Analyzing Rights Statements in Digital Object Aggregators

Mark A. Matienzo <mark@dp.la>
Digital Public Library of America

Society of American Archivists Research Forum — Cleveland, Ohio
August 18, 2015
Outline

• What is DPLA?
• Developing a global framework for rights statements
• Data model for rights statements
• Analysis of rights statements in DPLA
What is DPLA?
The problem ...

other human expression
IRSWG Deliverables

• Shared framework/data model for rights statements under common namespace external to partners

• Best practice guidelines for aggregators and cultural heritage institutions to adopt rights statements

• Governance model to maintain framework
• Paul Keller (co-chair, Kennisland)
• Marie-Claire Dangerfield (Europeana)
• Julia Fallon (Europeana)
• Ranu Gayadin (Europeana)
• Lucie Guibault (University of Amsterdam)
• Antoine Isaac (Europeana)
• Lyubomir Kamenov (Europeana)
• Patrick Peiffer (B.N. Luxembourg)
• Joris Pekel (Europeana)
• Henning Scholz (Europeana)
• Maarten Zeinstra (Kennisland)

• Emily Gore (co-chair, DPLA)
• Esmé Cowles (UCSD)
• Greg Cram (New York Public Library)
• Karen Estlund (University of Oregon)
• Dave Hansen (University of North Carolina)
• Tom Johnson (DPLA)
• Matt Lee (Creative Commons)
• Melissa Levine (University of Michigan)
• Mark Matienzo (DPLA)
• Diane Peters (Creative Commons)
• Amy Rudersdorf (DPLA)
• Richard Urban (Florida State University)
IRSWG Data Model

- Vocabulary modeled using RDF/Simple Knowledge Organization System
- Aligned with other rights vocabularies (e.g. PREMIS Copyright Status and Europeana rights framework)
Rights statements in DPLA

• Analysis undertaken in October 2014 on 8.1 million records in the DPLA Metadata Application Profile

• Reports (CouchDB view) generated from DPLA’s primary metadata repository to return JSON-encoded information

• Reports contained rights statements, Hub, and contributing institution, and count
Analysis process

- Transformed into CSV and imported into OpenRefine
- Associated institutions removed
- Normalized (whitespace and multipass clustering process) to reduce and aggregate near-duplicates
Findings

- Incredible diversity and absence of rights statements in 8.1 million DPLA records
- 87,610 “unique” rights statements after normalization
- ~1.01M (~12.5%) records missing rights statements
- ~2.4M (~29.9%) “in copyright/(c)/all rights reserved”
- ~1.6% under a Creative Commons license
- Rights statements matching multiple “categories” that could be confusing to end users
Next steps

- More analysis necessary to review rights statements in further detail
- Finalize IRSWG data model and implementation guidelines
Thank You!

Mark A. Matienzo <mark@dp.la>
Digital Public Library of America