Practical Innovations of the NIS and PASTA

Duane Costa
NIS Lead Programmer

2012 Mid-term Review

LTER Network Office

Four Key Practical Innovations

INTEROPERABILITY

PROVENANCE

ACCESS

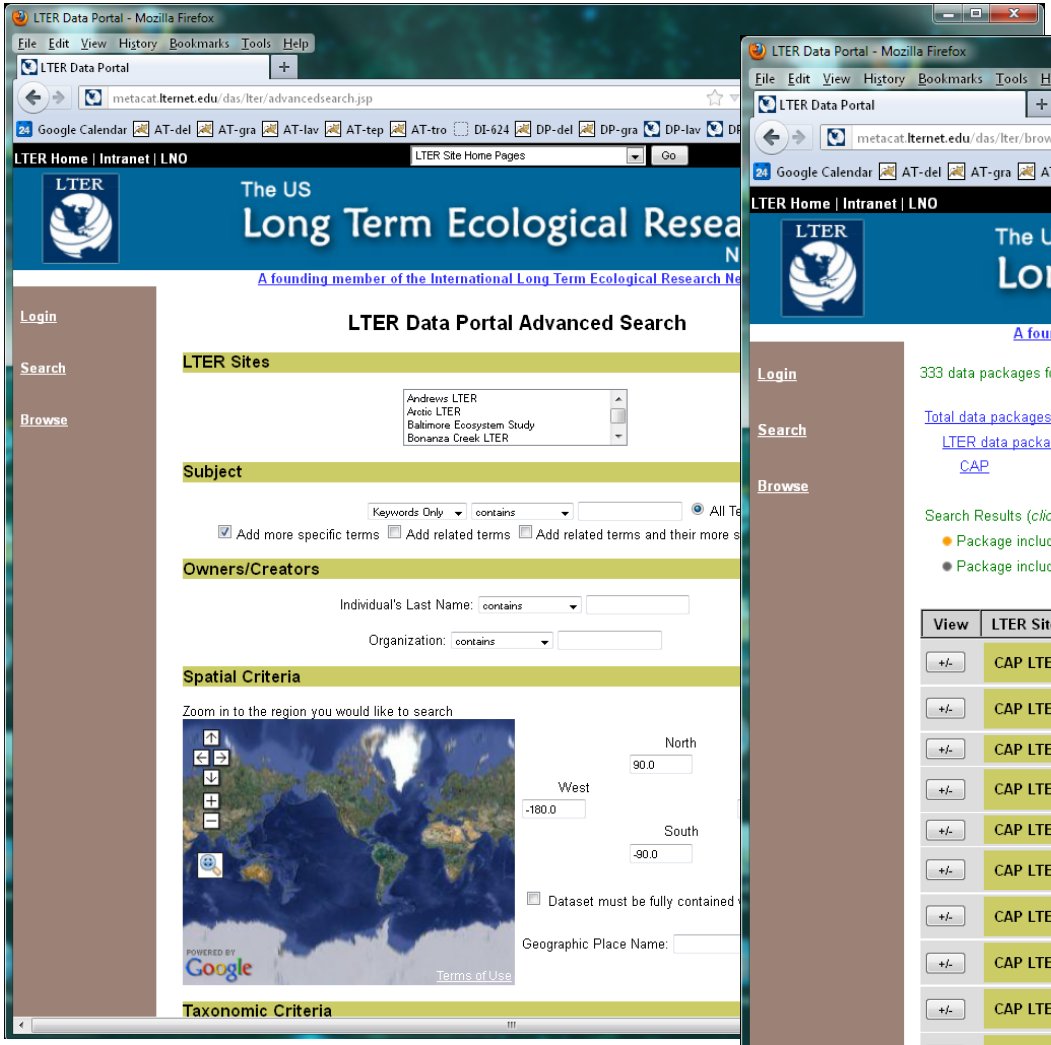
QUALITY

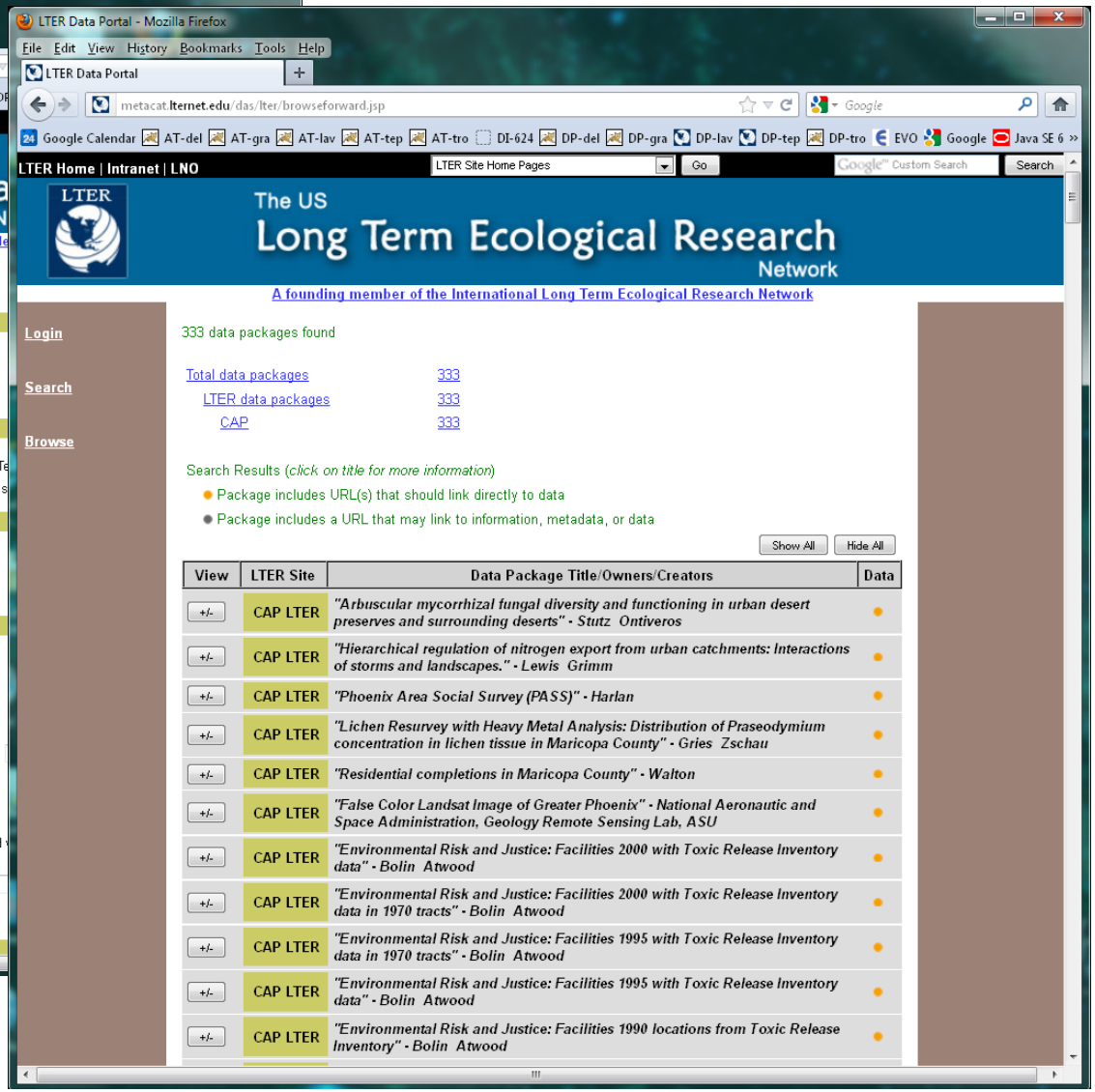
Why "practical innovations"?

- What's *practical* about them?
  - All four can be demonstrated in the NIS prototype
  - Add to the flexibility, usefulness, responsiveness, and data accessibility for individual LTER sites and users outside of the LTER community
- What's *innovative* about them?
  - All four capabilities are either deficient or absent from the legacy LTER Data Catalog
  - Innovative even in the context of the greater ecoinformatics community, in particular:
    - Provenance tracking
    - Quality reporting

Legacy LTER Data Catalog

http://metacat.lternet.edu





LTER Network Office

Legacy LTER Data Catalog: Does some things well

- 2004-Present
- 7000+ LTER metadata (EML) documents
- Good search capability
  - Simple and advanced search
  - Utilizes LTER Controlled Vocabulary
- Reasonable performance
- Recent improvements to UI
  - Search results presentation
  - Metadata presentation
  - Data access

http://metacat.lternet.edu

Legacy LTER Data Catalog: Where it needs improvement

- Performance could be better
- Mostly closed system:
  - Web application, not web services
  - Machine-to-machine interaction available via Metacat back-end, but not prominent
- Data access has improved but is still uneven
  - Historically, a large source of user frustration
- Lacks provenance tracking
  - Important to synthesis efforts
- Minimal quality control
  - Valid EML metadata is the only requirement for insertion

http://metacat.lternet.edu

Four Key Practical Innovations

INTEROPERABILITY

PROVENANCE

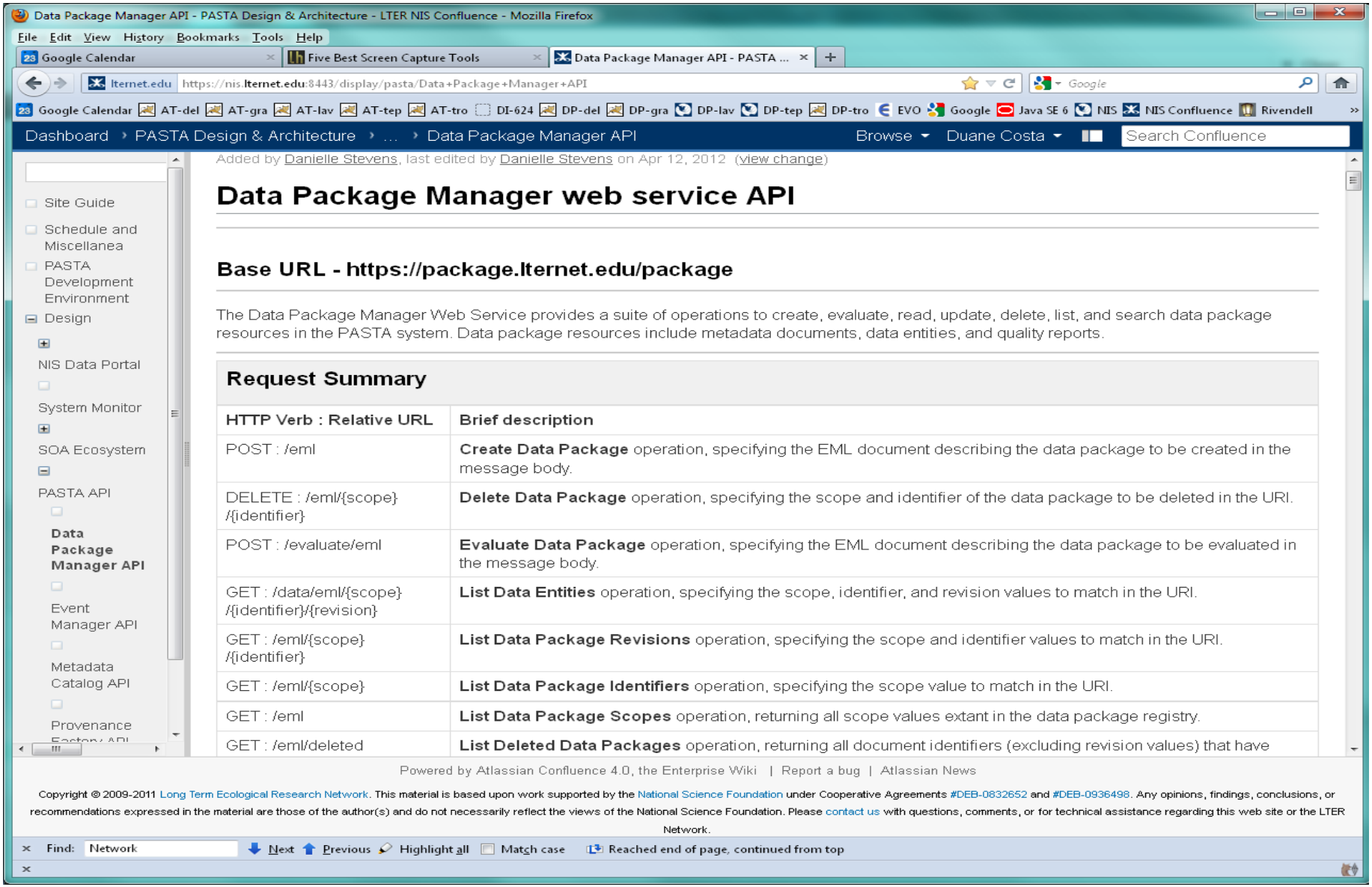
ACCESS

QUALITY

Interoperability

- Open system
  - Service Oriented Architecture, Web services API
- Flexibility to build the client application *you* want using the platform and programming language of *your* choice
- Could potentially be utilized by NEON, GLEON, CUASHI, etc.

PASTA Web Service API



Interoperability: Examples

- Example 1: EML Congruency Checker (Summer, 2011)
  - Margaret O'Brien, SBC Information Manager
  - Generated LTER-wide quality reports using an early implementation of the Quality Engine
  - Perl scripts and shell scripts
- Example 2: EML Pre-flight Checker (Winter, 2012)
  - Sven Bohm, KBS Information Manager
  - Updated version of the ECC using a newer PASTA API
  - Ruby on Rails
- Example 3: NIS Data Portal (Winter-Spring, 2012)
  - Serves as a reference implementation of a PASTA client application
- Example 4: Audit report web application for a particular LTER site
  - Example of a potential client application

"The PASTA web services were very powerful in their assessment of the data, but also simple to access using only Linux command line tools."

Margaret O'Brien

Information Manager, Santa Barbara Coastal LTER

Used with permission

Four Key Practical Innovations

INTEROPERABILITY

PROVENANCE

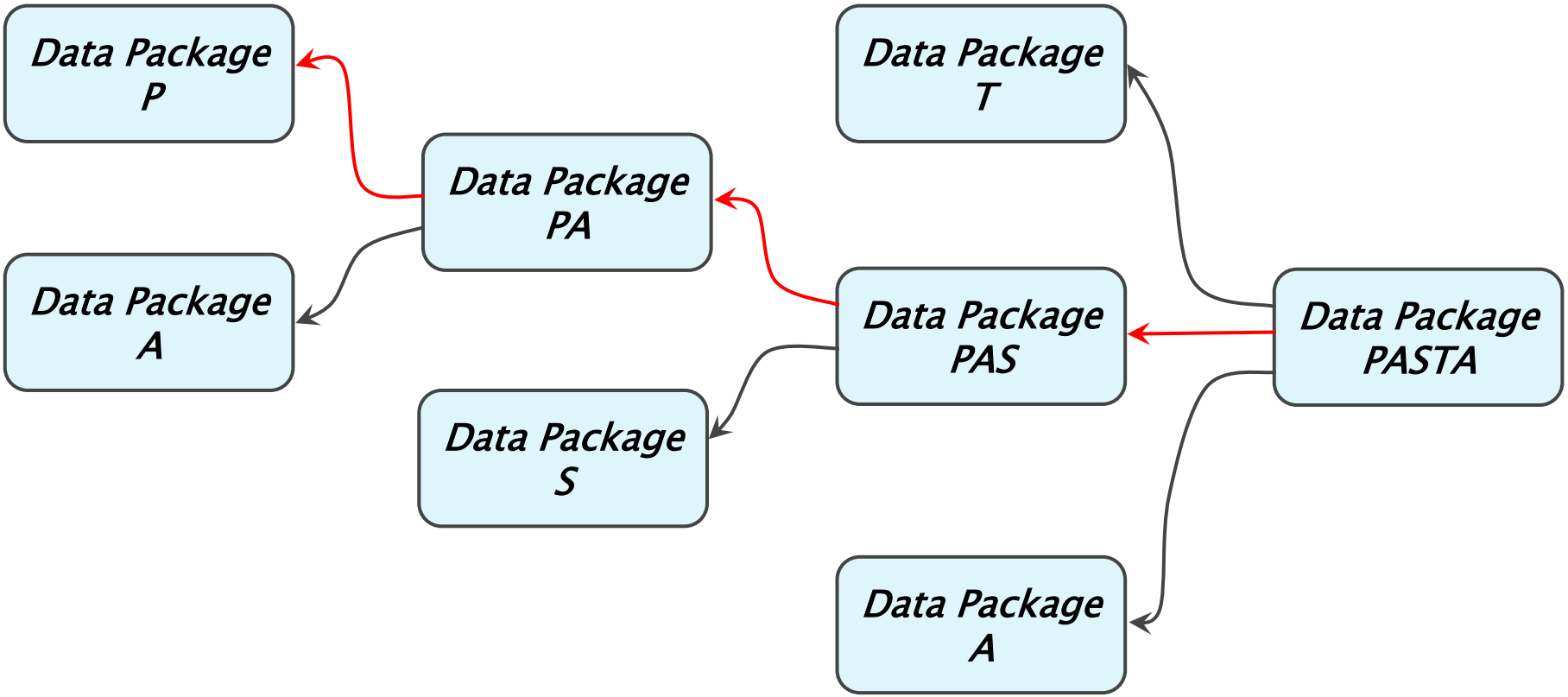
ACCESS

QUALITY

Provenance

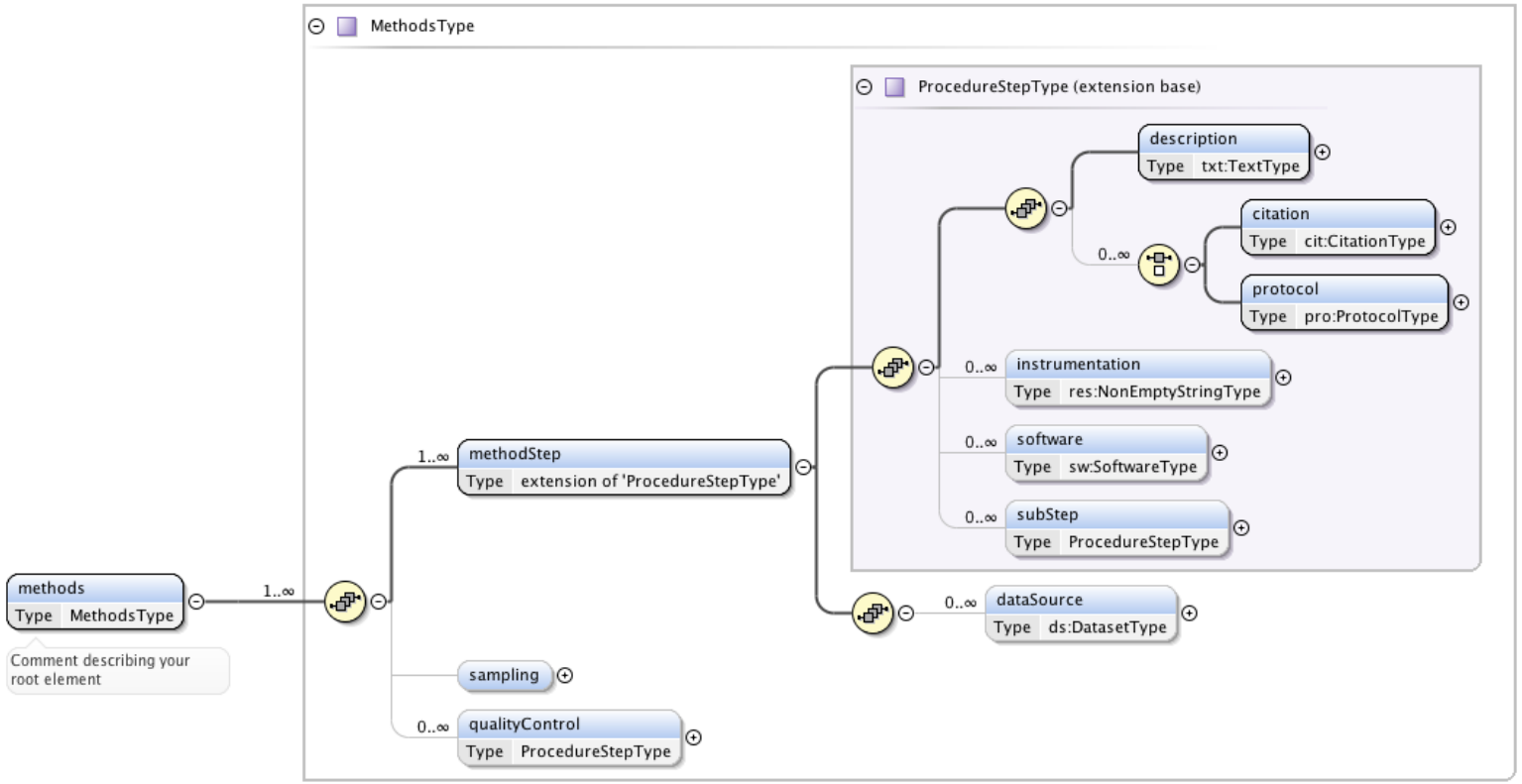
- It's the "P" in "PASTA“
- Critical for documentation and understanding of synthesis/derived products
- Innovative use of the "methods" section of EML to document provenance
- Provenance Factory generates provenance block in EML metadata

Provenance Chaining



Provenance Factory...

Generates provenance metadata for use in PASTA dependent derived products



Provenance in PASTA

Geographic Coverage	
Geographic Description:	North Inlet encompasses about 2,630 hectares of tidal marshes and wetlands near Georgetown, South Carolina, USA. North Inlet-Winyah Bay features high quality, ocean-dominated waters and salt marshes in North Inlet, contrasting with the brackish waters and marshes of Winyah Bay. The bay estuary is dominated by riverine discharges from a watershed impacted by agricultural, municipal and industrial development. Former rice fields and canals provide another system for study within the Reserve. The Debidue site is located at the confluence of Town Creek and Debidue Creek. The Bread and Butter site is located along the western shoreline of Town Creek adjacent to the mouth of Clambank Creek. Oyster Landing in Crab Haul creek within the NORTH INLET ESTUARY SYSTEM Georgetown, South Carolina. 33,20 lat. 79,11 long.
Bounding Coordinates:	-79.2936 W, -79.1042 E, 33.357 N, 33.2125 S
Geographic Coverage	
Geographic Description:	Oyster Landing in Crab Haul Creek 33.21'2" Lat., 79.11'27" Long.
Bounding Coordinates:	-79.1175 W, -79.1175 E, 33.2106 N, 33.2106 S
Methods	
Method Step 1 :	Computational methods The dew point values are computed based on the daily mean relative humidity and the current temperature using the following equation: Td = (b*gamma(T,RH))/(a-gamma(T,RH)), where gamma(T,RH)=(a*T)/(b+T)+LN(RH/100) and a=17.271 and b=237.7degC
Method Step 2 :	**The following data package was used in the creation of this product:** National Weather Service data for North Inlet Estuary, South Carolina, from 1986 to 1992, North Inlet LTER (Click here to view metadata)

Four Key Practical Innovations

INTEROPERABILITY

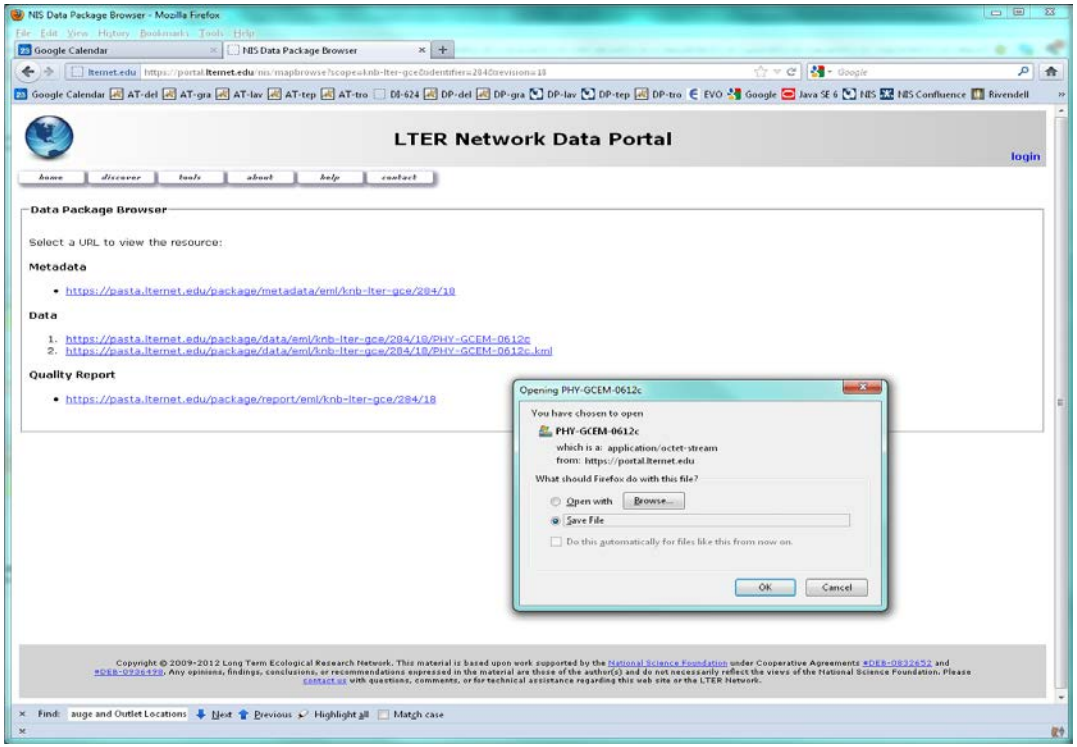
PROVENANCE

ACCESS

QUALITY

Access

- All NIS data packages link to data
  - A *quality check* guarantees this (more on this later)
- Local data storage
  - Snapshot of data entities at the time the data package was submitted by the LTER site
  - Consistent and reliable storage and retrieval



Four Key Practical Innovations

INTEROPERABILITY

PROVENANCE

ACCESS

QUALITY

Quality

- Could deliver a talk titled "Quality, Quality, Quality, and More Quality"
- LTER EML Metrics Working Group
  - ASM 2009 to present
  - Very active and productive group; includes DataONE participants
- NIS Tiger Teams
  - Data Manager, Data Package Manager, Metadata Quality

Quality Engine

- A subcomponent of the Data Package Manager
- Generates a quality report for each data package
- A quality report contains a set of quality checks
- Stored as XML but rendered in HTML for human readability
- 19 quality checks implemented in the NIS prototype
- 50+ quality checks documented by EML Metrics Working Group and Metadata Quality Tiger Team
- Quality Engine is available to the greater ecoinformatics community via the Data Manager Library (ecoinformatics.org)

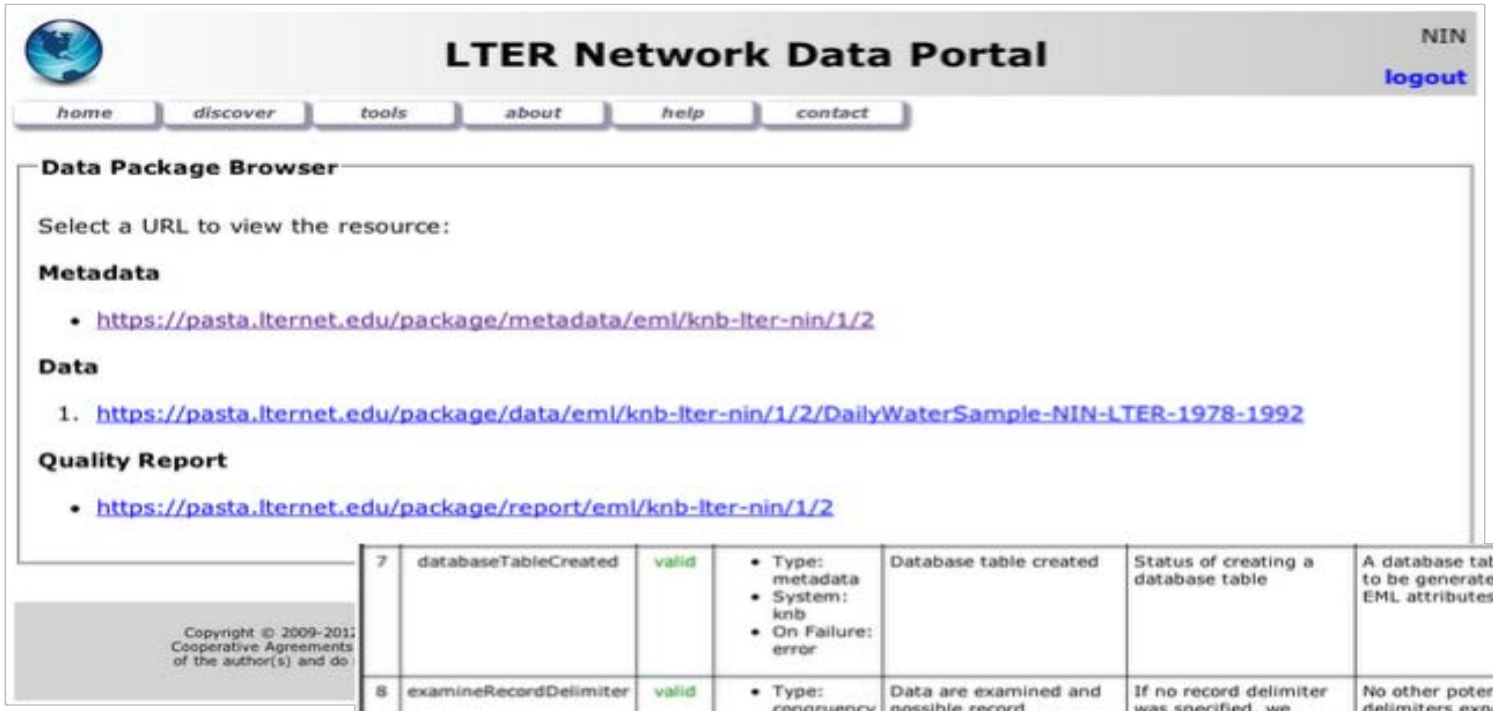
What's a Quality Check?

- An individual metric or a best practice
- It may involve looking at:
  - metadata (independent of data), *or*
  - data (independent of metadata), *or*
  - congruency between metadata and data
- Can result in one of four statuses
  - valid
  - info
  - warn
  - error

How is the Quality Engine used in the NIS?

- Used as a litmus test
  - Any **error** status reported by a quality check blocks insertion of the data package into PASTA
- Users can evaluate data packages before inserting them into PASTA
  - Key idea contributed by the Data Manager Tiger Team
  - Helps site Information Managers prepare their data packages for insertion
- Quality report is a resource of the data package
  - Persists and can be accessed alongside metadata and data resources

Data Package Resource Map and Quality Report



7	databaseTableCreated	valid	• Type: metadata • System: knb • On Failure: error	Database table created	Status of creating a database table	A database table is expected to be generated from the EML attributes.	A database table wa generated from the description
8	examineRecordDelimiter	valid	• Type: congruency • System: knb • On Failure: warn	Data are examined and possible record delimiters are displayed	If no record delimiter was specified, we assume that \r\n is the delimiter. Search the first row for other record delimiters and see if other delimiters are found.	No other potential record delimiters expected in the first row.	No other potential re delimiters were dete valid record delimite previously detected
9	tooManyFields	error	• Type: congruency • System: knb • On Failure: error	Data does not have more fields than metadata attributes	Compare number of fields specified in metadata to number of fields found in a data record	15 fields	17 fields
10	dataLoadStatus	warn	• Type: congruency • System: knb • On Failure:	Data can be loaded into the database	Status of loading the data table into a database	No errors expected during data loading or data loading was not attempted for this data entity	One or more errors during data loading

LTER Network Office

Four Key Practical Innovations

LTER NIS is now poised to utilize these four key practical innovations as enabled by PASTA

INTEROPERABILITY

PROVENANCE

ACCESS

QUALITY