REAGAN W. MOORE
DIRECTOR
DATA INTENSIVE CYBER ENVIRONMENTS CENTER
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
RWMOORE@RENCI.ORG

PRIMARY RESEARCH OR PRACTICE AREA(S):

- POLICY-BASED DATA MANAGEMENT

PREVIOUS EXPERIENCE

- SAN DIEGO SUPERCOMPUTER CENTER

RELATED WORK

- DATANET FEDERATION CONSORTIUM, HTTP://WWW.DATAFED.ORG
- IRODS, HTTP://IRODS.DICERESEARCH.ORG

CONTACT INFORMATION:

RENCI
100 Europa Drive, Suite 540
Chapel Hill, NC 27517

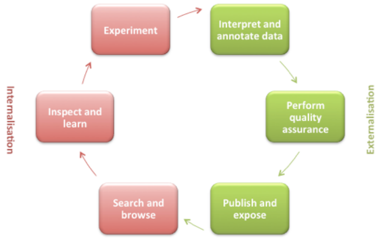
Examples of Data Management Systems

□ **Data Grids** **(data sharing)**

- Babar High Energy Physics data grid
- Broad Institute genomics data grid
- National Optical Astronomy Observatory data grid
- Ocean Observatories Initiative data grid
- The iPlant Collaborative data grid
- WellCome Trust Sanger Institute genomics data grid

□ **Digital Libraries** **(data publication)**

- French National Library
- Texas Digital Library
- UNC-CH SILS LifeTime Library

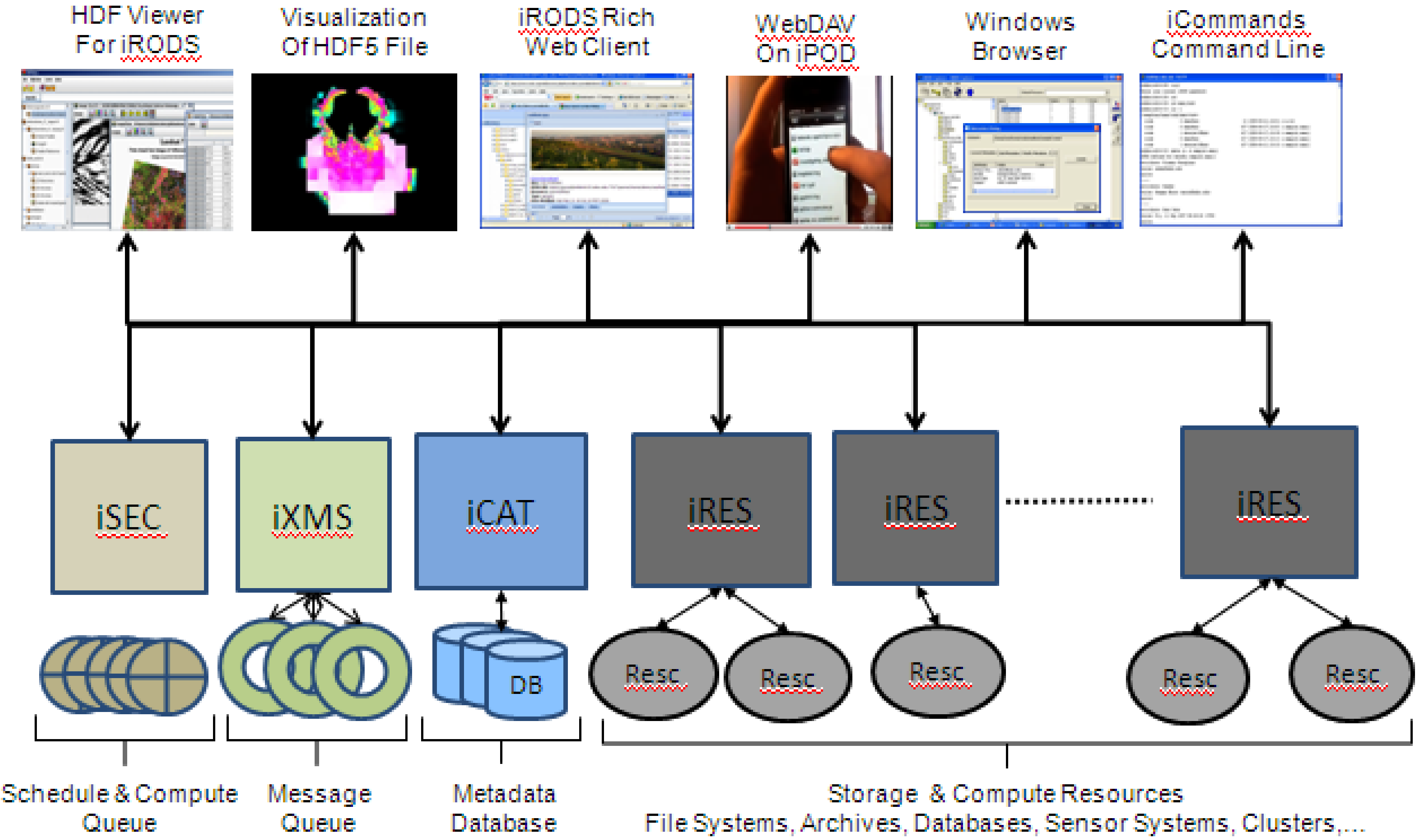


□ **Repositories / Archives** **(data preservation)**

- Carolina Digital Repository
- NASA Center for Climate Simulation
- NOAA National Climatic Data Center

▪ **Processing Pipelines** **(Reproducible data-driven research)**

iRODS Distributed Data Management



HDF Viewer For iRODS

Visualization Of HDF5 File

iRODS Rich Web Client

WebDAV On iPOD

Windows Browser

iCommands Command Line

iSEC

iXMS

iCAT

iRES

iRES

iRES

DB

Resc

Resc

Resc

Resc

Resc

Schedule & Compute Queue

Message Queue

Metadata Database

Storage & Compute Resources
File Systems, Archives, Databases, Sensor Systems, Clusters,...

Policy-Based Data Management (iRODS)

- ***Purpose***
  - Reason a data collection is created
- ***Properties***
  - Attributes needed to ensure the ***purpose***
- ***Policies***
  - Controls for enforcing desired ***properties***
  - **Mapped to computer actionable rules**
- ***Procedures***
  - Functions that implement the ***policies***
  - **Mapped to computer executable workflows**
- ***Persistent state information***
  - Results of applying the ***procedures***
  - **Mapped to system metadata**
- ***Property verification***
  - Validation that ***state information*** conforms to the desired ***purpose***
  - **Mapped to periodically executed policies**

Preservation Concepts
(data, information, knowledge)

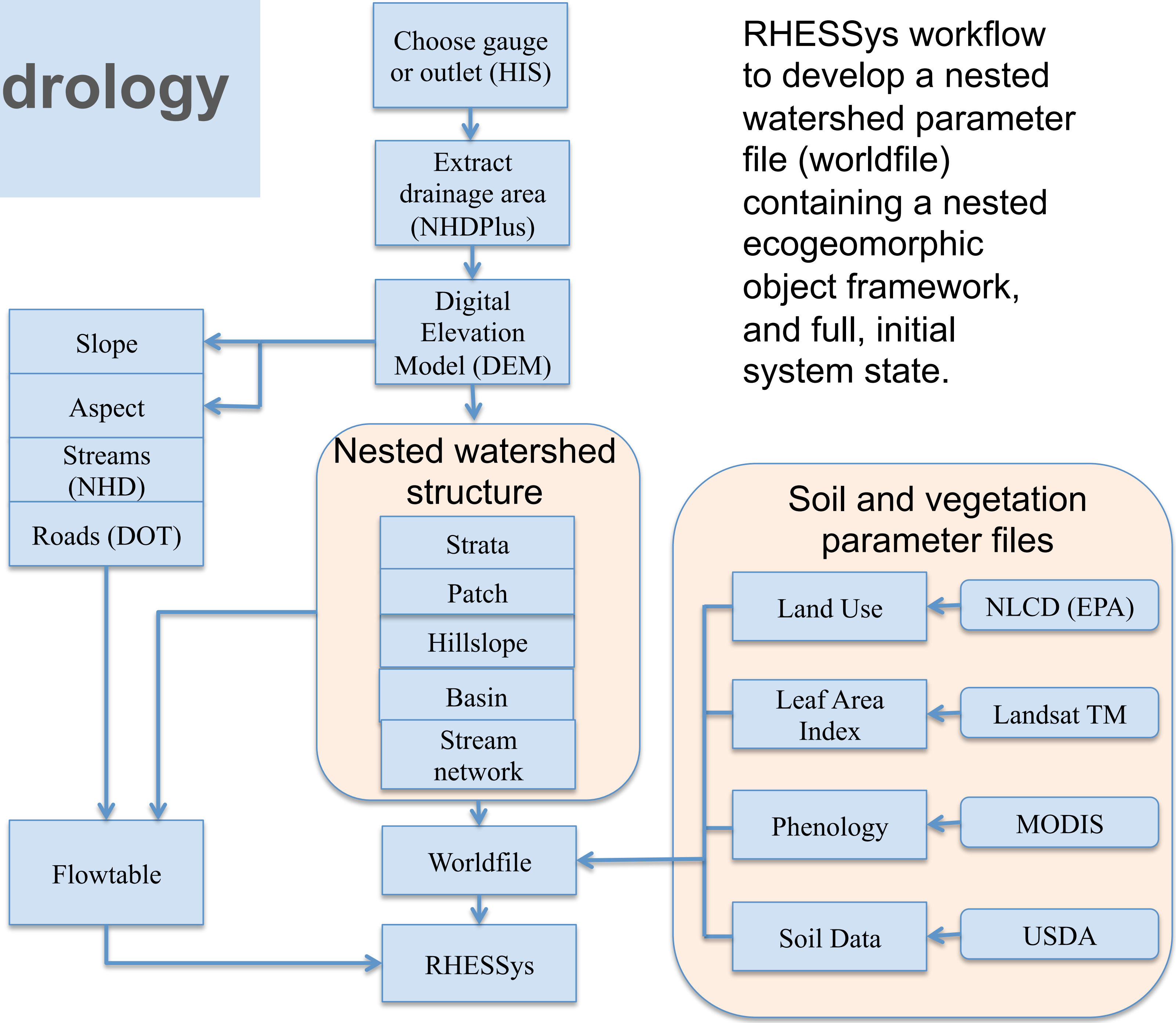
- ❑ **Preservation is communication with the future**
  - ▪ Provide a context for explaining relevance of records to a future archivist
    - Manage context as information associated with each record
    - Descriptive, provenance, and system metadata
- ❑ **Preservation is management of communication from the past**
  - ▪ Verify that assertions made by prior archivists are true
    - Encapsulate knowledge of preservation actions in procedures
    - Authenticity, integrity, chain of custody, arrangement
  - ▪ Verify application of preservation policies and procedures
    - Encapsulate assessment knowledge in procedures
    - Trustworthiness assessment criteria

Capturing Knowledge for Reproducible Data Driven Science

- ❑ **Knowledge required to interact with a remote resource for record ingestion**
  - ▪ Protocol, semantics, formats
  - ▪ Capture in micro-services
- ❑ **Knowledge embedded in analysis workflows**
  - ▪ Processing steps, input files
  - ▪ Register and share workflows
- ❑ **Knowledge associated with managing data**
  - ▪ Management policies, assessment criteria
  - ▪ Policy enforcement points

Eco-Hydrology

RHESSys workflow to develop a nested watershed parameter file (worldfile) containing a nested ecogeomorphic object framework, and full, initial system state.



Choose gauge or outlet (HIS)

Extract drainage area (NHDPlus)

Digital Elevation Model (DEM)

Slope

Aspect

Streams (NHD)

Roads (DOT)

Nested watershed structure

Strata

Patch

Hillslope

Basin

Stream network

Soil and vegetation parameter files

Land Use

NLCD (EPA)

Leaf Area Index

Landsat TM

Phenology

MODIS

Soil Data

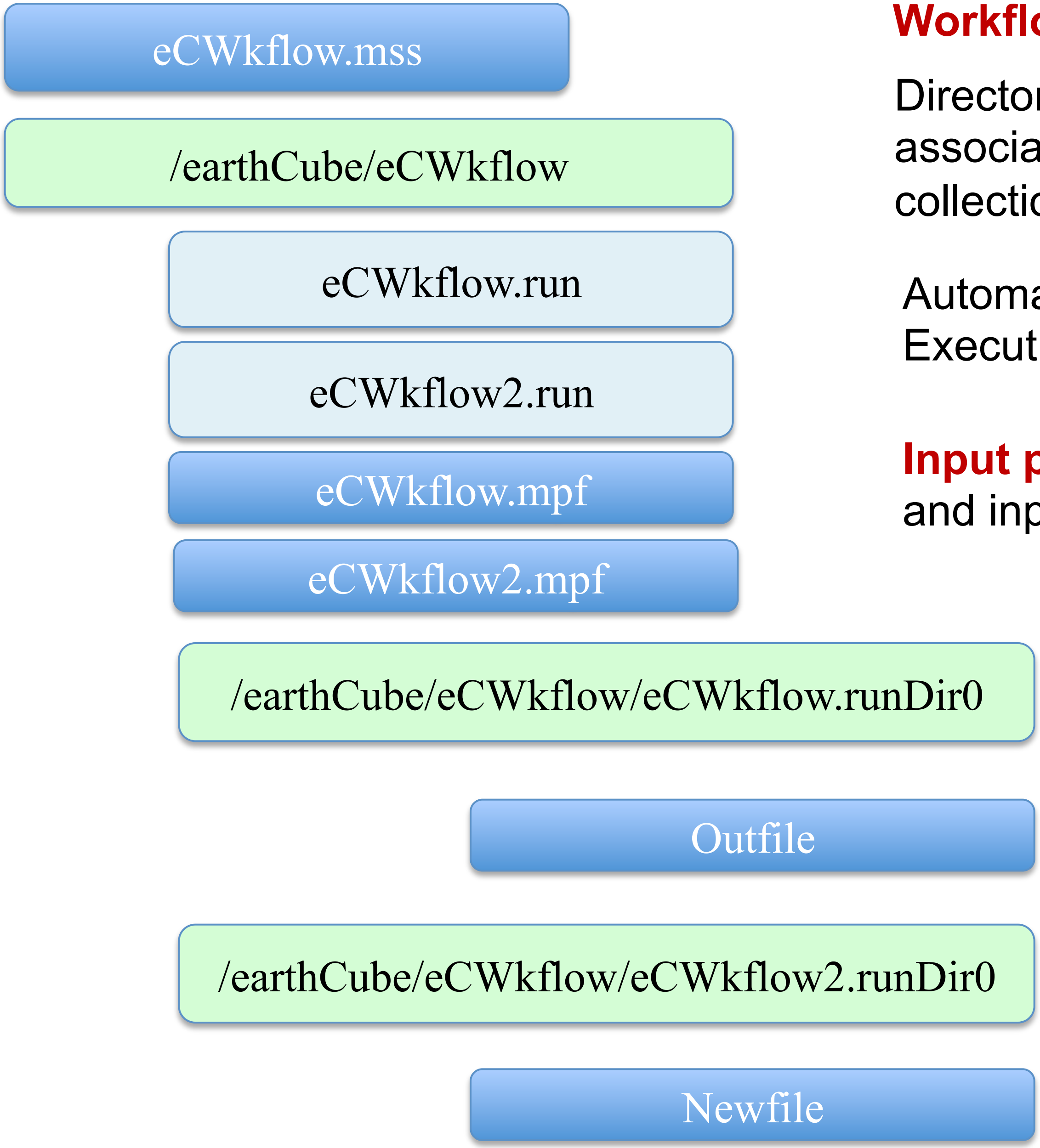
USDA

Flowtable

Worldfile

RHESSys

Workflow Management – Reproducible Research



eCWkflow.mss

Workflow file

/earthCube/eCWkflow

Directory holding all input and output files associated with workflow file (mounted collection that is linked to the workflow file)

eCWkflow.run

eCWkflow2.run

Automatically generated run file for
Executing each input file

eCWkflow.mpf

eCWkflow2.mpf

Input parameter file, lists parameters and input and output file names

/earthCube/eCWkflow/eCWkflow.runDir0

Directory holding all output files generated for invocation of eCWkflow.run, the version number is incremented

Outfile

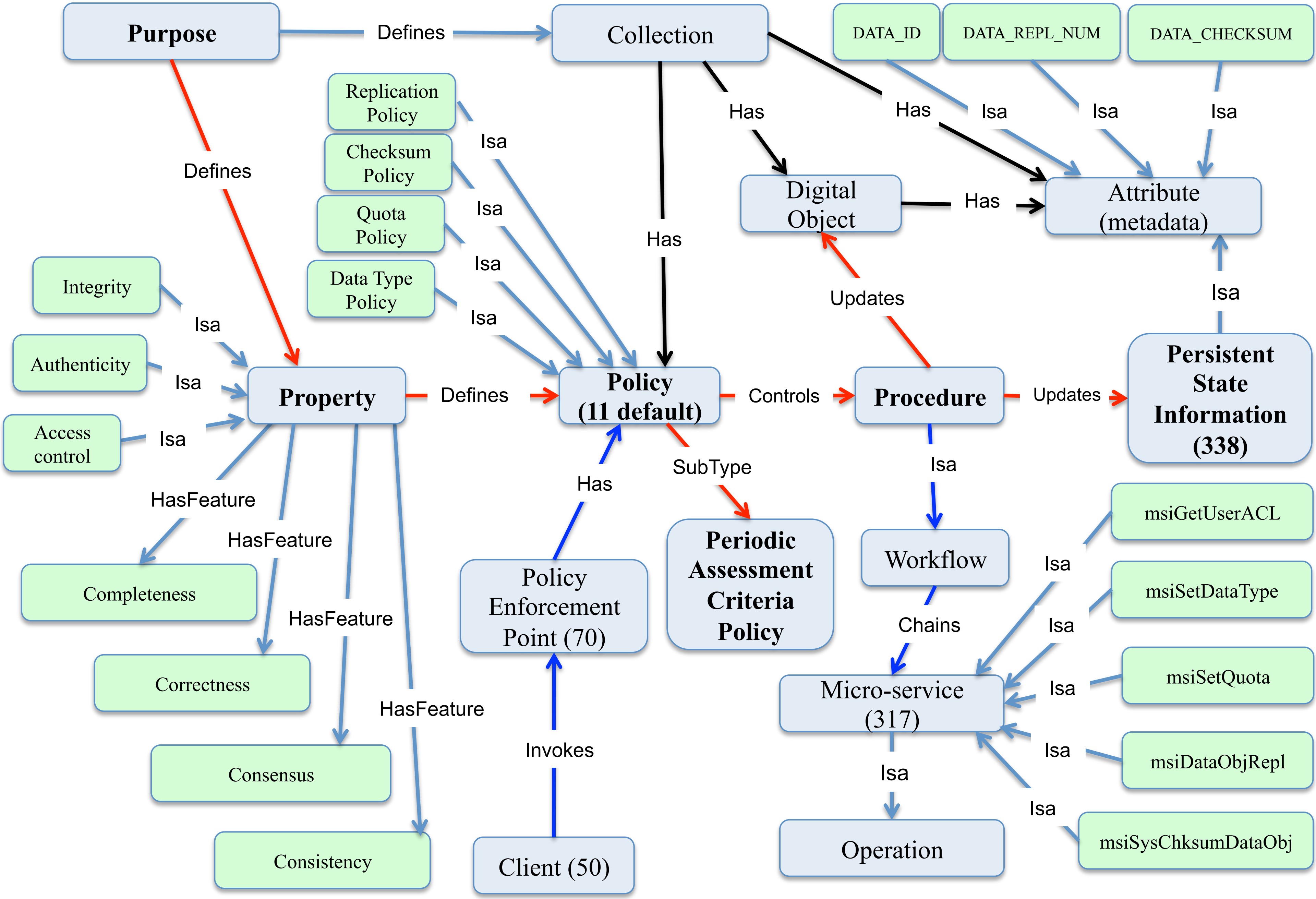
Output file created for eCWKflow.mpf

/earthCube/eCWkflow/eCWkflow2.runDir0

Newfile

Output file created for eCWKflow2.mpf

Policy-based Data Management



Technology

- ❑ **iRODS – integrated Rule Oriented Data System**
  - Open source software, source distribution
  - http://diceresearch.org
- ❑ **E-iRODS – Enterprise iRODS**
  - Open source software, binary distribution
  - http://e-irods.org

- ❑ ***NSF OCI-0940841 "DataNet Federation Consortium"***
- ❑ ***NSF* OCI-1032732 *"Improvement of iRODS for Multi-Disciplinary Applications"***
- ❑ ***NSF OCI-0848296 "NARA Transcontinental Persistent Archives Prototype"***
- ❑ ***NSF SDCI-0721400 "Data Grids for Community Driven Applications"***