What's New in DPLA Technology?

Mark A. Matienzo
Director of Technology, DPLA

DPLAFest – Indianapolis, IN – April 17, 2015

Overview

- Heiðrún: DPLA's new ingestion system
  - Motivation for development/overall goals
  - Development progress/current features
  - Differences in workflow
  - Future work

- New IMLS grant on Hydra development
  - Overall goals
  - Opportunities for feedback

Heiðrún:
DPLA's New Ingestion System

- Development began October 2014
- Project team: Audrey Altman, Mark Breedlove, Gretchen Gueguen, Tom Johnson, Mark Matienzo, Amy Rudersdorf
- Initial objective: implementation of a new ingestion and partner management system for DPLA's metadata aggregation activities
- Long-term objective: develop generalized "aggregation system in a box" for DPLA Hubs and others

Motivation for development

- Architectural & scale issues
- DPLA MAP is an RDF model, but previous system did not deal with data as such
- Workflow was opaque to non-developer and partner staff and was an unbalanced workload
- Workflow too tightly coupled
- Inadequate tools for process/partner management

Overall Goals

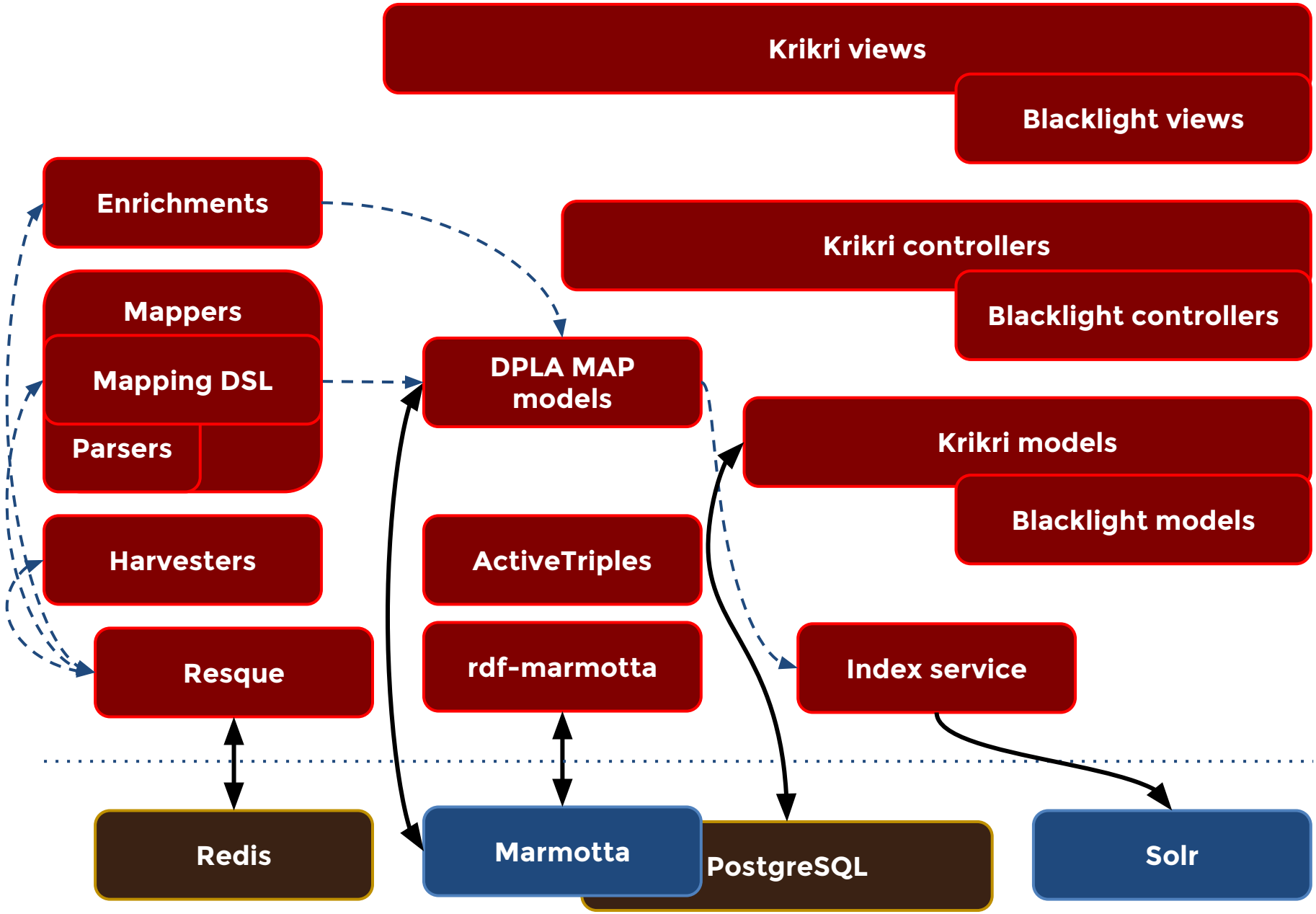
- New system for scalable harvesting, mapping, enrichment, storage, and indexing of metadata
- Native support for DPLA Metadata Application Profile v4 as RDF
- Web-based “dashboards” for DPLA and Hub staff, to support ingest scheduling, partner management, QA, and metadata mapping

Development Progress and Current Features

- Brand new modular architecture
- Metadata mapping language
- Better decoupled workflow
- QA improvements

Improved Architecture

- Real RDF: ActiveTriples & DPLA MAP modules
- Marmotta LDP server & triple store
- Easier to develop QA system using Blacklight, Solr, SPARQL queries
- Queueing and scheduling system


Krikri views
Blacklight views
Enrichments
Krikri controllers
Blacklight controllers
Mappers
Mapping DSL
Parsers
DPLA MAP models
Krikri models
Blacklight models
Harvesters
ActiveTriples
Resque
rdf-marmotta
Index service
Redis
Marmotta
PostgreSQL
Solr

Krikri

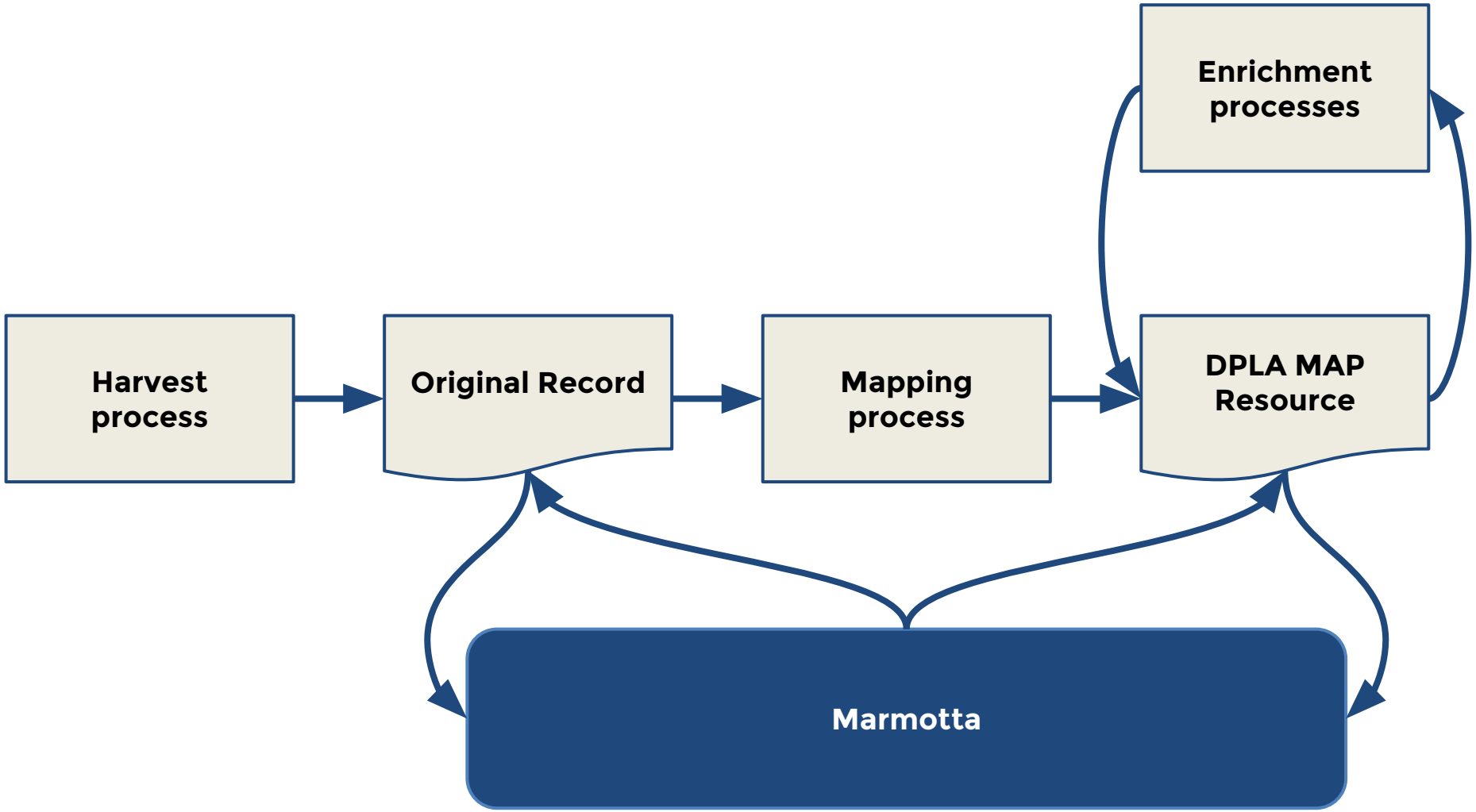
Krikri is:

- The core component of Heidrun.
- A Rails engine.

Krikri provides:

- Basic harvesters, mappers, and enrichments.
- User interfaces for managing workflows.
- CRUD interfaces for Marmotta, Solr/ElasticSearch, and a relational database (PostgreSQL or MySQL).

Ingestion Workflow



Metadata Mapping Language

- Used to define mappings from provider metadata to DPLA MAP properties and classes
- Allows for matching elements based on values, attributes, etc.
- Under continual improvement as complexity of mappings is addressed


```
Krikri::Mapper.define(:esdn_mods, :parser => Krikri::ModsParser) do
  provider :class => DPLA::MAP::Agent do
    uri 'http://dp.la/api/contributor/esdn'
    label 'Empire State Digital Network'
  end

  dataProvider :class => DPLA::MAP::Agent do
    providedLabel record.field('mods:note').match_attribute(:type, 'ownership')
  end

  isShownAt :class => DPLA::MAP::WebResource do
    uri record.field('mods:location', 'mods:url')
        .match_attribute(:usage, 'primary display')
        .match_attribute(:access, 'object in context')
  end

  preview :class => DPLA::MAP::WebResource do
    uri record.field('mods:location', 'mods:url')
        .match_attribute(:access, 'preview')
  end

  originalRecord :class => DPLA::MAP::WebResource do
    uri record_uri
  end

  sourceResource :class => DPLA::MAP::SourceResource do
    collection :class => DPLA::MAP::Collection, :each => header.field('xmlns:set_spec'), :as => :coll do
      title coll
    end

    contributor :class => DPLA::MAP::Agent, :each => record.field('mods:name')
                                                     .select { |name| name['mods:role'].map(&:value).include?('contributor') },
                :as => :contrib do
      providedLabel contrib.field('mods:namePart')
    end

    creator :class => DPLA::MAP::Agent, :each => record.field('mods:name')
                                                   .select { |name| name['mods:role'].map(&:value).include?('creator') },
            :as => :creator_role do
      providedLabel creator_role.field('mods:namePart')
    end

    date :class => DPLA::MAP::TimeSpan, :each => record.field('mods:originInfo'),
         :as => :created do
      providedLabel created.field('mods:dateCreated').match_attribute(:keyDate, 'yes')
                   .reject { |date| date.attribute? :point }
      self.begin created.field('mods:dateCreated').match_attribute(:point, 'start')
      self.end created.field('mods:dateCreated').match_attribute(:point, 'end')
    end

    description record.field('mods:note').match_attribute(:type, 'content')

    extent record.field('mods:physicalDescription', 'mods:extent')

    # non-DCMIType values from type will be handled in enrichment
    dcformat record.field('mods:physicalDescription', 'mods:form')

    genre record.field('mods:physicalDescription', 'mods:form')

    identifier record.field('mods:identifier')

    language :class => DPLA::MAP::Controlled::Language, :each => record.field('mods:language', 'mods:languageTerm'), :as => :lang do
      prefLabel lang
    end

    spatial :class => DPLA::MAP::Place, :each => record.field('mods:subject', 'mods:geographic'), :as => :place do
      providedLabel place
    end

    publisher :class => DPLA::MAP::Agent, :each => record.field('mods:originInfo'), :as => :publisher do
      providedLabel publisher.field('mods:publisher')
    end

    relation record.field('mods:relatedItem', 'mods:titleInfo', 'mods:title')

    rights record.field('mods:accessCondition')

    subject :class => DPLA::MAP::Concept, :each => record.field('mods:subject'), :as => :subject do
      providedLabel subject
    end

    title record.field('mods:titleInfo', 'mods:title')

    dctype record.field('mods:typeOfResource')
  end
end
```

QA improvements

- Better reporting, using search index
- Introduce Provenance Ontology (PROV-O) to track info about ingestion activities

Empire State Digital Network ▾ Dashboard Records Reports

Search... Search ☐

Limit your search

- Format ☐
- Place ☐
- Subject ☐
- Collection ☐
- Data Provider ☐
- Creator ☐

« Previous | **1 - 10** of **87,671** | Next »

10 per page ▾

1. http://ldp.dp.la/ldp/items/8eaa4c2a8a26819aa02426fddf320db0

Title: View of Geneva to the West taken from the Nestor Hotel

Description: View of Geneva to the West taken from the Nestor Hotel

Rights: This digital image may be used for educational purposes, as long as it is not altered in any way. Prior written permission is required from the Geneva Historical Society for any other use of the image.

2. http://ldp.dp.la/ldp/items/0fd6a7f6e06111c59e8c303227982d7b

Title: Empire state tax news. - WIN-2001

Rights: This document or image is provided for education and research purposes. Rights may be reserved. Responsibility for securing permissions to distribute, publish, reproduce or use it in any way rests with the user. For additional information, see the New York State Library's Copyright and Use Statement, available at http://www.nysl.nysed.gov/scandocs/rights.htm.

Validation Report: dataProvider_providedLabel

All Providers

The following records have missing values for `dataProvider_providedLabel`.

« Previous | **1** - **10** of **193** | Next »

10 per page ▾

ID:	http://ldp.dp.la/ldp/items/1d3b8694a8699d956476c40be62481d8
Title:	Street scene during a May Day celebration in Moscow
Is Shown At:	http://cdm16786.contentdm.oclc.org/cdm/ref/collection/eurasia/id/5620

ID:	http://ldp.dp.la/ldp/items/f213e252248be35437b6cf6c4ab1fdb2
Title:	Santa Barbara County Courthouse
Is Shown At:	http://cdm16786.contentdm.oclc.org/cdm/ref/collection/buildings/id/11092

ID:	http://ldp.dp.la/ldp/items/9e7836f70fa549da8bae13151eb94efc
Title:	Juniperus occidentalis
Is Shown At:	http://cdm16786.contentdm.oclc.org/cdm/ref/collection/plants/id/649

ID:	http://ldp.dp.la/ldp/items/b0e06c602b8d90035eda4517e45ed2f0
Title:	opp2-4
Is Shown At:	http://cdm16786.contentdm.oclc.org/cdm/ref/collection/ptec/id/1183

Record

http://ldp.dp.la/ldp/items/1d3b8694a8699d956476c40be62481d8

Random record from University of Washington

Enriched Record


```
        "identifier": [
          "5287",
          "http://cdm16786.contentdm.oclc.org/cdm/ref
/collection/eurasia/id/5620"
        ],
        "rights": "Augerot, James",
        "spatial": [
          {
            "@id": "_:b6",
            "@type": "edm:Place",
            "providedLabel": "Russia"
          },
          {
            "@id": "_:b5",
            "@type": "edm:Place",
            "providedLabel": "Tsentralnyy Ekonomiches
kiy Rayon"
          },
```


Original Record


```
c-1.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-i
nstance" xsi:schemaLocation="http://worldcat.org/xmls
chemas/qdc-1.0/ http://worldcat.org/xmlschemas/qdc/1.
0/qdc-1.0.xsd http://purl.org/net/oclcterms http://wo
rldcat.org/xmlschemas/oclcterms/1.4/oclcterms-1.4.xsd
">
      <dc:title>Street scene during a May Day celebra
tion in Moscow</dc:title>
      <dc:creator>Augerot, James</dc:creator>
      <dc:date>1962</dc:date>
      <dc:subject>City and town life ;  Celebrations
;  Parades and processions ;  Festive decorations ;
Flags ;  Streets</dc:subject>
      <dc:coverage>Russia</dc:coverage>
      <dc:coverage>Tsentralnyy Ekonomicheskiy Rayon</
dc:coverage>
      <dc:coverage>Moskovskaya Oblast</dc:coverage>
      <dc:coverage>Moskva</dc:coverage>
      <dc:contributor>Augerot, James</dc:contributor>
```

Future Plans

Tools to automate and ease workflow

- Responsive metadata mapping interface.
- QA reports that target common ingestion errors.
- Ingestion dashboard for scheduling and monitoring harvests.

A metadata aggregation toolkit for the community

- Define an RDF model for your data using ActiveTriples.
- Harvest metadata (RDF or non-RDF) from a various sources.
- Map harvested metadata to your RDF model.
- Enrich metadata by linking to other RDF sources.

Hub Dashboard

Past Ingest Events

Ingest Date	Start Time	End time	Records Retrieved	Errors	Live Records	Live Date
8/5/2014	8:15am	3:25pm	55,774	347	55,400	8/7/2014
12/7/2014	10:45am	3:32pm	25,117	17	-cancelled	-cancelled

Current Ingest

Ingest Date	Start Time	End time	Records Retrieved	Errors
1/9/2015	8:15am	3:25pm	55,774	347

QA

Test

Start New Ingest

More Information on Krikri/Heiðrún

- Project page:
  - http://bit.ly/heidrun

- GitHub projects:
  - http://github.com/dpla/krikri
  - http://github.com/dpla/heidrun

“Hydra-in-a-Box” Project

- Project partners: DPLA, Stanford University, DuraSpace
- $2M, 30 month grant awarded by Institute of Museum and Library Services, in their National Leadership Grants program
- Focuses on fostering a new, national, library network through a community-based repository
- Leverages successes and infrastructure from the Hydra project and community, while contributing back to it

“Hydra-in-a-Box” Goals

- Development of turnkey, Hydra-based application
- Improve and DPLA’s metadata ingestion system into an “aggregator system in a box”
- Connect components with pieces with DPLA hubs, current Hydra partners, and prospective Hydra adopters
- Work toward a hosted service

More information on “Hydra-in-a-Box”

Coming soon!
Contact mark@dp.la

Contributor Agreements & Licensing for DPLA Software Projects

- Addressing need to define terms under which code has been contributed to DPLA projects
- Development of draft guidelines for DPLA staff to select FLOSS licenses
- DPLA in the process of determining best options for contributor agreements, e.g.:
  - What is the state of the art?
  - Should we use contributor license agreements or copyright assignment agreements?

Planning for the future: DPLA and the Technical Advisory Committee

- Survey for feedback:
  - http://bit.ly/dpla-techadvisory-survey
  - Feel free to submit after DPLAFest
  - Open to all

- Small group discussion:
  - What should we be thinking about over the next year or more?
  - Who should be in the conversation?
  - How can the Technical Advisory Committee help DPLA?

- Report back and next steps