bioCADDIE Data Citation Implementation Pilot (DCIP) Status Report

Tim Clark, PhD

Harvard Medical School & Massachusetts General Hospital

bioCADDIE All Hands Meeting - Denver CO

September 11, 2016

DCIP Goals

- **Facilitate data citation in biomedical research as the standard practice, with common information models**.

- Coordinate efforts amongst publishers, repositories, identifier services, bioCADDIE & NIH.

- Integrate with & significantly support bioCADDIE prototype development.

- **bioCADDIE will be a major consumer of cited data**.

What is DCIP Based On?

- National Academies, CODATA & NIH recommendations.
- Joint Declaration of Data Citation Principles (JDDCP).
- Starr et al. 2015 “Achieving Human and Machine Accessibility of Cited Data”.
- Existing & emerging standards e.g. JATS, schema.org, DATS.
- Community participation by publishers, repositories, identifier and metadata services, standards groups.

DCIP Approach

- Coordinate early adopter best practices.
- Help establish standard benchmark implementations.
- Report on lessons learned to the community.
- Focus on primary biomedical research data.
- Make cited data discoverable and consumable.

DC[1]
Data Citation Principles

DCIP Major Expected Outputs

- Publishers: Develop a Publisher's Roadmap.
- Repositories: Standardize landing page metadata for data citation as a subset of DATS.
- Identifiers: Harmonize major ID prefix resolvers.
- FAQs: Guidance for common implementations.

DC1
Data Citation Principles

Publishers Roadmap Development

GROUP LEADER

Helena Cousijn

Elsevier

- Leads: Amye Kenall & Helena Cousijn
- Elsevier, SpringerNature, eLife, PLoS, etc.

GROUP LEADER

Amye Kenall

SpringerNature

- Workshop July 22 @ SpringerNature London campus, partially funded by NPG.
- Continuing work via Telcons.

Publisher's Roadmap Approach & Status

- Based on real experiences of publishers in implementing data citation.

- Organized based on "life of a publication" starting with Instructions to Authors continuing through final release of peer-reviewed publication.

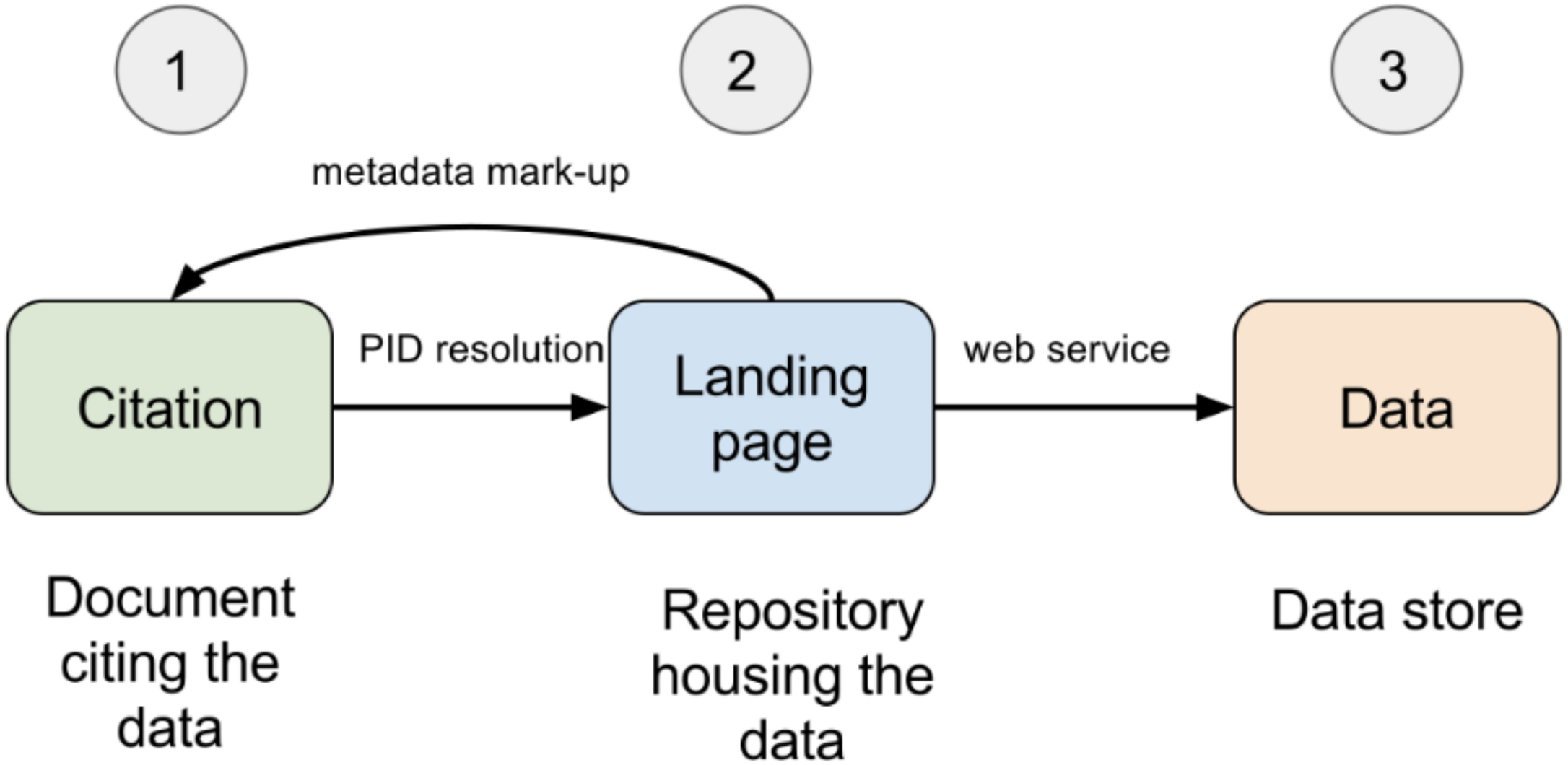
- Examples from real publishing situations with recommended approaches.

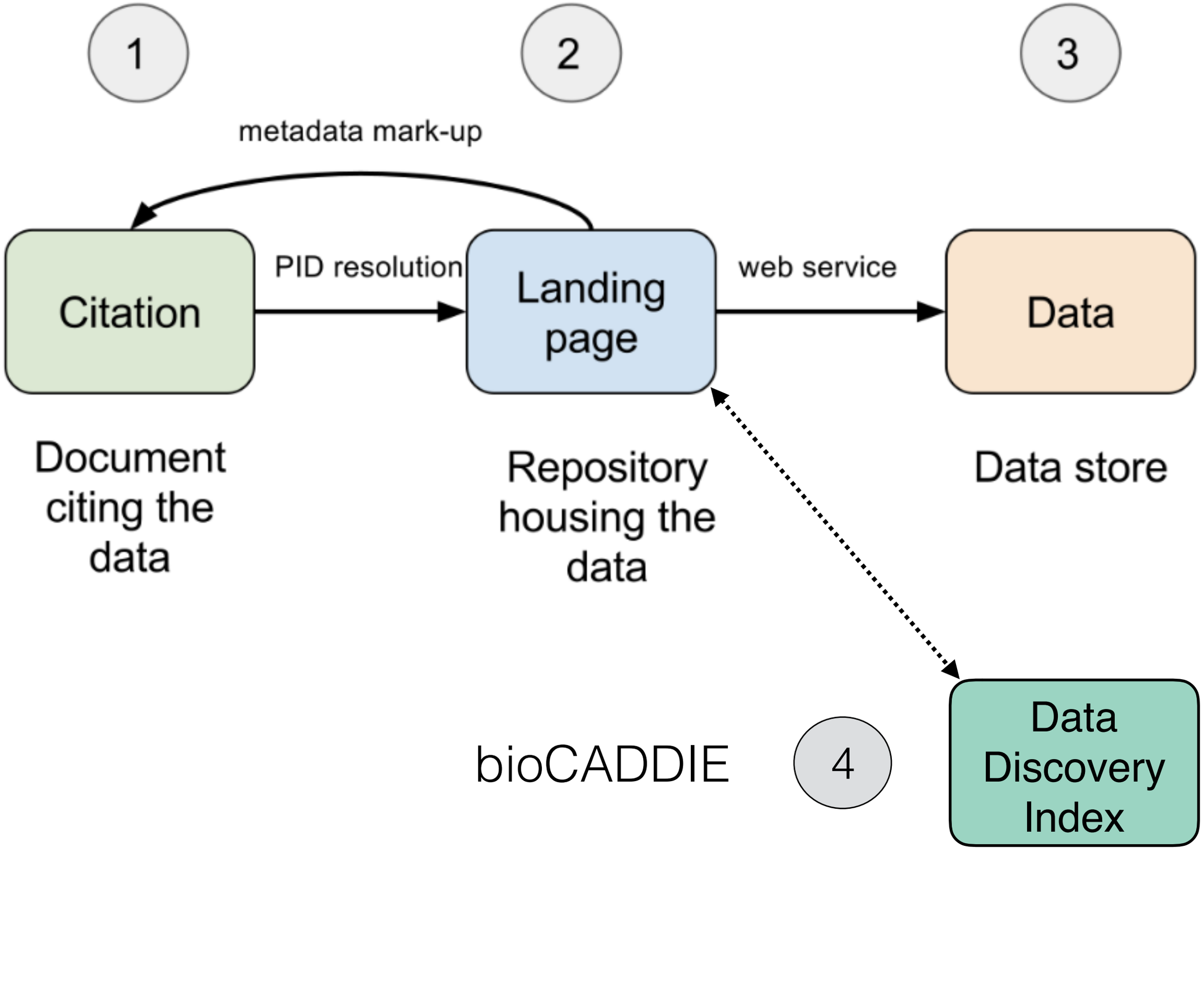
- Expected ready for external comment Oct / Nov 2016.

Repository Metadata Expert Group

Leads: Martin Fenner (DataCite), Merce Crosas (Dataverse)

Repository Metadata Expert Group


1
2
3
metadata mark-up
Citation
PID resolution
Landing page
web service
Data
Document citing the data
Repository housing the data
Data store


1
2
3
metadata mark-up
Citation
PID resolution
Landing page
web service
Data
Document citing the data
Repository housing the data
Data store
bioCADDIE
4
Data Discovery Index

Landing Page Metadata

Data Citation Metadata Element	**Dublin Core**	**Schema.org**	**DataCite**	**DATS**
Dataset Identifier	identifier	• @id • Resource • itemid*	identifier	identifier
Title	title	name	title	title
Creator	creator	author	creator	creator
Data repository or archive	publisher	publisher	publisher	publisher
Publication Date	date	datePublished	publicationYear	date
Version	<not defined>	version	version	version
Type	type	type	resourceTypeGeneral	type

* name of ID field depends on schema.org serialization format:**@id** in JSON-LD, **resource** in RDFa, and **itemid** in microdata;
* JSON-LD the preferred serialization for schema.org elements.

Landing Page Data Citation Metadata s.b. Human and Machine Readable

Cite this Dataset

Bilokapic, S; Schwartz, TU. 2015. "X-Ray Diffraction data for: Nup37-Nup120 full-length complex from Schizosaccharomyces pombe. PDB Code 4FHN", Structural Biology Data Grid, V1, http://dx.doi.org/10.15785/SBGRID/179.

Download Citation

```
$ curl https://data.sbgrid.org/api/bibtex?dataset=179

@data{ sbdg179,
 author = { Bilokapic, S and  Schwartz, TU   },
 publisher = { Structural Biology Data Grid },
 title = { X-Ray Diffraction data for: Nup37-Nup120 full-length complex from Schizosaccharomyces pombe. PDB Code 4FHN },
 year = { 2015 },
 doi = { 10.15785/SBGRID/179 },
 url = { http://dx.doi.org/10.15785/SBGRID/179 }
}
```

Repository Metadata Status

- Required and supplemental metadata defined with alternative vocabularies and serializations specified.
- Backward and forward compatibility modes defined.
- Integration w/ ref. managers (EndNote, Zotero, CSL).
- Document expected ready for review: Oct 2016
- Moving forward: outreach to repositories.

Identifier Harmonization Expert Group

DCIP Identifiers Workshop, June 2, 2016, Harvard University, Cambridge MA

John Kunze (CDL), Niall Beard (Manchester), Tim Clark (Harvard),Nick Juty (EBI), Ian Fore (NIH), Julie McMurry (UCSB), Jeff Grethe (UCSD), Rafa Jimenez (ELIXIR), Sarala Wimalaratne (EBI)

Data Citation Generic Example

example of a data citation as it would appear *in a reference list**

Principle 2: Credit and Attribution (e.g. authors, repositories or other distributors and contributors)

Principle 4: Unique Identifier (e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:** A persistent link to a landing page with metadata and access information

Author(s), Year, Dataset Title, Data Repository or Archive, [Accession], Global Persistent Identifier, version or subset

Principle 7: Version and granularity (e.g. a version number or a query to a subset) In addition, access to versions or subsets should be available from the landing page,

*Note that the format is not intended to be defined with this example, as formats will vary across publishers and communities **[Principle 8: Interoperability and flexibility]**.

Force11 Data Citation Working Group	N. Juty
Internet-Draft	European Bioinformatics Institute
Intended status: Informational	J. Kunze
Expires: December 9, 2016	California Digital Library
	June 7, 2016

Prefix-Based Collection Access

draft-kunze-prefixes-00

Abstract

This document specifies a YAML [YAML] file that serves as an open registry of unique collection prefixes. These prefixes can be used by meta-resolvers to redirect identifiers to appropriate collection resolvers.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 9, 2016.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD

Identifier Harmonization Status

- Technical approach for common prefix registry has been agreed and preliminary document (RFC) drafted.

- Current tasks:
  - Complete resolver rules definition
  - Explore further resolver system standardization
  - Light weight software engineering tasks.
- Document expected to be ready for outside review: Oct 2016

FAQ / Primer Group

GROUP LEADER

Maryann Martone

UCSD

GROUP LEADER

Joan Starr

California Digital Library

- Communicates DCIP outcomes.
- Major Deliverables:
  - FAQs for Repositories & Publishers
  - Data Citation Primer
- Status:
  - Repository FAQ done
  - Publishers FAQ v0.1 ready for comment.

Participants

And you!

DCIP

- Major publishers and repositories participating in developing common data citation technologies.

- DCIP deliverables now in late draft stage.

- DCIP is helping to enable the ecosystem around bioCADDIE for long-term success.