資料2

連携するデータ：Linked Data Hub

於:メタデータのオープン化等検討WG
2016-10-11 神崎正英

1 本日の視点

■ Linked Data（LD）とデータモデル

- 柔軟なグラフモデルとURIのグローバルな識別により、機関・分野を超えた連動が可能
- 識別のためのURIをリンクに用いることでデータのウェブを生み出す
- 提供されるのが表形式データ(CSVなど)であってもグラフモデルに変換できる

■ データのウェブ：分散とリンク

- URIによる識別により、データは分散して記述しても併合できる
- リンクによりデータのつながりをアプリケーションが辿っていける
- 名前＝識別子の共有がデータ連携の基本。ただし対象の捉え方は視点により異なる

■ LDのハブとしてのアーカイブ

- データ、そして資料の有益なつながりのために識別子のハブが重要
- メタデータは、求められる／提供する機能の要求分析にもとづき、無理のない収集を
- メタデータの利活用というよりも、資料(コンテンツ)自身とそれに関連する情報を結びつけ活性化する、その媒介となること

2 表形式データ

■ **多くのデータが表形式**

- エクセル、CSV
  - 一つのセルに複数値があることも珍しくない
- 複雑な関係や入れ子関係 → 表の正規化と関係データベース
- 表(列)の意味の共有(スキーマ)が必要。表内部でのみの識別(主キー)

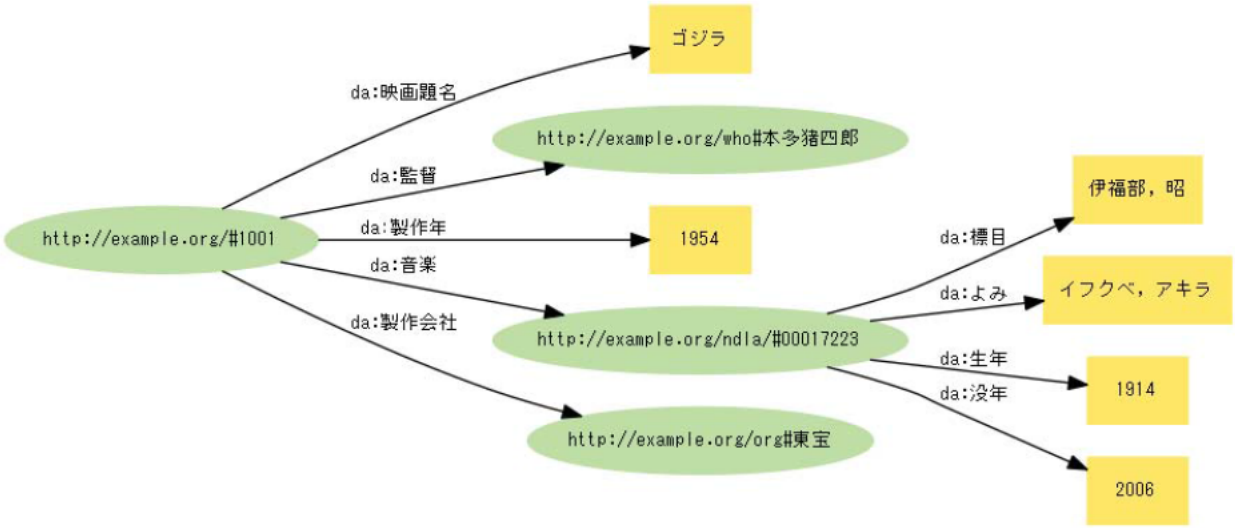
No	映画題名	監督	製作年	製作会社	音楽
1001	ゴジラ	本多猪四郎	1954	東宝	伊福部昭
1002	ゴジラの逆襲	小田基義	1955	東宝	佐藤勝
1003	シン・ゴジラ	庵野秀明 樋口真嗣	2016	東宝	伊福部昭 鷺巣詩郎

ID	標目	よみ	生年	没年
00017223	伊福部, 昭	イフクベ, アキラ	1914	2006

3 グラフモデルとURI

■ **柔軟なグラフモデル**

- 表も含め多様な情報形態を表現できる
- あらかじめスキーマを用意する必要がない → 異なる形の(異分野の)情報を容易に追加できる
- ノードやアークの識別にURIを用いる → どこで作られたグラフでもつながる
- CSVなどの表データをグラフモデルに変換するための情報記述CSVW[1]も標準化
  - データ提供側でグラフモデルを用意するのが難しければ、収集側で変換してもよい



4 URI：識別とリンク

■ 識別子Uniform Resource Identifier

- **Uniform**: allows different types of resource identifiers to be used in the same context, even when the mechanisms used to access those resources may differ. (RFC 3986[2])
- **Resource**: A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources.
- **Identifier**: in many cases, URIs are used to denote resources without any intention that they be accessed.

♣ URIは人や書物、概念なども「リソース」として識別する → 必ずしもネットワーク上でアクセスできる(リンクする)ものとは限らない

- Transcription: A URI often has to be remembered by people, and it is easier for people to remember a URI when it consists of meaningful or familiar components. (1.2.1) → 不必要に長く分かりにくいURIはできれば避けるほうが望ましい

■ URLとハイパーリンク

- ウェブは文書間のリンクによって発展
- HTMLのa要素、link要素に記述されるURIは、ほとんどの場合アプリケーションが辿っていける(リンクする)＝Uniform Resource Locator
- httpスキームによるURIは、少なくともサーバーとの対話方法が了解されている → 望んだ応答が得られるかどうかは別にして、アクションは可能

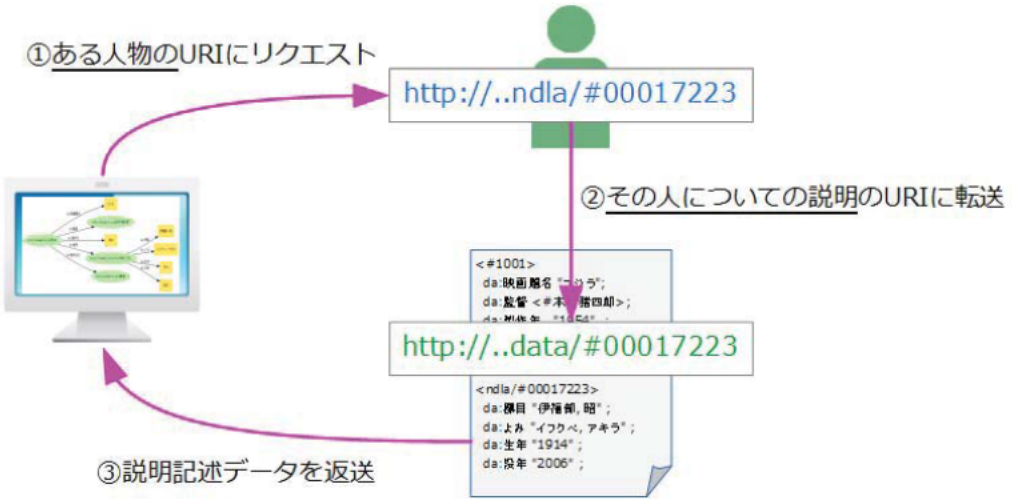
5 リンクするデータ

■ TimBLの4つのルール[3]（2006）

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information (using the standards).
- Include links to other URIs. so that they can discover more things.

■ データのURIを識別だけでなくリンクにも用いる

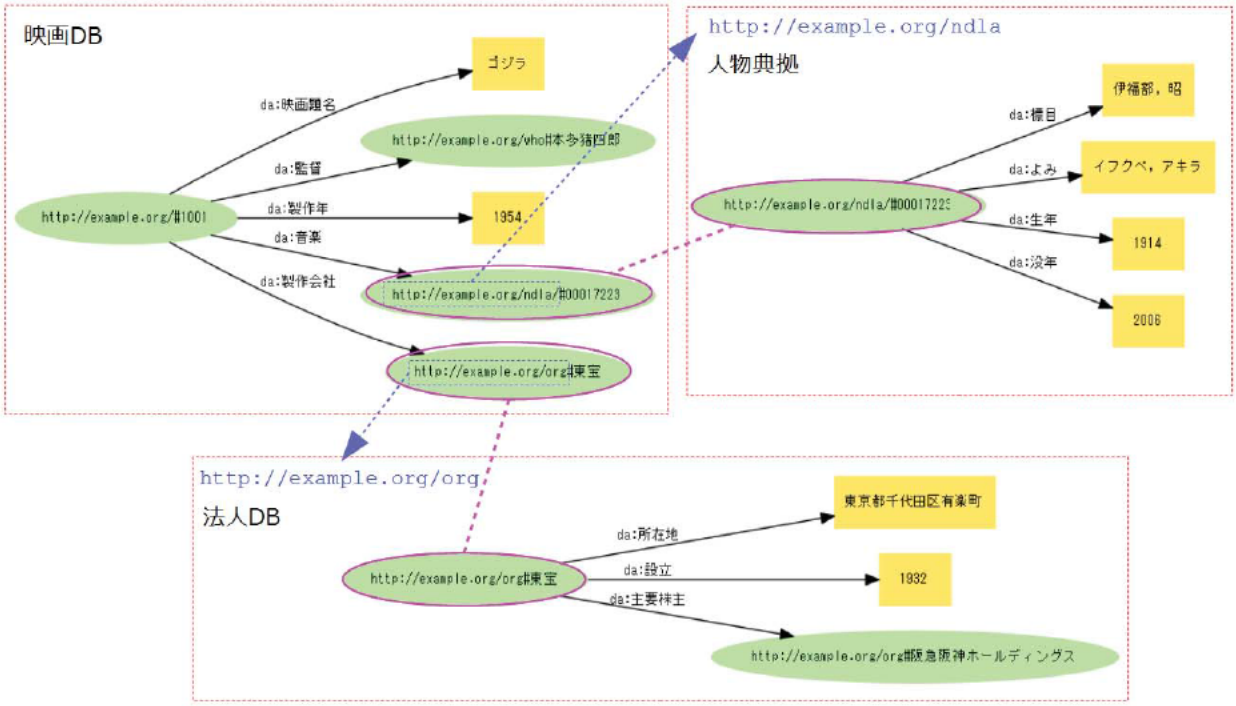
- データ(グラフ)が文書と同じようにリンクする → **データのウェブ**
  - 閉じたDB内のデータ → ウェブとなってつながるデータへ
- 人間を識別するURIにリンク？ → 識別子としてのURIにアクセスされたら、関連情報URIに転送する



6 データの分散とLD

■ **データの分散が可能なLD**

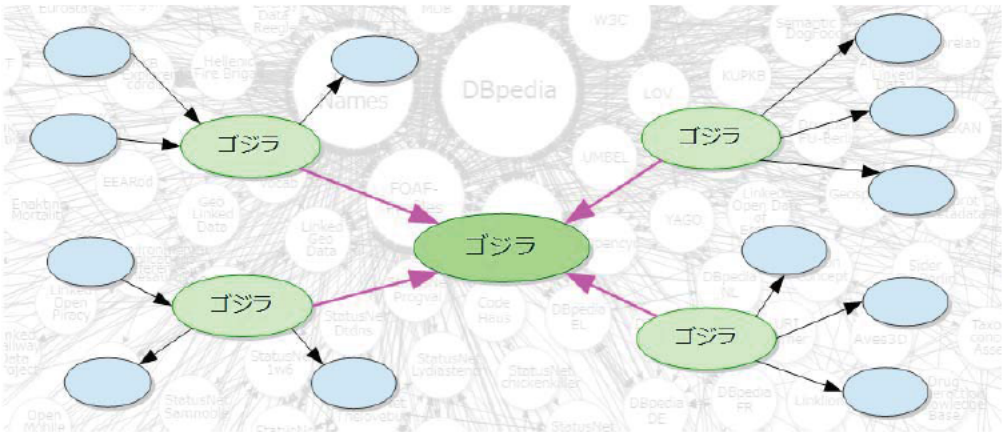
- 共通のURIがあれば、どこにあるデータでもつながる（リンクを辿って併合できる）
- アプリケーションがリンクを辿って関連情報を取得できる（自前ですべての情報を持つ必要がない）
- 無理にフォーマットやデータ構造を標準化しなくてもデータは連携できる
  - 明確な目的があるものはマッピング、それ以外はそのままで構わない



7 LDのハブ（LDH）と典拠データ

■ **分散したデータをつなぐためのハブ**

- 情報がそれぞれ個別につながるだけでは連携効果が出にくい
- 同じものごとを指し示す**名前の共有**によって関連情報が集約される
- リンクするデータのハブ（以下**LDH**）としてのDBpediaの重要な価値は、共通の名前を提供しているところに

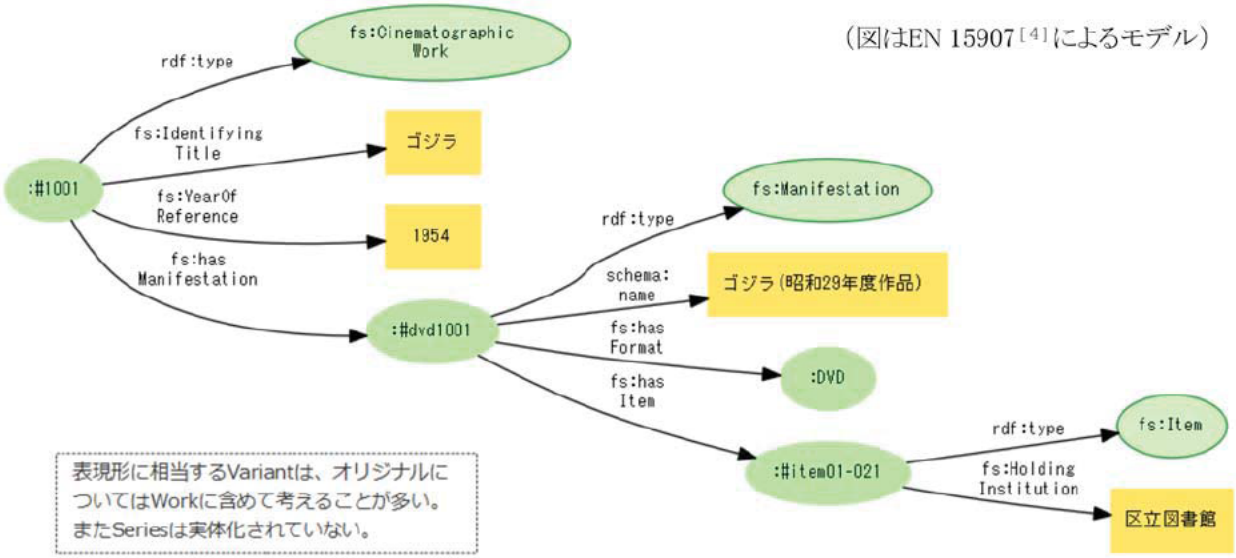


8 「おなじもの」について話すこと

■ 異なる実体レベル

- ゴジラの1954年版の話なのか、シリーズなのか、シン・ゴジラDVDのディレクターズ・カットなのか
  - 映画を扱う領域ではこの違いが重要
  - LDとして外部から利用するときこうした区別は必要なのか、曖昧であるほうが便利なのか

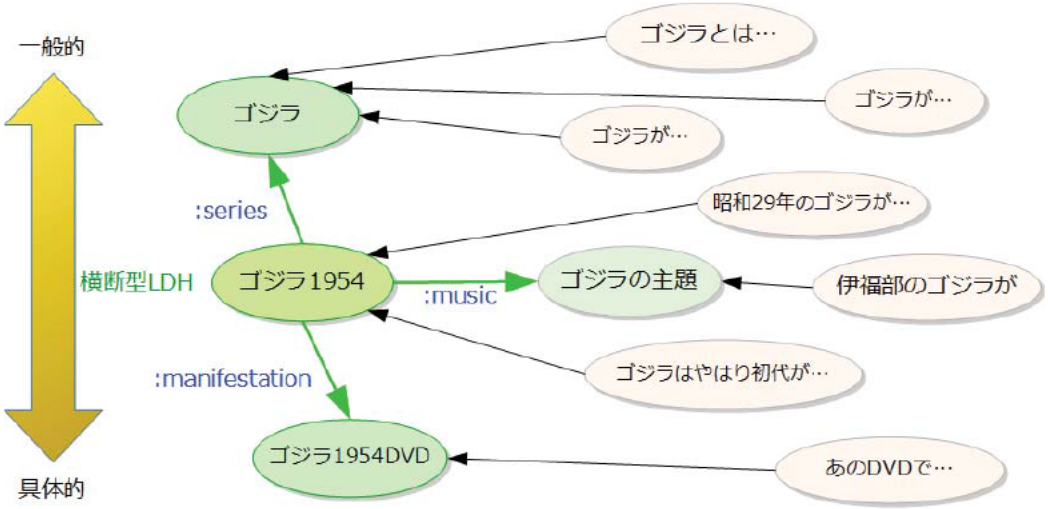
（図はEN 15907[4]によるモデル）



9 LDの識別子が指しているものは何か

■ LDHの役割分担

- 一般的な（分野を特定しない）名前のハブ（Wikipedia/DBpedia、NDLAなど）
  - たとえばhttp://ja.dbpedia.org/resource/ゴジラは映画シリーズと怪獣の両方
  - http://ja.dbpedia.org/resource/ひまわり_(絵画) はゴッホの一連の作品
- 具体的なものを示す分野別LDHと、橋渡しとしての横断型LDH
  - 個別の「版」や派生作品などを指し示すURIの定義とメタデータ管理は、分野に委ねる
  - 横断型LDHはこれらの個別URIと一般名URIとの橋渡しや発見をサポートする



10 メタデータ収集のための分析

■ 収集する情報範囲の検討

- LDHに求められる機能（要求分析）
  - 利用する視点で、何のためにどんな情報が必要なのかを整理
- LDHに必須の情報＝分野を超えた共通収集項目
  - メタデータが有効に機能するために最小限必要な情報の定義
  - たとえば2011年の総務省メタデータ共有ガイドライン[5]では、ラベルが優先度A、作者、日時、位置情報がB、キーワードが「可能ならば」としてB
- それ以上の詳細情報が提供されれば、機能に照らし公開。提供レベル差があってもかまわない
- プロバイダの負担やインセンティブを考慮。場合によってはCSVでの収集（CSVWで変換）も

■ FRBRでの利用者タスク分析

- カタログ利用者の目的は多様。その分野にどんな情報／資料があるのか、特定の資料／対象についての情報があるか、ある資料はどんな形態で／条件で利用可能なのか、など（FRBR[6] 2.2 Scope）
- → Find, Identify, Select, Obtainの4タスク

11 4タスクから見たLDH

■ Find：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）
  - たとえば、分野ごとに異なる構造の情報は、単純なテキスト情報項目に集約するなど
- 作者、時間（時代）、場所（地域）など：統制語彙や外部LDHリンク
  - リッチなAPIの提供、ポータルでの地図マッピングといった機能を求めるなら重要
- LDHが収集・保持すべきメタデータは求める検索の種類や精度によって違ってくる
  - たとえば地域活性化のために場所メタデータを重視するなど、目的からの選定

■ Identify：識別

- 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
  - ラベル、サムネイル、ソース／来歴など
- URIを持たない資料や複数存在する資料にもLDHとしてのURIを付与する
  - 回顧展図録などから全作品のURIを（資料所蔵の有無にかかわらず）一括付与するのも一案
  - プロバイダの人名、地名情報を集結する共同識別ハブ（Cooperative Identities Hub[7]）、その延長に資料の識別子のハブとしてのLDH。
- Wikidata、Wikipediaなどに識別リンクを提供する（LDクラウドとの双方向リンク）

■ Select：選択（記述）

- 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
  - 概要の記述、ライセンスなど。やはり機能要件により異なってくる

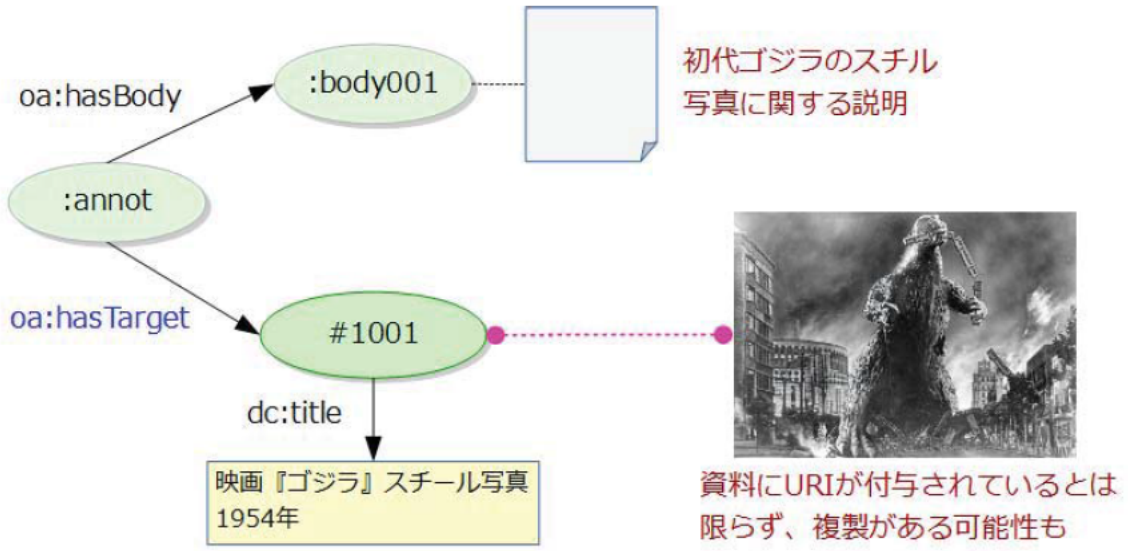
■ Obtain：取得

- 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
  - 資料のURI。場合によっては別手段の取得情報も
- 取得のために特化したメタデータ、たとえばIIIFマニフェストURIの集約など

12 利用の視点：識別子

■ 適切な引用・参照情報／URI体系による参照

- 提供元に恒久URIがない資料も含め、**一貫してURIで言及・リンク**できる
  - （データを正規化して）適切に整理された引用情報としても
- **その資料についての**研究、文献、関連情報などを集約する
  - 資料自身の情報記述（主語URI）だけでなく、資料への言及（目的語URI）の面が重要
  - 論文に引用文献を記載するのと同じく、対象資料URIの記載があれば（いわば資料のDOI）
  - 「そこに求められるのは、総体としての情報や、文化財を取りまく記憶や歴史を重層的に集積した姿」[8]
- 例：注釈（Web Annotation[9]モデル）のtargetを言及するためのURIとして利用



13 利用の視点：API

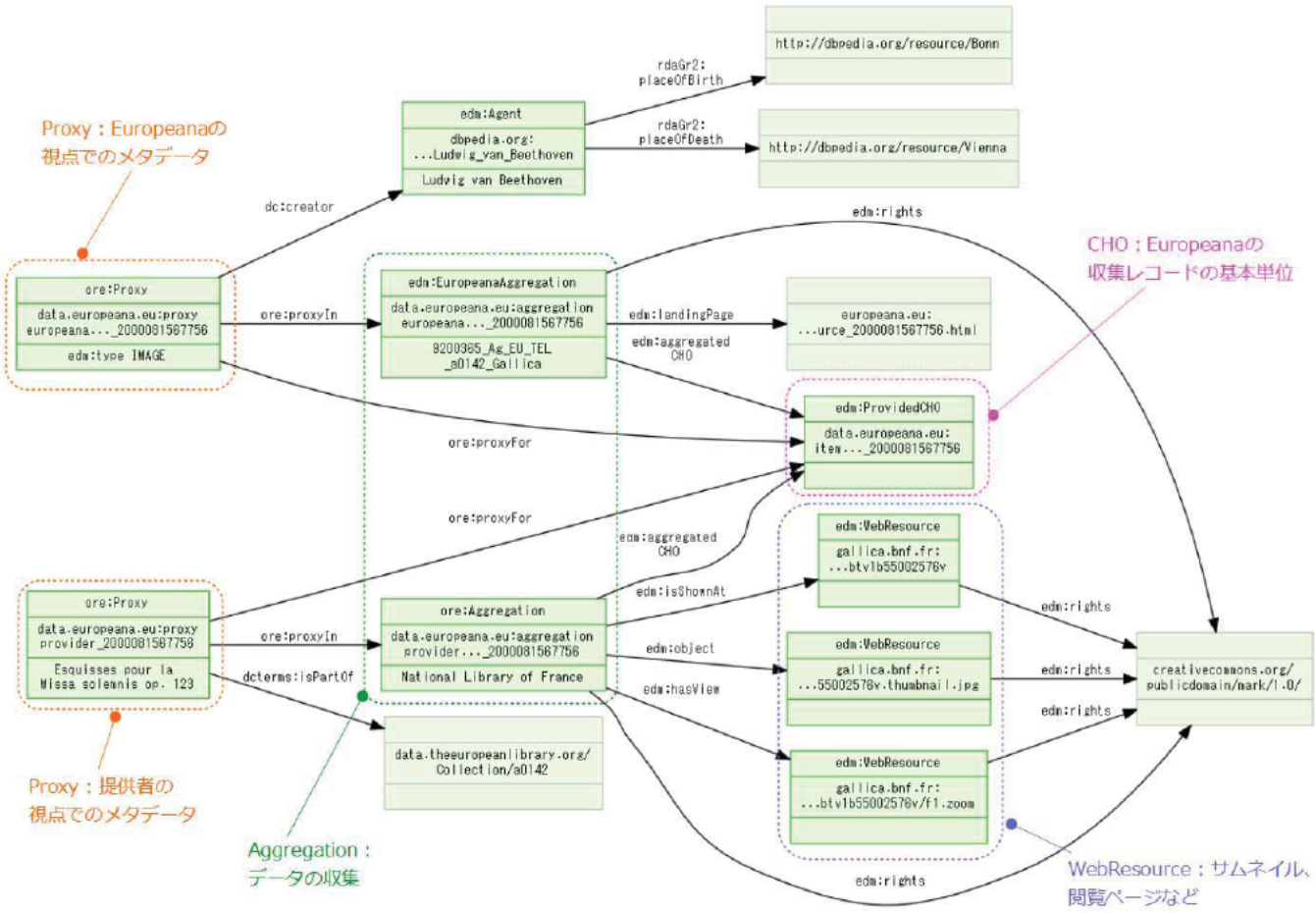
■ 先行事例のAPI

- 基本的な検索：自由キーワード検索、フィールドを指定したテキスト検索（DPLA:`title`, `contributor`など、Europeana:`query=who:...`など）
- **時間軸**の検索：タイムラインに表示（DPLA:`date`, `temporal`、Europeana:ファセット `qf=YEAR:....`）
- **空間軸**の検索：地図へのマッピング（DPLA:`spatial`プロパティ、Europeana:ファセット `qf=where::....`、緯度経度`query=pl_wgs84_pos_lat:...`）
- **特徴**指定（Europeana:`reusability`, `colourpalette`など）

■ APIの提供

- 時間、空間、特徴などでの検索を可能にするためには、それに応じたメタデータ項目の整備が必要
- SPARQLエンドポイント：任意のクエリで検索（DPLAは未提供、Europeanaは2016年8月で一旦停止）
- 利用しやすくするためには、検索・選択された識別子をIIIFマニフェストとして返す方法も
  - DPLAでも、画像の扱い方がポータル、APIともにユーザに分かりにくい"last mile"問題解消にIIIF提供を模索[10]
- ショーケースとしての「ポータル」
  - テーマに沿ってAPIを用い「展示」を実施する
  - EuropeanaもDPLAもポータルサイトは内部的にAPIを用いて構築されている

14 （参考）EDMについて



■ EuropeanaでのEDM[11]データ構造

- **ProvidedCHO**：収集レコードの基本単位
- **WebResource**：個別資料の具体的リソース（サムネイル／閲覧ページなど）
- **Aggregation**：提供者への収集アクションの結果
- **Proxy**：提供者ごとの作品情報を体現するための「代理」リソース
  - Aggregation、Proxyでは、Europeana自身もそのメタデータなどの「提供者」。可能なものは正規化や外部リンクなど付加価値を加えている。

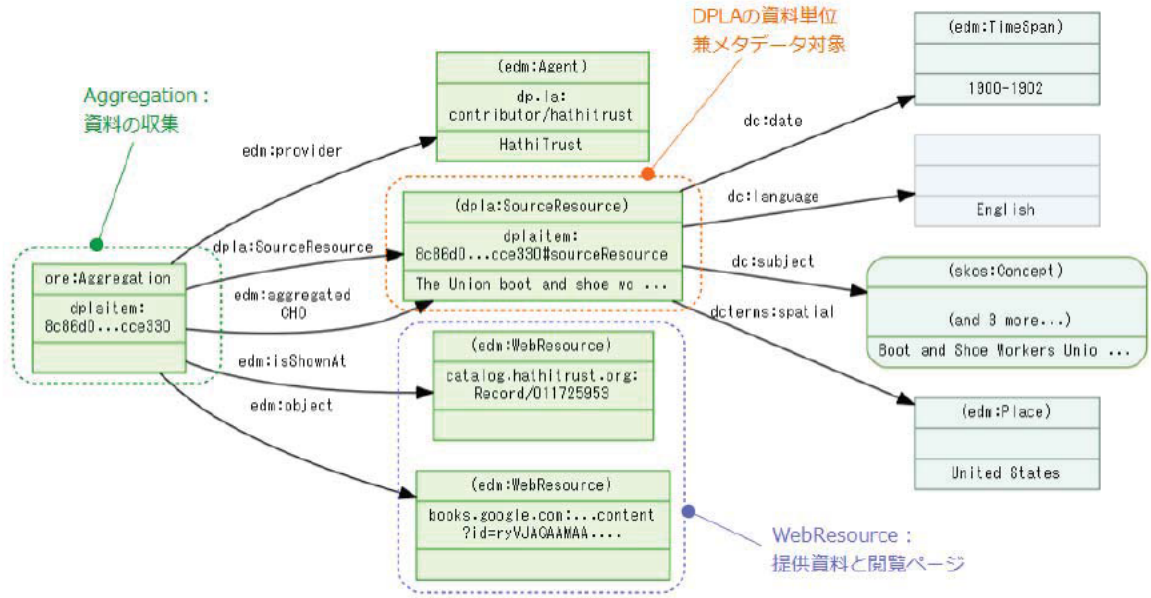
■ 資料（Proxy）のメタデータ項目

- プロバイダProxyは、Dublin Coreの`title`, `description`, `date`, `creator`, `source`, `subject`, `type`など（例ではすべてリテラル）
  - 提供されたものをほぼそのまま公開している模様（資料によって項目はかなり異なる）
- Europeana Proxyは、（正規化ができたものは？）DBpedia、GeoNamesなどを介して外部LDにリンク。生没年なども（独自に？）付加

■ URIの使いやすさ

- 複数の実体URIの使い分けはEDMを理解していないと難しい（どれを注釈のtargetにすれば？）
- URIの構造規則は明確で分かりやすいが、ID部分が長く機械的で、「名前」としては使いにくい（transcribeしにくい）

15 （参考）DPLA MAPについて



■ DPLA MAP[12]のデータ構造

- EDMの基本型による。ProvidedCHOを中心リソースとし、WebResourceで具体的な資源を示す。
- Aggregationは、Europeanaとは違ってアーカイブ自身のものと分けずに1本化している。
- Proxyを置かずレコード（SourceResource）に直接メタデータを付与している。

■ メタデータ項目とURI

- メタデータ項目は、EuropeanaのプロバイダProxyとほぼ同様で、Dublin Coreの`title`, `description`, `date`, `contributor`, `format`, `subject`, `type`など、および地域（`dct:spatial`）、時代（`dct:temporal`）
- 主題、地域、時代、日付（時期）、言語を構造化して記述している
  - enrichmentでURIを与えLD化するとしているが、JSON-LDで得られるものは今のところ空白ノード＋ラベルで、識別子もなくリンクもしない。
- URI構造はシンプルで明快だが、UUIDによるID付与はtranscribeしやすいとはいえない。

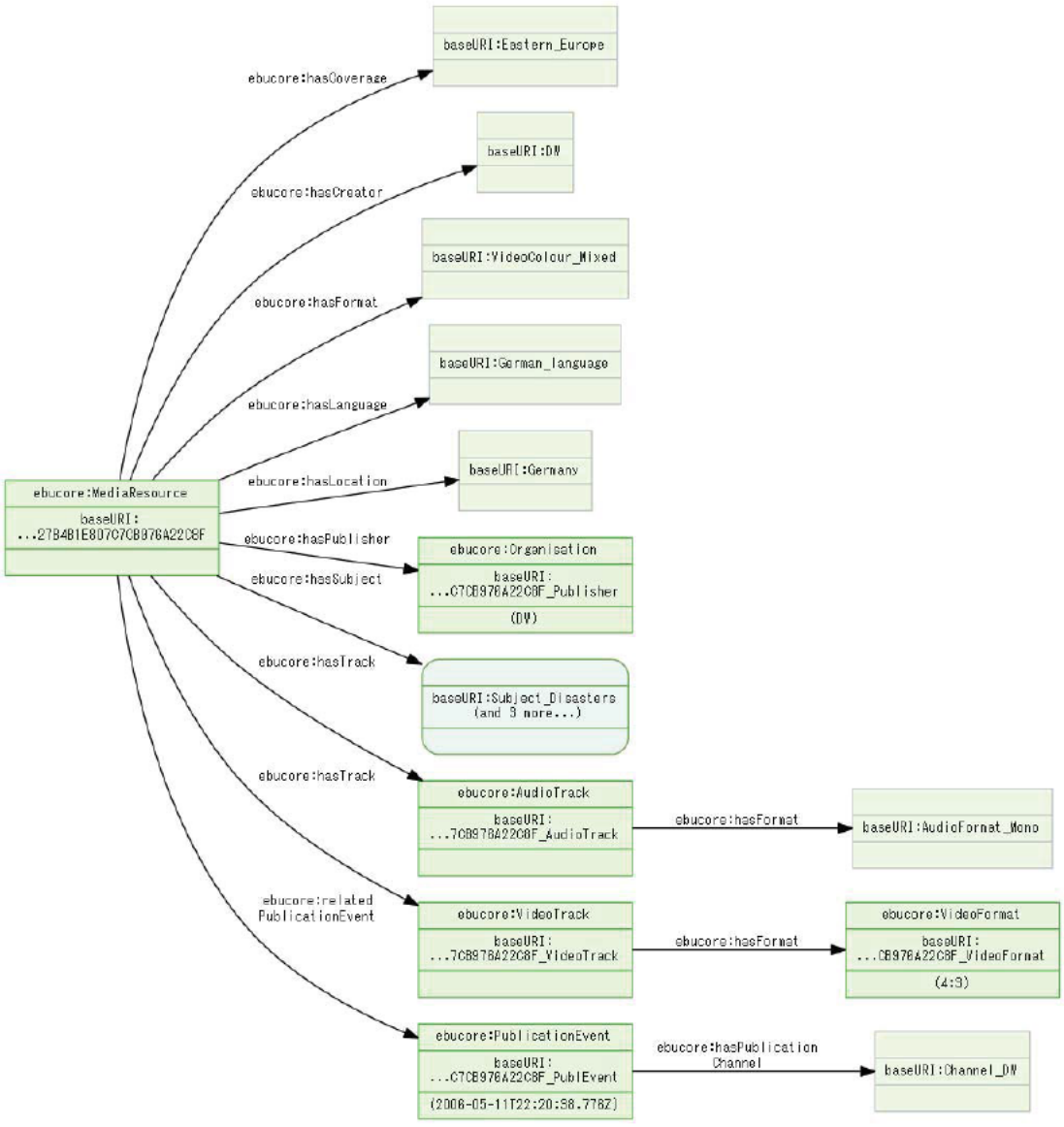
16 （参考）映像資料のデジタルアーカイブについて

■ 映画表現の基本メタデータ

- 映画のグローバルな**識別子登録システム**としてのEIDR[13]、ISAN[14]
  - アーカイブというよりは映画関連の流通管理や権利を登録するレポジトリとしての利用
- 動画目録のためのメタデータ基準としてFIAF目録マニュアル[15]、および前述のEN 15907
  - アーカイブ内部での記述には利用（を検討）されていると思われるが、この形でのデータが公開されている事例は見当たらない。
- どのモデルも階層的な実体レベルを構成している。国ごとのメタデータの違い（権利含む）を重視。また作成、配給などをそれぞれイベントとして扱う志向もある。
- カタログページHTMLに埋め込む形でSchema.orgの語彙を用いる例が複数あり、注目される

■ アーカイブの公開と連携

- 20程度の映画アーカイブを調査した範囲では、ウェブ画面でのカタログ的情報提供が中心で、連携可能なデータは図書館系のみ
- EBU (European Broadcasting Union) の定義するメタデータ仕様EBUcore[16]が、Europeanaのアグリゲータであるの EU Screenで中間標準に用いられている（下図）。
  - 134のクラスと500以上のプロパティで構成されるラジオ・テレビ番組記述の仕様で、番組の内容とそのフォーマットを記述する



17 参照先

■ 参照したリソース

1. Model for Tabular Data and Metadata on the Web, 2015-12-17, W3C Recommendation <https://www.w3.org/TR/tabular-data-model/>
2. RFC 3986 Uniform Resource Identifier (URI): Generic Syntax, by T. Berners-Lee, R. Fielding, L. Masinter, 2005-01 <https://tools.ietf.org/html/rfc3986>
3. Linked Data - Design Issues, by Tim Berners-Lee, 2006-07-27, rev.2009-06-18 <https://www.w3.org/DesignIssues/LinkedData html>
4. EN 15907 - Film identification, 2011-04 <http://filmstandards.org/fsc/index.php/EN_15907>
5. メタデータ情報共有のためのガイドライン, 2011-03-28 <http://www.mi3.or.jp/item/A03.pdf>
6. Functional Requirements for Bibliographic Records -- Final Report, by International Federation of Library Associations, 1998 <http://archive.ifla.org/VII/s13/frbr/frbr1 htm#2.2>
7. Networking Names, by Karen Smith-Yoshimura, 2009, OCLC Research <http://www.oclc.org/content/dam/research/publications/library/2009/2009-05.pdf>
8. 「コレクションデータベースから文化財アーカイブズへ」『MLA連携の現状・課題・将来』p.145, 宮崎幹子, 2010, ,水谷長志 編著,勉誠出版
9. Web Annotation Data Model, by R. Sanderson e.a., 2016-09-06, W3C Candidate Rec. <https://www.w3.org/TR/annotation-model/>
10. DPLA and the International Image Interoperability Framework, 2016-05-10 <https://dp.la/info/2016/05/10/dpla-and-iiif/>
11. Definition of the Europeana Data Model v5.2.7, 2016-04-25 <http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Definition_v5.2.7_042016.pdf>
12. DPLA Metadata Application Profile, version 4.0, 2015-03-04 <https://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf>
13. EIDR SYSTEM VERSION 2.0 Best Practices Guide, by Entertainment ID Registry Association, 2015-09-20 <http://eidr.org/documents/EIDR_2.0_Best_Practices.pdf>
14. ISAN (International Standard Audiovisual Number) <http://www.isan.org/>
15. FIAF Moving Image Cataloguing Manual, by International Federation of Film Archives, 2016-04 <http://www.fiafnet.org/pages/E-Resources/Cataloguing-Manual html>
16. EBU Technology & Innovation - Metadata Specifications <https://tech.ebu.ch/MetadataEbuCore>