The ARK Identifier Scheme at Ten Years Old

7 May 2012

John Kunze
University of California Curation Center
California Digital Library

California Digital Library

Serving the University of California

- 10 campuses
- 360K students, faculty, and staff
- 100's of museums, art galleries, observatories, marine centers, botanical gardens
- 5 medical centers
- 5 law schools
- 3 National Laboratories

CDL supports the research lifecycle

- Collections
- Digital Special Collections
- Discovery & Delivery
- Publishing Group
- UC Curation Center (UC3)

UC3

University of California
CDL
California Digital Library

California Digital Library (CDL)

Today's journey

- What are ARKs?
- Separation of concerns
  - Naming ≠ hosting
  - Scheme ≠ resolution
  - Syntax ≠ persistence
- Inflections and metadata
- EZID (easy identifiers) and N2T (name-to-thing)
- Data citation, passthrough

UC3

University of California
CDL
California Digital Library

What's an ARK identifier?

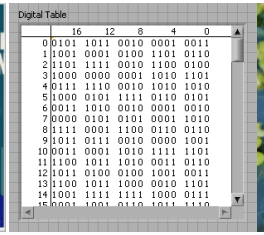
ARK = Archival Resource Key

ARKs support long-term access to information objects

ARKs identify objects of any type:

- digital objects – data, documents, images, software, ...
- physical objects – books, bones, statues, ...
- groups & living beings – people, animals, orchestras, ...
- Intangibles – places, chemicals, diseases, terms, ...





The URL is dead, long live the URL!

Fallacy #1: URLs are unreliable, so instead use this... um... well... ah ... (shhh!) "URL"

Some of your best friends are URLs:

http://dx.doi.org/10.1234/98765

http://hdl.handle.net/10.1234/98765

http://purl.org/10.1234/98765

http://n2t.net/ark:/101234/98765

UC3

University of California CDL California Digital Library

Persistence is about service

- Imagine the “perfect” golden identifier
- Apply bankruptcy, disk crash, human error, or war, and there’s nothing that syntax, scheme, or resolver can do to prevent identifier breakage.

UC3

University of California
CDL
California Digital Library

What's an ARK identifier? (take 2)

An ARK *is a URL*, with some extra rules

ARK reserves / and . for what we often assume

- A/B/C means C is *contained* in A/B, and B in A
- A.pdf, A.html, and A.docx are all *variants* of A

Could drastically improve search result display

- No need to lookup relationships

UC3

University of California
CDL
California Digital Library

ARK inflections (declinations)

An ARK is a special URL with access to 3 things

1. An information object
2. Its metadata, by appending '?' inflection
3. A provider's promise, by appending a '??'

An *inflection* changes a name ending for a purpose

- Reduces the number of different names needed
- Use semantic web without hiring a programmer

UC3

University of California
CDL
California Digital Library

‘?’ Inflection returns Dublin Kernel

Same machine-readable information as before:

```
erc:
who:   National Research Council
what:  The Digital Dilemma
when:  2000
where: http://books.nap.edu/html/digital%5Fdilemma
```

Even shorter:

```
erc: National Research Council
     | The Digital Dilemma | 2000
     | http://books.nap.edu/html/digital%5Fdilemma
```

See http://dublincore.org/groups/kernel/ for more information

UC3

University of California CDL California Digital Library

Why use ARKs?

ARKs are assigned for a variety of reasons:

- affordability – there are no fees to assign or use ARKs
- self-sufficiency – can host ARKs on your own web server
- portability – can move ARKs without change of identity

```
http://cdlib.org/ark:/12025/654xz321
http://rutgers.edu/ark:/12025/654xz321
http://n2t.net/ark:/12025/654xz321
```

- global resolvability – can host ARKs at N2T resolver
- density – mixed case means CD, Cd, cD, cd are all distinct

UC3

University of California CDL California Digital Library

Some unique advantages of ARKs

- simplicity – uses only ordinary "redirects" & "get" requests
- versatility – with "inflections" (different endings), an ARK should access data, metadata, promises, and more
- transparency – no identifier can guarantee stability, and ARK inflections help users make informed judgments
- visibility – syntax rules make ARKs easy to extract and to compare for containment and variant relationships
- reserved characters: - (hyphen), / (slash), . (period)

UC3

University of California CDL California Digital Library

What's an ARK identifier? (take 3)

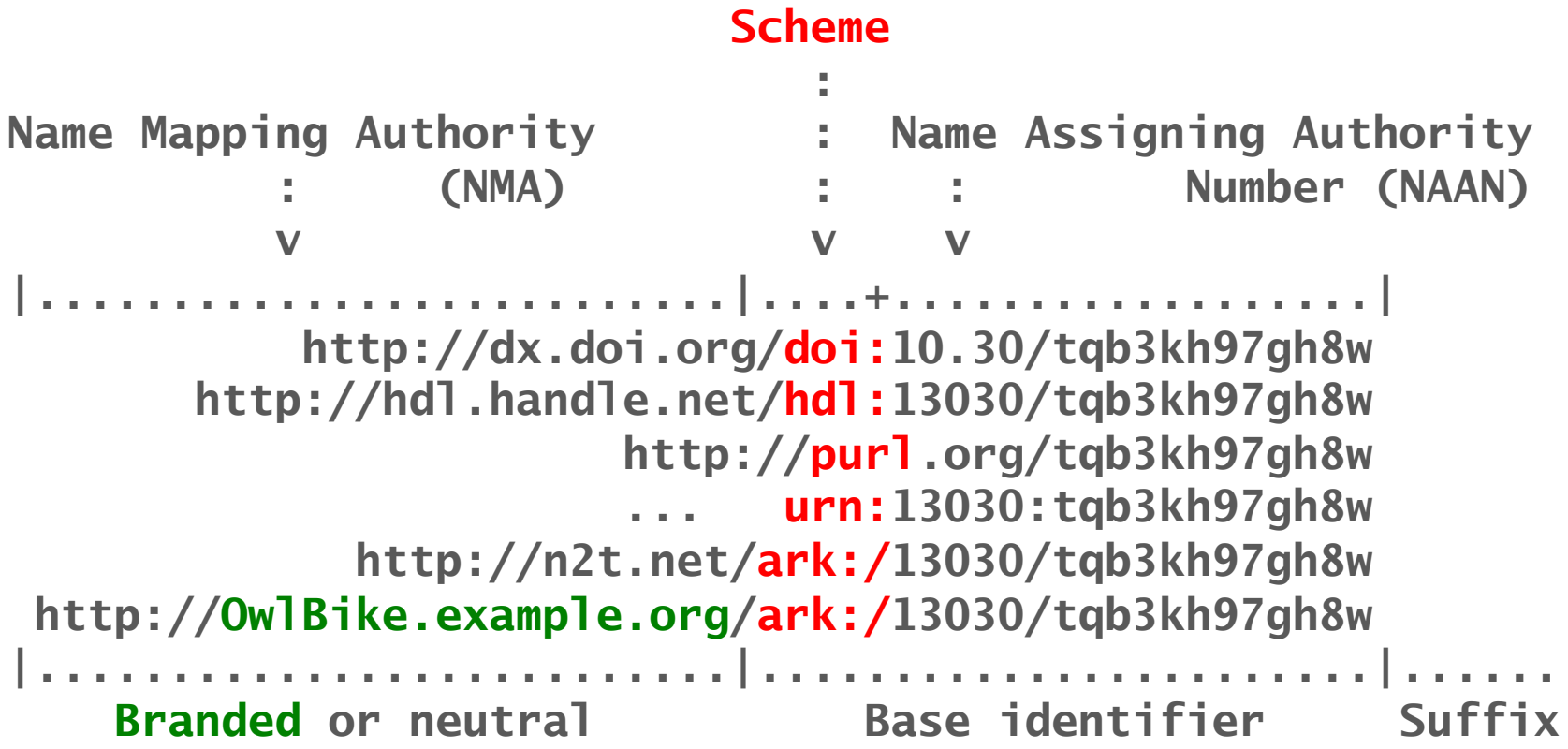
ARK is a collection of good ideas

- Separates scheme syntax from resolver rules
  - *Resolution* is a process of mapping an id to a thing
- Separates name assigning from name mapping
- All schemes encouraged to use these ideas, even ordinary URLs
- N2T resolver can support them for any scheme

UC3

University of California CDL California Digital Library

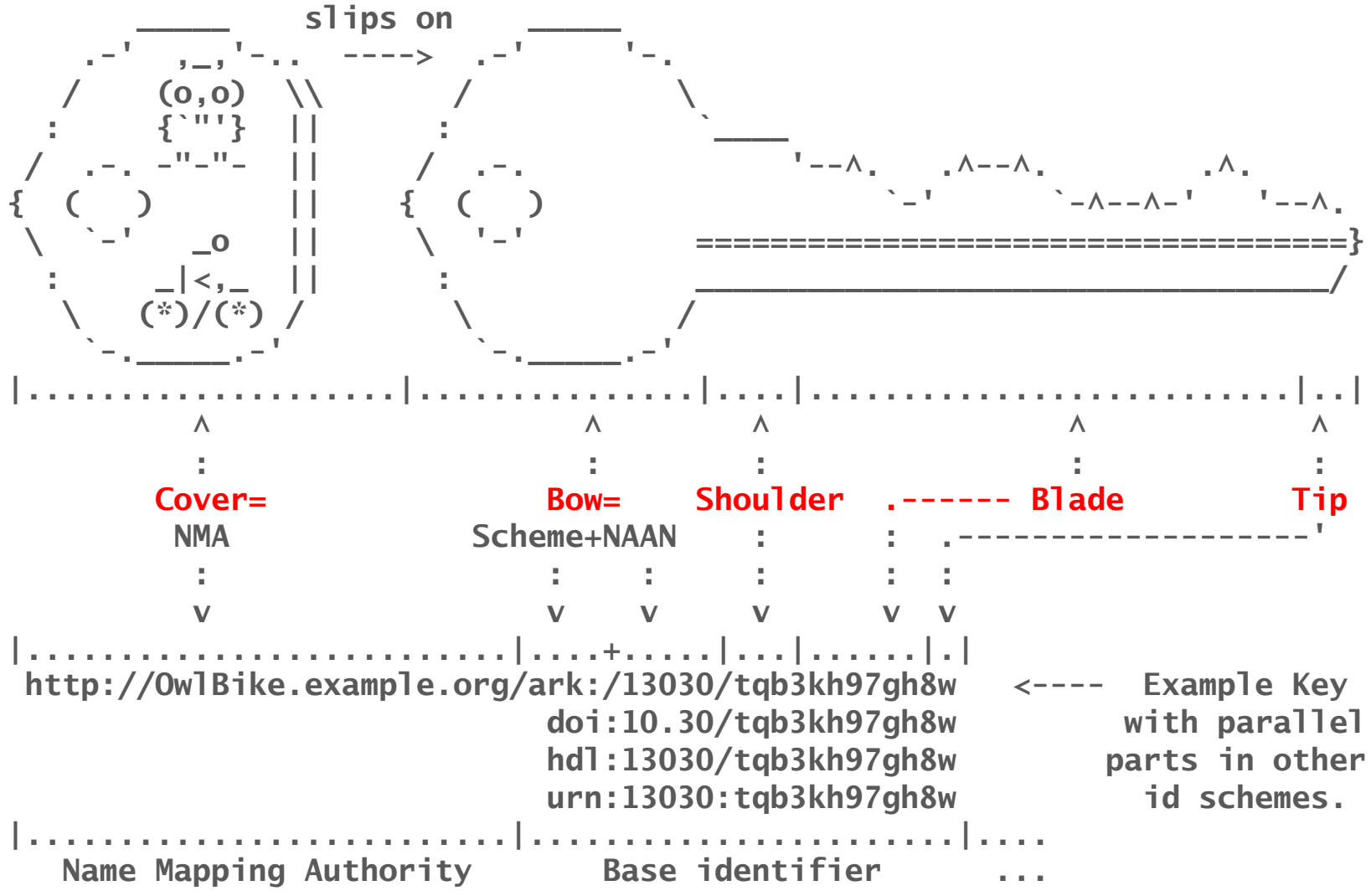
Identifier schemes are highly parallel



UC3

University of California CDL California Digital Library

Locksmith jargon: shoulder, blade, tip, bow, cover



UC3

University of California
CDL
California Digital Library

ARK usage in 10 years

- In 2001-2011 ~100 organizations registered for ARKs
- Registry is replicated at BnF and NLM
- Some of the largest users are
  - The California Digital Library
  - The Internet Archive
  - Bibliothèque nationale de France
  - Portico Digital Preservation Service
  - University of California Berkeley
  - University of Chicago

UC3

University of California CDL California Digital Library

Some other ARK registrants

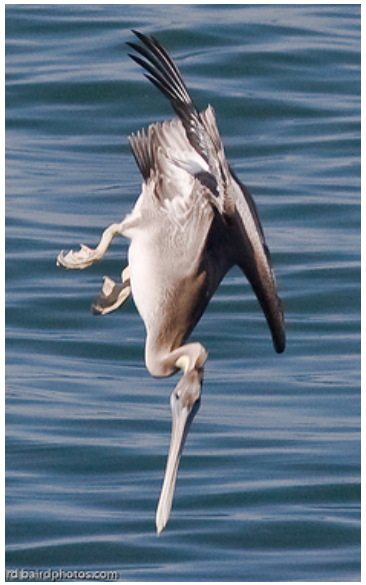
12025	US National Library of Medicine
86077	Cornell Institute for Social and Economic Research
26677	Library and Archives Canada
77635	Humboldt-Universität zu Berlin
13038	World Intellectual Property Organization
78319	Google
61001	University of Chicago
28722	University of California Berkeley
64269	UK Digital Curation Centre
87895	Centre Informatique National de l'Enseignement Supérieur
61903	Family Search
52327	National Library and Archives of Quebec
10261	Jüdisches Museum Berlin
71479	Spanish National Research Council
32833	Massachusetts Institute of Technology
81055	British Library
80713	Biblioteca Nacional de Portugal

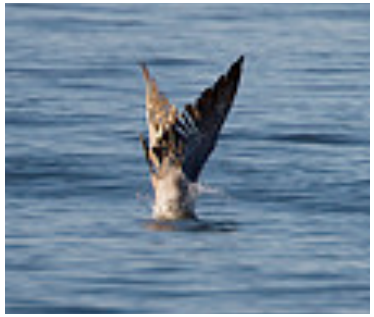
UC3

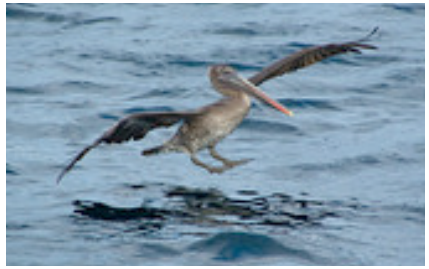
University of California
CDL
California Digital Library

Immersion vs landing page

What do you mean by “get the data”?
What inflections might distinguish these?







- *Immersion* – a consumptive experience or

- *Landing page* – a menu-study experience?

UC3

University of California
CDL
California Digital Library

The UC Curation Center at CDL

Logged in as dryad | Logout | Contact UC3

Recent identifiers

EZID

long-term identifiers made easy

Home | **Create** | **Manage** | **Help**

Create a long-term identifier. The identifier will be the concatenation of the prefix and remainder; leave the latter blank to have an identifier generated for you.

Namespace	Identifier		
	Prefix	Remainder	
(•) Dryad	doi:10.5061/		Create
( ) CDL ARK	ark:/13030/c7		

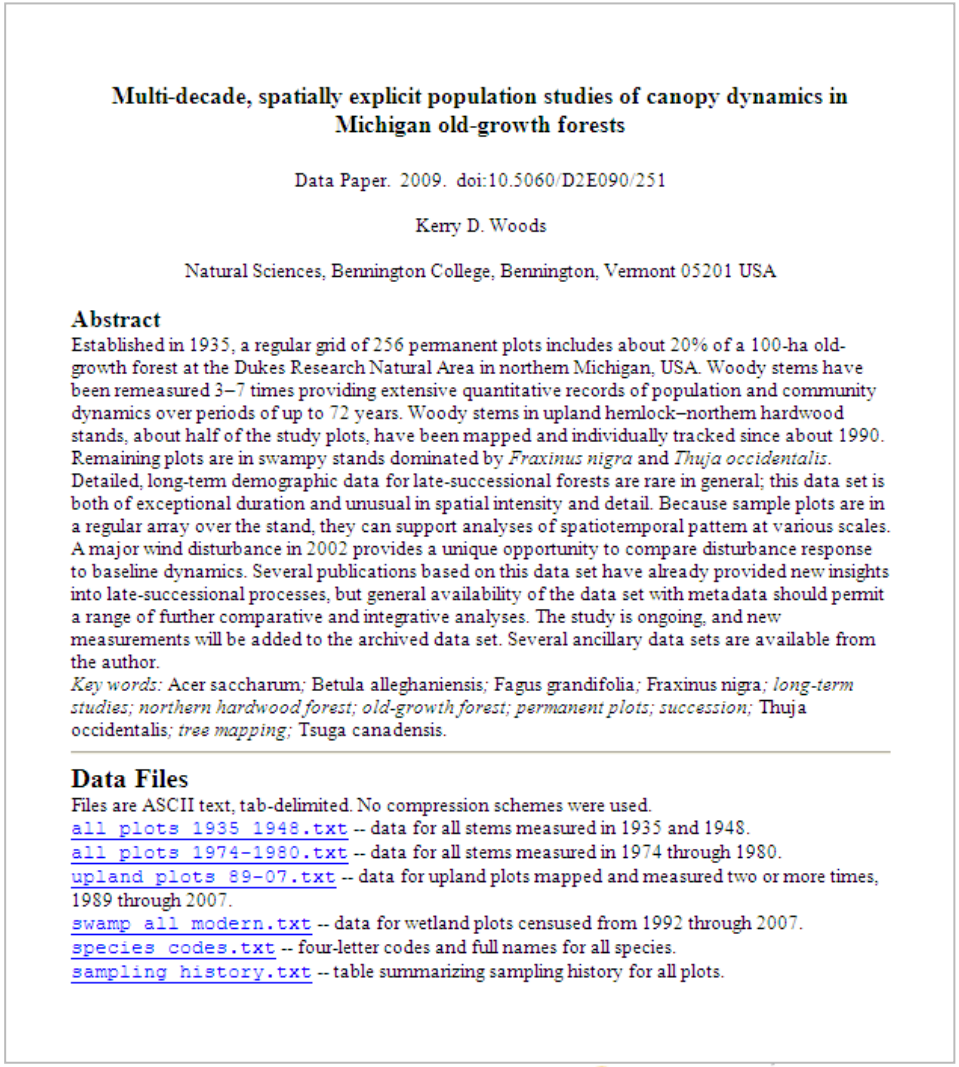
Once created, an identifier cannot be renamed or deleted.
Consult help for considerations in naming identifiers and to create test identifiers.

UC3

University of California CDL California Digital Library

Vision for a “data paper”

- *Wrap the unfamiliar in a familiar façade*
- A “data paper” is minimally a cover sheet and a set of links to archived artifacts
- Cover sheet contains familiar elements: title, date, authors, abstract, and persistent identifier (DOI, ARK, etc.)
- Just enough to permit basic exposure and discovery
  - Building a basic data citation
  - Indexing by services such as Web of Science, Google Scholar
  - Instilling confidence in the identifier’s stability



Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests

Data Paper. 2009. doi:10.5060/D2E090/251

Kenny D. Woods

Natural Sciences, Bennington College, Bennington, Vermont 05201 USA

Abstract

Established in 1935, a regular grid of 256 permanent plots includes about 20% of a 100-ha old-growth forest at the Dukes Research Natural Area in northern Michigan, USA. Woody stems have been remeasured 3–7 times providing extensive quantitative records of population and community dynamics over periods of up to 72 years. Woody stems in upland hemlock–northern hardwood stands, about half of the study plots, have been mapped and individually tracked since about 1990. Remaining plots are in swampy stands dominated by *Fraxinus nigra* and *Thuja occidentalis*. Detailed, long-term demographic data for late-successional forests are rare in general; this data set is both of exceptional duration and unusual in spatial intensity and detail. Because sample plots are in a regular array over the stand, they can support analyses of spatiotemporal pattern at various scales. A major wind disturbance in 2002 provides a unique opportunity to compare disturbance response to baseline dynamics. Several publications based on this data set have already provided new insights into late-successional processes, but general availability of the data set with metadata should permit a range of further comparative and integrative analyses. The study is ongoing, and new measurements will be added to the archived data set. Several ancillary data sets are available from the author.

Key words: Acer saccharum; Betula alleghaniensis; Fagus grandifolia; Fraxinus nigra; *long-term studies; northern hardwood forest; old-growth forest; permanent plots; succession;* Thuja occidentalis; *tree mapping;* Tsuga canadensis.

Data Files

Files are ASCII text, tab-delimited. No compression schemes were used.

`all_plots_1935_1948.txt` -- data for all stems measured in 1935 and 1948.

`all_plots_1974-1980.txt` -- data for all stems measured in 1974 through 1980.

`upland_plots_89-07.txt` -- data for upland plots mapped and measured two or more times, 1989 through 2007.

`swamp_all_modern.txt` -- data for wetland plots censused from 1992 through 2007.

`species_codes.txt` -- four-letter codes and full names for all species.

`sampling_history.txt` -- table summarizing sampling history for all plots.

GORDON AND BETTY MOORE FOUNDATION

UC3

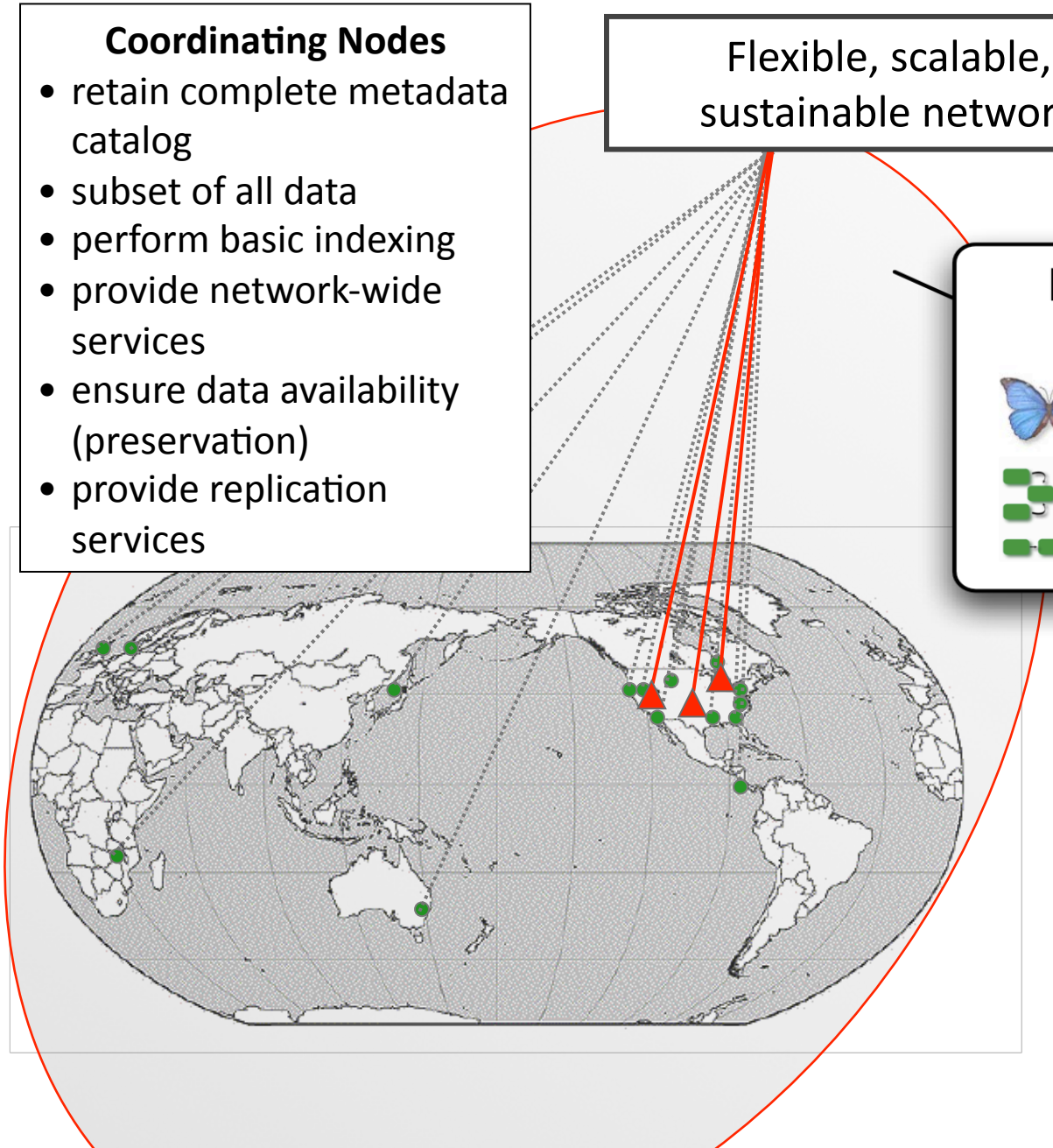
CDL California Digital Library

New distributed framework

Coordinating Nodes

- retain complete metadata catalog
- subset of all data
- perform basic indexing
- provide network-wide services
- ensure data availability (preservation)
- provide replication services

Flexible, scalable, sustainable network



Investigator 1..N Toolkit

MERCURY

Kepler

Sp

MetaCat

University of California
CDL
California Digital Library

ARKs – coming soon

- Community forum
- Standardization as an Internet RFC
- New inflections for landing page & immersion

UC3

University of California
CDL
California Digital Library

N2T/EZID – coming soon

- Indexing by A&I vendors
- Suffix pass-through
  - Register Name -> target T
  - Resolve Name/a/b/c -> T/a/b/c automatically
  - Greatly reduce number of ids to manage
- URNs

UC3

University of California
CDL
California Digital Library