Forecasting the Costs for Preserving and Promoting Access to Biomedical Data

NAS Workshop
July 11-12, 2019
John Chodacki

University of California
CDL
California Digital Library

UC3


UNIVERSITY OF CALIFORNIA
University of California
CDL
California Digital Library
Open Access & Publishing
Collections
UC3
Discovery & Delivery

UC3 - University of California Curation Center

Digital Curation: maintaining, preserving, and adding value to digital research data throughout its lifecycle

UC3 currently focuses on:

- Research data management
- Data publishing
- Data/software skills training for librarians
- Digital preservation
- Persistent identifiers

...direct and indirect support for researchers

University of California
CDL
California Digital Library

UC3

Current Situation...across campuses

What are the costs of preservation?

How do we effectively communicate this to researchers?

How do we work with campuses to formalize and normalize processes?

Current Situation...across campuses

- Best practices and policies require we capture and preserve research outputs
- There are consistent structures to support research computing but ad hoc structures for long-term research data preservation
- There are specialists that can offer consistent stewardship of research data outputs but lack funding for (and expertise regarding) underlying storage
- Scope: data that underlies publications, long tail data, projects at end of lifecycle, etc.

Also...competitive advantage for institutions

Data Preservation Pilot

We wanted to solidify our connections. Leverage each other's strengths

- Research: policy, incentives, scholarship
- IT: capacity, reliability, technical hub of campus
- Libraries: stewardship, memory institutions

Data Preservation Pilot

Policy + Storage + Stewardship

University of California
CDL
California Digital Library

UC3

Our goals - make data FAIR

- ***Findable:*** assigned persistent identifiers & descriptive metadata that are registered or indexed in public catalogs, finding aids, & search engines
- ***Accessible***: openly retrievable using common protocols and tools
- ***Interoperable:*** associated metadata should provide vital description and context meaningful to the appropriate domain of scholarly discourse.
- ***Reusable***: provide transparent access to their provenance and change history. made available under the terms of permissive licenses, subject to appropriate ethical, legal, campus guidelines

Can’t do it alone

We want to leverage each other’s strengths

- Research: policy, incentives, scholarship
- IT: capacity, reliability, technical hub of campus
- Libraries: stewardship, memory institutions

We are all focused on the same successes. Building capacity for researchers.

Data Preservation Pilot

tackle, at an institutional level, the storage cost of long term preservation

- We got Research IT, VCRs, libraries to talk, understand, and commit
- We found our niche: reduce hurdles by making upfront capital investments in storage and making this available for IT teams, etc.
- Our focus: long tail data and orphaned projects
- Three campuses joined pilot in 2018.
- Determined to sunset pilot in 2019

University of California
CDL
California Digital Library

UC3

Lessons Learned

We need to make preservation a more compelling story for researchers

It was difficult to demonstrate the value of long term preservation to researchers.

We were piloting a service that focused on the back-end storage costs for back-end preservation services.

This was not an easy story to tell and quite often our outreach to campuses and researchers was lost when describing this relationship.

Lessons Learned

Smaller scale ≠ smaller scope

Our original premise was that a systemwide effort at data preservation would be the most efficient approach.

However, as the pilot progressed, we realized that the wider academic community was also grappling with similar cost issues.

Pilot team members realized that appropriate economies of scale should actually come from collaborations beyond the UC system.

Lessons Learned

We need to keep our eyes on the prize

Our original goal was to remove the cost barriers to data preservation and increase the number of quality data sets preserved.

The pilot team remained focused on this as our goal and the pilot experience gave us the space to brainstorm alternative approaches to tackling this issue.

This consistent focus on our ultimate goal eventually led to the partnership we forged with Dryad (described further below).

Partnerships help

- External partnerships can offer success stories
- Tackling digital preservation for long tail data
- Alignment in organizational values & mission
- Driving adoption of curated, FAIR, research data publishing
- Direct integrations with publishing platforms, preprints, computing environments

University of California
CDL
California Digital Library

UC3

Dryad

- Institutional membership: No DPCs for UC researchers
- Institutional Single Sign On
- CoreTrust Seal certified
- Up to 300GB per DOI
- Assigning DataCite DOIs
- Networked metadata with funder information and clean institution affiliation
- Consistent APIs for deposit and retrieval
- Helping control and model costs for a large set of our data issues

University of California
CDL
California Digital Library

UC3

Interconnected relationships

University of California
CDL
California Digital Library

UC3

DMPTool

dmptool.org

- Platform for DMP creation and guidance with 43 templates for 17 US funders (NSF, NIH, DOE, DOT, etc.) and international funders
- 31k+ users with 28k+ data management plans at 237 participating institutions
- Open to anyone. 250+ campuses around the world have custom DMPtool

- NSF-funded research project to prototype machine-actionable DMPs
- Need more connections to the wider ecosystem

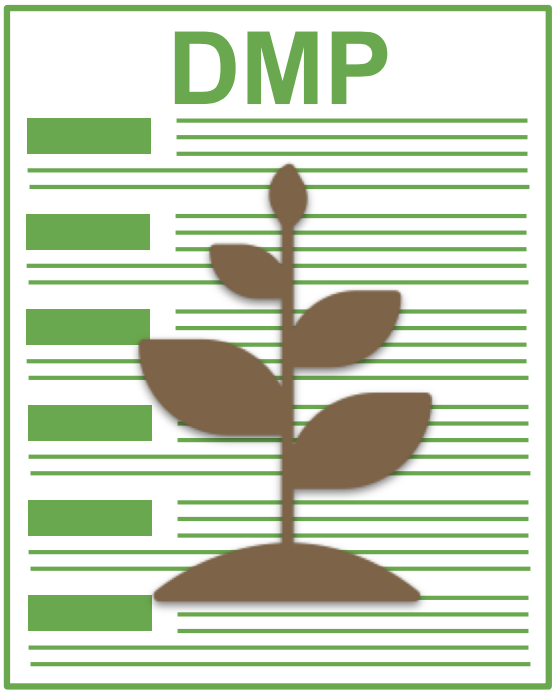
University of California
CDL
California Digital Library

EAGER Grant

DMPs become an active document

To help with forecasting costs, we need DMPs to:

- expose **structured information** as a project progresses over time including data volume
- make info available to right parties **over time** (i.e., respecting privacy until it can be public)
- be update-able over time by multiple parties in a **decentralized** fashion



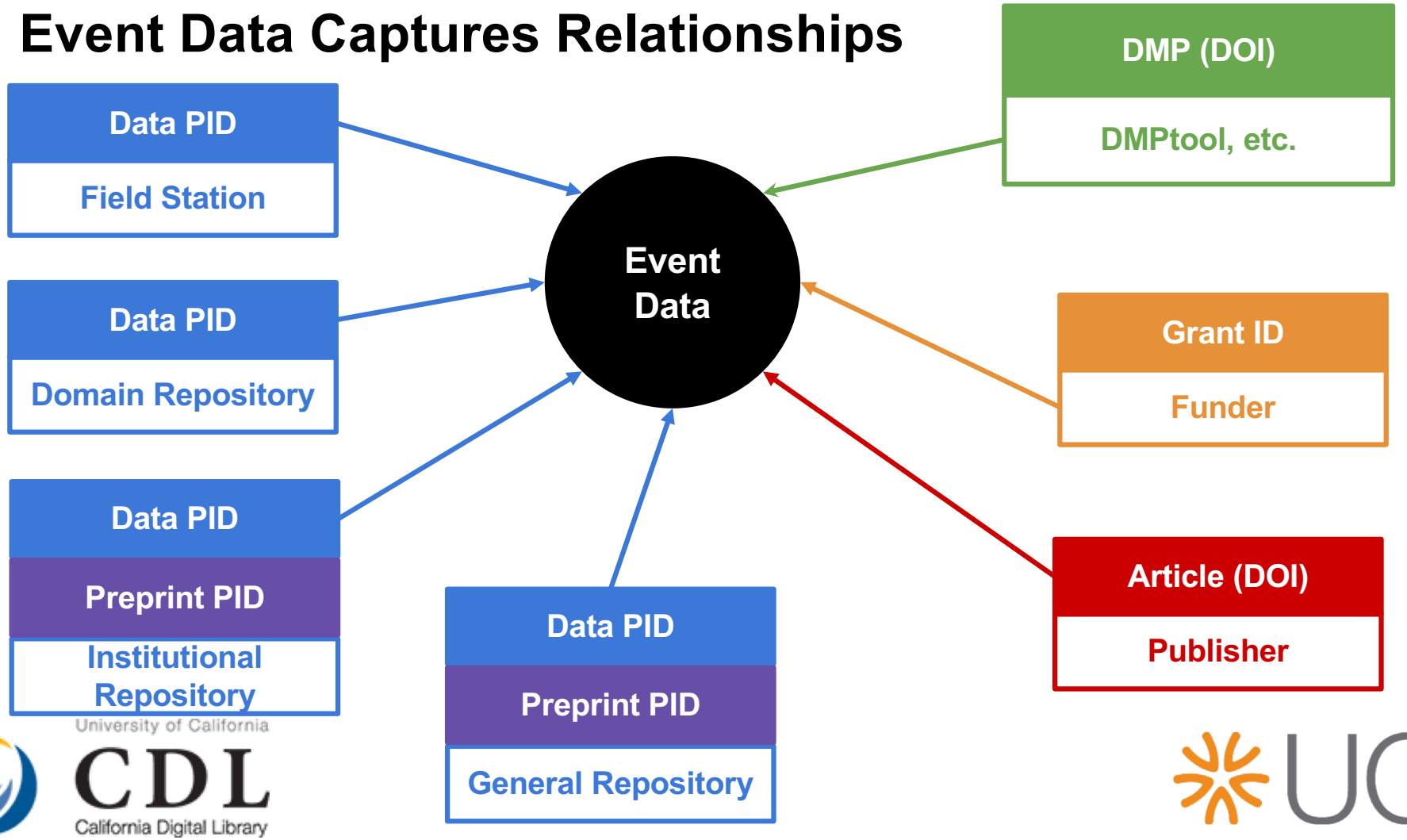
University of California
CDL
California Digital Library

UC3

Event Data

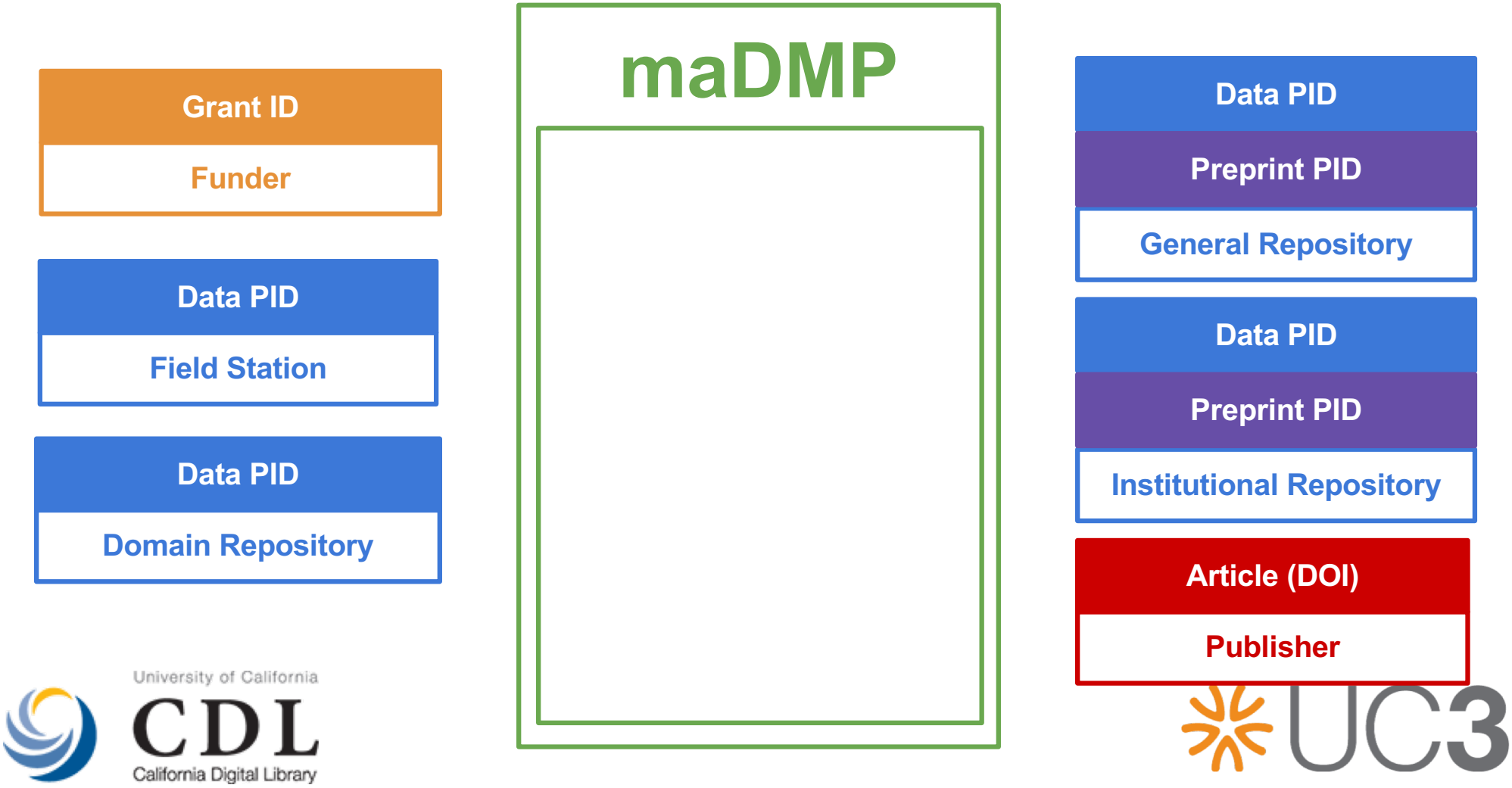
With access to DataCite's Event Data service, we can tap into the potential of the PID Graph. **Event Data is an open scholarly infrastructure service run jointly by:**

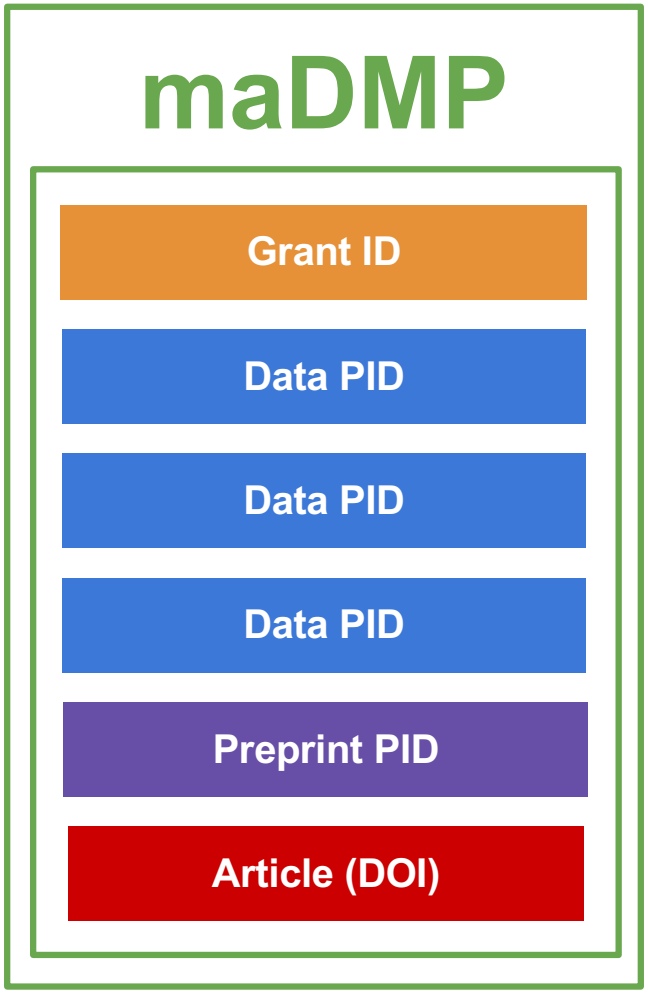
Event Data Captures Relationships



University of California
CDL
California Digital Library

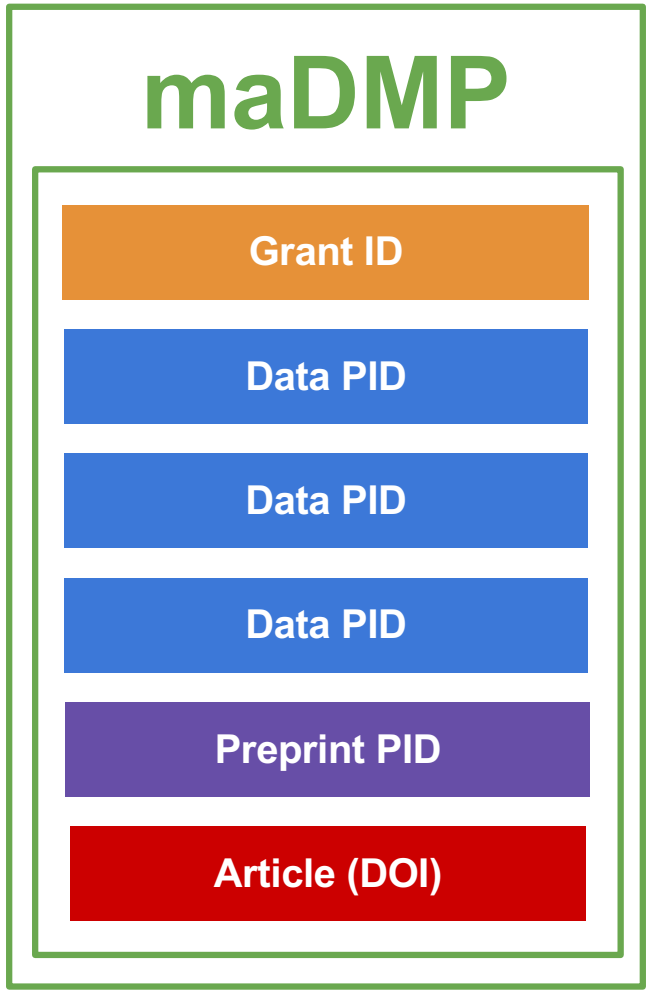
UC3


Grant ID
Funder
Data PID
Field Station
Data PID
Domain Repository
maDMP
Data PID
Preprint PID
General Repository
Data PID
Preprint PID
Institutional Repository
Article (DOI)
Publisher
University of California
CDL
California Digital Library
UC3



University of California
CDL
California Digital Library

UC3



Represents RDA maDMP Common Standards **in aggregate**

University of California CDL California Digital Library

UC3

Modular approach to Common Standards

In RDA model, the maDMP can be the container

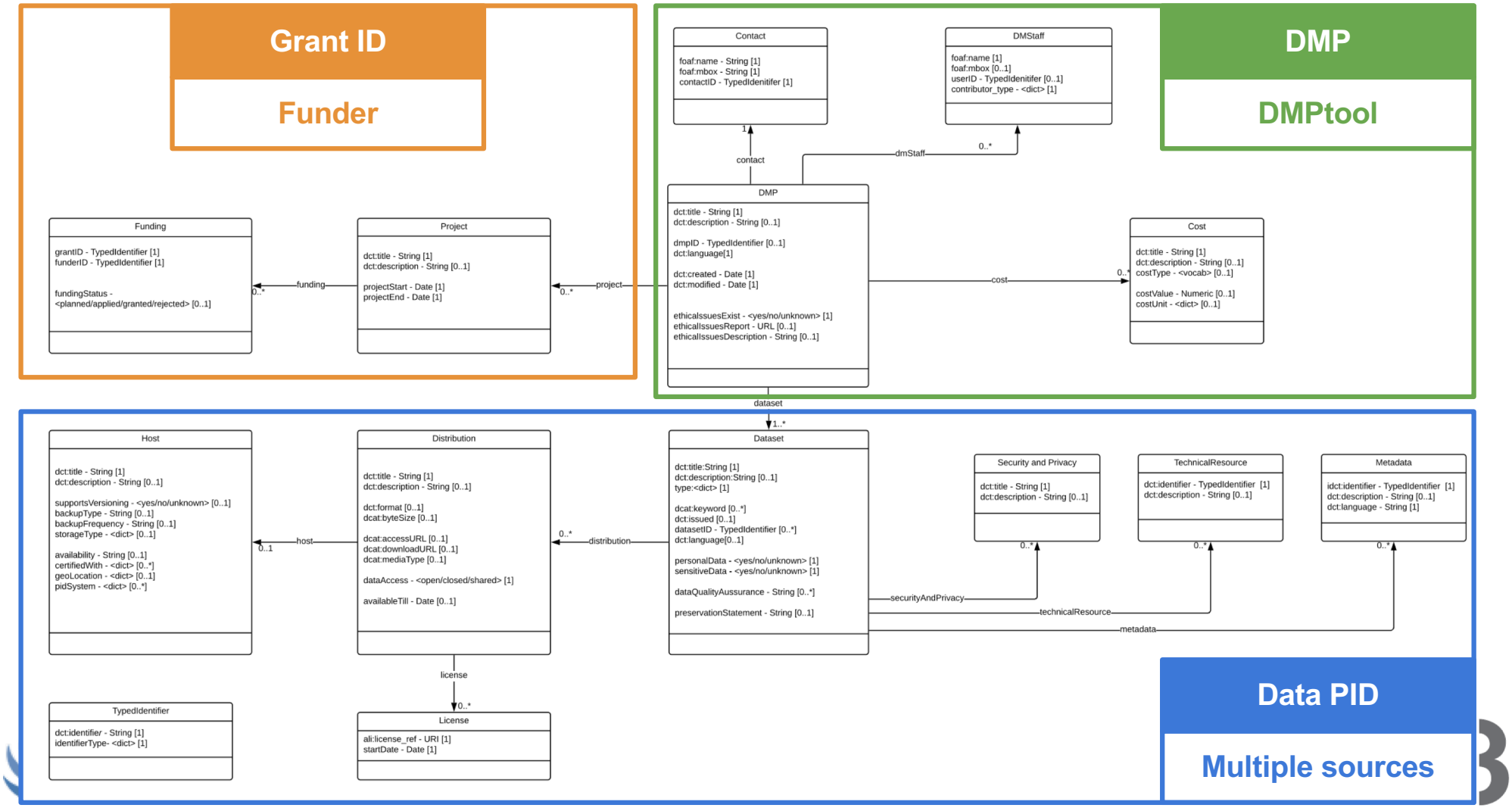
RDA Common Standards metadata describes the maDMP itself

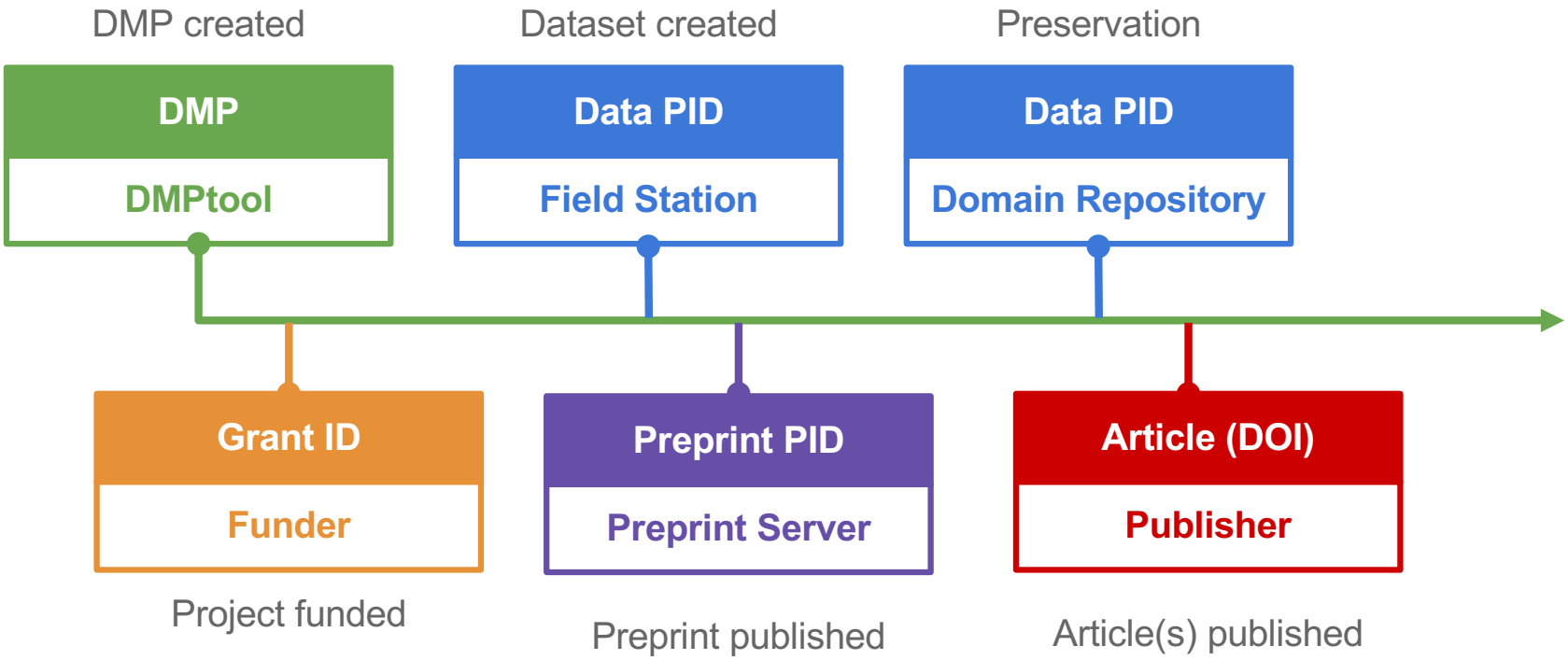
RDA Common Standards metadata describe entities connected to the maDMP

RDA Common Standards metadata describe the data volume, destination, versioning

University of California
CDL
California Digital Library

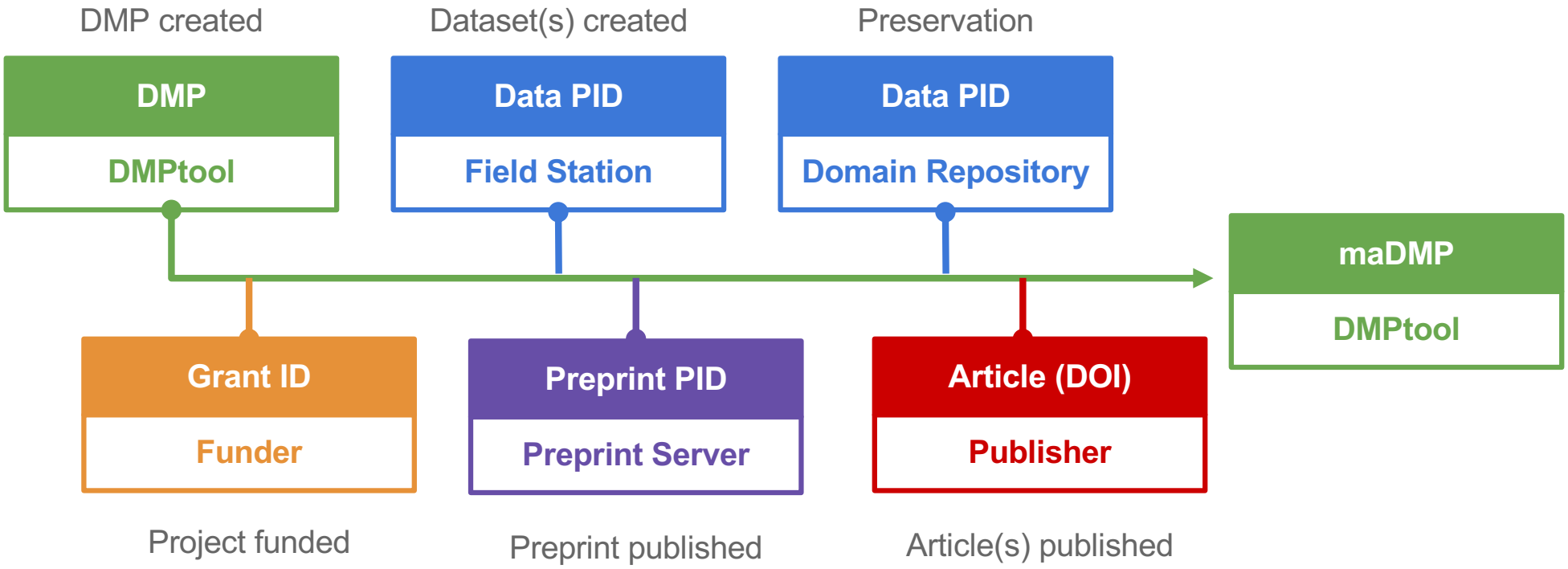
UC3


Grant ID
Funder
DMP
DMPtool
Data PID
Multiple sources
Contact
foaf:name - String [1]
foaf:mbox - String [1]
contactID - TypedIdenitifer [1]
DMStaff
foaf:name [1]
foaf:mbox [0..1]
userID - TypedIdenitifer [0..1]
contributor_type - <dict> [1]
contact
dmStaff
DMP
dct:title - String [1]
dct:description - String [0..1]
dmpID - TypedIdentifier [0..1]
dct:language[1]
dct:created - Date [1]
dct:modified - Date [1]
ethicalIssuesExist - <yes/no/unknown> [1]
ethicalIssuesReport - URL [0..1]
ethicalIssuesDescription - String [0..1]
cost
Cost
dct:title - String [1]
dct:description - String [0..1]
costType - <vocab> [0..1]
costValue - Numeric [0..1]
costUnit - <dict> [0..1]
Funding
grantID - TypedIdentifier [1]
funderID - TypedIdentifier [1]
fundingStatus - <planned/applied/granted/rejected> [0..1]
funding
Project
dct:title - String [1]
dct:description - String [0..1]
projectStart - Date [1]
projectEnd - Date [1]
project
dataset
Host
dct:title - String [1]
dct:description - String [0..1]
supportsVersioning - <yes/no/unknown> [0..1]
backupType - String [0..1]
backupFrequency - String [0..1]
storageType - <dict> [0..1]
availability - String [0..1]
certifiedWith - <dict> [0..*]
geoLocation - <dict> [0..1]
pidSystem - <dict> [0..*]
host
Distribution
dct:title - String [1]
dct:description - String [0..1]
dct:format [0..1]
dcat:byteSize [0..1]
dcat:accessURL [0..1]
dcat:downloadURL [0..1]
dcat:mediaType [0..1]
dataAccess - <open/closed/shared> [1]
availableTill - Date [0..1]
distribution
Dataset
dct:title:String [1]
dct:description:String [0..1]
type:<dict> [1]
dcat:keyword [0..*]
dct:issued [0..1]
datasetID - TypedIdentifier [0..*]
dct:language[0..1]
personalData - <yes/no/unknown> [1]
sensitiveData - <yes/no/unknown> [1]
dataQualityAussurance - String [0..*]
preservationStatement - String [0..1]
Security and Privacy
dct:title - String [1]
dct:description - String [0..1]
securityAndPrivacy
TechnicalResource
dct:identifier - TypedIdentifier [1]
dct:description - String [0..1]
technicalResource
Metadata
idct:identifier - TypedIdentifier [1]
dct:description - String [0..1]
dct:language - String [1]
metadata
license
TypedIdentifier
dct:identifier - String [1]
identifierType- <dict> [1]
License
ali:license_ref - URI [1]
startDate - Date [1]



University of California
CDL
California Digital Library

UC3



University of California
CDL
California Digital Library

UC3

Interconnected relationships

University of California
CDL
California Digital Library

UC3