DataONE: Providing Solutions to Complex Science and Data Challenges

Mark Servilla
University of New Mexico

NDSC5 at RENCI
Chapel Hill, NC. 4-6 April 2016
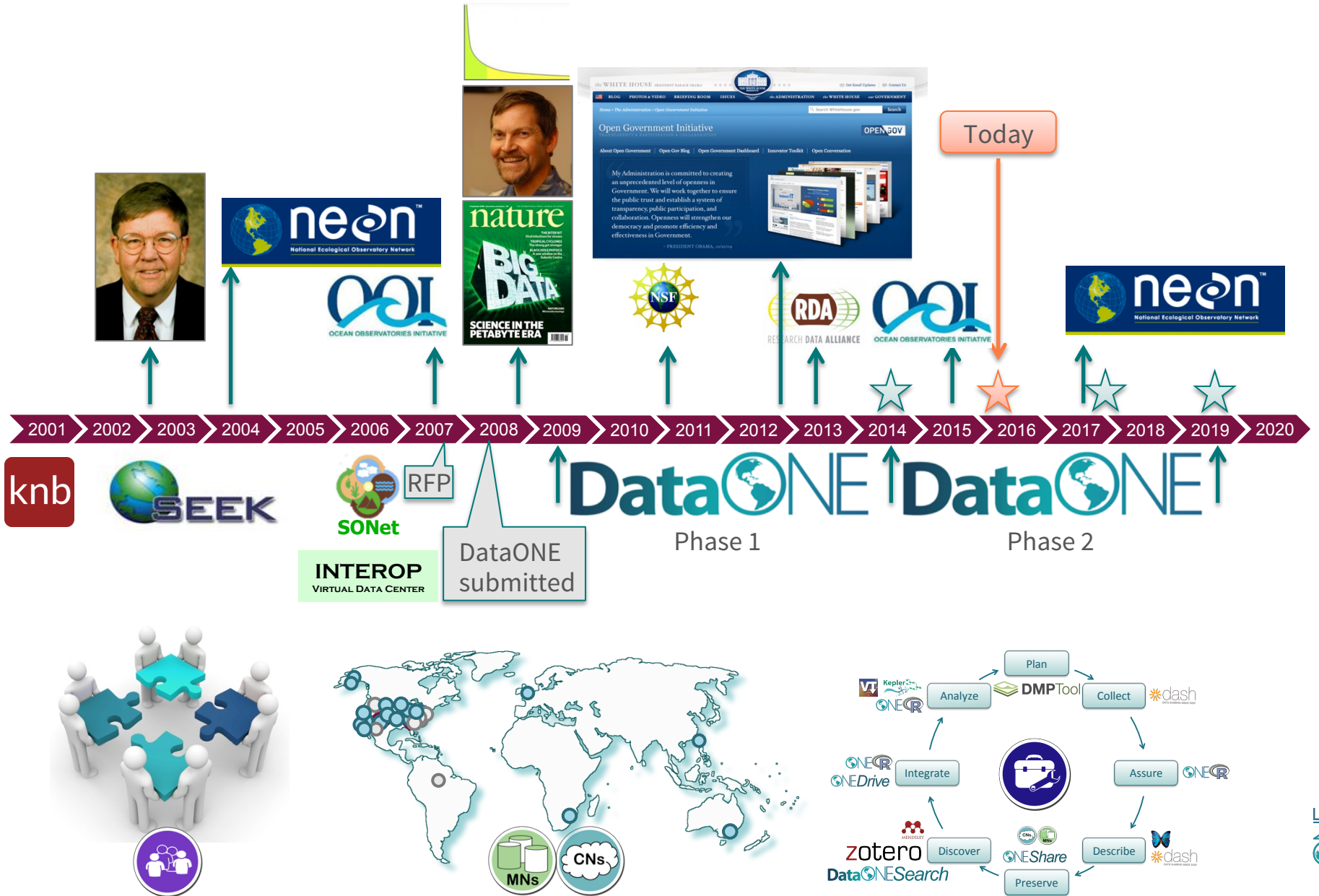
DataNet and DataONE
Objectives

DataNet: "The new types of organizations envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to:
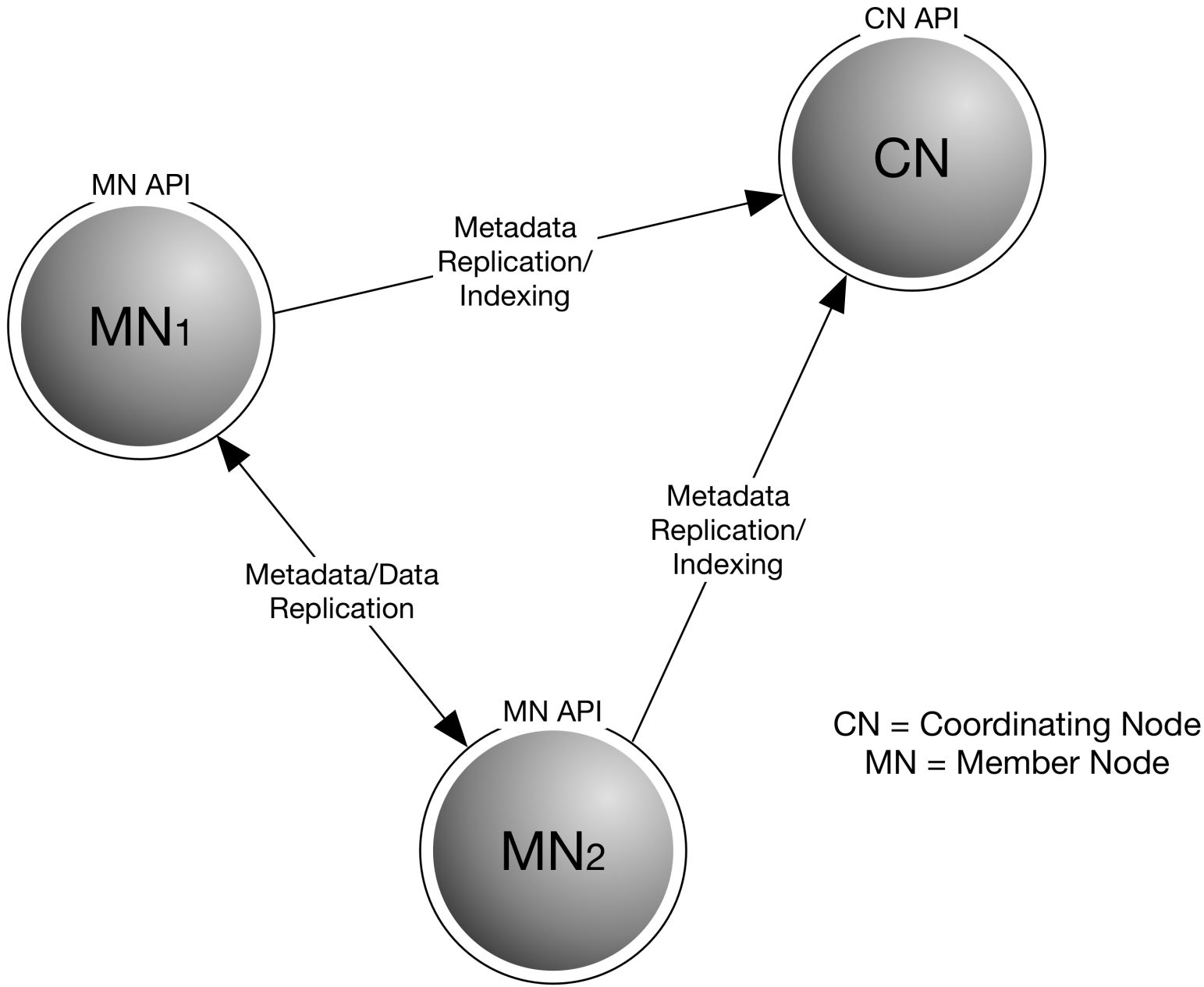
- *provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;*
- *continuously anticipate and adapt to changes in technologies and in user needs and expectations;*
- *engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and*
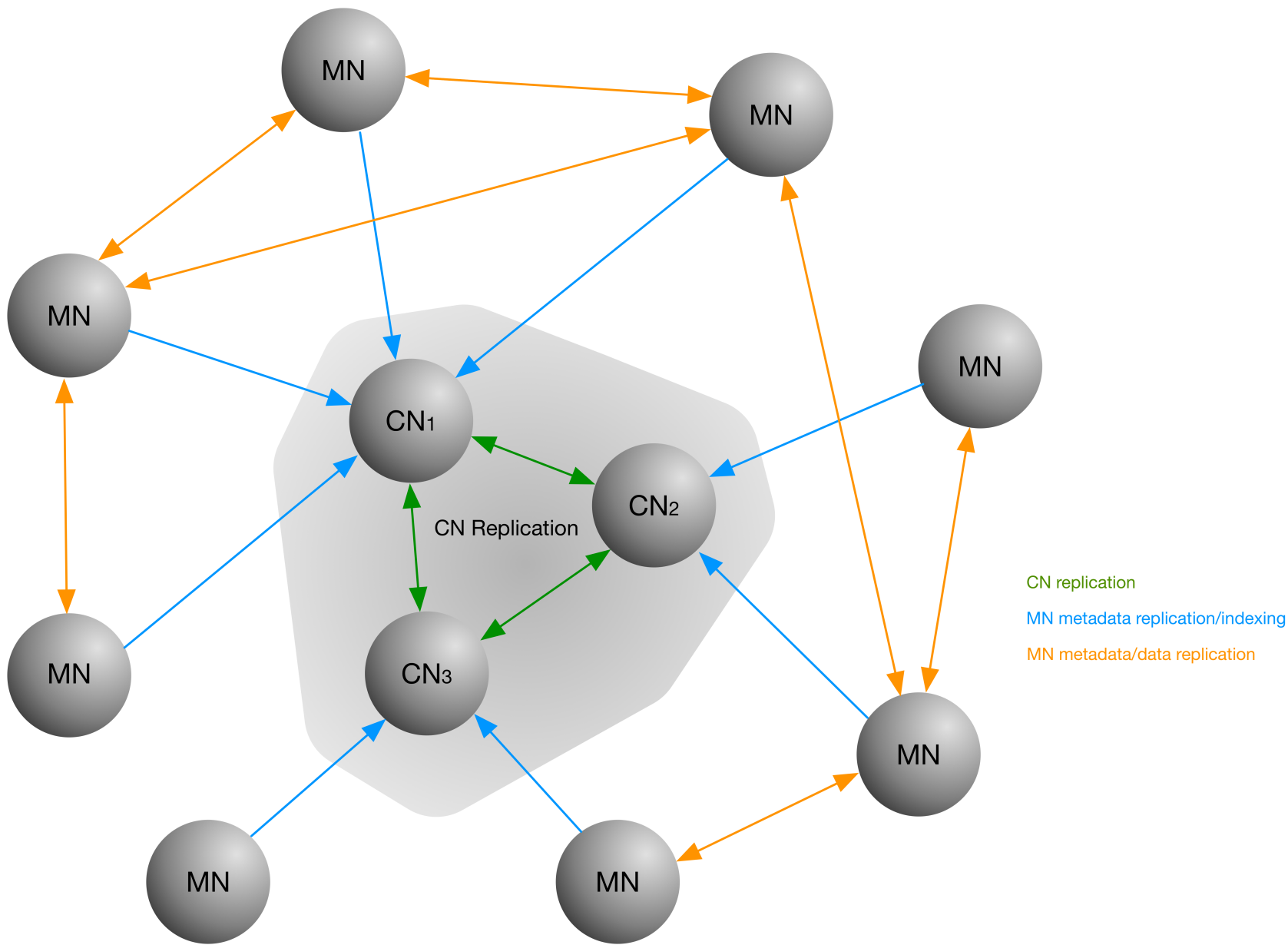- *serve as component elements of an interoperable data preservation and access network."*

DataONE works with library, archival, computer, information and domain scientists to:

- *provide a distributed framework and sustainable cyberinfrastructure that **meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data**;*
- *assess and adapt to user needs and continuously improve the usability of DataONE products and services;*
- *incorporate cutting-edge computer science R&D into DataONE CI;*
- *engage and support users through training, webinars and online resources; and*
- *develop investment-worthy elements of a long-term, sustained, interoperable data preservation and access network.*

DataONE

Timeline and Context

Today

Phase 1

Phase 2

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

knb

SEEK

SONet

RFP

INTEROP
Virtual Data Center

DataONE submitted

DataONE

DataONE

CN API

MN API

MN$_1$

Metadata
Replication/
Indexing

CN

Metadata/Data
Replication

Metadata
Replication/
Indexing

MN API

MN$_2$

CN = Coordinating Node
MN = Member Node

CN Replication

CN replication
MN metadata replication/indexing
MN metadata/data replication

Cyberinfrastructure Component 1
Coordinating Nodes

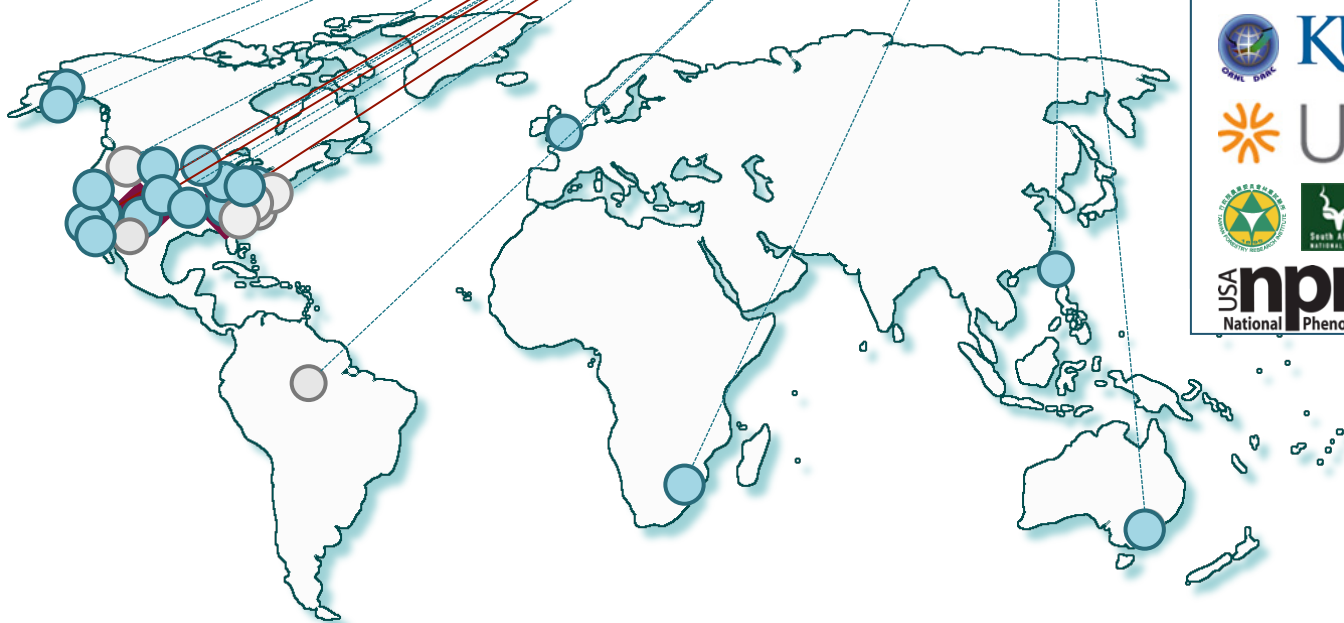Components for a flexible, scalable, sustainable network
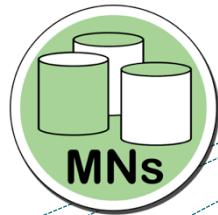
CNs

Coordinating Nodes
- retain complete metadata catalog
- indexing for search
- network-wide services
- ensure content availability (preservation)
- replication services

THE UNIVERSITY of NEW MEXICO

OAK RIDGE National Laboratory

NCEAS

UCSB

UT

www.dataone.org/coordinating-nodes

DataONE

Cyberinfrastructure Component 2
Member Nodes

Components for a flexible, scalable, sustainable network
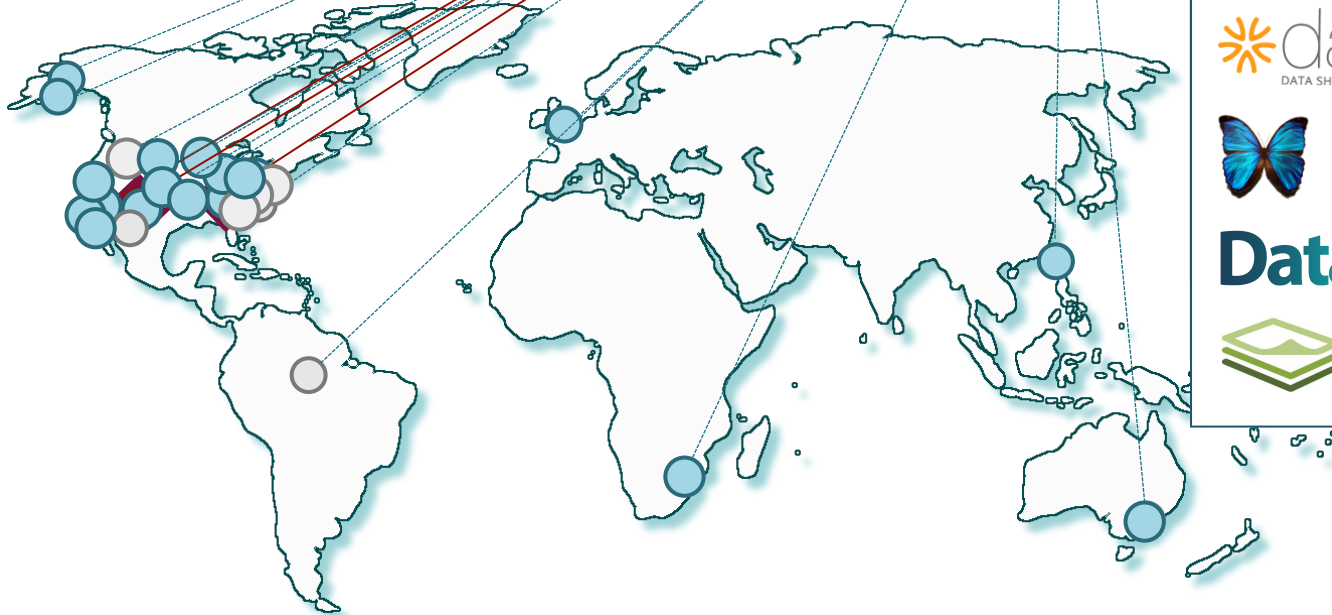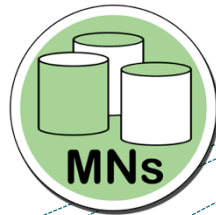

Coordinating Nodes

Member Nodes
- diverse institutions
- serve local community
- provide resources for managing their data
- retain copies of data

www.dataone.org/member-nodes

Cyberinfrastructure Component 3
Investigator Toolkit

Components for a flexible, scalable, sustainable network
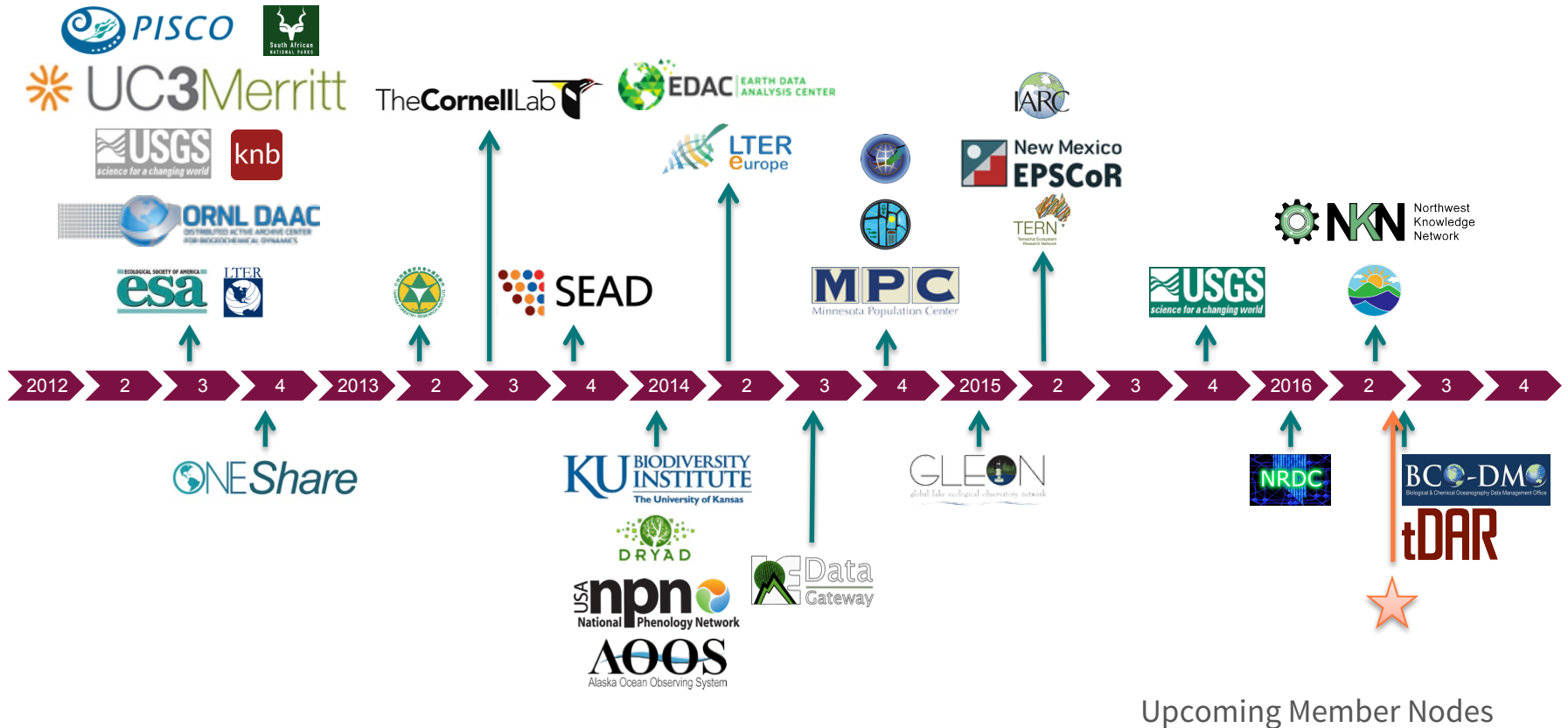

Coordinating Nodes

Member Nodes

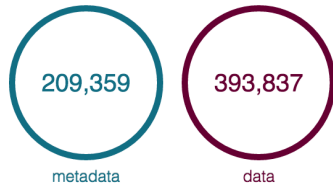**Investigator Toolkit**

>> command line interface

DataONESearch

DMPTool

www.dataone.org/investigator-toolkit

DataONE Member Nodes
Current and Upcoming

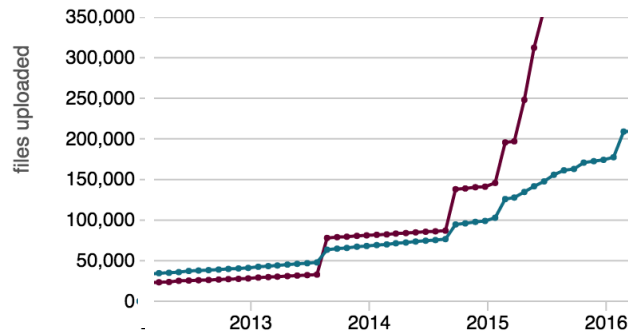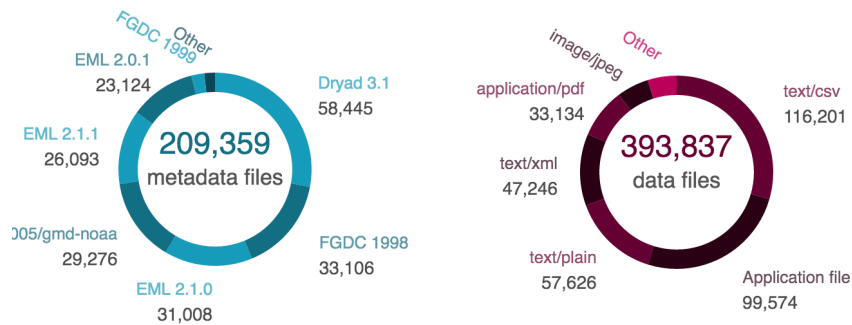Upcoming Member Nodes

9

Data Holdings

209,359 — metadata

393,837 — data

Uploads

The number of individual metadata and data files uploaded over time. Only the first version of each file is counted.

files uploaded: 350,000 / 300,000 / 250,000 / 200,000 / 150,000 / 100,000 / 50,000 / 0

2013   2014   2015   2016

File formats

We breakdown the types of metadata and data files uploaded. Only the most recent version of each file is included.

209,359 metadata files

- FGDC 1999
- Other
- EML 2.0.1 — 23,124
- EML 2.1.1 — 26,093
- 005/gmd-noaa — 29,276
- EML 2.1.0 — 31,008
- FGDC 1998 — 33,106
- Dryad 3.1 — 58,445

393,837 data files

- image/jpeg
- Other
- application/pdf — 33,134
- text/xml — 47,246
- text/plain — 57,626
- text/csv — 116,201
- Application file — 99,574

DataONE

Cyberinfrastructure

Data Services: Extraction, sub-setting etc

Provenance

Semantics-enabled Discovery