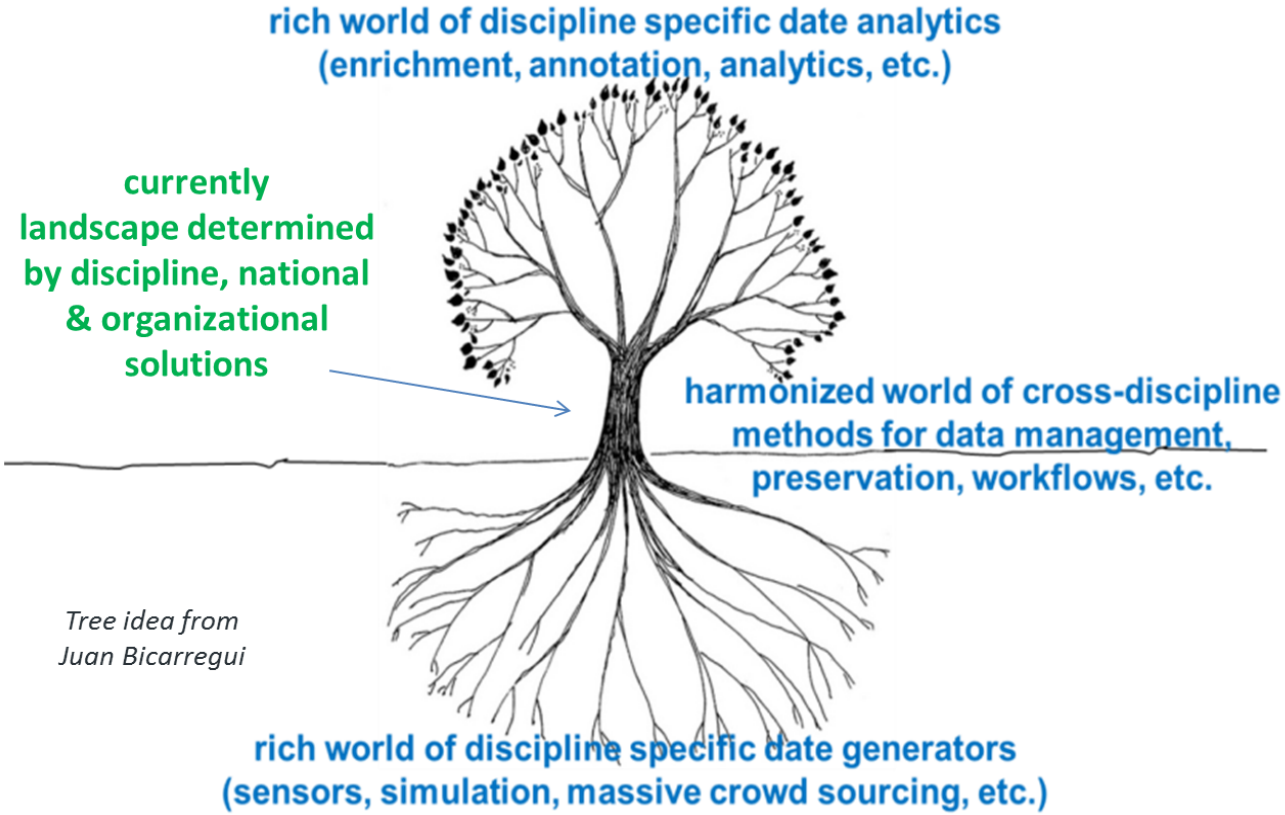
Data Fabric : the pragmatic potential (aka, let's start small)

Beth Plale
Indiana University
RDA Technical Advisory Board co-Chair

research data sharing without barriers
rd-alliance.org

Current landscape is lacking a lifeblood trunk



Until now in RDA

- started with a number of WG activities in RDA P1
  - DTR, PP, PIT, MD, DFT – the old ones
  - some people found urgent and interesting topics
  - almost at the end and some questions: what now, how does it all fit into landscape, etc.?
  - they started this Data Fabric brainstorming
- even more groups started
- are there general themes in the data landscape
  - the whole issue of data publishing/citation/etc.
  - the whole issue of scientific culture/legal & ethical aspects
  - our daily data work in the departments – the Data Fabric
  - may be more

Scope and Characteristics of RDA's DF

- DF is about
  - making departments' data science reproducible
  - creating the conditions for trust in the anonymous data domain
  - identifying mechanisms, components and interfaces making data science efficient and cost effective
  - discussing cross-disciplinary approaches
  - defining a framework that allows to include new components or component variants in a flexible way

- Example:
  - DF will state necessity of a worldwide available machinery to register & resolve DOs, we will say something about registered attribute types and specify an API
  - but we will not say how to implement and use such a system

Scope and Characteristics of RDA's Data Fabric

- Data Fabric is NOT about
  - prescribing an overarching architecture we need to follow
  - specifying an implementation of such an architecture
  - discussing specific technologies and tools
  - more than discussing the processing machinery (not publication, citation, I & e, etc.)

- Data Fabric is about highly automated procedures or at least guidance to follow such procedures.

DF IG way of acting

- DF IG must be an inclusive open platform for interaction
- DF IG needs to place the various WGs/IGs on the landscape
- DF IG needs to identify barriers across groups
- DF IG can work as umbrella to maintain WG results
- open position papers will summarize the state of discussions and provoke convergence debates
- it will NOT take place of existing coordination mechanisms

Snapshot of state of DF IG from RDA P4 Amsterdam Sep 14

1. What are DF's Scope and Characteristics?
2. What are DF's components, interfaces, mechanisms?
3. How should DF act?
4. Who will chair DF IG?

Data Fabric : low hanging fruit

- Example of type of activity discussed at P4.
- First step:
  - Show RDA/US Fellow work over Summer 2014 with one of the finishing Working Groups
  - Illustrates what can be done with the Data Type Registry
- Second step:
  - Suggest how can be stitched into larger, and still larger demonstration use case

Data Type Registry (DTR) : a first product of RDA

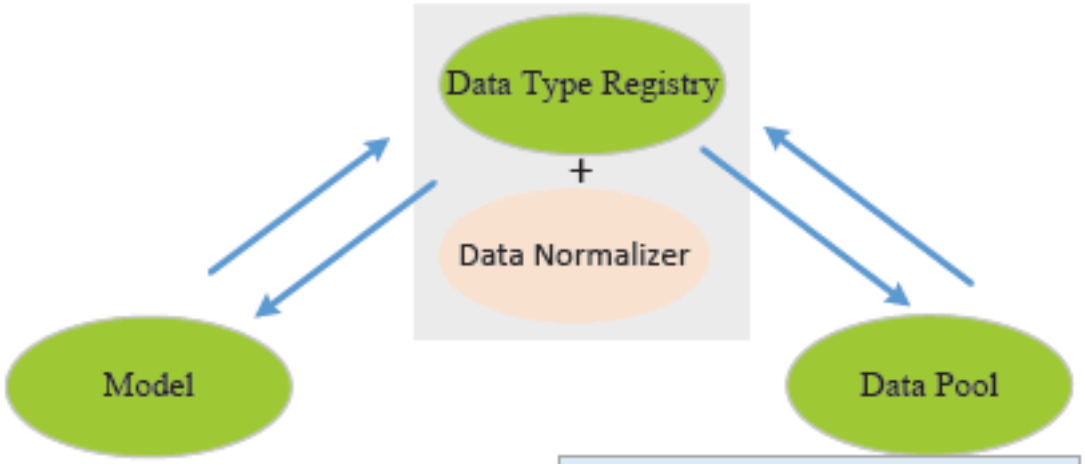
- Data type has a unique and resolvable identifier
  - Resolves to characterization of structures, conventions, semantics, and representations of data
  - Serves as a shortcut for humans and machines to understand and process data
- File formats and mime types have solved the representation problem at a unit level
- DTR data types aims to solve other problems
  - It is a number in cell A3, but is it a temperature? If so, in Celsius?
  - It is a dataset consisting of location, temperature, and time, but what variable names should I look for?
  - It is all packaged in CSV or NetCDF? And as a single unit or a collection of units?
- Prototype registry: http://typeregistry.org

Repre-sentation of data type for stream gauge

```
identifier : "11314.3/6debc53338e99ff15731",
 name: "Stream Gauge",
 description: "Information that defines stream discharge at a specific location
and time interval. Useful for the geosciences community."

standards :
   issuer : "ISO"; name: "4375:2000"; natureOfApplicability: "depends"

provenance:
 contributors
       identifiedUsing: "Text"; name: "Mostafa Elag"; details: "A Researcher in
the geosciences community from UIUC."
       identifiedUsing : "Text"; name: "Giridhar Manepalli"; details: "A data
infrastructure expert from CNRI."
 creationDate : "2014-08-07T04:28:21.479Z",
 lastModificationDate : "2014-09-08T15:28:00.733Z"
 expectedUses : "Used for comparing outputs of surface runoff discharge
models as applied to data pertaining to\na specific watershed.",

representationsAndSemantics:
 expression : "Measurement Unit",
 value: "Cubic Meter per Second"
 properties:
       name: "value"; identifier: "11314.3/f0f2c4382dcf8d257462";
       name:"coordinate"; identifier : "11314.3/4102c3ebe68bed21d644"
       name: "timestamp"; identifier: "11314.3/6386f4ebd23e9baace50"
 relationships:
       name: "Primary Key"; relativeNames : [ "value" ]
```

Credit: Mostafa Elag, UIUC

Building blocks: RDA Fellow proposed to introduce normalizing service to normalize data for use by a model using data types and DTR



Data Normalizer

-GetExpectedModelInputType()
- GetDataTypeFromData()
- GetTypeDefinitionFromRegistry()
- Normalize()
  - units()
  - variable-names()
  - Scale()
-ApplyModelToNormalizedData()
- AddNewDataType()

Example Model: Hargreaves-Samani

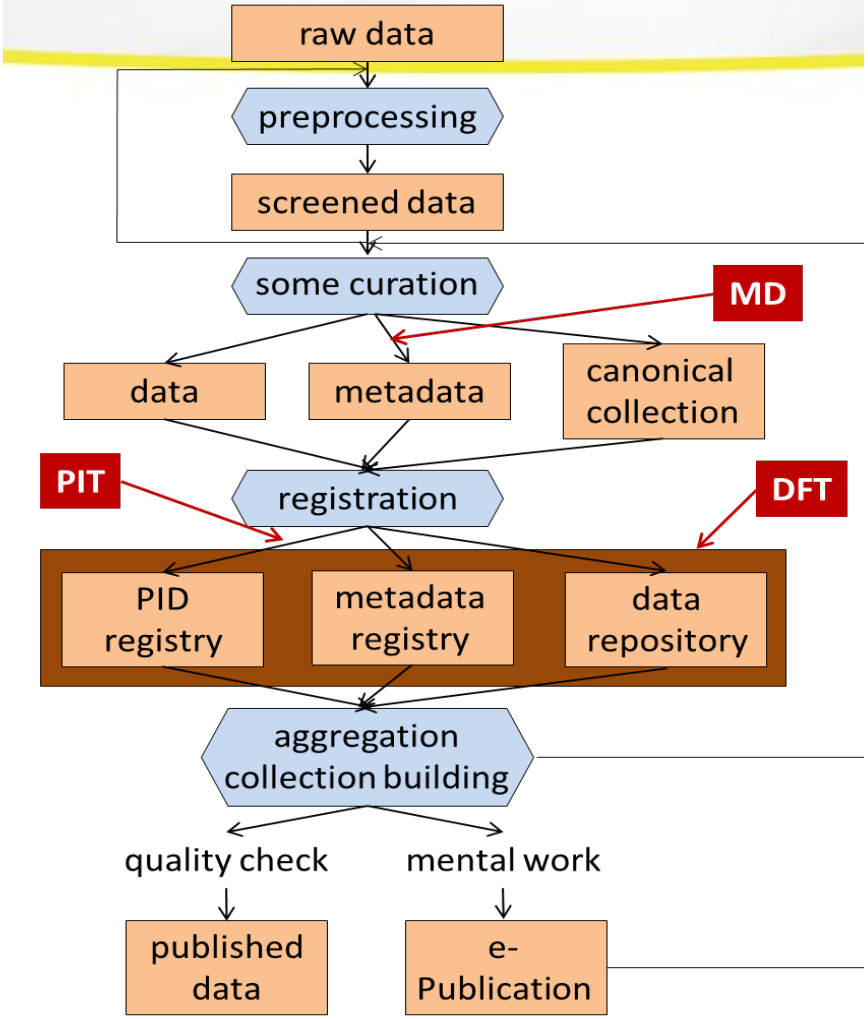
- Required Model Inputs: Radiation and Temperature
- Data A has Radiation, Location
- Data B has Temperature, Location
- Extract Data C to include Radiation and Temperature for a given Location
- Normalize Data C to match Model Input forms using "Data Normalizer"
- Apply Model
- Model Output is registered as new Data Type if applicable

Credit: Mostafa Elag, UIUC, Dept of Civil Engineering

- We have outlined the conceptual development of Data Normalizer using Data Types.
- Potential uses of DTR + Data Normalizer combination include:
  - Overcoming the syntactic heterogeneity in the exchange items between models and data in different component- based modeling frameworks (e.g. data in CUASHI used in HydroModeler).
  - Guiding a user in choosing appropriate data from long-tail data collection (e.g. SEAD).
  - Verifying the consistency of data structures across related scientic domains.

Sketch of Data service / data facility

(inverted that is)



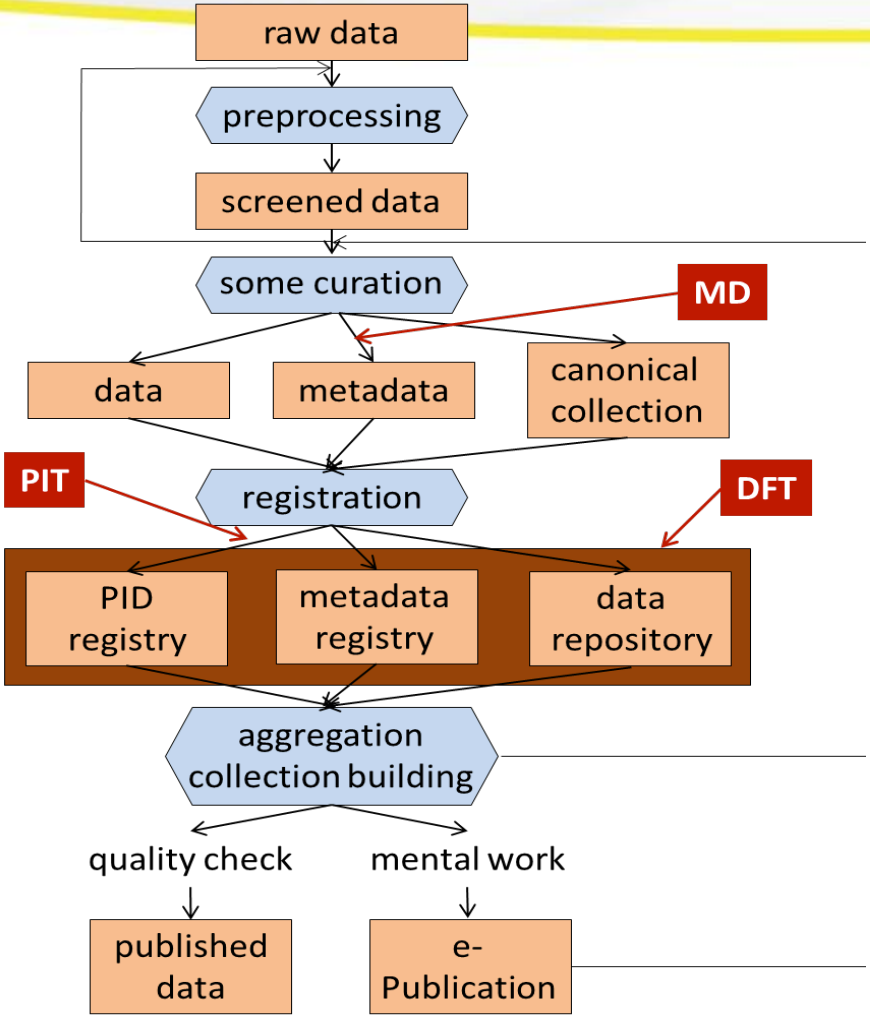
Red boxes show select WG activity

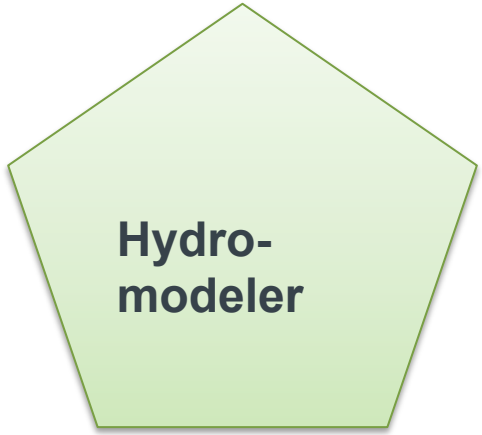
MD: Metadata Registry WG

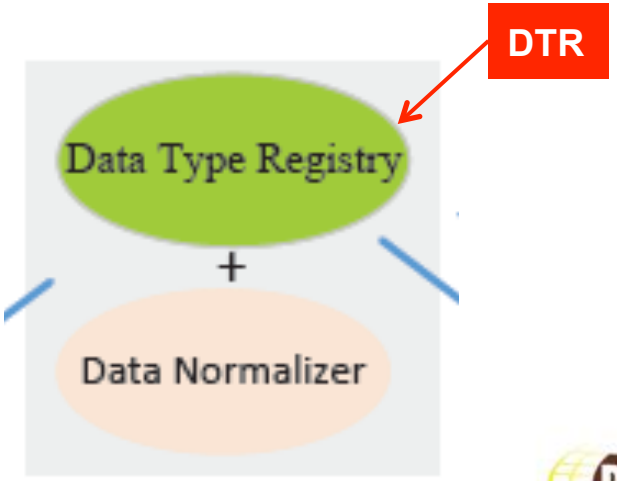
PIT: Persistent Identifier Type WG

DFT: Data Foundations and Terminology WG

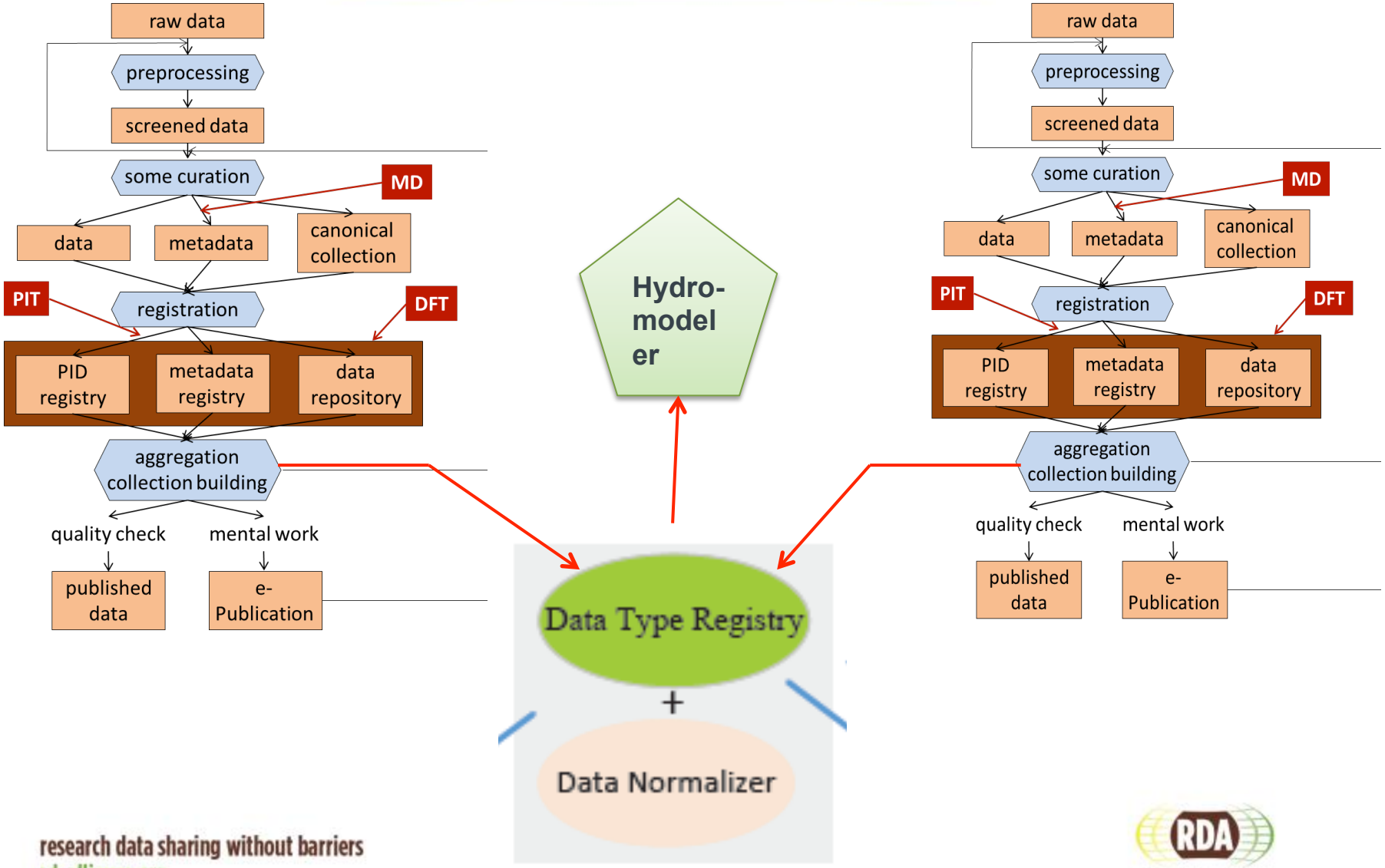
The demonstration: data in CUASHI used in HydroModeler







Data in CUASHI and SEAD and used in HydroModeler



Summary Thoughts

- RDA has 12-18 month working groups. This means outcomes are going to be small (but real) and built on consensus.
- Stitching these together has to be possible
- Data Fabric is an IG that emerged summer prior to P4 to discuss issues of common fabric. The thought in everyone's mind in the group is that many of the WG products can and should work together.
- Today's talk shows one example of how Data Fabric interoperability could be achieved.

Sincere thanks to

- Mostafa Elag, Univ of Illinois Urbana Champaign Dept of Civil Engineering, and RDA/ US Fellow, 1st Cohort
- Peter Wittenberg, Max Planck Institute of Psycholinguistics
- Giridhar Manepalli, CNRI
- National Science Foundation
- … and everyone worldwide who is working to reduce the technical and social barriers to sharing research data