Chapter 3

Three Case Studies in Linked Open Data

Abstract

In chapter 3 of Library Technology Reports *(vol. 49, no. 5) "Library Linked Data: Research and Adoption" we explore three current LOD-aware services that focus on serving cultural heritage and memory communities: the Europeana digital library, museum, and archive; the Digital Public Library of America; and the BIBFRAME initiative, guided by Library of Congress. The chapter explores the platforms according to the building block model of chapter 2 and uses a case study approach. The examination of high-level similarities and differences reveals a development path for LODLAM services.*

Europeana
http://europeana.eu

Digital Public Library of America (DPLA)
http://dp.la

BIBFRAME
http://bibframe.org

Introduction

In this chapter, we will explore three current LOD-aware services. These services are the Europeana digital library, museum, and archive; the Digital Public Library of America (DPLA); and the BIBFRAME initiative being guided by the Library of Congress. These platforms were selected given the current activity in the community and because of their use of LOD/LOV techniques to aggregate and publish data. As with the LOD/LOV exploration in chapter 2, this chapter uses our metadata building blocks model (data model, content rules, metadata schema, data serialization, and data exchange) as a guideline for exploring the platforms.

This chapter uses a case study approach that is grounded in a content analysis of the publications and metadata specifications of the services. This focus impacts the data analysis in two ways. First, because these systems are all technically different, they do not give equal weight to each of our five building blocks. Second, because these systems are continuing to develop, these case studies can be considered as snapshots of the services and not necessarily representative of what they will look like in the months and years following this issue.

Our case studies each consist of six sections. Each case study begins with a metadata specification overview, continues with a discussion of each of the five building blocks, and concludes with a community activities section that discusses the current direction of the service. Following the presentation of case study data, our discussion section compares and contrasts these services by exploring the advantages, challenges, and motivations behind the data choices made in these communities.

BIBFRAME

Metadata Specification Overview

The BIBFRAME initiative grew from a long line of metadata re-envisioning work in the library community that had its first signals in a 2008 report on the future of bibliographic control.[1] These reports helped define the vision of the new bibliographic data exchange standard, and in 2012 the Library of Congress (LoC) partnered with Zepheria to produce a data model that met this vision.[2] The resulting initiative, called BIBFRAME, seeks to translate bibliographic data to a linked data model while also incorporating emerging data standards and models including Functional Requirements for Bibliographic Records (FRBR) and Resource Description and Access (RDA).

Like many of the standards in the LAM community, BIBFRAME seeks to enable cross-domain use and interoperability. BIBFRAME accomplishes this using a linked data framework to describe bibliographic and authority entities as well as relationships among these entities. The model also differentiates between concepts and the physical and digital objects that these concepts describe.[3] The goals and scope of BIBFRAME are quite large, seeking to accommodate "different content models and cataloging rules, exploring new methods of data entry, and evaluating current exchange protocols."[4] The overarching goal of this work is to support a transition to metadata work and services that support engagement and querying of a network of data.[5] The BIBFRAME model is still under active development, and the technical products of this work center largely on tools that facilitate the exploration of BIBFRAME data.

Building Block 1: BIBFRAME Data Model

BIBFRAME is designed using a graph-based data model grounded in RDF using classes and properties developed to represent bibliographic-related entities (e.g., Creative Work, Instance). The data model conforms to our understanding of linked data because it emphasizes both the deconstruction of bibliographic description into discrete and unambiguous statements as well as the use of URIs instead of text-based or literal values. Resources available on the BIBFRAME.org site discuss the idea that the atomization of description will lead to a more flexible platform. In addition, it is clear that the BIBFRAME specification seeks to use URIs from Library of Congress LOV services. An overview of the model as implemented with the core BIBFRAME vocabulary can be found in figure 3.1.

Building Block 2: BIBFRAME Content Rules

Documentation on the BIBFRAME site indicates an intent to use RDA rules as a source for content rules as well as the rules associated with the core data model. Some initial work in aligning BIBFRAME classes and properties with RDA and MARC content has been completed and is available on the BIBFRAME site. The current BIBFRAME documentation makes it clear that the standard is designed as an open interchange format and as such should not prescribe specific cataloging principles as these are often tied to a specific domain.

BIBFRAME Vocabulary Updates
http://bibframe.org/vocab

BIBFRAME Vocabulary example: Work
http://bibframe.org/vocab/Work.html

The BIBFRAME model uses four main entity types to create metadata, Creative Works, Instances, Authorities, and Annotations (figure 3.1). In figure 3.1 authority values are represented with the resources identified by the creator, subject, format, publisher, and publishedAt properties. Figure 3.1 does not include any reference to annotations, but the intended use of annotations in the BIBFRAME model is to support the attachment of holdings and other information to BIBFRAME resources. The model is intentionally generalized to support creation of objects from a number of resource domains, including museum and archival settings. Outside of the rules governing the definition of these four main classes and the properties that connect them, the model does not have built-in content rules but rather defers to external standards and guidelines.

Building Block 3: BIBFRAME Metadata Schema and Vocabularies

The core metadata schema and vocabularies for BIBFRAME are represented in figure 3.1. This figure shows the direct relationship between a Creative Work and an Instance (a Work `hasInstance` Instance) and between an Instance and its descriptive metadata (`publisher`, `publishedAt`, and `format`). These concepts are documented in a number of publications,[6] and there have been discussions on cataloging e-mail lists about the interoperability between FRBR and this Work/Instance model.

While the Work entity is roughly analogous to the FRBR work concept, retaining the abstract focus of the FRBR concept, the BIBFRAME Instance concept can roughly be described as a conflation between Expression and Manifestation. This has been a focus of discussion within the community as it signals a potential shift away from the delineated FRBR model used in RDA.

Authority entities describe real-world and

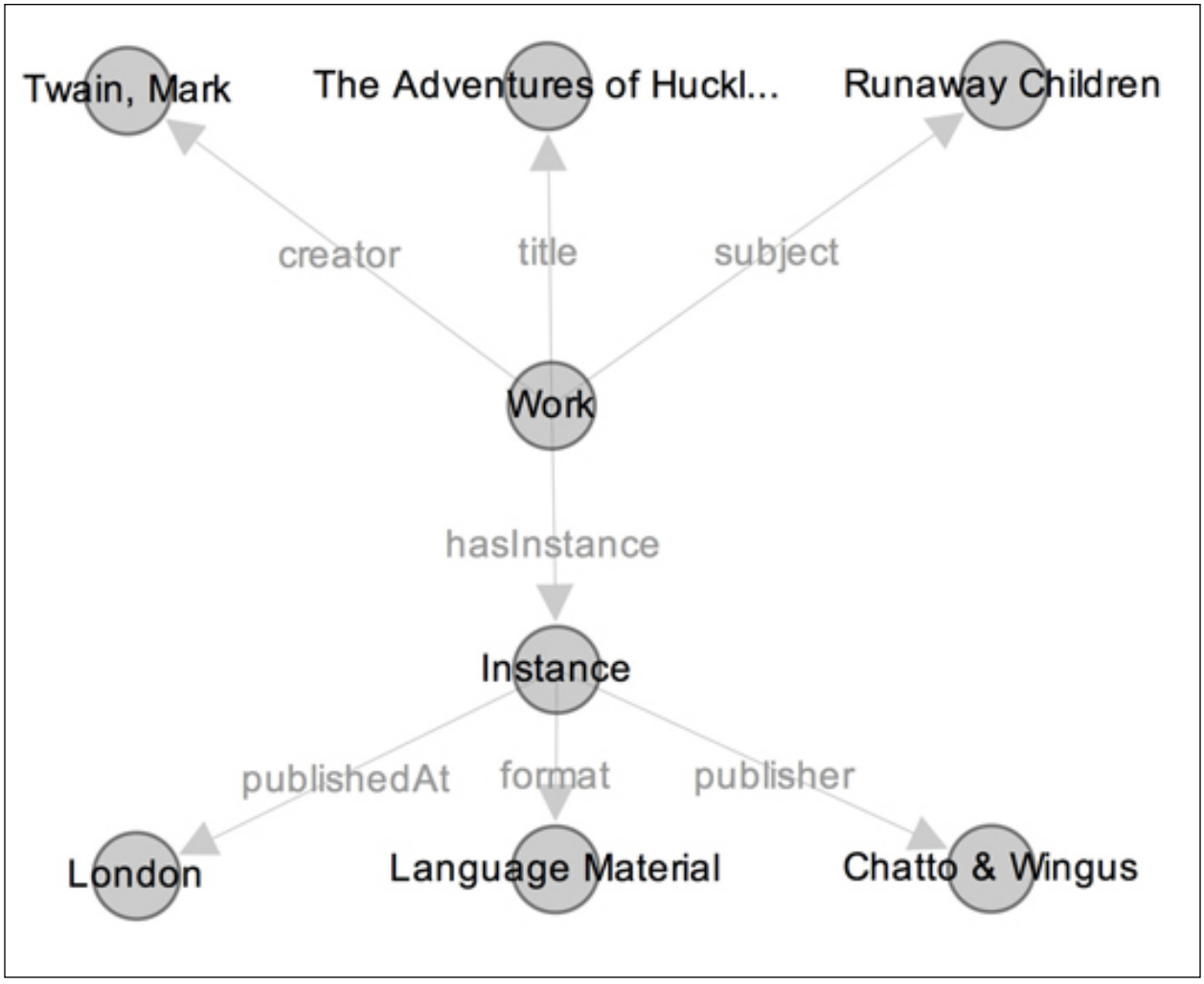

Figure 3.1
BIBFRAME entity and relationship model

conceptual resources, including people, places, subjects, and classifications, and can be attached to Creative Works or Instances to provide context and descriptive attributes. Authority classes are closely aligned with existing roles of authorities, with a key difference being that Authority entities should be unambiguously referenced using URIs rather than literal values. Annotations are broadly defined as entities that enhance the core description of a Creative Work and Instance. Examples given at BIBFRAME.org include cover art, library holdings, and reviews.

Conforming to the W3C definition of a resource, the BIBFRAME model characterizes Works, Authorities, and instances of Annotations as resources, addressable by URIs and contextualized by other web-based data. As we recall from our exploration of RDF in chapter 2, this design approach enables a wide range of resource representation techniques and also enables the expression of complex resource relationships using graph structures. In order to get a complete understanding of the BIBFRAME entities, we will explore each entity in detail.

For a detailed view of the RDF/XML and JSON serializations of the Mark Twain record used as an example in chapter 2, please refer to the online appendixes located at the GitHub repository for this issue. Although all of the figures and tables mentioned in this issue of *LTR* are available both online at GitHub and in print, the appendixes associated with this issue are available only online. Appendix 1 shows MARCXML; appendix 2, RDF/XML; and appendix 3, JSON. Appendix 4 contains the record as represented in the Library of Congress online web catalog.

GitHub LTR repository
https://github.com/mitcheet/ltr

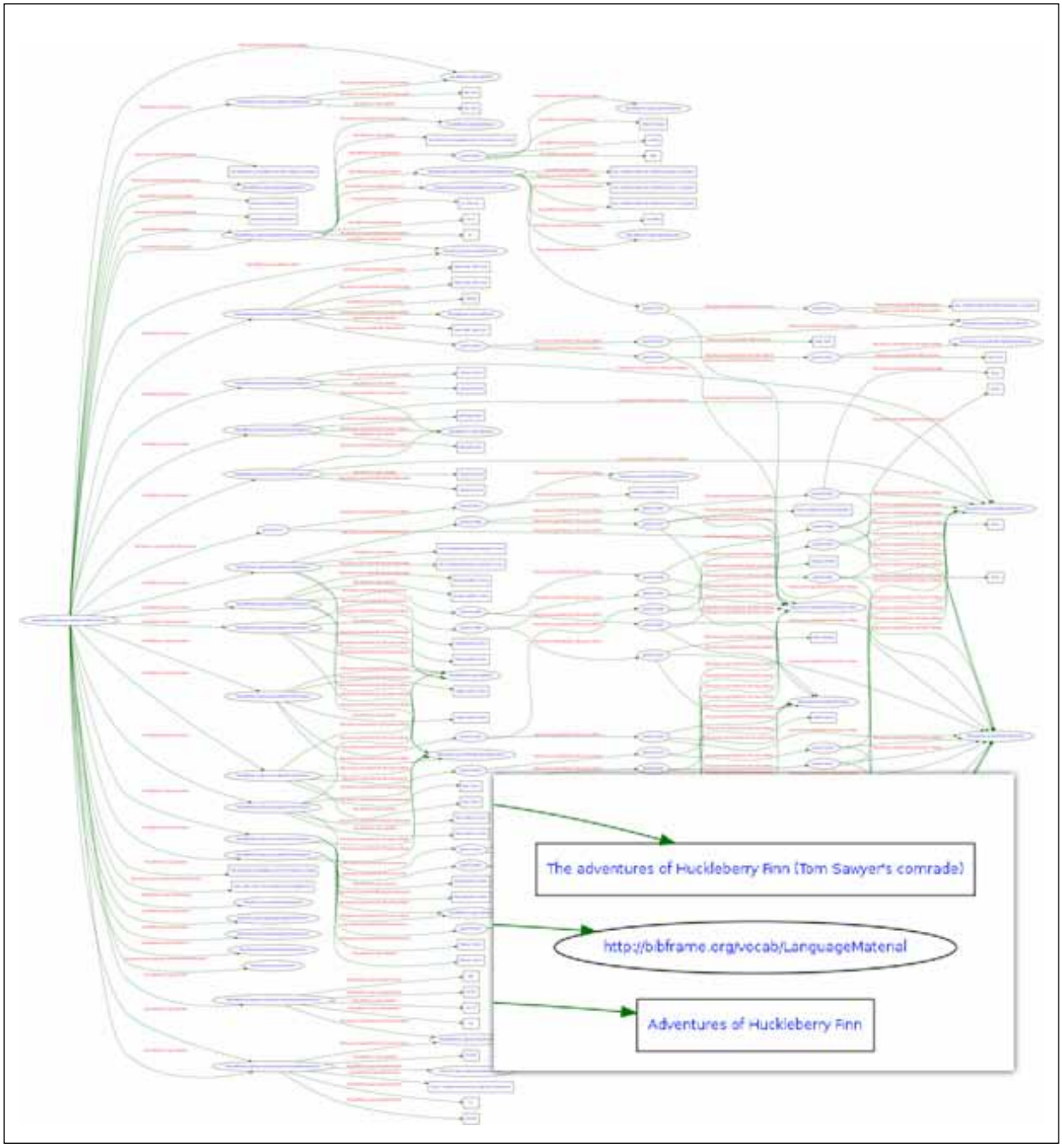

Figure 3.2
A graph visualization of the book *The Adventures of Huckleberry Finn* in the BIBFRAME specification

In the following sections we explore these four entities with our example title *The Adventures of Huckleberry Finn* in mind. A visualization of this title using BIBFRAME-conforming RDF/XML is represented in figure 3.2. This example includes 219 triples and represents what would be, in bibliographic environments, a fully cataloged MARC record. While the graph is illegible in the figure, it does help demonstrate the difficulty associated with visualizing even simple RDF-based data. This figure can be regenerated at larger size by submitting the RDF/XML file in appendix 2 to the RDF validator at the W3C. For more information about using the RDF/XML appendixes located at GitHub along with the RDF validator at the W3C, please see the tutorial located on the GitHub site.

Library of Congress catalog listing for Adventures of Huckleberry Finn
http://lccn.loc.gov/35020965

W3C RDF validator
www.w3.org/RDF/Validator

While figure 3.2 represents a complete BIBFRAME resource, the following sections explore the individual classes of BIBFRAME and provide more legible snapshots of portions of this broad visualization.

BIBFRAME Entity 1: Creative Work

The Creative Work in BIBFRAME has twenty-seven core properties, including title, classification, content scope, credits, audience, identifiers, language, and related work. These properties are available at the BIBFRAME Vocabulary example on Creative Work and are still in draft form. Creative Work types include thirteen predefined types such as Audio, Cartography, Dataset, Language Material, Mixed Material, Still Image, and Moving Image. These types are used in conjunction with the `rdf:type` element and are adapted from those defined in the RDA 336 field and MARC 21 Leader (006, 007) fields.

In addition to these core properties and types, the Creative Work entity includes properties to indicate relationships with Annotations (`hasAnnotation`, `annotationOf`), Expressions (`hasExpression`, `expressionOf`), and Instances (`hasInstance`, `instanceOf`). A visualization of just the Creative Work entity from figure 3.2 with selected data is represented in figure 3.3. The RDF/XML metadata for this figure is shown in list 3.1. This Creative Work example includes eight triples from the full record and is still too small to be legible. In order to generate the graph, you can submit the XML in list 3.1 to the RDF validator at W3C.

List 3.1
RDF/XML representation of a Creative Work class with selected properties

```
1 <rdf:RDF xmlns:rdf="http://www.
  w3.org/1999/02/22-rdf-syntax-
  ns#">
2 <bf:Work xmlns:bf="http://
  bibframe.org/vocab/"
  rdf:about="http://bib-
  frame.org/resources/
  aNJ1367336414/746732">
3 <rdfs:label xmlns:rdfs="http://
  www.w3.org/2000/01/rdf-schema#"
  xmlns:dcterms="http://purl.org/
  dc/terms/">The adventures of
  Huckleberry Finn (Tom Sawyer's
  comrade)</rdfs:label>
4 <rdf:type
  xmlns:dcterms="http://
  purl.org/dc/terms/"
  rdf:resource="http://bibframe.
  org/vocab/LanguageMaterial"/>
5 <bf:uniformTitle
  xmlns:dcterms="http://purl.org/
  dc/terms/">Adventures of Huck-
  leberry Finn</bf:uniformTitle>
6 <rdfs:label xmlns:rdfs="http://
  www.w3.org/2000/01/rdf-schema#"
  xmlns:dcterms="http://purl.org/
  dc/terms/">Adventures of Huck-
  leberry Finn</rdfs:label>
7 <bf:title
  xmlns:dcterms="http://purl.org/
  dc/terms/">The adventures of
  Huckleberry Finn (Tom Sawyer's
  comrade)</bf:title>
8 <madsrdf:authoritativeLabel
  xmlns:madsrdf="http://
  www.loc.gov/mads/rdf/
  v1#" xmlns:dcterms="http://
  purl.org/dc/terms/">Twain,
  Mark, 1835-1910. Adven-
  tures of Huckleberry Finn</
  madsrdf:authoritativeLabel>
9 <bf:instance
  rdf:resource="http://bibframe.
  org/resources/aNJ1367336414/
  instance1"/>
10 </bf:Work>
11 </rdf:RDF>
```

It is worth noting in list 3.1 that the BIBFRAME Creative Work class includes vocabulary very familiar to traditional bibliographic description (e.g., `bf:uniformTitle`, `LanguageMaterial`, and `mads:authoritativeLabel`). The XML data in list 3.1 was created using a metadata transformation tool and as such uses internal URIs that have meaning only within the converted dataset (see line 9). It is anticipated that in production releases, more durable endpoint URIs would be used.

Instances

Instances are discussed in BIBFRAME as "material embodiments" of works[7] and include both physical and digital manifestations. There are currently seventy-nine properties associated with BIBFRAME Instances, including many of the shared properties from Creative Work (e.g., Title, Alternative Title,

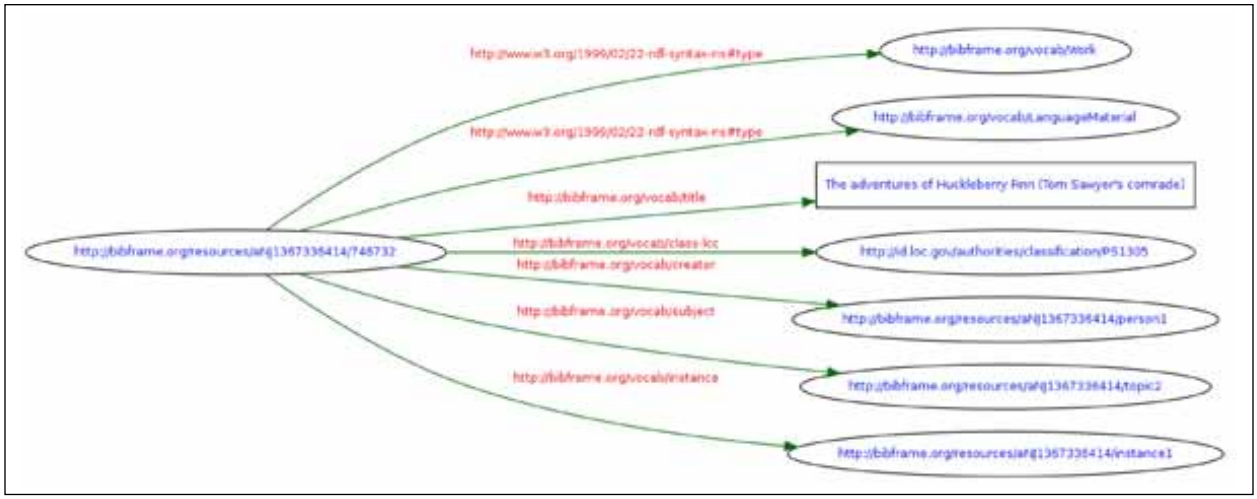

Figure 3.3
A graph visualization of a BIBFRAME Work entity

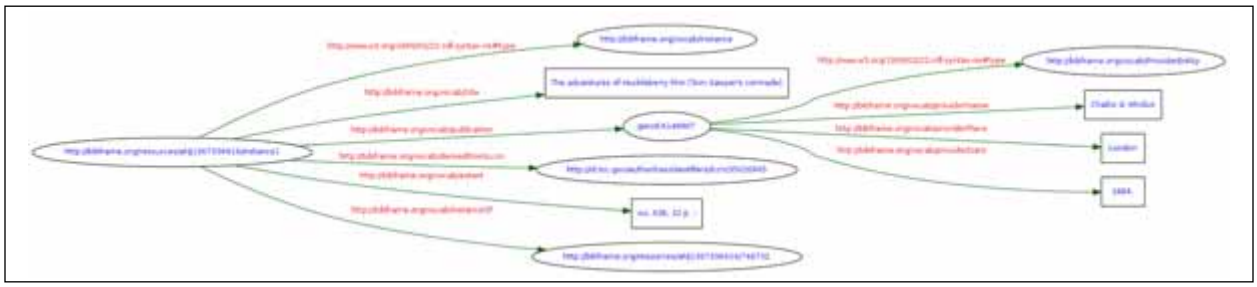

Figure 3.4
Graph of a BIBFRAME Instance

ISSN) as well as manifestation-specific properties such as `modeOfIssuance`, `publication`, `pubDate`, and `upc`. Instances have specific properties designed to show relationships including `hasAnnotation` and `instanceOf`. Confining itself just to descriptive elements, the graph of the printed book instance of our example record returns fourteen triples that describe the printing, repeat some properties associated with the Creative Work (e.g., Title), and show relationships between the Creative Work, Annotations, and Authorities. Figure 3.4 shows the graph of this Instance.

Authorities

Authority entities include people, places, things, topics, organizations, and events as defined in some external vocabulary or ontology. The Authority entity in BIBFRAME is designed as a container that highlights the relationship and context of an authority entry to the Creative Work, Instance, or Annotation being described. The actual authority reference may point to any linked data endpoint that contains authority data. For simplicity, the examples pulled from the BIBFRAME test site for this issue employ the Metadata Authority Description Schema (MADS). Figures 3.5, 3.6, and 3.7 show BIBFRAME Authority entities for personal, topical, and classification authorities. While these examples point to MADS references, they could just as easily point to other external linked data vocabularies.

It is worth noting that these entities are quite different from the full MADS record for these authorities and that the converted BIBFRAME record does not reconcile authority data with the endpoints available at Library of Congress Linked Data Service. The MADS record there for Mark Twain includes 369 triples that capture the descriptive, versioning, and administrative metadata associated with this record. MADS records also include links to the Virtual International Authority File (VIAF), which, in addition to representing the author's authority data, includes aggregated information about related works, coauthors, and publication dates. A small subset of the MADS record for Mark Twain, excluding variant name forms and administrative metadata, is shown in figure 3.8. This figure and a view of the MADSRDF

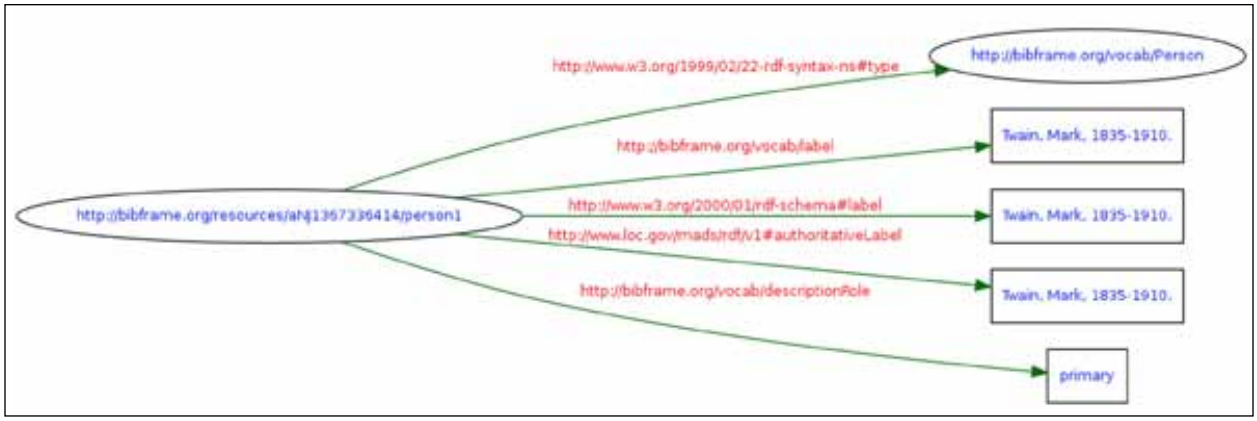

Figure 3.5
BIBFRAME Authority graph for personal authority reference

Figure 3.6
BIBFRAME Authority graph for topical authority reference

file help show exactly how much data can be accessed when URIs are used to link bibliographic records to the Library of Congress Linked Data endpoint.

Library of Congress Linked Data Service Authorities and Vocabularies
http://id.loc.gov

Annotations

BIBFRAME Annotations include only three properties (`annotates`, `annotationAssertedBy`, `annotationBody`) and provide a container for the inclusion of annotations either by value or by reference. Annotation types currently documented in BIBFRAME include CoverArt, Holdings, and Reviews, but can be expanded to include a much wider range of information. Each annotation type has its own properties. For example, the Holdings Annotation includes the `callno`, `callno-ddc`, `callno-lcc`, and `callno-udc` properties. Figure 3.9 shows a simple call number Holding entity.

The BIBFRAME model stores holdings information as Annotations in recognition that these entries represent statements made by libraries about resources[8] and provide an equal platform through which library and nonlibrary statements can be asserted about a resource.

Building Blocks 4 and 5: BIBFRAME Serialization and Data Exchange

Current work in BIBFRAME is focusing on RDF/XML serializations, but the specification is compliant with any RDF-based serialization including N3/Turtle and N-Triples. No public information is available regarding back-end system storage, but the BIBFRAME test software that is used to generate test instances produces data using the JavaScript Object Notation (JSON) serialization as well as RDF/XML data. Sample output data from this tool can be found at GitHub in appendixes 2 and 3, and tutorials at BIBFRAME.org can guide users through the process of converting test data.

BIBFRAME test software
http://bibframe.org/demos

The BIBFRAME specification is being conceptualized as a metadata exchange format rather than an internal data model, and serialization bit project is early

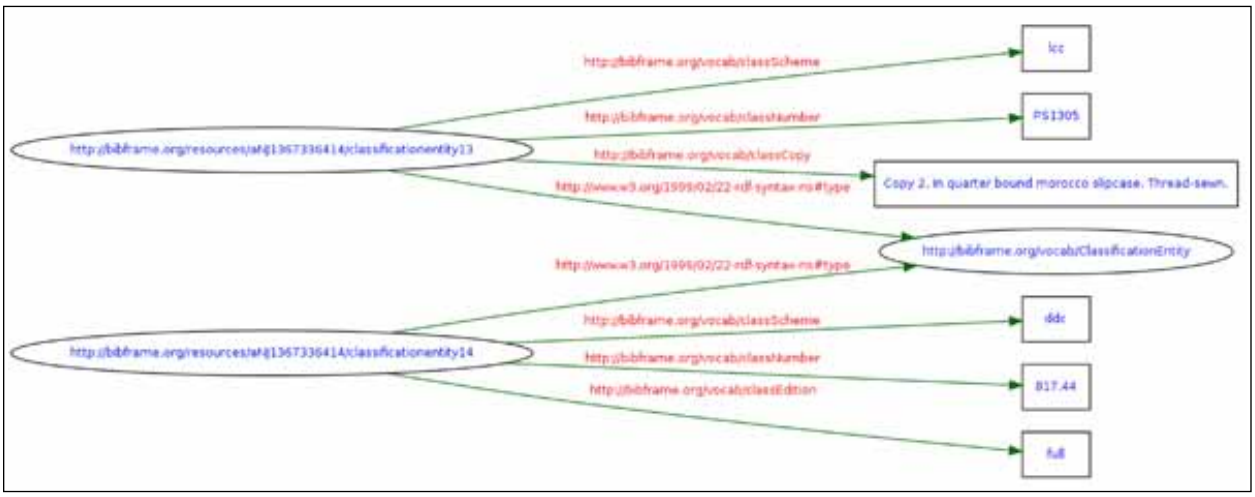

Figure 3.7
BIBFRAME Authority graph for classification reference

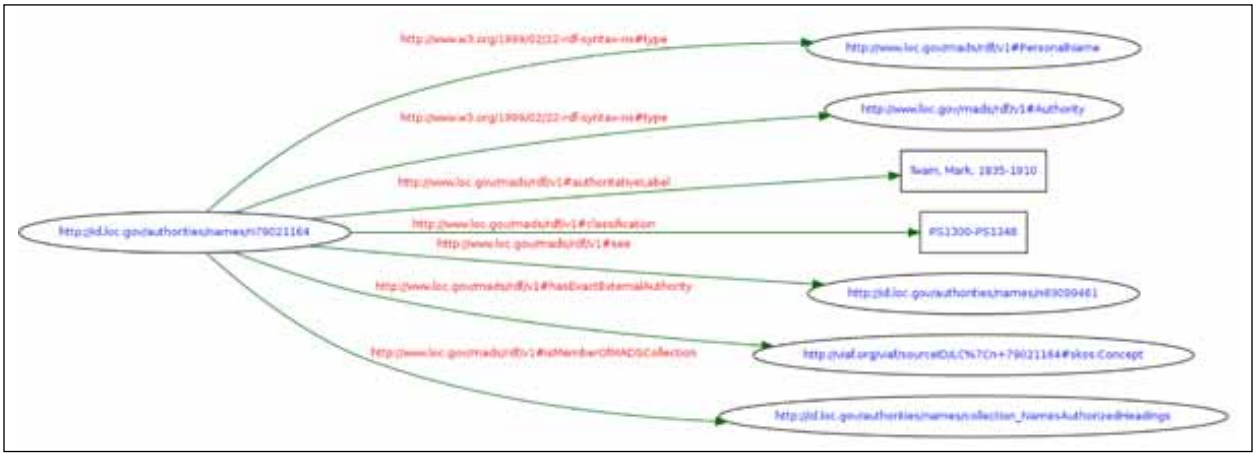

Figure 3.8
Graph of selected content from MADSRDF record for Mark Twain

Figure 3.9
BIBFRAME Holdings entity—call number

enough in its work that the mechanics and details of this process have yet to be made public. In addition to the data model and metadata specification, the project has resulted in demonstration tools that interested institutions can use to explore the BIBFRAME specification. This suite of tools was cocreated by Zepheria and the Library of Congress and provides conversion of MARCXML data into BIBFRAME data represented in JSON- and XQUERY-compliant data. In addition, the BIBFRAME.org site hosts two transformation tools, one of which mirrors the downloadable toolkit and the other of which will return a single bibliographic record

from the Library of Congress website represented using MARCXML and BIBFRAME principles.

Utilities to transform MARCXML bibliographic records to BIBFRAME resources
https://github.com/lcnetdev/marc2bibframe

BIBFRAME transformation tools
http://bibframe.org/tools/transform/start

These tools produce output in a web-based user interface (figure 3.10). This interface provides a faceted navigation scheme as well as multiple serialization formats (MARC/XML, BIBFRAME RDF/XML, and Exhibit JSON). It also provides an easy-to-implement method for libraries to explore the BIBFRAME model and test their own data conversion. This tool was used to extract all of the metadata in online appendixes 1–3 as well as in figures 3.2–3.10.

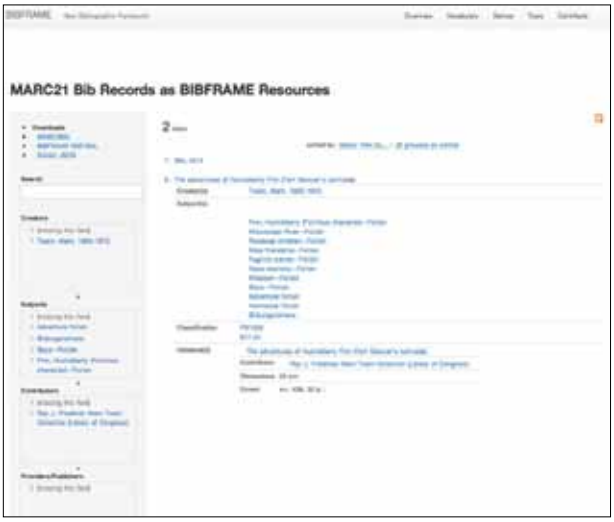

Figure 3.10
BIBFRAME transform tool output

Community Direction and Activities

The BIBFRAME standard is still a work in progress, and many of the public products of this work have been released only recently. The BIBFRAME community is developing the specification as a data interchange format, much as MARC was envisioned during its development, and as a result, BIBFRAME does not seek to provide metadata schema and data representation models for authority and community data in the way that MARC did.

The primary avenue for contribution to the BIBFRAME discussion is through e-mail discussion list interaction, and at the time of writing, the Library of Congress was engaging the community in a discussion of model features. At this time, there is no adoption timeline or migration plan, but the effort has considerable backing and momentum. More information about the direction and implementation plan of BIBFRAME can be found on the BIBFRAME.org website.

The Digital Public Library of America (DPLA)

Metadata Specification Overview

The Digital Public Library of America (DPLA) is a broad initiative geared toward the development of a unified digital library of materials in the United States. Formed in 2010 and launched in 2013, the DPLA is supported by a number of funding agencies, including the Alfred P. Sloan Foundation, the Arcadia Fund, the Institute for Museum and Library Services, the John S. and James L. Knight Foundation, and the National Endowment for the Humanities.[9] The DPLA states that it "brings together the riches of America's libraries, archives, and museums."[10] Two of the initial products of this work are a portal for discovery and resource access and a platform through which data and services can be accessed programmatically (APIs). In addition, the DPLA community embraces the tenets of open data and takes an advocacy stance in support of open-access policies.

On April 18, 2013, the DPLA officially launched a discovery platform that provides access to the initial set of data contributed by eighteen partners, over 3,200 collections, and over two million records.[11] The discovery platform includes both an end-user–oriented website and an API that allows direct access to aggregated data. Since this launch, the number of partner institutions and the size of the collection have continued to grow.

Building Block 1: DPLA Data Model

The DPLA internal data model is based on RDF but also employs an RDF-inspired serialization called JSON-LD that is disseminated via API output. The data model of the API includes broad information about the request object, an array-based list of documents or collections, and an array-based list of facets. While the internal structure of the DPLA employs RDF, the data-harvesting method employs the OAI-ORE standard. As with the BIBFRAME and Europeana data models, an emphasis is placed on supporting the creation of graph structures. More information about the data model is available on the DPLA site, although at the time of this writing, the documentation is not yet complete.

DPLA Metadata Application Profile
http://dp.la/info/map

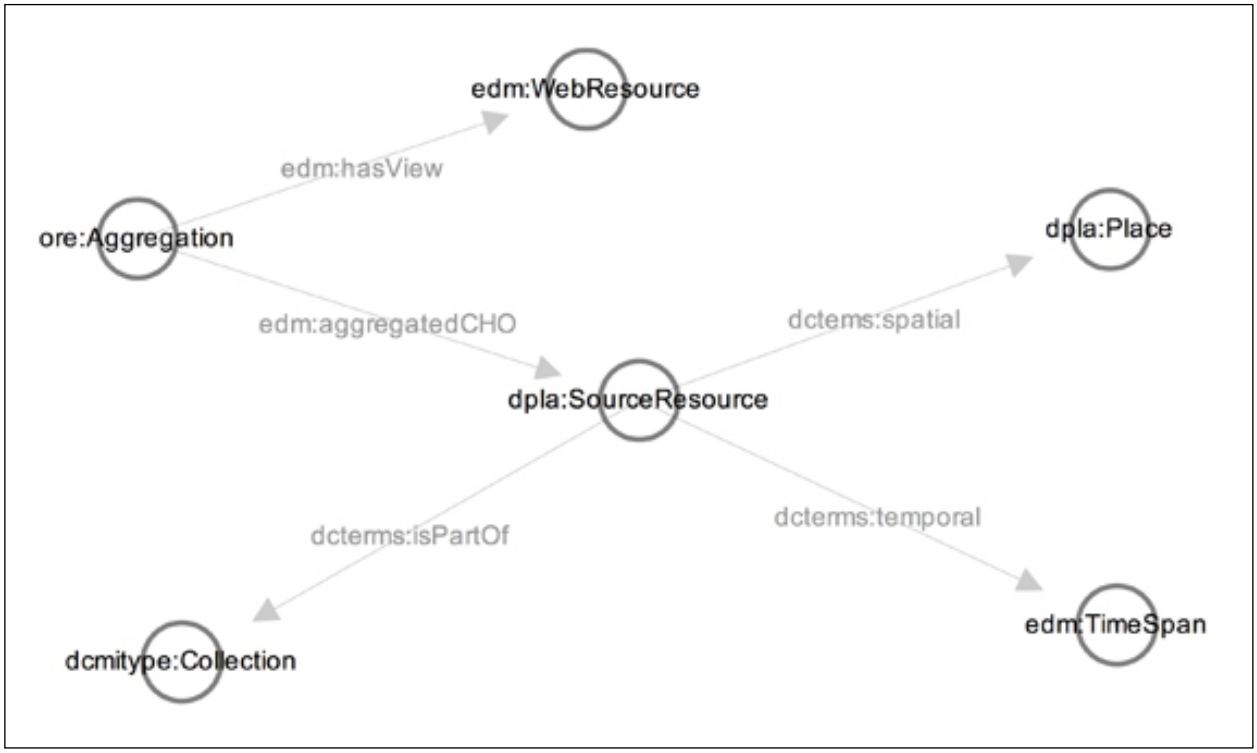

Figure 3.11
DPLA class and property model as adapted from DPLA metadata documentation

Building Block 2: DPLA Content Rules

The DPLA standard is essentially a data aggregation and sharing service and as such is not designed for manual cataloging work. The data model does have content formatting rules for certain properties but typically adopts those rules from other vocabularies. For example, the property `dpla:country` employs the IOS 3166-1 country code, and properties borrowed from the Europeana Data Model (EDM) use EDM content types. Because the DPLA model focuses on storing harvested data, it does store some data gathered from a provider and some data generated or extracted during the data collection process. For example, the EDM property `edm:currentLocation` conforms to the ISO3166-1 standard but is generated from the data provider value in the gathered dataset.

Building Block 3: DPLA Metadata Schema and Vocabularies

The DPLA metadata model is currently in its third iteration[12] and builds on the Europeana data model. The metadata model is based on RDF and uses Dublin Core as a central descriptive metadata standard. The core DPLA class or entity `dpla:SourceResource` is similar to an `ore:AggregatedResource` in that it has a relationship to an `ore:Aggregation` using the property `edm:aggregateCHO` of the Europeana Data Model. The `dpla:SourceResource` class employs DC, DCTERMS, and EDM properties to gather descriptive and some administrative metadata (e.g., contributor, date). Figure 3.11 shows a simple representation of the class and property structure implemented in Gephi and adapted from the DPLA metadata specification.

Gephi
http://gephi.org

Within the DPLA metadata schema, vocabulary and content rules for faceted data are given special attention. In addition, administrative data regarding the providing repository is used to support a base level of provenance data as well as to support discovery services.

Building Blocks 4 and 5: DPLA Serialization and Data Exchange

DPLA data is available as bulk downloaded files using the JavaScript Object Notation for Linked Data (JSON-LD) serialization, is stored internally as RDF/XML, and is available via API access as JSON-LD.

DPLA Bulk Download
http://dp.la/info/developers/download

DPLA API Codex
http://dp.la/info/developers/codex

JSON-LD is an extension of JSON that includes a method for identifying data through IRIs, supports disambiguation of JSON objects when combining datasets from different documents, provides a method for identifying language and data types of literals, and provides a method for expressing graph relationships using JSON structures.[13] JSON-LD's parent specification, JSON, is an increasingly popular data serialization format that is seeing wide use given its ability to work natively with a number of programming languages, its relative simplicity and readability, and its lightweight encoding structure. JSON uses a key/value data structure implemented in an ordered list or array structure that closely resembles triple statements (subject, predicate, object). JSON supports hierarchical definitions within a given object and as such can be traversed using Document Object Model (DOM) techniques. Figure 3.12 shows the basic structure of JSON, which employs French braces, colons, commas, and quotes to structure data. More information on JSON is available at W3Schools and JSON.org.

W3Schools: JSON Tutorial
www.w3schools.com/json

JSON
http://json.org

JSON-LD extends standard JSON structures by introducing unique identifiers (`@id`) as well as data types (`@type`), vocabularies (`@vocab`), and a graph (`@graph`) data structure. The DPLA utilizes these structures to return data that can be cross-aggregated with data extracted using other API calls. This data may contain collection, item, or facet information depending on the structure of the API call. Figure 3.12, for example, contains a high-level view of the JSON-LD data returned in response to an API call that requested information for all items in the database along with faceting information for two fields (`sourceResource.publisher` and `sourceResource.creator`). The API URL for this request is `http://api.dp.la/v2/items?facets=sourceResource.publisher,sourceResource.creator&api_key=[personalapikey]`. A personal API key, which is required to experiment with the DPLA API, is available for free from the DPLA developer site. For a tutorial on how to get your own API key and work with the DPLA API, check out the online resources on GitHub that supplement this issue.

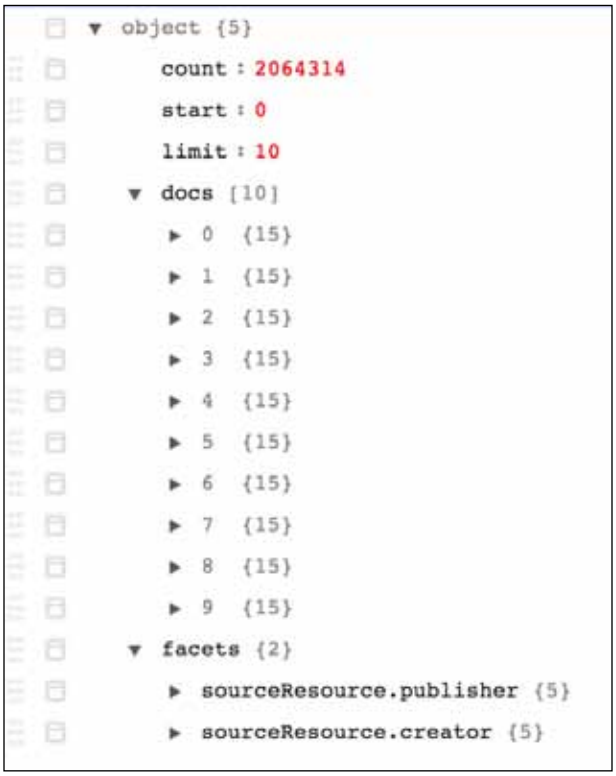

Figure 3.12
DPLA JSON-LD output for an API request for all items with two facets

Figure 3.13
DPLA JSON-LD output featuring facet information

DPLA: For Developers
http://dp.la/info/developers

Figure 3.12 shows the broad count of resources

Figure 3.14
DPLA JSON-LD output featuring document information

Figure 3.15
DPLA JSON-LD output featuring `@context` information

available with the API call (2,064,314), the actual resources returned (resource 0–9), the actual documents (docs), and information about the facets requested. Figure 3.13 shows the facet information returned in the JSON-LD document. This figure shows the JSON-LD data element `@type` (terms) and the top fifty facets along with their document counts. The faceting information can be used to generate new API requests that focus just on specific facet information. For example, a request to get all documents that had as a facet value "ala" for the field `sourceresource.Publisher` would be `http://api.dp.la/v2/items?sourceResource.publisher=ala&api_key=`.

The docs stanza of the JSON-LD document includes source, administrative, and descriptive metadata on the object returned. Figure 3.14 shows a subset of information from a single document in the results set. This includes a context section that provides information about the object within the DPLA environment, a provider section that includes the source of the object, an admin section that documents the status of the object, a `sourceResource` section that contains detailed metadata harvested from the source repository parsed into DPLA fields, and an `originalRecord` section that contains metadata as it was structured when harvested.

The `@context` information returned as part of each document (figure 3.15) maps the terms used in the internal DPLA and EDM properties, including utilized vocabularies and schemas, object types, and collection types to IRIs for those terms. By providing IRI mapping to external schemas, the `@context` object contextualizes data in JSON-LD and enables computational analysis in support of semantic services.

Figures 3.16 and 3.17 show the outcome of the data processing that occurs during resource ingest. In these figures, data from the source record on object geospatial location, which is represented in the coverage property in the original record, is mapped to the DPLA spatial property and enhanced with county, state, geocoordinate, and country data. Likewise, subjects that were returned as a text block in the original record have been parsed into discrete entries within the subject property. Additional linked data processing is possible for this record, including the resolution of corporate, publisher, and personal name authorities as well as subject heading authorities in external linked data endpoints.

An examination of this record reveals issues with specificity that can occur when aggregating and converting metadata. For example, the inclusion of geocoordinate data for what was originally described as a broad location gives the impression that the data points to the actual street address of the building in question. The DPLA public interface properly handles this by showing the data in context, but the data returned via the API enables subsequent systems to use the data in ways that do not reflect this comprehension of the limitation of this particular element. A full sample record from DPLA following the JSON-LD serialization is available online in appendix 5.

A key element of the data exchange functionality of the DPLA site is implemented using the API service discussed above. There are a number of projects underway that demonstrate how the API and DPLA data can be used. The University of Illinois, for example, has published a prototype application that aggregates collections from cultural heritage institutions. Likewise, the MINT platform, a metadata ingest and transformation service, has been configured to work with DPLA data. The development of tools like these

Figure 3.16
DPLA JSON-LD output featuring original document metadata

Figure 3.17
DPLA JSON-LD output featuring parsed document metadata

was a key part of the DPLA exploration phase, and it is expected that this work will continue as the project continues. As of this writing, the DPLA does not publish a SPARQL endpoint for LOD querying.

University of Illinois DPLA prototype
http://dpla.grainger.illinois.edu

MINT Ingestion Server—DPLA
http://mint-projects.image.ntua.gr/dpla

Community Direction and Activities

The DPLA community is best described as newly formed and undergoing transition. Following the launch in late April 2013, the community undertook some reorganization steps to organize documentation and bring together discussion groups. While there is still much to be determined about the future of DPLA, it has already gained a number of contributing organizations and has several applications developed that capitalize on its data API.

Europeana

Metadata Specification Overview

The Europeana Digital Library is a large-scale community effort to bring together collections across European libraries under a common metadata schema and using a common indexing and dissemination platform. The Europeana portal is a key product of the Europeana Foundation, whose overarching goal is to bring together and publish data from Europe's cultural heritage and

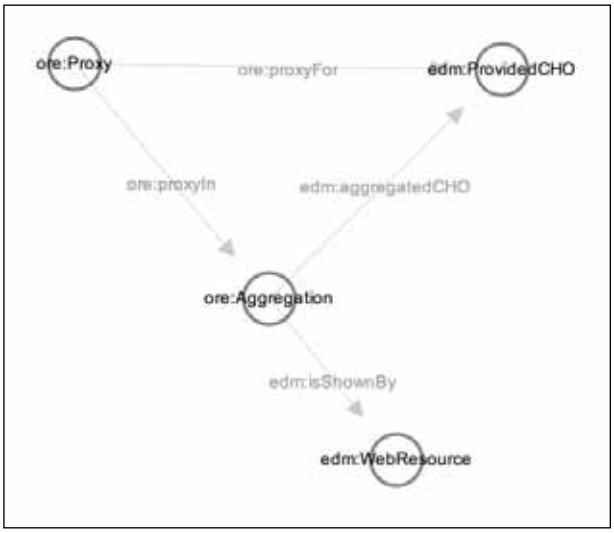

Figure 3.18
Europeana Data Model core classes

scientific communities.[14] It is the most mature of the metadata services discussed in our case study review.

The Europeana Digital Library contains over two hundred million records and ten million digital objects that have been contributed from over 1,500 institutions.[15] The Europeana Data Exchange Agreement stipulates that all metadata shared with Europeana be publishable as Creative Commons CCO 1.0, allowing reuse without restriction.[16] The data agreement also stipulates that a rights statement describing terms of use must accompany digital objects.

The Europeana Digital Library offers a selected dataset in LOD form. The repository consists of twenty million records available via both file downloads and a SPARQL endpoint and is structured using the EDM. The EDM supports complex record disambiguation and connection across metadata providers and includes constructs to help maintain versions and provenance information about the resources and metadata contributed. The resulting metadata model provides detailed structures for tracking resources by centering on objects as well as events in the object's life cycle.

Europeana Digital Library selected LOD dataset
http://data.europeana.eu

Our review of the EDM goes into some detail on the specification, but the model is sufficiently detailed that interested readers would be well served by consulting other resources to supplement their understanding. A good source for this information is available at the Europeana Professional site.

Europeana Professional
http://pro.europeana.eu

Building Block 1: Europeana Data Model

Like the other projects discussed in this chapter, the EDM is built using RDF and graph-based data models, structures, and schemas. Given its maturity, the EDM is documented to a larger extent than DPLA and BIBFRAME, and there is a robust data harvesting and transformation service that employs concepts and vocabularies from the OAI-ORE specification to provide a base model for contributed resources. The use of OAI-ORE vocabularies and data structures enables the representation of complex relationships between providers, objects, and versions of metadata statements, making it possible, for example, to reconcile different versions of resources and track conflicting descriptive statements about those resources. The RDF-based EDM was developed following an initial specification called the Europeana Semantic Elements (ESE), which employed an XML Schema model.

Building Block 2: Europeana Content Rules

The EDM application guidelines provide cataloging guidance for most if not all classes and properties in the data model. The key document is the Europeana Data Model Mapping Guidelines, which details the use of EDM and elements from other schemas.[17] These cataloging guidelines do not entirely override other cataloging rules as the model is largely silent on what other metadata schemas can be included in the supplied metadata.

The generation of linked data in the EDM involves considerable linking to external vocabularies and endpoints. This includes creating geographic name links to GeoNames, subject links to GEMET, personal links to DBpedia, and linked data services of partner institutions.[18] Because of the emphasis on these connections, the content rules governing metadata in EDM focus more on adhering to proper RDF syntax for the establishment and management of links than creating literal values.

GeoNames
http://geonames.org

GEMET
www.eionet.europa.eu/gemet

DBpedia
http://dbpedia.org

Building Block 3: Europeana Metadata Schema and Vocabularies

The EDM makes considerable use of the aggregation, aggregated resource, and proxy properties of the OAI-ORE environment. The EDM uses a locally defined property to track contributed cultural heritage objects, often referred to with the acronym CHO (`edm:ProvidedCHO`) and its associated web representation (`edm:WebResource`). In order to connect these classes, the EDM also depends on an `ore:Aggregation` class that defines the relationships between these resources. This focus on providing a means for tracking both the cultural heritage object itself and the descriptive metadata surrounding it reflects a firm adherence to the one-to-one principle that has been a challenge in metadata record-based digital library systems.

Figure 3.18, adapted from the EDM mapping guidelines,[19] shows the simple relationship between an ORE

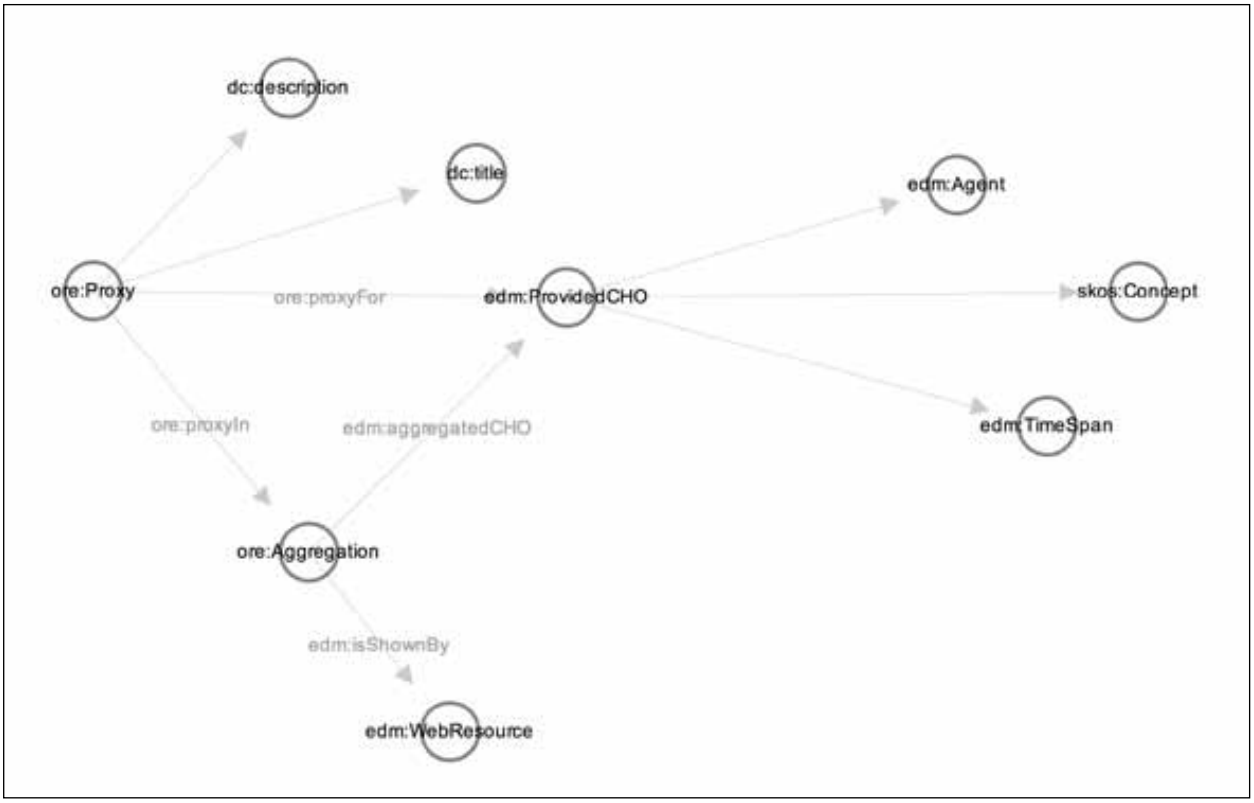

Figure 3.19
Europeana Data Model enhanced classes

aggregation and EDM cataloged resources. This model uses a common exchange standard (ORE) as well as a complex internal object tracking approach that distinguishes between real resources, the web-based representations of them, and the descriptive metadata about these resources. For a full record example, see appendix 6 online. This appendix shows a sample Europeana record from the LOD dataset serialized in RDF/XML.

The distinction between an object and its digital surrogate is part of a larger focus on metadata design principles in the EDM model that seeks to accurately capture the provider and versioning information of metadata. This is one goal of the seven EDM design guidelines, which are included (paraphrased) here:[20]

1. Differentiate between a real-world object and its digital surrogate.
2. Distinguish between an object and the metadata describing that object (the one-to-one principle).
3. Support multiple metadata records about an information object, and support the presence of conflicting information in these records.
4. Support data models that allow an information object to be aggregated from other information objects.
5. Support metadata that conforms to domain-specific abstract models (e.g., FRBR).
6. Reuse elements from other standards where possible.
7. Be flexible in how the standard supports the description of concepts and contextual resource.

These design principles reflect many of the fundamental issues in complex metadata work that we touched on in chapter 1. The disambiguation of duplicated resources from multiple repositories, for example, has proven to be a considerable issue in cultural heritage aggregation environments. In implementing these principles, however, care must be taken to ensure that the rules of property inheritance and association are followed.

The adherence to these principles has led to a relatively complex schema that takes care to associate metadata about a resource with a proxy object for that resource. In an attempt to show some of these relationships, figure 3.19 provides an expanded but still incomplete view of the relationship of EDM classes, ORE classes, and individual metadata. This model employs the `ore:Proxy` class to collect metadata that may be shared across different instantiations or representations of a `providedCHO`. In sample records available at the Europeana data site, this is commonly

the implementation used. The proxy class is then connected with aggregations, items, and other digital or physical object containers.

While the use of proxies to store descriptive metadata puts an additional level of complexity into the system, this structure also enables the recording of different versions of metadata, a key functional requirement identified by the Europeana community.[21] This documentation does discuss an alternative model, however: the use of named graphs or quads. Quads are discussed in more detail in the next chapter, but broadly speaking they enable statements to be identified as belonging to another resource (e.g., subject, predicate, object, quad name), largely duplicating the role of the proxy class in the EDM. At the time of this writing, quads are still a W3C working group and are not a part of the RDF recommendation.

Building Blocks 4 and 5: Europeana Serialization and Data Exchange

Like the DPLA, Europeana makes data available via file download and API access. The EDM API is designed to output data as JSON files, but the datasets contributed by individual institutions are available as EDM data serialized in RDF/XML and RDF/NT files. In addition to these two dissemination methods, data can also be queried through the Europeana SPARQL endpoint. This endpoint can return data serialized in RDF/XML, JSON, N3/Turtle, and N-Triples. The EDM follows a native graph design and as such can be stored in any RDF-compliant triple store.

Europeana data download
http://data.europeana.eu/download/2.0

There are a number of tools available to process metadata in relation to the EDM. The EDM data is generated from the Europeana data store and the Europeana Semantic Elements (ESE) model. This code converts data from an XML Schema–based structure to RDF. The toolkit for this conversion is open source, hosted at GitHub, and documented in journal articles.[22] In addition, the MINT platform can process EDM data, and Europeana itself supports an API test environment that supports open exploration of the Europeana API. The online materials for this issue of *LTR* include a brief tutorial about how to create an API key and access Europeana data using API requests. More information is available in the online materials.

ESE2EDM Converter on GitHub
https://github.com/behas/ese2edm

Europeana API test environment
http://preview.europeana.eu/portal/api/console.html

Community Direction and Activities

The Europeana foundation is the most mature of the three services discussed in this chapter and has developed a robust foundation model to drive the addition of new community members. The early success of the LOD pilot, as well as the robust data model, positions the EDM to make contributions to other communities as well. This was seen, for example, in the use of EDM data models in the BIBFRAME and DPLA data and poses opportunities for collaboration between these communities.

The Europeana community, broadly defined, is open to any European cultural heritage and memory institution. The current content in the discovery platform trends towards museum and archival collections but does include some bibliographic information.

Case Study Discussion

Our case study exploration used the metadata building blocks model (table 1.2) to categorize features of the metadata components of these services and showed remarkable similarity in how these services approach metadata design questions. For example, each environment relied primarily on RDF as a data model, de-emphasized content rules in favor of a big-umbrella approach, and used a mix of API and end-user access points. In addition, how these services approached technical and policy issues related to metadata, the tools they provide to their community, and the intended direction of their work shared similar features. For example, both the DPLA and Europeana platforms sought to publish open data and used JSON serialization in their API as a technical means for disseminating data. DPLA's choice to implement JSON-LD, as well as its use of a subset of EDM classes in its data model, shows variation but also enough overlap as to support interoperability with other LOD environments. In addition, an informal exploration of the external vocabularies referenced in the sample records located in the online appendixes shows a considerable amount of overlap in descriptive and structural metadata vocabularies (e.g., OAI-ORE, DC, QDC) but also little use of some other popular linked data vocabularies like the Bibliographic Ontology.

Table 3.1 provides a brief summary of unique features of each system evaluated using the metadata building blocks model as a guide.

Our exploration of these services also found different data services (APIs vs. SPARQL), different

Table 3.1
Comparison of unique features of case study examples using the metadata building block model

Metadata building block	DPLA	BIBFRAME	Europeana
Data model	RDF	RDF, FRBR-inspired	RDF
Content rules	Largely undefined except for faceted elements, embraces 'big-tent' approach to metadata aggregation	RDA and other bibliographic standard foundation but claims to be extensible	Europeana Data Model defines relationships between resources but is not overly prescriptive for content, embraces 'big-tent' approach to metadata aggregation
Metadata schema / vocabularies	Simplified EDM model with DPLA-specific elements geared towards faceted browsing	Relies heavily on the BIBFRAME vocabularies and Library of Congress endpoints	Europeana Data Model is highly prescriptive with structure, extensive use of external endpoints where possible
Data serialization	JSON-LD	RDF/XML, JSON	RDF/XML
Data exchange	Primarily API-based right now	Data transformation tools, no data aggregation or dissemination tools	SPARQL, APIs

serializations (JSON-LD, JSON, RDF/XML), and varying scope and detail in the data disseminated. Given the commonalities of RDF and many shared vocabularies in the schema, interoperability is not a significant issue as a DPLA/Europeana test application has already shown. In fact, the presence of shared vocabularies and administrative data models may make it possible to build virtual collections using APIs and SPARQL queries on demand. At the same time, however, Europeana was the only environment to offer a SPARQL endpoint and, as we will discuss in chapter 4, the use of SPARQL is an important aspect to growing linked data interoperability and Semantic Web technologies.

DPLA/Europeana Query
www.digibis.com/dpla-europeana

Conclusion

In focusing on high-level similarities and differences between these services, this set of case studies shows a development path for LODLAM services. Phase 1, as exemplified by BIBFRAME, focuses on specification definition, data modeling, and metadata quality. Phase 2, as exemplified by the DPLA, focuses on resource aggregation, harmonization, and dissemination and implements end-user and API services. Phase 3 implementations, as exemplified by Europeana, focus on issues of scale, add additional dissemination methods, and feature an enhanced data model required given the growth of the community. It turns out that LAM institutions are part of a much larger movement towards semantic data and services in the commercial and open-source information sectors (see, for example, Mondeca Labs).

Mondeca Labs
http://labs.mondeca.com

At all levels, issues of policy, licensing, community engagement, and project support are important factors that have an impact in metadata design but were not part of our exploration. In order to understand these issues as well as to project what a phase 4 development model might look like, additional analysis is needed that takes as its source other data sources. For readers interested in diving more into the technical details of the specifications and LOD platforms discussed in this chapter, Coyle's issue of *Library Technology Reports*[23] is an excellent and still largely current starting point for available vocabularies. In chapter 4, we will ask these broader questions about adoption, scale, issue, and opportunities in regard to LOD in libraries, archives, and museums by looking back at the published literature as well as studying the content of current conversations in this arena.

Notes

1. Deanna Marcum, "A Bibliographic Framework for the Digital Age," Library of Congress, Bibliographic Framework Initiative website, October 31, 2011, www.loc.gov/bibframe/pdf/bibframework-10312011.pdf; *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control* (Washington, DC: Library of Congress, January 9, 2008).

2. "The Library of Congress Announces Modeling Initiative," Bibliographic Framework Initiative website, May 5, 2012, www.loc.gov/marc/transition/news/modeling-052212.html.
3. Library of Congress, "BIBFRAME Frequently Asked Questions," Bibliographic Framework Initiative website, accessed April 20, 2013, www.loc.gov/marc/transition/news/faqs.
4. Ibid.
5. Eric Miller, Uche Ogbuji, Victoria Mueller, and Kathy MacDougall, *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services* (Washington, DC: Library of Congress, November 21, 2012), 5.
6. Library of Congress, "BIBFRAME Frequently Asked Questions"; Library of Congress, "The BIBFRAME Model: Vocabulary Updates," Bibliographic Framework Initiative website, accessed April 20, 2013, http://bibframe.org/vocab.
7. Miller et al., *Bibliographic Framework as a Web of Data*, 8.
8. Ibid., 11.
9. "About the Digital Public Library of America," DPLA website, accessed April 21, 2013, http://dp.la/info.
10. Ibid.
11. "Welcome to the Digital Public Library of America," DPLA website, accessed May 29, 2013, http://dp.la/info/2013/04/18/message-from-the-executive-director.
12. Digital Public Library of America, *Metadata Application Profile, Version 3*, February 8, 2013, DPLA website, http://dp.la/info/wp-content/uploads/2013/04/DPLAMetadataApplicationProfileV3.pdf.
13. Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, and Niklas Lindström, "JSON-LD 1.0: A JSON-Based Serialization for Linked Data," W3C Last Call Working Draft, April 11, 2013, www.w3.org/TR/json-ld-syntax.
14. "The Europeana Foundation," Europeana Professional website, accessed April 22, 2013, http://pro.europeana.eu/web/guest/foundation.
15. "FAQs," Europeana Professional website, accessed April 22, 2013, http://pro.europeana.eu/web/guest/europeana-faq; Bernhard Haslhofer, Elaheh Momeni, Bernhard Schandl, and Stefan Zander, "Europeana RDF Store Report," Europeana Connect, Results and Resources, March 8, 2011, www.europeanaconnect.eu/documents/europeana_ts_report.pdf.
16. "Data Exchange Agreement," Europeana Professional website, accessed April 22, 2013, http://pro.europeana.eu/web/guest/data-exchange-agreement.
17. "Europeana Data Model Mapping Guidelines v1.0.1," Europeana Professional website, February 24, 2012, http://pro.europeana.eu/documents/900548/ea68f42d-32f6-4900-91e9-ef18006d652e.
18. Haslhofer et al., "Europeana RDF Store Report," 97.
19. "Europeana Data Model Mapping Guidelines," 8.
20. Ibid., 5.
21. "Technical Details," Europeana Professional website, accessed April 23, 2013, http://pro.europeana.eu/tech-details.
22. Haslhofer et al., "Europeana RDF Store Report."
23. Karen Coyle, "Linked Data Tools: Connecting on the Web," *Library Technology Reports* 48, no. 4 (May/June 2012).