Policies and Standards Community of Practice

Drs. Helen Tibbo and Christopher Lee
Ph.D. students: Jewel Ward and Deborah Maron

School of Information and Library Science
University of North Carolina at Chapel Hill

DFC DataNet FEDERATION CONSORTIUM

NSF

Original Charge

- Identify common policies and procedures across the multiple science and engineering domains for use within the DFC prototype
- Coordinate development of federation policies that will be delivered in the national prototype
- Coordinate development of a standard set of policies and procedures that can serve as an NSF data management plan

Initial Goals

1. To elicit grand challenges and significant issues from domain scientists
2. To understand how scientists think improved data management technologies can help to produce new and better science
3. To understand how to improve the retention and sharability of data across domains in order to maximize output from analysis of the data.
4. To maximize taxpayers' funding of research by improving the use and re-use of government-sponsored data and making data widely available
5. To understand current workflows and find commonalities across similar workflows to create multi-disciplinary, replicable data analysis across domains

In Other Words

Understand the workflows and data management needs of scientists in such a way as to provide the DFC programmers with insights that would allow them

- to make meaningful and useful additions and alterations to the federation infrastructure and
- to build tools that would support the work of all our domain partners and science as a whole.

Initial Tasks

- **DFC-PS1:** Elicited data needs, data practices and workflows from scientists and engineers
- **DFC-PS2:** Analyzed data and translated user needs and practices into informal "policies"
- **DFC-PS3:** Analyzed the informal "policies" and themes that emerged from the data

What We've Done

- In Years 1 and 2 we:
  - Talked with hydrologists
  - Talked with engineers
    - Probably wrong group
  - Talked with social science repository managers
  - Interviewed as many people in each category as we could, transcribed the interviews, coded the data, and analyzed what the scientists said.

What We Learned

1. **Repositories** and **networks** have official and unofficial policies.
2. **Individuals**, e.g., creators and users of data, rarely have official policies but:
   - They do have attitudes, practices and behaviors that can be seen as **unofficial "policies."**
   - The unofficial policies may **thwart** the best of system designers' intentions and the best technology, thus
3. System designers **must design** to meet not only the information needs of their customers, but also the ways in which they work and understand the ways in which these individuals will not work.

Hydrology Interviews

- 24 individuals participated in 20 interviews
- We asked about their practices, perceptions and aspirations concerning their use of data and how IT could help them
- Results revealed implicit “policies” that influence the use of technology.
- Interviews were not workflow analysis interviews/observations (Raja and his team did these).

Hydrologists' Data Sources

1. USGS stream flow data
2. NCDC/NOAA precipitation data sets
3. EPA
4. USDS Forest Service
5. FEMA
6. NASA
7. Weather stations
8. Radars
9. National Geologic survey
10. Meteorology and flux tower data
11. Water quantity and quality
12. National elevation data
13. RCS web soil survey
14. University/State Climatology Offices
15. US Army Corp of Engineers
16. US Department of Agriculture
17. NAIP/National Agricultural Imagery Project
18. Network Climate and Hydrology Databases/LTER Network Office
19. NHD/national hydrology dataset
20. Southern Nevada Water Authority
21. Various State Climate Offices

Hydrologists' File Formats & Software

- **Sample File Formats:** ASCII/binary files, NetCDF, CSV/ comma delimited files, XLS, HDF, WML, XMRG, NOAA format (?), IMG, LGS, Vector GIS, KML, ESRI, GEO TIFFs
- **Sample Software Used:** MS Office (Word, Excel), MATLAB, GIS (arcGIS), R, Fortran, ENVI, Adobe (PDF)

Common Data Operations

- Data discovery
- Data access from a workflow
- Data manipulation (parsing of a data format)
- Data transformation (converting to a new coordinate system)
- Data transformation (creating new physical variables by combining other variables)
- Data transformation (converting to new physical units)
- Data subsetting (extracting a sub-region)
- Data registration (GIS co-registration)
- Data visualization
- Creation of derived data products

What We Didn't Learn

- Despite strong interview protocol and lots of hours coding and analyzing, we did not learn anything from hydrologists and engineers that was particularly useful to our developers.
  - Scientists do not have policies as do repositories
  - We couldn't interview at a low enough level or watch individuals' workflows in order to translate this activity for the system developers
  - We were not scientists and thus had limited access
  - Widespread surveys could not elicit what we wanted to know
  - Scientists' activities were more disparate that we could manage with 20-30 interviews in each "discipline."

Year 3

- We interviewed social science repository managers/data curators and came to a revelation
  - These surrogates for various scientific communities understand data and data management in such a way that they can provide meaningful insights to the DFC project.
  - Data curators could tell us what scientists could not.
  - However, not all the fields have repositories or curators.

Where We Are

- NSF wants DFC and all the DataNet projects to be user- (i.e. scientist) driven.
- We are supposed to be eliciting “policies” from scientists but with few exceptions they do not have policies and cannot share their workflows at a low-level.
- We tried to observe workflows with no success.
- What we learn from the scientist interviews has been at a very high level.

What Is Next

- Interview data curators for science collections, starting with TDLC, iPlant, and Odum
- We will interview other science data curators as appropriate
- **Project requirements:** We will elicit policies from each project for how they want to manage their collections. **Example:** the requirement from TDLC for compliance with federal regulations on student and privacy .

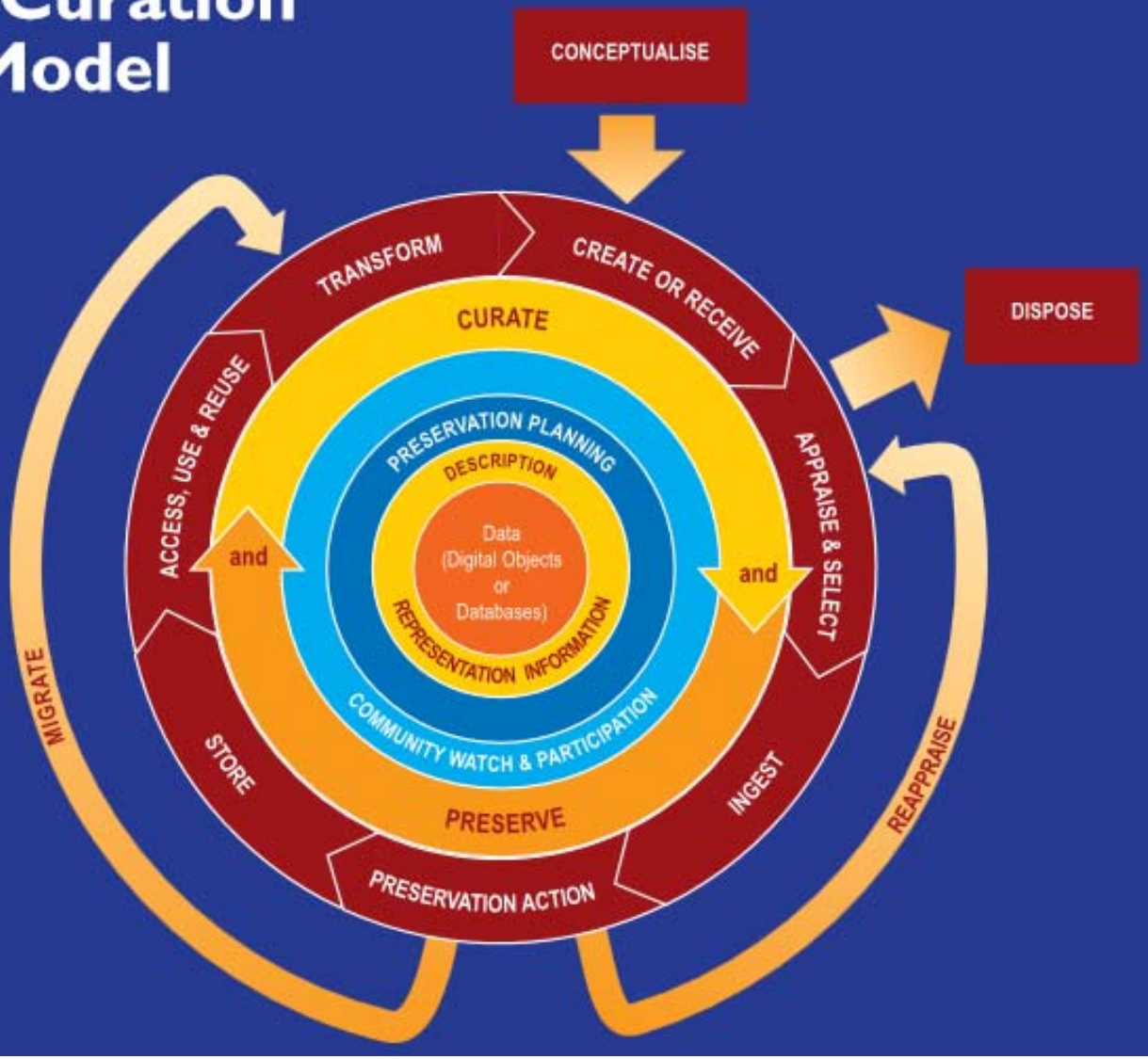
Not All Data Managers Are Curators

- From Jewel Ward's dissertation research we know that:
  - Very few iRODS "power users" have made significant alterations to the Core RE files.
  - The Core RE files do not currently contain rules supporting preservation vis-à-vis ISO 16363.
  - Just because there is a repository and someone is using DFC software does not mean the data are being managed for long-term preservation.

Infusing Preservation in DFC Microservices, Rules and Polices

- We will analyze international data curation standardsat a level by which we can translate these best practices, mandates, and policies into DFC rule sets
- Top-down approach, although these standards were created with bottom-up approaches.
- Looking at the preservation of access for data sharing.

The DCC Curation Lifecycle Model



Need for Early Intervention

- Systems we build must:
  - Support scientific workflows
  - Support long-term preservation
- Curation literature repeatedly notes that curators (or their automated systems) must intervene early in the data lifecycle if data is to be preserved for the long-term
- Very hard to go back and curate retrospectively
- Need to involve **data creators** in the process in as effortless and seamless a fashion as possible
- This calls for more **data curators**

Metadata, Metadata, Metadata

- Best to gather metadata (automatically or as easily as possible) at the point of data creation
- Very expensive and hard (impossible??) for curator to do this after the fact
- Without metadata there is no data reuse or preservation
- This is a major impediment to data sharing.

Curation, Curation, Curation

- Data reuse and preservation are not natural acts and do not happen on their own
- Data sharing and preservation require effort and someone must oversee these activities
- The data curator must work with the data creators, data users, and IT staff across the data curation lifecycle to ensure long-term reusability

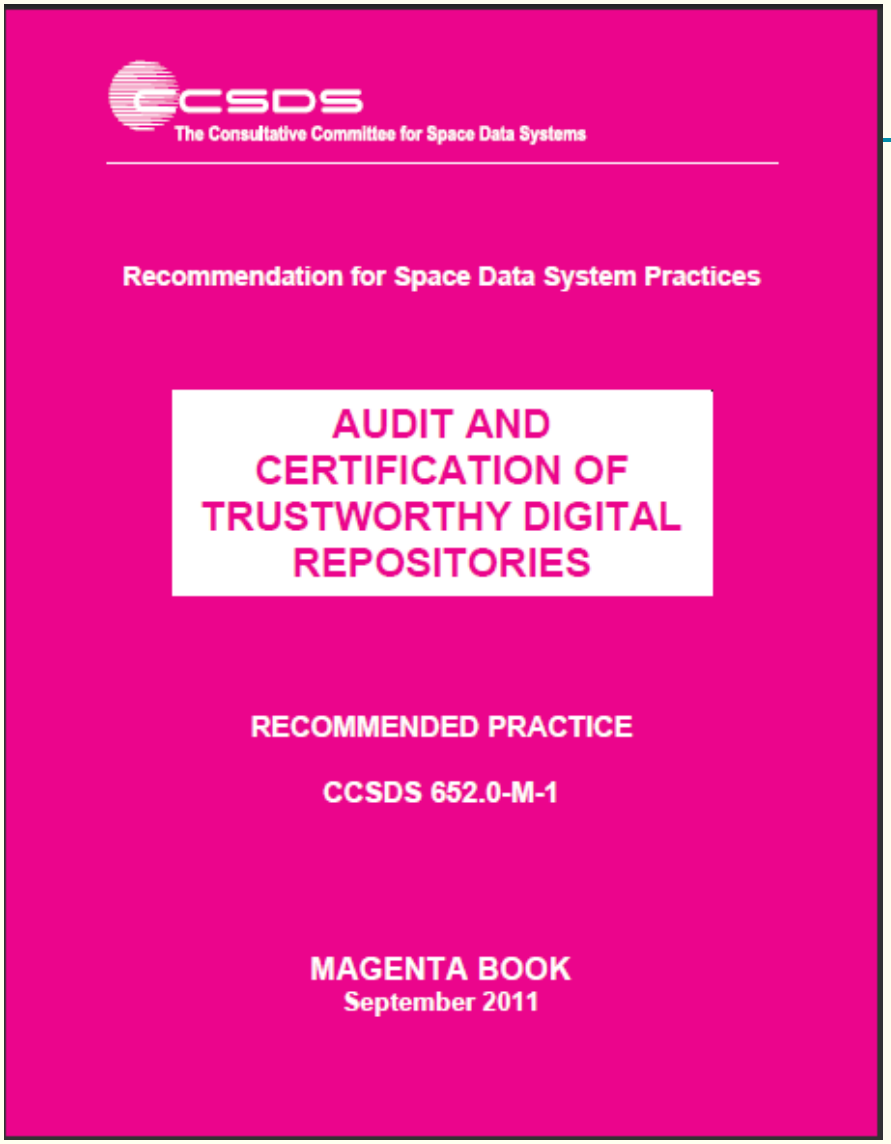
ISO 16363: Audit and Certification of Trustworthy Digital Repositories

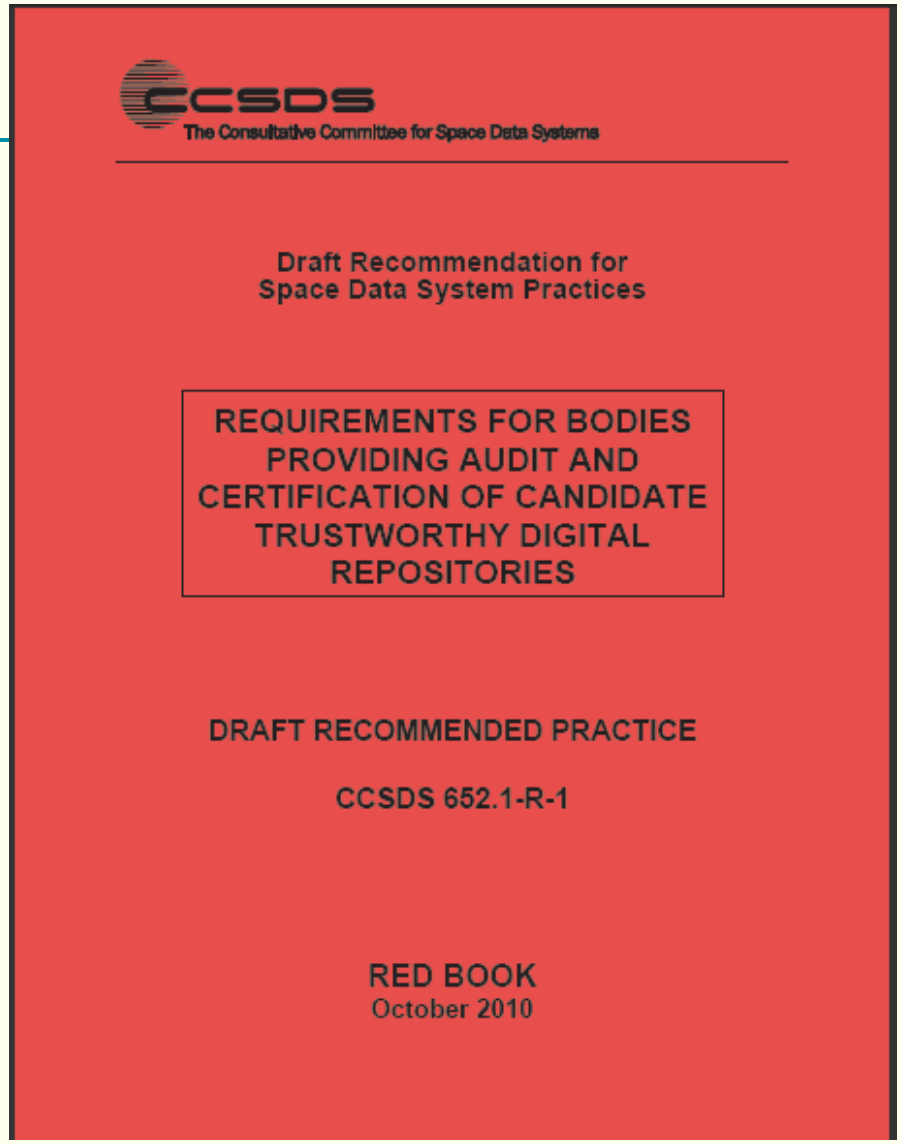
Several different audit and certification processes now developed:

- Basic Certification
  - Data Seal of Approval (DSA)
  - World Data System Certification (WDS)
- "Formal" Certification
  - Trustworthy Repositories Audit and Certification (TRAC)/ISO 16363 (includes site visit)
- Other alternatives
  - Self-audits against TRAC,
  - Peer reviews
  - Digital Repository Audit Method Based On Risk Assessment (DRAMBORA)

ISO 16363 Background

- Operating under authority of International Standards Organization, a Working Group has developed the ISO 16363, Audit and Certification of Trustworthy Digital Repositories. (CCSDS 652.0-M-1)
- And a companion draft International Standard 16919, Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories
- The latter draft standard was developed by the Primary Trustworthy Digital Repository Audit and Certification Accreditation Board (PTAB)


CCSDS
The Consultative Committee for Space Data Systems
Recommendation for Space Data System Practices
AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES
RECOMMENDED PRACTICE
CCSDS 652.0-M-1
MAGENTA BOOK
September 2011


CCSDS
The Consultative Committee for Space Data Systems
Draft Recommendation for Space Data System Practices
REQUIREMENTS FOR BODIES PROVIDING AUDIT AND CERTIFICATION OF CANDIDATE TRUSTWORTHY DIGITAL REPOSITORIES
DRAFT RECOMMENDED PRACTICE
CCSDS 652.1-R-1
RED BOOK
October 2010

ISO 16363- Organizational Infrastructure

- 3.1 Governance and Organizational Viability
- 3.2 Organizational Structure and Staffing
- 3.3 Procedural Accountability and Preservation Policy Framework
- 3.4 Financial Sustainability
- 3.5 Contracts, Licenses, and Liabilities

ISO 16363: Digital Object Management

- 4.1: Ingest: Acquisition of Content
- 4.2 Ingest: Creation of the AIP
- 4.3 Preservation Planning
- 4.4 AIP Preservation
- 4.5 Information Management
- 4.6 Access Management

ISO 16363: Infrastructure & Security Risk Management

- 5.1 Technical Infrastructure Risk Management
- 5.2 Security Risk Management

DFC Focus on Section 4: Data Object Management

- 4.1 INGEST: ACQUISITION OF CONTENT
  - 4.1.1The repository shall identify the Content Information and the Information Properties that the repository will preserve.
    - 4.1.1.1 The repository shall have a procedure(s) for identifying those Information Properties that it will preserve.
    - 4.1.2 The repository shall clearly specify the information that needs to be associated with specific Content Information at the time of its deposit
    - 4.1.4 The repository shall have mechanisms to appropriately verify the identity of the Producer of all materials

From "Creation of the AIP"

- 4.2.5 The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains.
  - 4.2.5.1 The repository shall have tools or methods to identify the file type of all submitted Data Objects.
  - 4.2.5.2 The repository shall have tools or methods to determine what Representation Information is necessary to make each Data Object understandable to the Designated Community.
  - 4.2.5.3 The repository shall have access to the requisite Representation Information.
  - 4.2.5.4 The repository shall have tools or methods to ensure that the requisite Representation Information is persistently associated with the relevant Data Objects.

To Recap

- For the most part, scientists appear not to care about long-term preservation and certainly do not know how to carry this out (well, it's not their job…)
- IT staff (at least those whom Jewel interviewed) do not seem focused on preservation and have done little to ensure long-term reuse of data
- Data repository staff (Odum, ICPSR, UKDA, etc.) do understand and care about long-term data preservation

Preservation Rules

- Appears to be a gap regarding preservation across content creators (scientists), IT staff who manage IRODS systems, and data curators
- We cannot expect preservation rules to come from anyone except preservation repository curators and preservation experts
- Preservation rules must follow from international standards such as ISO 16363
- Implementation of rules must be tested with data creators and IT staff

Next Steps

- We will deconstruct ISO 16363
- Take those mandates that are concrete (not “The repository will have a mission statement”)
- Work with DFC staff to translate these into microservices and rules
- Bundle these microservices into various sets with preservation-appropriate default values
- Ship these bundles with their defaults in the standard iRODS install
- Educate repository curators and IT managers regarding the usefulness of employing these bundles

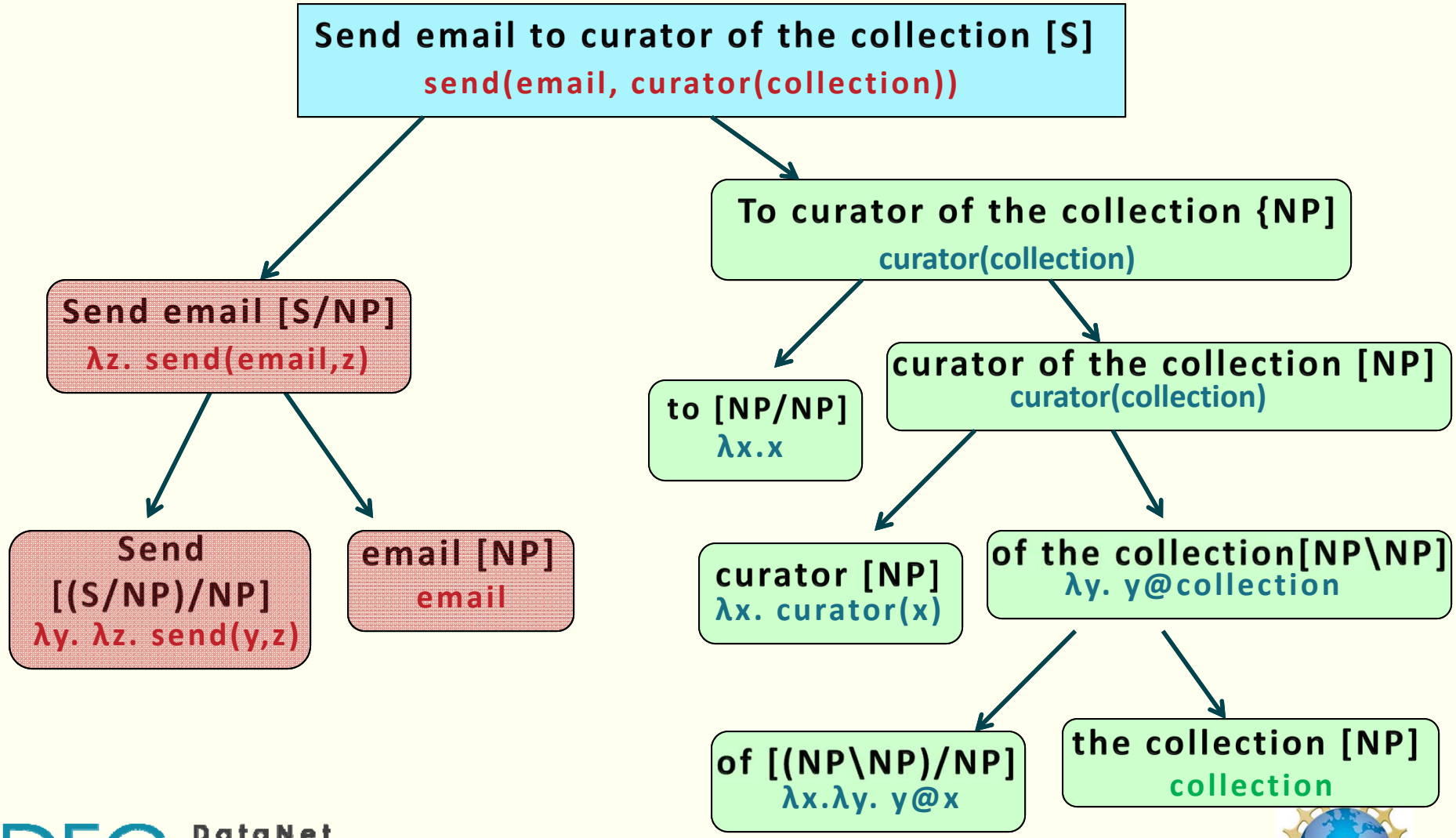
Improving Infrastructure for Writing Policies

Work by:

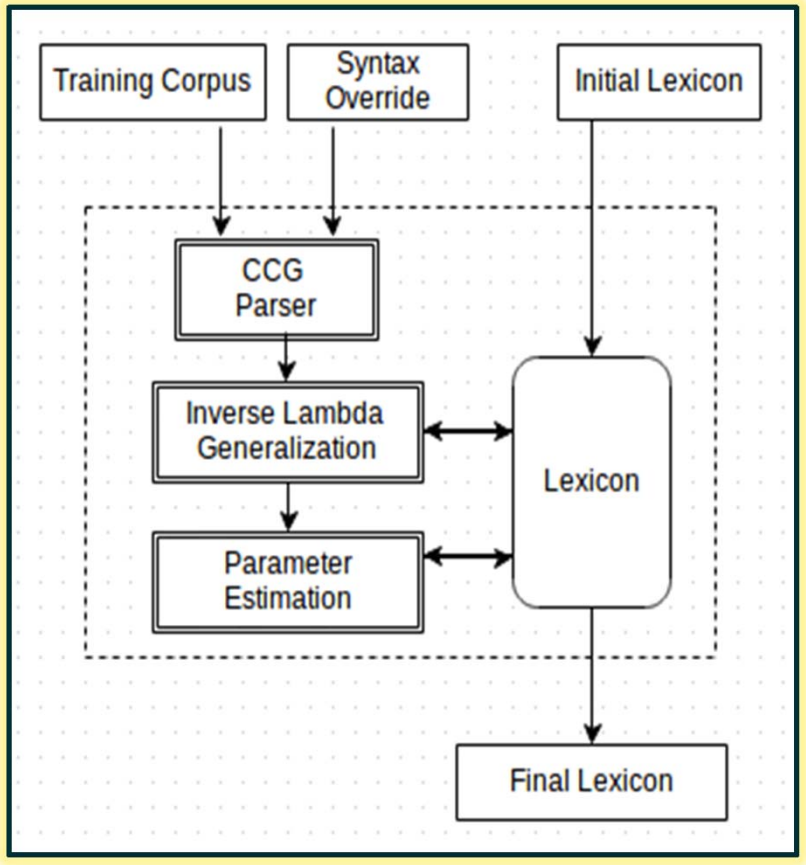
Chitta Baral
Arizona State University

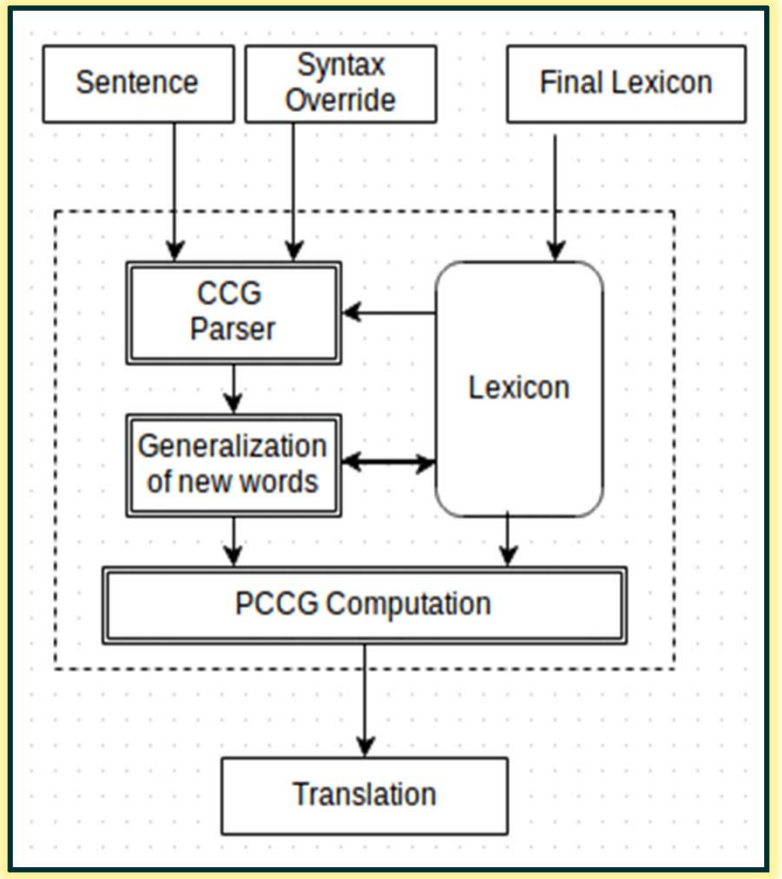
DFC DataNet FEDERATION CONSORTIUM

An Example


Send email to curator of the collection [S]
send(email, curator(collection))
Send email [S/NP]
λz. send(email,z)
To curator of the collection {NP]
curator(collection)
to [NP/NP]
λx.x
curator of the collection [NP]
curator(collection)
Send
[(S/NP)/NP]
λy. λz. send(y,z)
email [NP]
email
curator [NP]
λx. curator(x)
of the collection[NP\NP]
λy. y@collection
of [(NP\NP)/NP]
λx.λy. y@x
the collection [NP]
collection

NL2KR System Architecture





NL2KR-L **NL2KR-T**

DFC DataNet FEDERATION CONSORTIUM

www.datafed.org

www.irods.org

National Science Foundation Cooperative Agreement: OCI-0940841