことはないが、諸条件によっては、更新頻度を落としての運用も十分考えられる。たとえば DBpedia は半年程度の間隔で全件を更新する運用を行なっている。Europeana は 20 万点以下のレコードは月次、それ以上は四半期ごとの提供を上限としている。

間隔を置いての全件更新は、メタデータのバージョン管理（ある時点でのメタデータセットの保存）という点でも有利であり、検討の価値がある。

5.2.3　データベースの考え方

メタデータをウェブ画面（GUI）、REST API、SPARQL エンドポイントなど複数の方法で提供するにあたり、それぞれを別のデータベースに格納するのではなく、一つのデータベースに対して複数のインターフェイスを設けて実装する。

データベースを RDF（トリプルストア）として REST API を SPARQL にマッピングしてもよいし、データベースには RDB を用いて R2RML などで RDF にマッピングする、あるいはその他のグラフデータベースを用いるのでも、提供できる情報が同等であればどれでも構わない。ただし、複数レベルを組合せた段階的導入には、トリプルストアが適しているのではないかと考えられる。

5.3 マッピング運用事例

5.3.1 Europeana

Europeana への参加機関は、公開提供ガイド（Europeana Publishing Guide）に記された基準に沿ったデータ提供が求められる。

提供から公開への手順 公開までに次の手順を踏む。

1. 参加要望書、データ交換合意書で前提条件の合意
2. データ提供フォームでデータの概要（規模、内容のタイプほか技術的な情報）を伝えた上で、ガイドに従った形でのデータを提供
3. Europeana がデータをチェックし、問題があればフィードバックして修正、再提供
4. 必要ならば提供者がプレビューで確認し、OK であれば公開

提供データは EDM 仕様に基づいて妥当性検証が行なわれる。加えて、データ値が適切であるか、意味的構造が正しいかも個別にチェックされる。

必須提供データ また次の 10 の必須項目が設けられており、これらを提供者側で記述することが求められる。

- タイトル（`title`）もしくは説明（`description`）。データセット内で識別ができ、意味があるもの（同じタイトルを異なるレコードで用いることはできない）。
- 書籍、写本などテキストオブジェクトの場合はその言語情報（`language`）
- デジタル化オブジェクトの型（`type`：TEXT, IMAGE, SOUND, VIDEO, 3D のいずれか）
- オブジェクトに関する文脈もしくは詳細情報（`subject`, `type`, `temporal`, `spatial`）
- 利用者の貢献によって提供されるオブジェクト（例えばクラウドソーシングによるデジタル化）は `edm:ugc` を `true` として専門家によるキュレーションを経たものと区別する。
- オブジェクトに関するデータを作成（アグリゲータに提供）した機関の情報（`dataProvider` アクセス提供者）
- データを Europeana に直接提供する機関の情報（`provider` ＝ソース情報提供者）
- デジタル化オブジェクトの URL を少なくとも 1 つ。できればファイルを直接 DL 可能な URL（`isShownBy`）と目録や閲覧（ビューア）ページ URL（`isShownAt`）の両方が望ましい
- 全てのデジタル化オブジェクトの権利記述（`rights`）
- 各リソースの恒久的 URI

Semantic Enrichment さらに、提供メタデータを元にシソーラスや DBpedia、GeoNames などの LOD との自動マッチングを行ない、リテラル値実体化、外部リンク、多国語ラベルなど意味補強したメタデータを Europeana の Proxy として加え、EDM 全体を構成している（提供メタデータは提供元の Proxy で表現される）。

5.3.2 DPLA

DPLA は、Aggregation を起点として、対象リソース（SourceResource）の基本的な（ほぼ Dublin Core）記述と、EDM による提供情報（isShownAt、provider など）という比較的シンプルなメタデータ構造を採っている（図 14）。

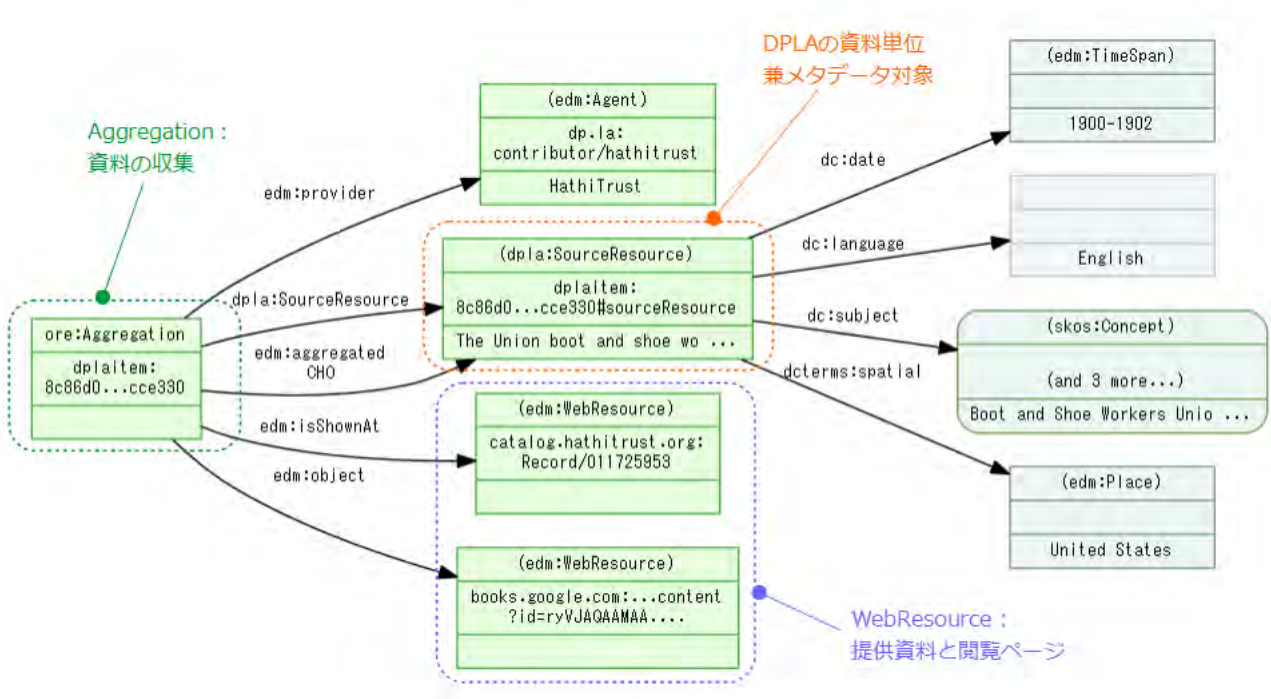

図 14: DPLA のモデル。この他に元データを originalRecord として保持／提供する

受入データと補強　提供元からのデータは Dublin Core、MODS、MARC XML などそれぞれのものを受け入れており、DPLA においてデータ正規化などの補強を加えた上で記述メタデータとしている。補強内容としてあげられているものは

- 名前や地名などに対応する、VIAF、GeoNames などの LOD で普及している URI を加える
  - DPLA は提供データをこうした典拠と照合するためのサービスを開発している
- データ値からの不要なスペースや区切り記号の除去
- 日付フォーマットの標準化

などがある。これらの補強データは、元データとは別に保持される。

必須提供データ　提供データにおいて必須とされているのは次の項目。

- Aggregation として扱う `isShownAt`（閲覧ページ URL）、`preview`（サムネイル URL）、`provider` および `rights`
- SourceResource に記述する `title` および `rights`

- ウェブリソース、デジタル化オブジェクトの rights

さらに SourceResource の記述メタデータとして isPartOf（コレクション）、creator、date、description、format、language、spatial、publisher、type が推奨となっている。

データ受入フォーマットは、OAI-PMH が最も多いが、TSV や XML も受け入れている（更新頻度についての記述はない）。

分野横断統合プラットフォームのメタデータフォーマット仕様(案)

2018-03-26

利用者タスクに対応して整理した分野横断モデルのデータ（共通アーカイブ情報）を記述するためのプロパティを定義する。

利用者の発見タスクを容易にするためには、プロパティ項目は単純であることが望ましい。一方で識別や選択タスクのためには、プロパティは適切な詳細度が必要である。そこで、(1)基本的な記述には単純なプロパティを用いるとともに、(2)「いつ」「どこで」「だれが」（および上位コンテンツ）について提供データの詳細を構造化記述するプロパティを併用する。(3)またコンテンツの提供情報とソースデータ情報の関係を整理し、それぞれを構造化する。

以下の見出しに挙げるプロパティは、共通アーカイブ情報を主語にした記述に用いる。接頭辞rdf:、rdfs:、schema:はそれぞれRDF、RDFスキーマ、Schema.orgの名前空間、v:は本仕様案で独自に定義する語彙を表す。

1 基本記述プロパティ

データモデルの予備知識無しに利用できるシンプルな記述のため、Schema.org語彙を中心にした基本記述プロパティを定める。

1.1 基本的な型と識別

1.1.1 rdf:type

説明	文書、絵図、音声、工芸などのコンテンツの基本区分に相当する基本情報。共通アーカイブ情報型のサブクラスとして扱う。
値	URI（別途、分野横断統合プラットフォームで定義する基本タイプを用いる）

1.1.2 rdfs:label

説明	一覧などに表示しコンテンツを識別する名前。提供されない場合はIDなどから生成する。言語タグは用いない。
値	文字列

1.1.3 schema:name

説明	タイトル、別名、読みなど検索対象とする名前。読み、英語名称は言語タグで区別する。同じ言語タグの名前が複数ある場合、リテラル値がrdfs:labelと一致しないものは「別名」に相当する。
値	文字列（必要に応じて言語タグ付き）

1.1.4 schema:identifier

説明	ISBNなど広く共有されている識別子
値	文字列（導入句、もしくはデータ型付き。あるいはURN）

※RDFグラフにおいては、共通情報に対応する「コンテンツ自身」は共通情報URIに#workを加えて識別し、そこからfoaf:isPrimaryTopicOf（※要検討）で共通情報と関連付ける（項目の複雑化を避けるため当面は共通情報のプロパティとはしない）。

1.2 「いつ」「どこで」「だれが」の汎用記述

次のプロパティは制作、発行などの役割の違いを区別しない汎用プロパティとし、役割を示す構造化記述とセットで用いる。

1.2.1 schema:contributor

説明	コンテンツに寄与した人／組織。出演者等も含む。creator｜publisherを構造化記述：寄与者[関係=制作|出版]のショートカットとして記述した場合は、このcontributorは省略する。
値	URI（v:contributionのschema:agent値と一致）

1.2.2 schema:temporal

説明	コンテンツの制作、出版、内容などに関する時間。時間の意味の違いは構造化プロパティで判断する。
値	URI（v:temporalのv:temporalValue値と一致）

1.2.3 schema:spatial

説明	コンテンツの制作、出版、内容などに関する場所。場所の意味の違いは構造化プロパティで判断する。
値	URI（v:spatialのv:spatialValue値と一致）

1.3 慣用的なプロパティ

「だれが」「いつ」について、多くのスキーマで用いられる項目（役割）は個別プロパティを用意し、可能な範囲で併記するものとする。

1.3.1 schema:creator

説明	作品制作の中心となった人／組織。構造化記述：寄与者[関係=制作]のショートカットとして位置づける。
値	URI（v:contributionのschema:agent値と一致）

1.3.2 schema:publisher

説明	出版者、製作会社、配給など。構造化記述：寄与者[関係=出版]相当のショートカットとして位置づける。
値	URI（v:contributionのschema:agent値と一致）

1.3.3 schema:dateCreated

説明	制作時の時間リテラル。構造化記述：時間[関係=制作]のリテラル値。
値	文字列

1.3.4 schema:datePublished

説明	出版・発行時の時間リテラル。構造化記述：時間[関係=出版]相当のリテラル値。
値	文字列

1.4 識別と選択

次のプロパティは主として識別、選択のために用い、構造化記述の対象としない（上位コンテンツは、コンテンツ内ページなど部分指定が必要な場合は構造化記述と併用しても良い）。

1.4.1 schema:about

説明	主題、分類、区分（キーワード含む）。主題や内容のジャンル、指定区分など、共通認識があり検索結果の絞込などに利用できる用語
値	URI（典拠URI、もしくは特別名前空間によるキーワードURI）

1.4.2 schema:inLanguage

説明	作品の記述言語
値	URI（id.loc.govによるISO639-2のURI）

1.4.3 schema:image

説明	コンテンツの特徴を確認するための画像。提供元の画像とは別に、統合プラットフォームが一定サイズの画像を複製して保管する。
値	URI

※Europeanaでも詳細情報ページ用のサムネイル（400px幅）は独自に保存している模様。

1.4.4 schema:description

説明	簡潔な説明文ほか、個別項目に収録できないが有益な情報
値	文字列（元項目名を導入句として加えた値）

1.4.5 schema:isPartOf

説明	上位コンテンツ、コレクションなどへのリンク
値	URI（v:partOfのv:source値と一致）

2 構造化記述プロパティ

基本プロパティで記述した内容について、より詳細な情報を提供するため、並行して構造化記述のプロパティを用意する。構造化記述プロパティの値は空白ノードとし、その空白ノードに「構造化値」で示すプロパティを記述する。relationTypeの値は、「制作」「発行」など標準的なものはあらかじめ定義して用い、同じ役割の「いつ」「どこで」「だれが」を結び付けられるようにする。

なお、この構造化記述および次の提供・ソース情報記述については、参考としてTurtle構文によるプロパティの用例を示す。値の記述に用いている接頭辞は、便宜的に付与した仮のものである。

2.1 v:contribution

説明	寄与関係：制作に関与した人／組織の関係	
構造化値	v:relationType	寄与のタイプ。領域を超えて一般化できる制作、編集、翻訳など〔URI〕
	schema:agent	寄与した人物・団体〔典拠URI／IDなどから生成するURI〕
	schema:description	領域固有の役割名、キャストの配役など補足記述〔文字列〕
	schema:url	関連リンク〔URI〕

※人物・団体の名前は典拠URIから辿ることができるようにする。

```
:id3919
    schema:name "朝日のあたる家"@ja ;
    schema:creator dbpedia:太田隆文 ;
    v:contribution [
        schema:agent dbpedia:太田隆文 ;
        v:relationType d:制作 ;
        schema:description "監督"
    ] ...
```

2.2 v:temporal

説明	時間関係：作成、公開、発見、主題などコンテンツに関する時間	
構造化値	v:relationType	時間関係のタイプ。制作、主題など〔URI〕
	v:temporalValue	時間関係における時間（範囲）を示す〔西暦年単位で正規化したURI〕
	v:era	時間関係における時代を示す〔時間オントロジーで定義するURI〕
	schema:description	月日ほか補足情報〔文字列〕

```
:id113315
    schema:name "絹本淡彩鷹見泉石像〈渡辺崋山筆／〉"@ja ;
    schema:temporal td:1837 ;
    v:temporal [
        v:era te:江戸時代 ;
        v:temporalValue td:1837
    ]
    ...
td:1837
    schema:name "天保8年" ;
    ...
te:江戸時代
    schema:name: "江戸時代", "1615〜1869年" ;
    owl:sameAs td:1615_1869 ;
    ...
```

2.3 v:spatial

説明	場所関係：作成、公開、主題などコンテンツに関係する場所・地域	
構造化値	v:relationType	場所関係のタイプ。制作、主題など〔URI〕
	v:spatialValue	場所関係における場所を示す〔国／都道府県レベルで正規化したURI〕
	schema:description	より詳細な地名・住所など補足記述〔文字列〕
	schema:geo	緯度経度などの位置情報〔schema:GeoCoordinates〕

```
:id4652721
    schema:name: "Phtheochroa pistrinana"@en, "セジロホソハマキ"@ja ;
    schema:spatial d:静岡県 ;
    v:spatial [
        v:spatialValue d:静岡県 ;
        v:relationType d:採集 ;
        schema:description "採集場所：都道府県（日本語）:静岡県",
            "採集場所：詳細（日本語）:梨本 南伊豆"
    ] ...
```

2.4 v:partOf

説明	上位コンテンツとの関係（掲載誌のページ）などを記述する	
構造化値	v:relationType	上位コンテンツとの関係のタイプ。掲載など〔URI〕
	v:source	上位コンテンツ関係における上位コンテンツ本体を示す〔URI〕
	v:selector	上位コンテンツ内の特定部分を示す〔文字列〕
	schema:description	補足記述〔文字列〕

※上位コンテンツの一部（ページ）が対象であることを示す必要がある場合に構造化する。

```
:idI5983376
    schema:name "ひらがな日本美術史(82)ルネサンスになる前のもの--喜多川歌麿筆「婦人相學十躰 ポッピンを吹く娘」"@ja ;
    schema:isPartOf m:芸術新潮200112 ;
    v:partOf [
        v:source m:芸術新潮200112 ;
        v:relationType d:掲載 ;
        v:selector "掲載ページ:118〜124"
    ]
    ...
m:芸術新潮200112
    schema:name: "芸術新潮2001年12月号",
    schema:description "巻:52", "号:12", "通号:624" ;
    schema:isPartOf dbpedia:芸術新潮 ;
    ...
```

3 提供・ソース情報記述プロパティ

コンテンツの提供・アクセス情報とソースデータ情報の関係を構造化して記述する。

3.1 v:accessInfo

説明	コンテンツの提供・アクセスに関する情報	
構造化値	schema:provider	コンテンツ（に関する情報）の提供者〔別テーブルで管理する識別URI〕
	v:contentHolder	コンテンツの保管者（提供者と別の場合）〔URI〕
	schema:url	コンテンツの紹介ページやアクセス情報が記載されたページ〔URI〕
	v:digitalObject	コンテンツのデジタル画像、音声動画など。全てライセンス情報を持つ〔URI〕
	v:contentRights	コンテンツ自身の権利記述〔文字列〕
	v:contentId	提供元が付与する識別子〔文字列〕
	schema:description	補足記述〔文字列〕

※提供者のURIは別テーブルで管理し、提供者の画像のデフォルトライセンスなどを記述する。
url、digitalObjectの値には、さらにそれを主語として画像、音声などの型（rdf:type）を付与する。
可能であればフォーマット（schema:fileFormat）も示したいが、提供情報の範囲では難しいかもしれない。
デジタル画像、音声動画などを主語として、schema:licenseでそのライセンスを記述する。

```
:id12247
    schema:name "色絵唐花福寿文反皿"@ja ;
    v:accessInfo [
        schema:provider dbpedia:九州国立博物館 ;
        schema:description "機関管理番号:G51" ;
        v:digitalObject
            <https://colbase.nich.go.jp/.../103dbc63d6af71b21cf57bae0204b75c.jpg> ;
        schema:url <http://collection.kyuhaku.jp/gallery/16428.html>
    ] ;
    ...
<https://colbase.nich.go.jp/.../103dbc63d6af71b21cf57bae0204b75c.jpg>
    a dcmitypes:Image ;
    schema:license <http://creativecommons.org/license/...> ;
    ...
```

3.2 v:sourceInfo

説明	収集したデータ及びその提供元に関する情報	
構造化値	schema:provider	ソース情報の提供者（アグリゲータ）〔別テーブルで管理する識別URI〕
	schema:url	集約元情報の公開ページ〔URI〕
	v:sourceData	プラットフォームが保持・提供するソースデータ〔URI〕
	rdfs:seeAlso	ソース情報がRDFである場合〔URI〕
	schema:description	補足記述〔文字列〕
	schema:dateModified	収集元データの更新日、もしくは収集日〔日付型〕

※提供者のURIは提供・アクセス情報と同じテーブルで管理する。

```
v:sourceInfo [
    schema:provider <https://colbase.nich.go.jp/> ;
    schema:url <https://colbase.nich.go.jp/collectionItems/view/.../16428> ;
    schema:dateModified "2018-03-01" ;
    v:sourceData <https://.../colbase-12247.json>
]
```