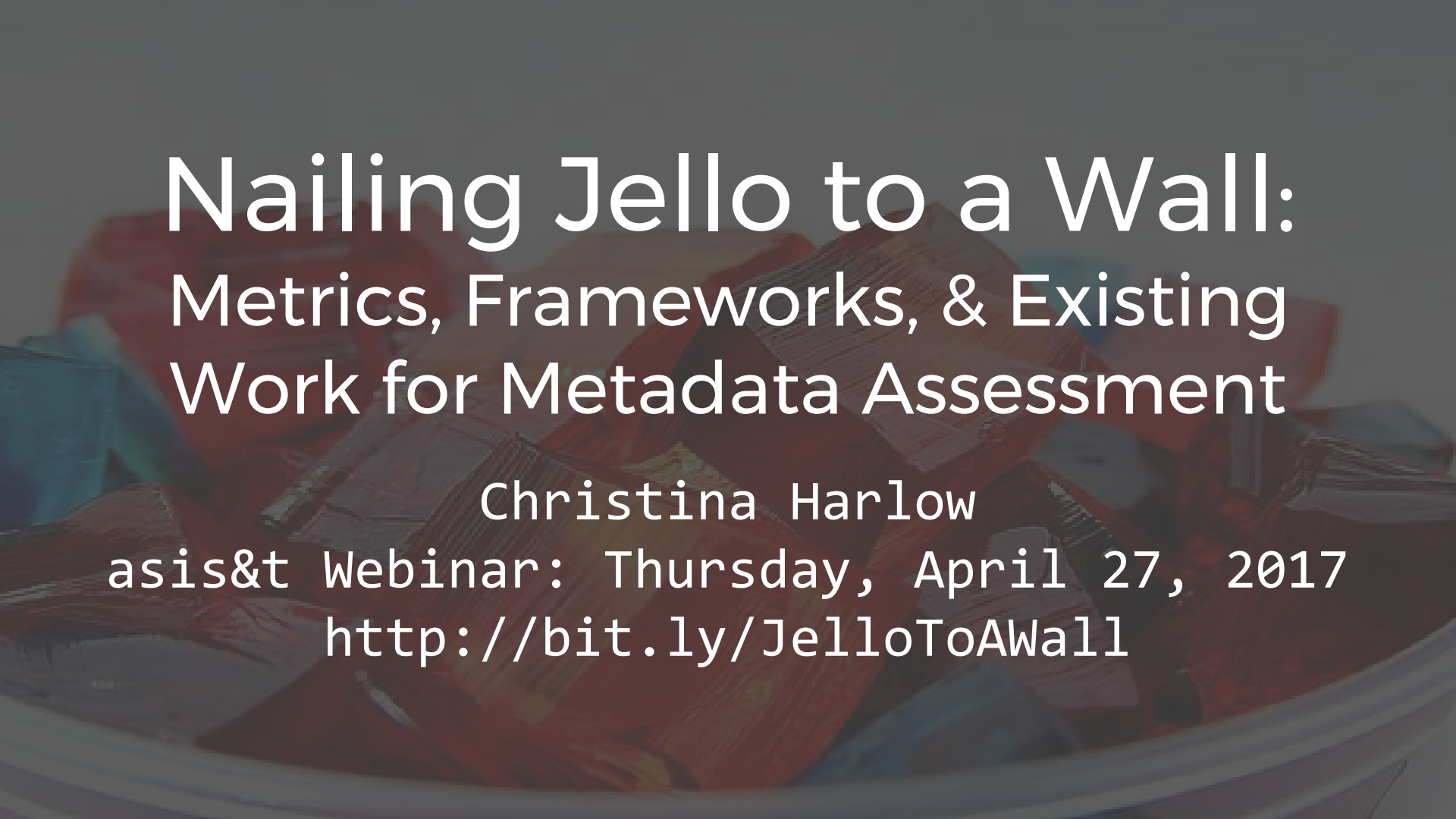


Nailing Jello to a Wall:
Metrics, Frameworks, & Existing Work for Metadata Assessment

Christina Harlow

asis&t Webinar: Thursday, April 27, 2017

http://bit.ly/JelloToAWall


http://bit.ly/JelloToAWall

About Your Speaker

Metadata Librarian
Cornell University Library

cmh329@cornell.edu
@cm_harlow

http://bit.ly/JelloToAWall

About Your Speaker

~~Metadata Librarian~~
~~Cornell University Libraries~~

Repository Specialist, Data Operations
Stanford University Libraries

~~cmh329@cornell.edu~~
cmharlow@stanford.edu
@cm_harlow

http://bit.ly/JelloToAWall

Topics in Today's Webinar

I. Use Cases for Metadata Assessment

http://bit.ly/JelloToAWall

Topics in Today's Webinar

I. Use Cases for Metadata Assessment

II. Metrics, Context, & “Quality”

http://bit.ly/JelloToAWall

Topics in Today's Webinar

I. Use Cases for Metadata Assessment
II. Metrics, Context, & “Quality”
III. Guidelines for Performing Assessment

http://bit.ly/JelloToAWall

Topics in Today's Webinar

I. Use Cases for Metadata Assessment
II. Metrics, Context, & “Quality”
III. Guidelines for Performing Assessment
IV. Examples of Analysis Workflows & Tools

http://bit.ly/JelloToAWall

Topics in Today's Webinar

I. Use Cases for Metadata Assessment
II. Metrics, Context, & “Quality”
III. Guidelines for Performing Assessment
IV. Examples of Analysis Workflows & Tools
V. Further Resources & Engagement

http://bit.ly/JelloToAWall

I. Use Cases for Metadata Assessment

http://bit.ly/JelloToAWall

Moving Beyond Discovery Interfaces Checking as Metadata Assessment

Why Do We Assess Metadata?

Handling New Object Types

Impact of Metadata Work

Migrations & Data Sharing

Profile Generation

Standards Choice

System Design Aid

Targeted Enhancement

Validation & Expectations

http://bit.ly/JelloToAWall

Handling New Object Types

Surfacing needs of special or unique types of materials that either are not sufficiently captured for current metadata usage, do not fit well within existing profiles or standards.

RIFF chunks (fields defined in Embedded Metadata Profile above are shown in bold text below):

field	external NS mapping	expected value	obligation	notes
RIFF:IARL		example (default): Cornell U Library example: Kheel Library	{1,1}	Specified by client organization
RIFF:IART		()	{0,1}	Blank - unless specified by client organization
RIFF:ICMS		example (default): Cornell U	{0,1}	If not Cornell University, leave blank or use client specif
RIFF:ICMT		example: () example: Damaged cassette, tape rehoused. One minute and twenty seconds of silence between separate interviews removed from file. example: Optimized extremely poor azimuth in original recording. example: Captured circulation copy	{0,1}	Digitization notes unique to file. Usually blank, indicatin copy" example notes when circulation copy is captured is assumed otherwise.
RIFF:ICOP		()	{0,0}	Blank - In catalog record
RIFF:ICRD		()	{0,0}	Blank - In catalog record
RIFF: ICRP		()	{0,0}	Blank - not applicable
RIFF: IDIM		()	{0,0}	Blank - not applicable
RIFF: IDPI		()	{0,0}	Blank - not applicable
RIFF:IENG		example (default): Fitzke, Karl E.	{1,1}	Supervising Engineer, e.g. person responsible for QC.

http://bit.ly/JelloToAWall

Impact of Metadata Work

Broad area to both measure the impact of metadata in discovery or other systems (through analytics or other), as well as to link metadata assessment to other areas of work, such as training/reskilling.

http://bit.ly/JelloToAWall

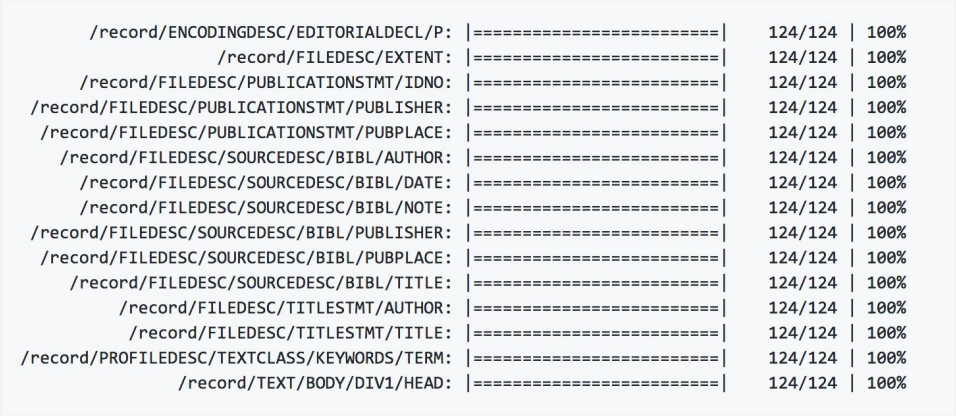
Migrations & Data Sharing

Assessment work done to support or enable the sharing, lossless conversion, or migration of metadata and data between data systems, standards, and repositories.

```
<#ConceptIdentifier>
    rml:logicalSource [
      rml:source "/Users/Christina/Tools/lts-vitro/lts-vit
      rml:referenceFormulation ql:JSONPath;
      rml:iterator "$"
    ];

    rr:subjectMap [
      rr:template "http://vitropilot.internal.library.corn
    ];

    rr:predicateObjectMap [
      rr:predicate rdfs:label;
      rr:objectMap [ rml:reference "$.conceptDetailInfo.di
    ];

    rr:predicateObjectMap [
      rr:predicate foaf:name;
```

http://bit.ly/JelloToAWall

Profile Generation

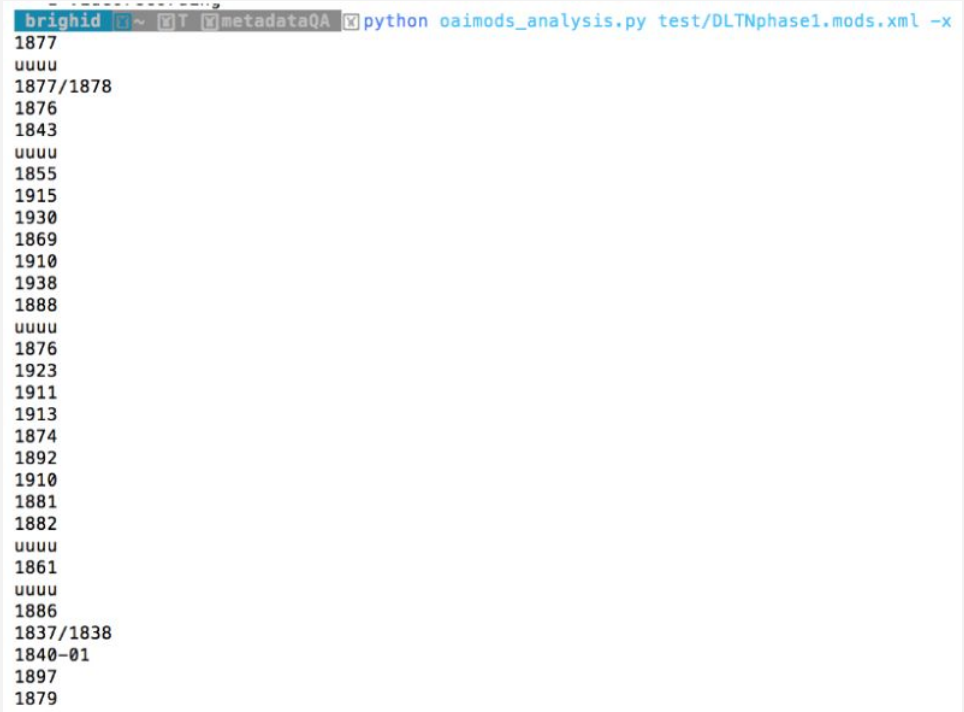


Metadata Application Profile: resource that defines the expected, recommended, & optional fields, as well as proposed values sources & standards, for metadata in particular application.

http://bit.ly/JelloToAWall

Standards Choice

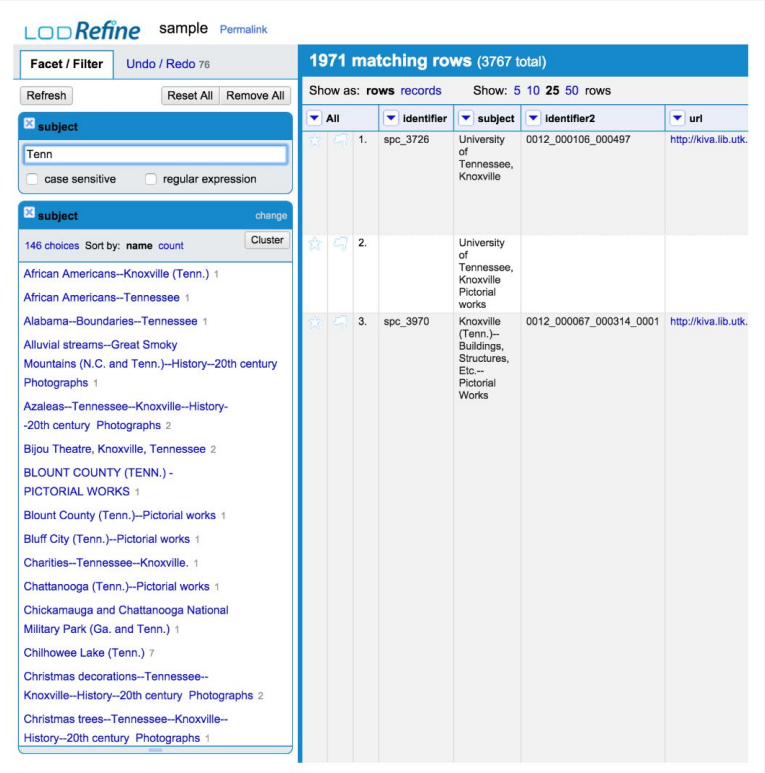
Decision of which standards- metavocabs, controlled vocabularies, encoding, formats, or other - best fit the current needs, the proposed needs, and the existing & proposed instance metadata.



http://bit.ly/JelloToAWall

Targeted Enhancement

Assessing metadata for areas of work at intersection of most impactful according to context, but also most efficient to perform normalization or enhancement work with given resources.



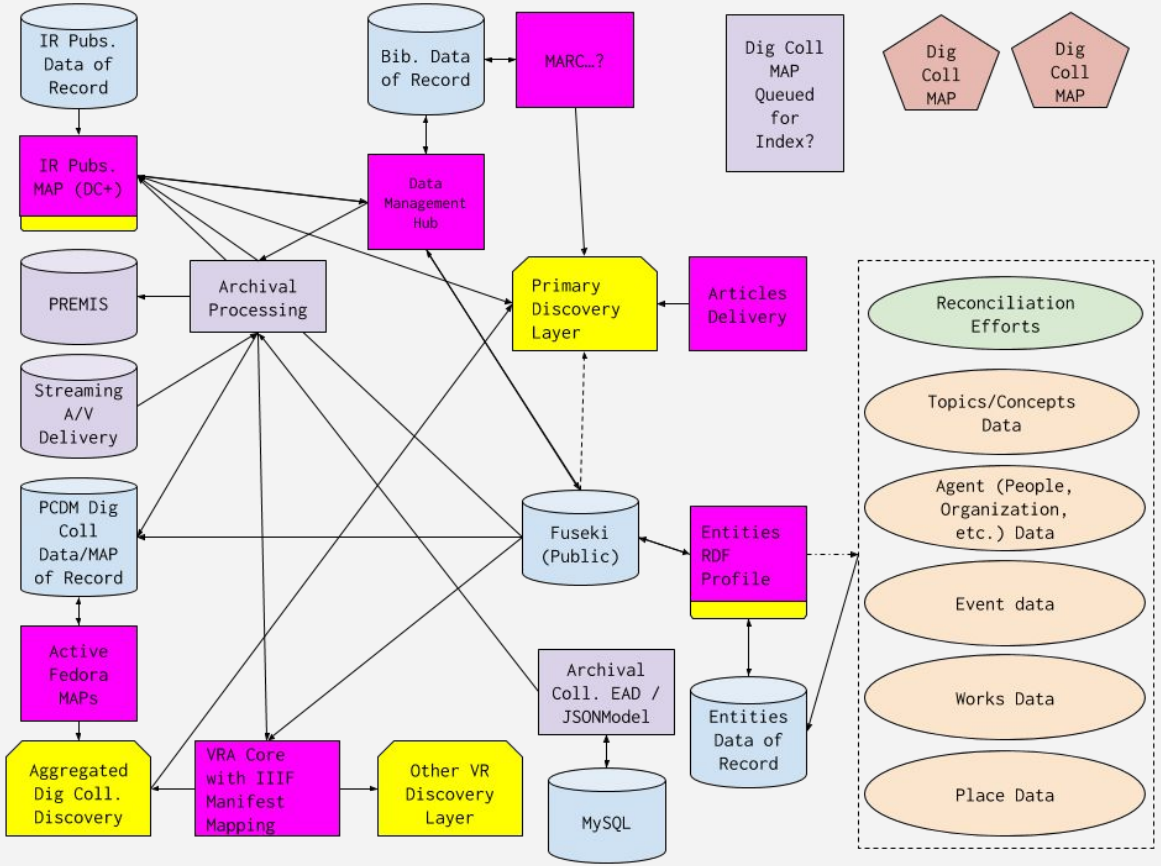
http://bit.ly/JelloToAWall

Validation & Expectations

```
"type": "object",
"required": [ "collection" ],
"properties": {
    "/": {},
    "collection_name": {"$ref": "#/definitions/collection"},
    "items": {"$ref": "#/definitions/item"},
    "item_subdir": {"$ref": "#/definitions/subdir"},
    "filename": {"$ref": "#/definitions/file"}
},
"definitions": {
    "collection": {
        "type": "object",
        "properties": {
            "phys_coll_id": { "type": "string" },
            "number_files": { "type": "integer" },
            "steward": {
                "type": "string",
                "pattern": "^[a-zA-Z]{1,4}[0-9]{1,6}$" },
            "date_s3_ingest": {
                "type": "string",
                "pattern": "^[0-9]{4}-[0-9]{2}-[0-9]{2}$" },
            "date_s3_update": {
                "type": "string",
                "pattern": "^[0-9]{4}-[0-9]{2}-[0-9]{2}$" },
            "items": {
                "type": "object",
                "$ref": "#/definitions/item"
            }
        },
        "required": [ "date_s3_ingest", "items" ]
```

Checking metadata follows a certain standard, profile, schema, or other meta-vocabulary, &/or conforms to the defined structure, usage, & expectations.

http://bit.ly/JelloToAWall



Metadata Assessment & Systems

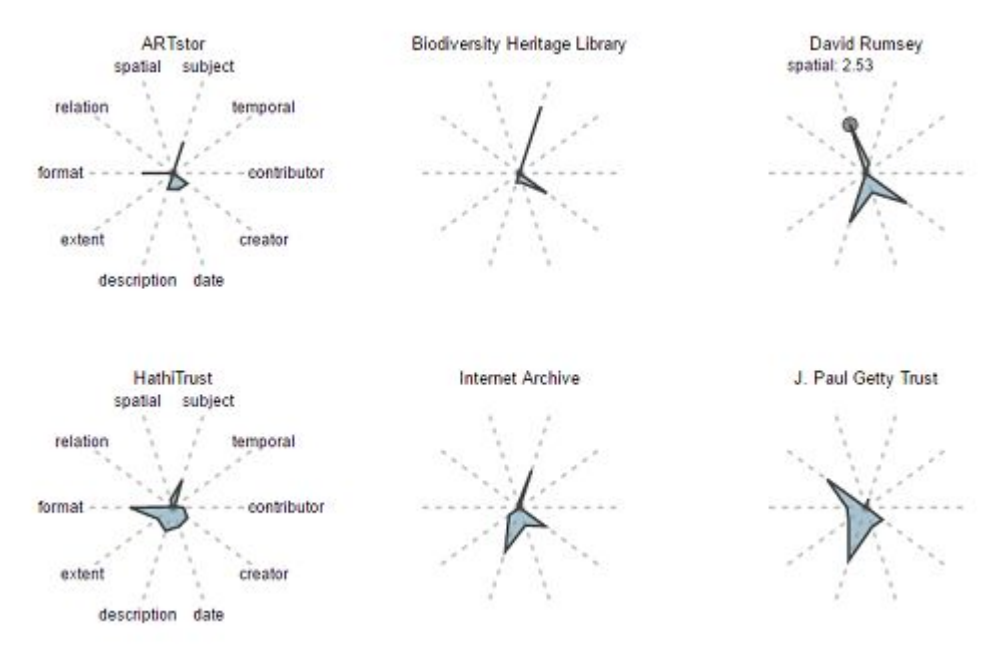
http://bit.ly/JelloToAWall

Other Reasons for Assessment…

Metadata “Quality”

Alternate Discovery?

Metadata Assessment as Research



http://bit.ly/JelloToAWall

Metadata Assessment First Involves Setting Context & Scope

Otherwise…

Nailing Jello to a Wall: U.S. English idiom that describes a task that is difficult because the parameters keep changing (like how Jello/Jell-o moves).

http://bit.ly/JelloToAWall

II. Metrics, Context, & “Quality”

http://bit.ly/JelloToAWall

Some Writing & Research...

- Bruce, Thomas R. & Hillmann, Diane I. (2004). The Continuum of Metadata Quality
- Bruce, Thomas R. & Hillmann, Diane I. (2013). Metadata Quality in a Linked Data Context.
- Europeana Tech. Evaluation and Enrichments Task Report Outcomes.
- Zavalina, Oksana; Kizhakkethil, Priya; et al. (2015). Building a Framework of Metadata Change to Support Knowledge Management.
- Zaveri, Amrapali, et al. (2015). Quality Assessment for Linked Data: A Survey. *(Not Available Online/OA)*

http://bit.ly/JelloToAWall

Some Practice...

- Harper, Corey A. (2016). Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA).
- Hochstenbach, Patrick (2016). Metadata Analysis at the Command-Line.
- Király, Péter (2015). A Metadata Quality Assurance Framework.
- Harlow, Christina (2015). Metadata Quality Analysis: Tools & Scripts to Check Your Data.
- Phillips, Mark (2013). Metadata Analysis at the Command-Line.

http://bit.ly/JelloToAWall

Some Proposed Metadata Quality Metrics

Accessibility

Accuracy

Availability (Technical)

Completeness

Conciseness

Conformance to expectations

Consistency & Coherence

Interlinking

Interoperability

Licensing

Normalization & Enhancement

Performance

Provenance

Timeliness

http://bit.ly/JelloToAWall

Metadata allows multiple access points via language, shared understanding of concepts, indication of accessibility, or other.

Accessibility

http://bit.ly/JelloToAWall

Correct use of the field;

Appropriate values captured;

Correctness of metadata.

Accuracy

http://bit.ly/JelloToAWall

Data server response;

Presence of data dumps;

Correct content types.

Availability

http://bit.ly/JelloToAWall

Obligations of fields;

Required or recommended;

Data retrieval & capture in fields.

Completeness

http://bit.ly/JelloToAWall

Avoid redundancy of fields, whether through multiple fields usage that have same meaning, or through annotations & schema usage.

Conciseness

http://bit.ly/JelloToAWall

Use of standards and standard data formatting;

Obligations for fields are fulfilled.

Conformance to expectations

http://bit.ly/JelloToAWall

Consistency & Coherence

Field values are normalized as applicable;

Fields are used consistently across instance data.

Yes

- A property not used by any other data
- A specific instance of a property that is used multiple times (i.e. first or last instance) that is consistently found in EVERY RECORD
- In the same property or small subset of properties in EVERY RECORD (including attribute variations)

In other words, something that can be logically predicted.

NO

- Must be parsed out of a data value (e.g. all the ones that start with “http://… etc.)
- Sometimes occurs in a specific instance of a repeated field but not in EVERY RECORD
- Occurs in a variety of properties, or in the same property with a variety of attributes

In other words, something that requires human intelligence or sophisticated logic
to find.

http://bit.ly/JelloToAWall

Good quality interlinks;

Links to external datasets, data publishers;

Check for link rot.

Interlinking

http://bit.ly/JelloToAWall

Reuse of external schema, terms, vocabularies;

Clear indication of source of terms & fields.

Interoperability

http://bit.ly/JelloToAWall

Presence of license;

License assigned is machine-readable;

Assigned license is correct.

Licensing

http://bit.ly/JelloToAWall

Previous cleanup, enhancement, or normalization jobs have been run on the metadata;

Values or scores present from enhancements.

Normalization & Enhancement

http://bit.ly/JelloToAWall

Low latency where applicable;

High throughput (able to handle many HTTP requests);

Scalability of data publication.

Performance

http://bit.ly/JelloToAWall

History of metadata creation/edits;

Originating source of metadata & metadata additions.

Provenance

http://bit.ly/JelloToAWall

Currency of the data captured;

Connection between changing resources & updated metadata.

Timeliness

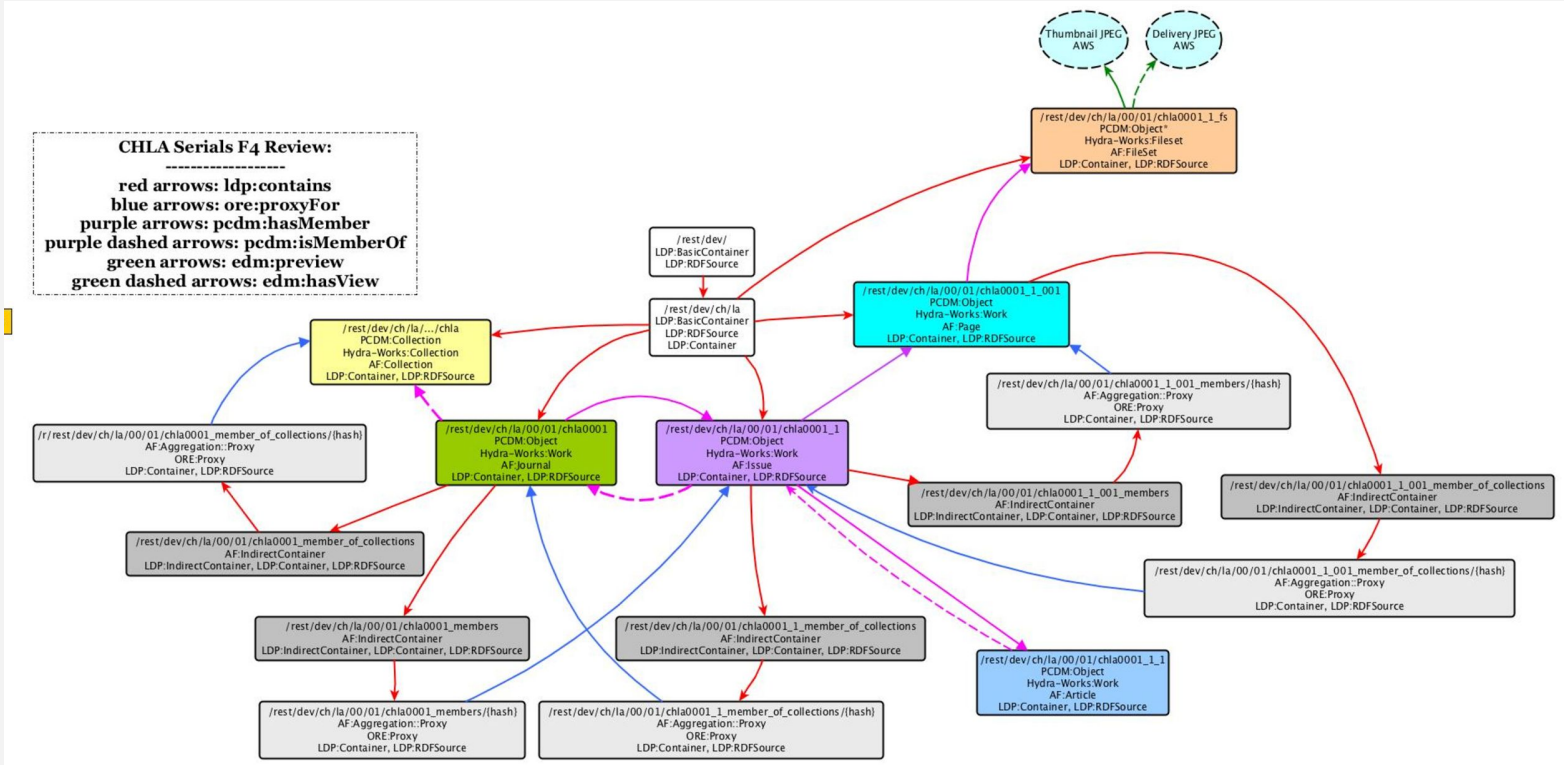
http://bit.ly/JelloToAWall

More Diverse, Interconnected Metadata Require Defining of Edges for Assessment

Metadata Assessment Also Includes Data Management Practices Review

III. Guidelines for Performing Assessment

Define & Document Your Context



http://bit.ly/JelloToAWall

Metadata Application Profiles

1. What are you describing with this metadata?
2. What do you intend to do with this metadata?
   a. Share with or generate from other systems?
   b. Enable some sort of discovery, lookup, resource management, or other functionality?
   c. Use within a particular system?
3. How will this metadata be generated, managed, and exposed? By whom or what processes?

Generic MAP Starter Template

http://bit.ly/JelloToAWall

Metadata Application Profiles

PCDM:Object > HydraWorks:Work | Article

Articles that belong to a particular Issue. This seems to be the level that files and OCR sit at (thus nullifying the need for page objects - but need to confirm).

Will require a lot of metadata improvements.

Descriptive Profile

field name	predicate	mapping or collection-wide static value [range]	notes
id	**dcterms:identifier**	HEADER/FILEDESC/PUBLICATIONSTMT/IDNO [literal]	n/a
author	**marcrel:author?**	TEXT/BODY/DIV1/DI V2/HEAD/AUTHOR	Needs checking, normalization.
title	**dcterms:title**	TEXT/BODY/DIV1/DI V2/HEAD/TITLE	n/a
OCR	**dcterms:relations**	TEXT/BODY/DIV1/DI V2/P/P	This is at article level?

Structural Profile

http://bit.ly/JelloToAWall

Build Out Your Data Documentation with Your Assessment Tools

Machine-Actionable Mappings

```
marc_map("020a",isbn.$append)
join_field(isbn," | ")
split_field(isbn, " | ")

marc_map("245ab",title, join:" ")

# Subject Identifiers?
if marc_match('6**0', "\w+")
  set_field("subjectRecon", "exists")
else
  set_field("subjectRecon", "does not exist")
end

# Name Identifiers?
if marc_match('1**0','\w+')
        set_field("nameRecon", "exists")
else
        set_field("nameRecon", "does not exist")
end
```

```
<#ConceptIdentifier>
    rml:logicalSource [
      rml:source "/Users/Christina/Tools/lts-vitro/lts-vit
      rml:referenceFormulation ql:JSONPath;
      rml:iterator "$"
    ];

    rr:subjectMap [
      rr:template "http://vitropilot.internal.library.corn
    ];

    rr:predicateObjectMap [
      rr:predicate rdfs:label;
      rr:objectMap [ rml:reference "$.conceptDetailInfo.di
    ];

    rr:predicateObjectMap [
      rr:predicate foaf:name;
```

http://bit.ly/JelloToAWall

Validation Profiles & “Continuous Testing”

```
"type": "object",
"required": [ "collection" ],
"properties": {
    "/": {},
    "collection_name": {"$ref": "#/definitions/collection"},
    "items": {"$ref": "#/definitions/item"},
    "item_subdir": {"$ref": "#/definitions/subdir"},
    "filename": {"$ref": "#/definitions/file"}
},
"definitions": {
    "collection": {
        "type": "object",
        "properties": {
            "phys_coll_id": { "type": "string" },
            "number_files": { "type": "integer" },
            "steward": {
                "type": "string",
                "pattern": "^[a-zA-Z]{1,4}[0-9]{1,6}$" },
            "date_s3_ingest": {
                "type": "string",
                "pattern": "^[0-9]{4}-[0-9]{2}-[0-9]{2}$" },
            "date_s3_update": {
                "type": "string",
                "pattern": "^[0-9]{4}-[0-9]{2}-[0-9]{2}$" },
            "items": {
                "type": "object",
                "$ref": "#/definitions/item"
            }
        },
        "required": [ "date_s3_ingest", "items" ]
```

```
3  # Fields 200-24X, except 240 - Other title fields - R1
4
5  describe 'bf:originDate from Title, 245 subfield f' do
6    context '$f - Inclusive dates', :bf do
7      let!(:graph) {
8        marcxml = '<record xmlns="http://www.loc.gov/MARC21/slim">
9          <leader>00956nam 2200229 4500</leader>
10         <controlfield tag="001">catalogKeyID</controlfield>
11         <controlfield tag="005">19911001004553.0</controlfield>
12         <controlfield tag="007">hd-afa014bacu</controlfield>
13         <controlfield tag="008">870616r19761854ctu a 000 0 eng d</controlfield>
14           <datafield tag="245" ind1="1" ind2="0">
15             <subfield code="a">Diaries,</subfield>
16             <subfield code="f">1854-1921.</subfield>
17             <subfield code="h">[Microform]</subfield>
18           </datafield>
19       </record>'
20       self.send(MARC2BF_GRAPH_METHOD, marcxml, '245_subfield_f_title')
21     }
```

http://bit.ly/JelloToAWall

Semi-Automated / Targeted Human Review

```
(venv) $ python analysis/oaidc_analysis.py data/carli_bra_jack.oai.qdc.xml -i -p -e 'date' | grep 'False'
      oai:collections.carli.illinois.edu:bra_jack/2200 False
      oai:collections.carli.illinois.edu:bra_jack/2201 False
      oai:collections.carli.illinois.edu:bra_jack/2202 False
      oai:collections.carli.illinois.edu:bra_jack/2203 False
      oai:collections.carli.illinois.edu:bra_jack/2204 False
      oai:collections.carli.illinois.edu:bra_jack/2205 False
      oai:collections.carli.illinois.edu:bra_jack/2206 False
      oai:collections.carli.illinois.edu:bra_jack/2207 False
      oai:collections.carli.illinois.edu:bra_jack/2208 False
      oai:collections.carli.illinois.edu:bra_jack/2209 False
      oai:collections.carli.illinois.edu:bra_jack/2210 False
      oai:collections.carli.illinois.edu:bra_jack/2211 False
      oai:collections.carli.illinois.edu:bra_jack/2212 False
      oai:collections.carli.illinois.edu:bra_jack/2213 False
      oai:collections.carli.illinois.edu:bra_jack/2214 False
...
```

http://bit.ly/JelloToAWall

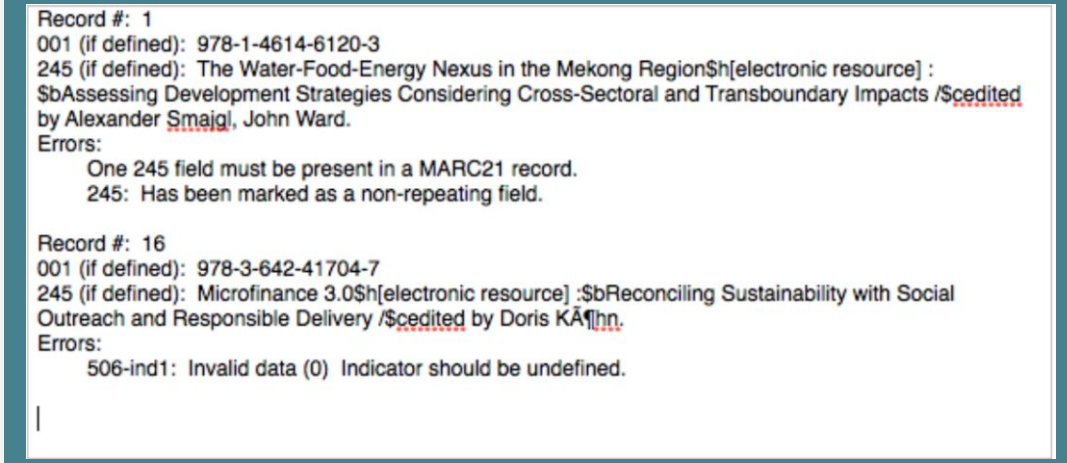
Metadata Assessment Will Sometimes Require Derivative Datasets

IV. Examples of Analysis Workflows & Tools

http://bit.ly/JelloToAWall

Using the Tools You Got

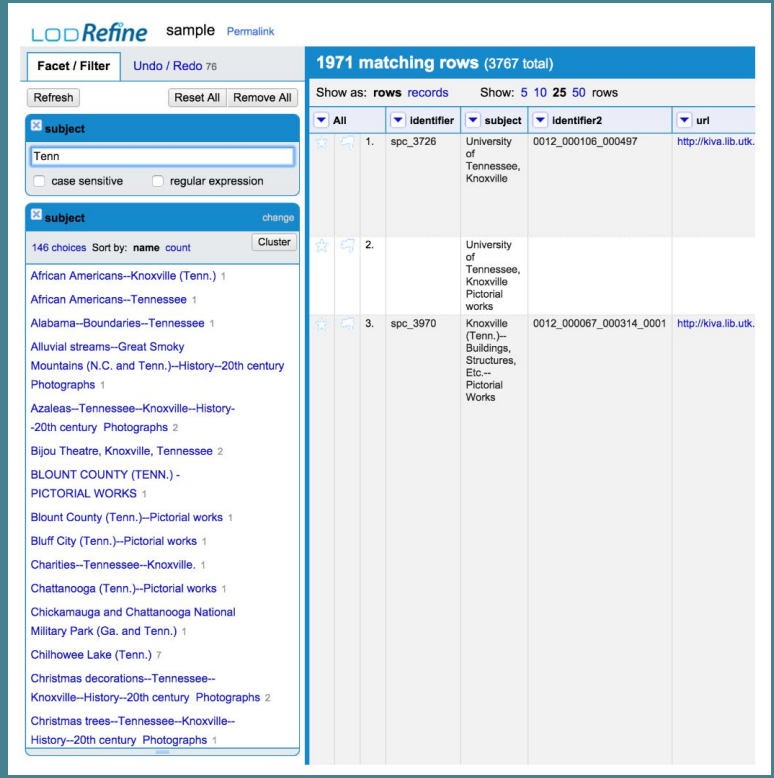
MARCEdit



http://bit.ly/JelloToAWall

Using the Tools You Got

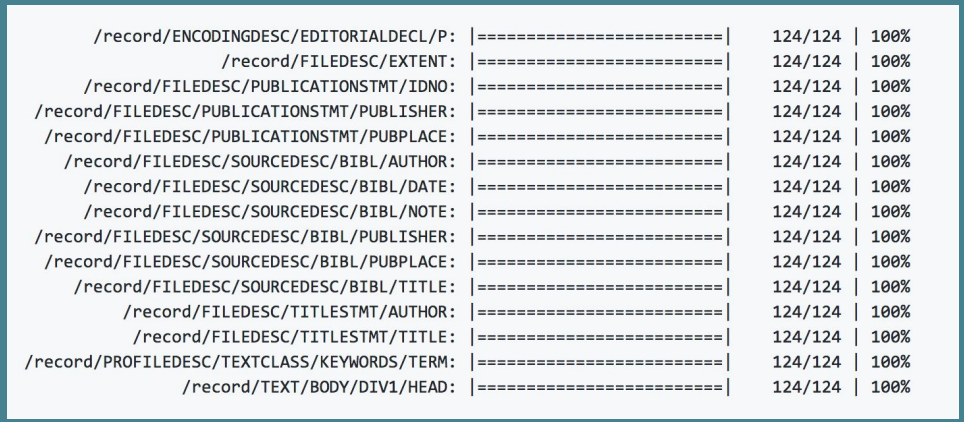
OpenRefine



http://bit.ly/JelloToAWall

Building Out the Duct Tape You Need

Python Metadata Breakers



http://bit.ly/JelloToAWall

Building Out the Duct Tape You Need

Catmandu Metadata Breakers

```
$ catmandu convert MARC to Breaker --handler marc < t/camel.usmarc >
result.breaker
$ catmandu breaker result.breaker
| name | count | zeros | zeros% | min | max | mean | median | mode | variance |
stdev | uniq | entropy |
|------|-------|-------|--------|-----|-----|------|--------|--------|----------|-
------|------|---------|
| 001 | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 10 | 3.3/3.3 |
| 003 | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.0/3.3 |
| 005 | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 10 | 3.3/3.3 |
| 008 | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 10 | 3.3/3.3 |
| 010a | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 10 | 3.3/3.3 |
| 020a | 9 | 1 | 10.0 | 0 | 1 | 0.9 | 1 | 1 | 0.09 | 0.3 | 9 | 3.3/3.3 |
| 040a | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.0/3.3 |
| 040c | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.0/3.3 |
| 040d | 5 | 5 | 50.0 | 0 | 1 | 0.5 | 0.5 | [0, 1] | 0.25 | 0.5 | 1 | 1.0/3.3 |
| 042a | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.0/3.3 |
| 050a | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.0/3.3 |
| 050b | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 10 | 3.3/3.3 |
| 0822 | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.0/3.3 |
| 082a | 10 | 0 | 0.0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 3 | 0.9/3.3 |
| 100a | 9 | 1 | 10.0 | 0 | 1 | 0.9 | 1 | 1 | 0.09 | 0.3 | 8 | 3.1/3.3 |
| 100d | 1 | 9 | 90.0 | 0 | 1 | 0.1 | 0 | 0 | 0.09 | 0.3 | 1 | 0.5/3.3 |
```

http://bit.ly/JelloToAWall

Selective Querying

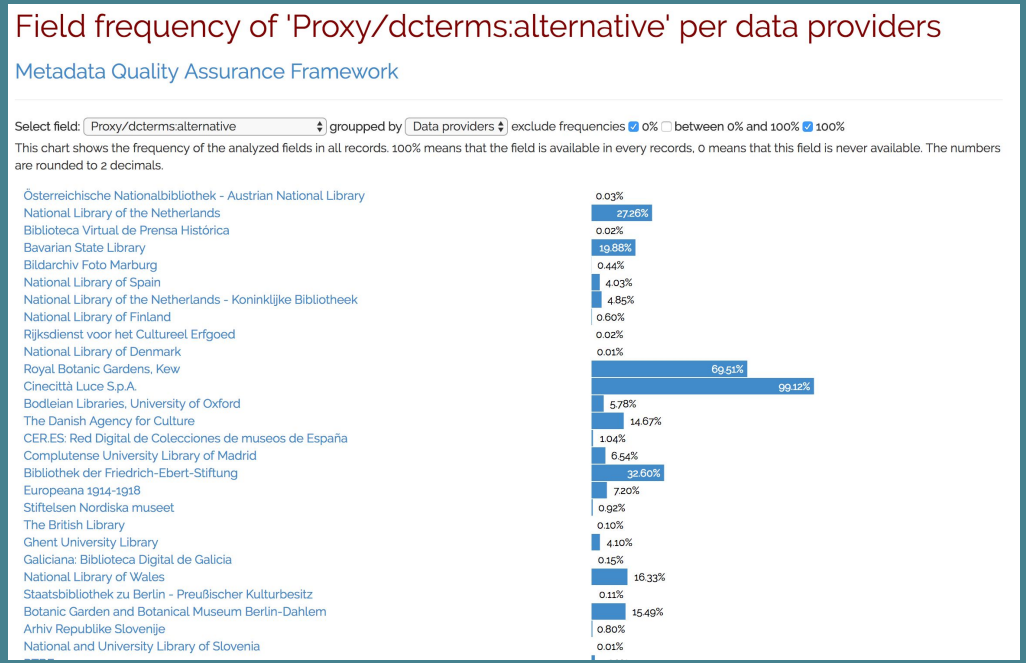
SQL/SPARQL & Response Checks

```
SELECT DISTINCT ?obj (COUNT(DISTINCT ?subject) as ?count)
WHERE {
  ?subject fedorasys:hasModel "Book" ;
           [prefix] ?obj .
}
GROUP by ?obj
```

http://bit.ly/JelloToAWall

Metadata MetaProfiling

Europeana QA Hadoop / Lucene / Interface



http://bit.ly/JelloToAWall

V. Further Resources & Engagement

http://bit.ly/JelloToAWall

DLF AIG Metadata Working Group

(Digital Library Federation Assessment Interest Group)

dlfmetadataassessment.github.io

www.zotero.org/groups/metadata_assessment

http://bit.ly/JelloToAWall

Europeana Task Force on Metadata Quality

pro.europeana.eu/publication/metadata-quality-task-force-report

http://bit.ly/JelloToAWall

DPLA Quality Assessment Working Group

(Digital Public Library of America)

bit.ly/dpla-metadata-bootcamp

github.com/dpla/Metadata-Analysis-Workshop

http://bit.ly/JelloToAWall

Metadata Assessment Needs Your Involvement!

Acknowledgements

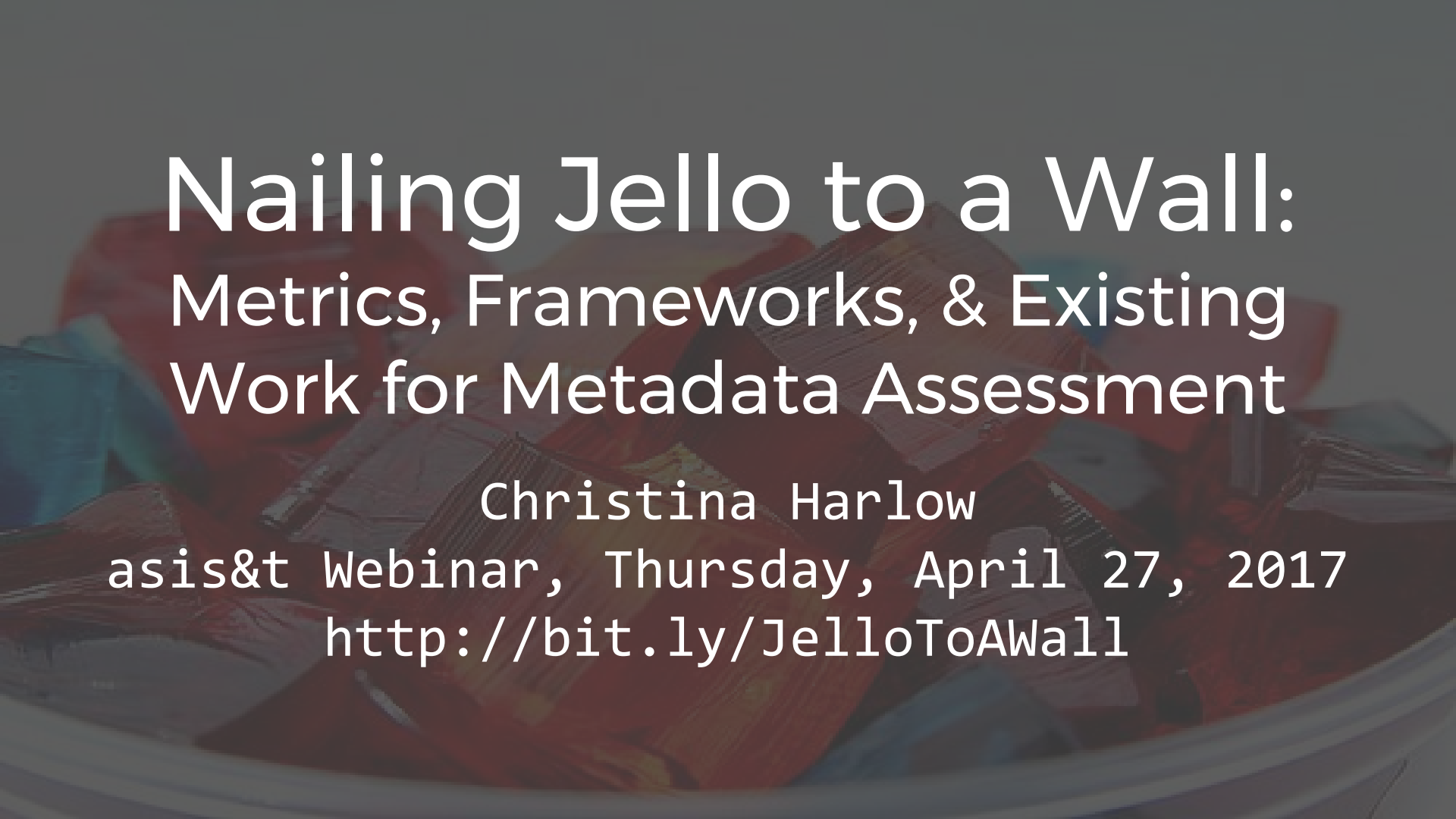
Members of DLF AIG Metadata Working Group

Members of Europeana / DPLA QA Efforts, Special Nod to Péter Király, Antoine Isaacs, & Gretchen Gueguen

Members of Open Library Technology Development Communities, especially Mark Phillips, Corey Harper & Patrick Hochstenbach

Everyone who has sat through my evolving set of workshops around this topic

http://bit.ly/JelloToAWall



Nailing Jello to a Wall:
Metrics, Frameworks, & Existing Work for Metadata Assessment

Christina Harlow

asis&t Webinar, Thursday, April 27, 2017

http://bit.ly/JelloToAWall