Data packages and EML Data Manager

Ben Leinfelder[1]

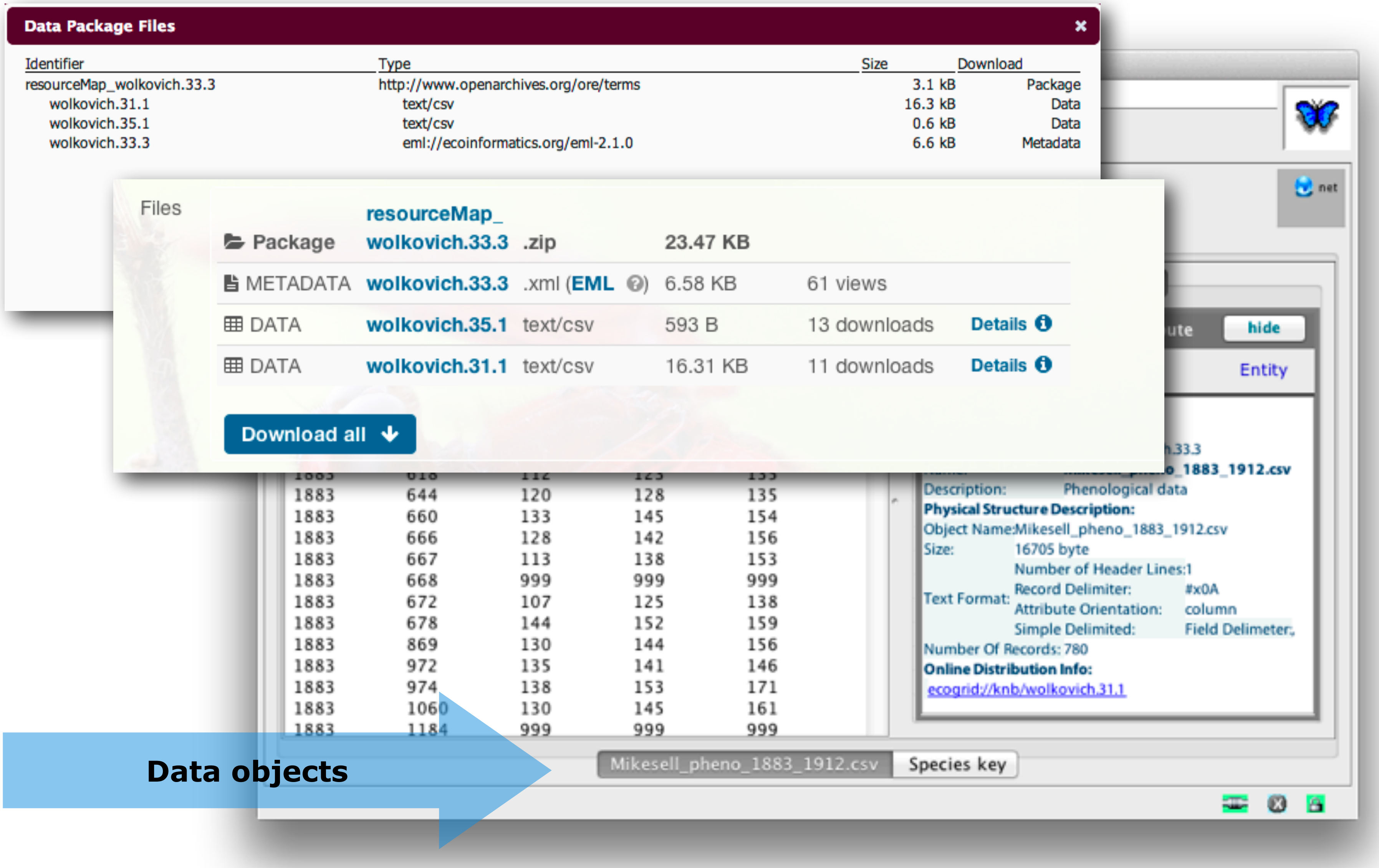
Jing Tao[1], Duane Costa[2], Matthew B. Jones[1], Mark Servilla[2],
Margaret O'Brien[3], Chad Burt[3]

[1] *National Center for Ecological Analysis and Synthesis, University of California Santa Barbara*
[2] *Long Term Ecological Research Network, University of New Mexico*
[3] *Santa Barbara Coastal LTER, University of California Santa Barbara*

Originally @ EIM 2008
September 10th, 2008

DATA PACKAGES

Data packages



Data Package Files ✖

Identifier	Type	Size	Download
resourceMap_wolkovich.33.3	http://www.openarchives.org/ore/terms	3.1 kB	Package
wolkovich.31.1	text/csv	16.3 kB	Data
wolkovich.35.1	text/csv	0.6 kB	Data
wolkovich.33.3	eml://ecoinformatics.org/eml-2.1.0	6.6 kB	Metadata

Files

Package	resourceMap_ wolkovich.33.3	.zip	23.47 KB		
METADATA	wolkovich.33.3	.xml (EML ?)	6.58 KB	61 views	
DATA	wolkovich.35.1	text/csv	593 B	13 downloads	Details ℹ
DATA	wolkovich.31.1	text/csv	16.31 KB	11 downloads	Details ℹ

Download all ↓

1883	618	112	123	133
1883	644	120	128	135
1883	660	133	145	154
1883	666	128	142	156
1883	667	113	138	153
1883	668	999	999	999
1883	672	107	125	138
1883	678	144	152	159
1883	869	130	144	156
1883	972	135	141	146
1883	974	138	153	171
1883	1060	130	145	161
1883	1184	999	999	999

.33.3

Name: ...o_1883_1912.csv

Description: Phenological data

Physical Structure Description:

Object Name:Mikesell_pheno_1883_1912.csv

Size: 16705 byte

Number of Header Lines:1

Text Format: Record Delimiter: #x0A

Attribute Orientation: column

Simple Delimited: Field Delimeter:,

Number Of Records: 780

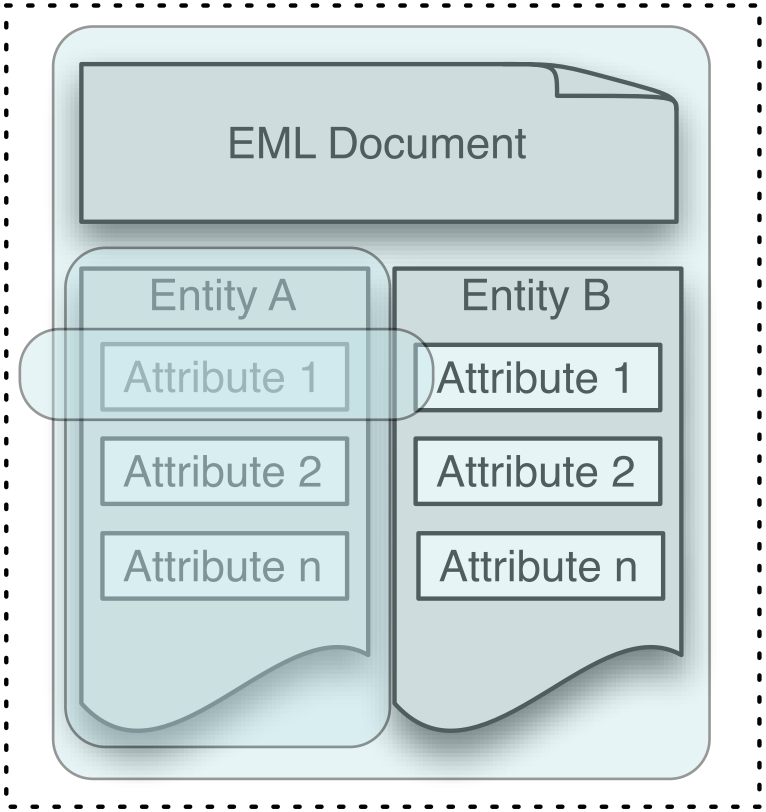
Online Distribution Info:

ecogrid://knb/wolkovich.31.1

Data objects

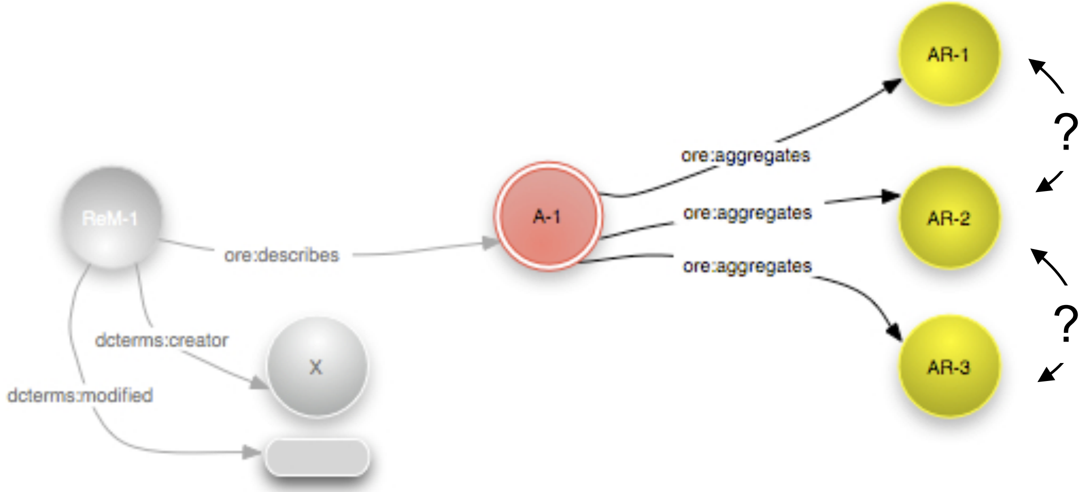
EML Data Model

- Data Package (eml-dataset)
  - collection of data Entities
- Entity (eml-dataTable)
  - tabular data
  - other
- Attribute
  - data column



OAI-ORE packages

- Resource map ***describes*** an aggregation
  - aggregation ***aggregates*** objects
- metadata **<--?-->** data

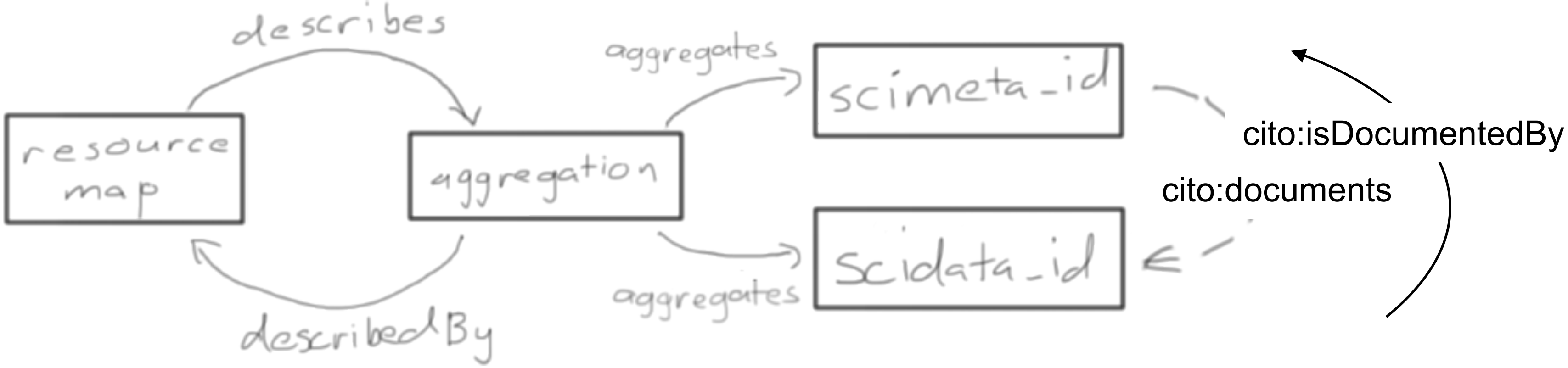


<A-1>	<ore:aggregates>	<AR-1>
<A-1>	<ore:aggregates>	<AR-2>
<A-1>	<ore:aggregates>	<AR-3>

A-1	http://www.dlib.org/dlib/february06/smith/aggregation
AR-1	http://www.dlib.org/dlib/february06/smith/02smith.html
AR-2	http://www.dlib.org/dlib/february06/smith/pg1-13.html
AR-3	http://www.dlib.org/dlib/february06/smith/pg1-13.pdf

DataONE packages

- Add relationships
  - metadata ***documents*** data
  - data ***isDocumentedBy*** metadata



cito:isDocumentedBy

cito:documents

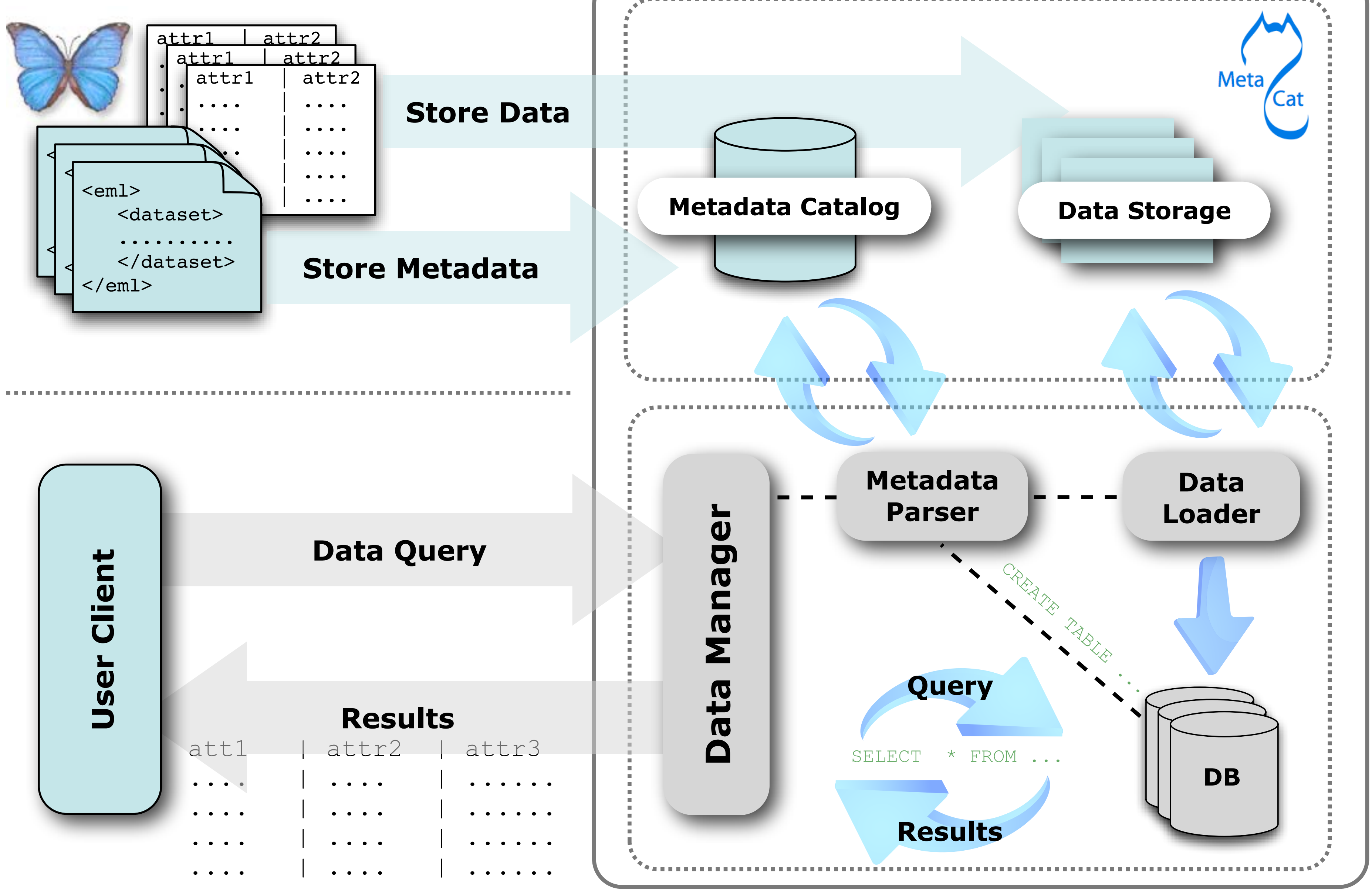
EML DATA MANAGER

Data Manager Library

“Provide metadata-based query access to data with the ability to filter, join, and concatenate across data sets.”

NCEAS

Dynamic Data Retrieval



```
attr1 | attr2
....  | ....
....  | ....
....  | ....
....  | ....
....  | ....
```

```
<eml>
    <dataset>
    ..........
    </dataset>
</eml>
```

Store Data

Store Metadata

Meta Cat

Metadata Catalog

Data Storage

User Client

Data Query

Results

```
att1    | attr2    | attr3
....    | ....     | .......
....    | ....     | .......
....    | ....     | .......
....    | ....     | .......
```

Data Manager

Metadata Parser

Data Loader

CREATE TABLE ...

Query

SELECT * FROM ...

Results

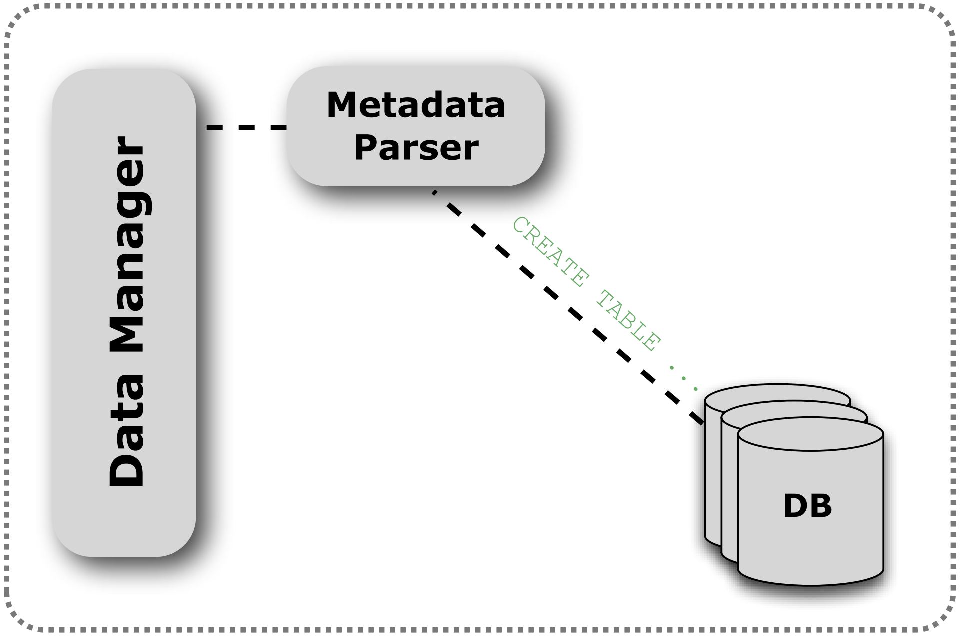
DB

Data Manager Library

- Uses EML metadata to:
  - Automatically create database tables
  - Download data from remote sites
  - Load data into the database
  - Manage table space via caching
- Client applications can:
  - Inspect table structures
  - Pose SQL-like queries
  - Join and concatenate data packages

Parse and Create Tables

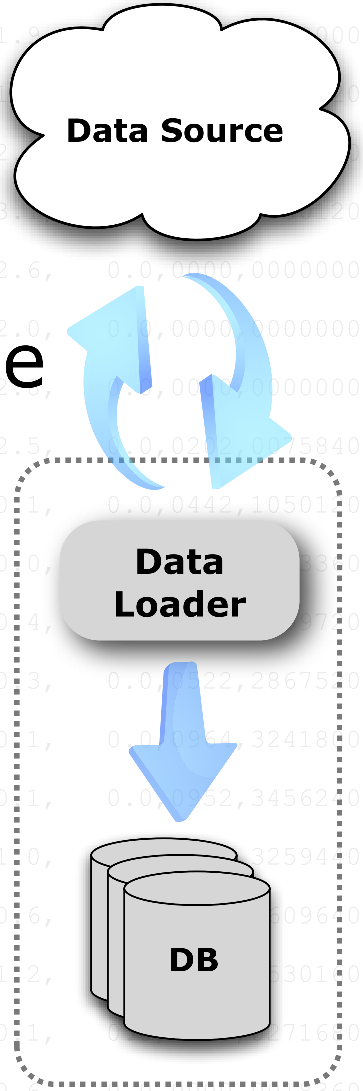
- Create underlying database table
  - Schema derived from EML
- Well-described attributes
  - type
  - precision
  - range
  - format



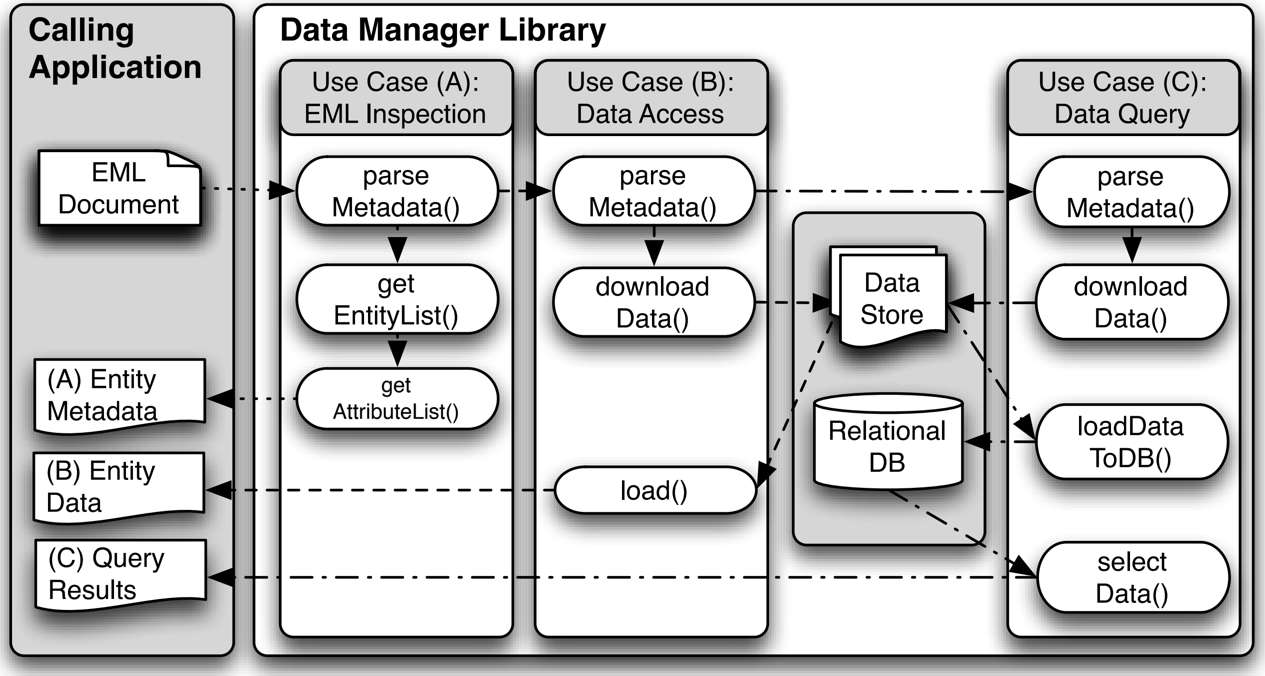
Download and Stage Data

- Retrieve data from source
  - Metacat with web services
  - Other external [accessible] source
- Insert data rows
  - Date formatting

Date: 09/10/2008
September 10th?
October 9th?



Use Cases

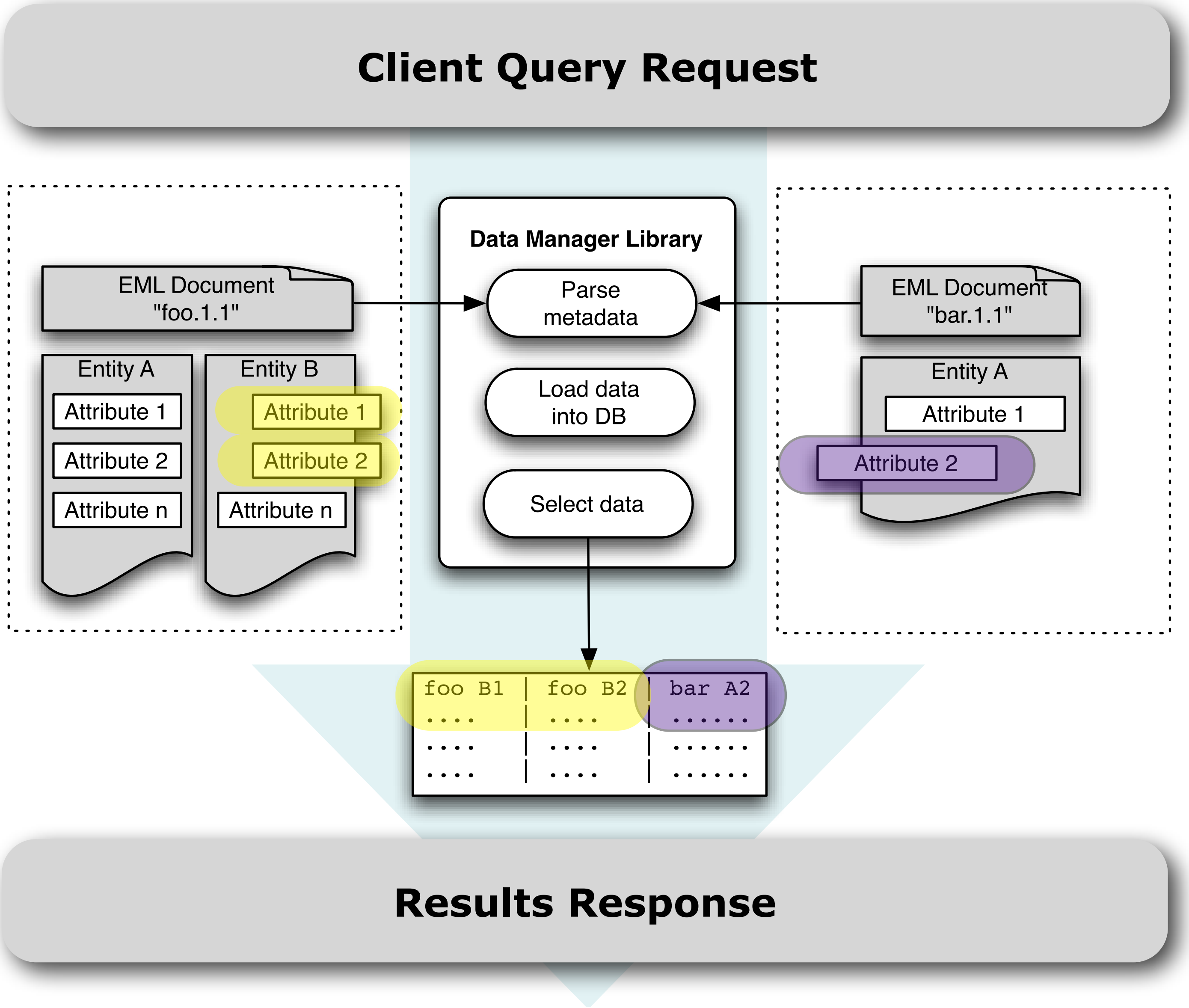


(A) Inspect metadata

(B) Download data files

(C) Query and join tabular data files

Join Query



Wednesday, August 27, 14

Data Query Specification

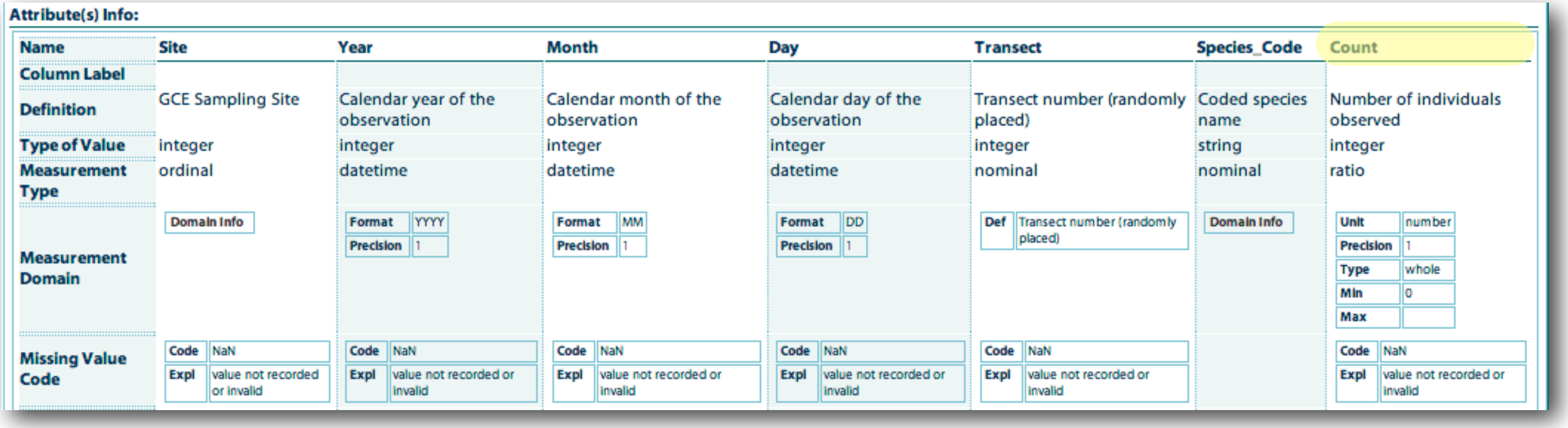
- XML syntax for describing queries

```xml
<?xml version="1.0" encoding="UTF-8"?>
<dataquery>
   <query>
      <selection>
         <datapackage id="tao.1.1">
            <entity index="0">
               <attribute index="0"/>
               <attribute index="1"/>
            </entity>
         </datapackage>
      </selection>
      <where>
         <condition type="condition">
            <left>
               <datapackage id="tao.1.1">
                  <entity index="0">
                     <attribute index="0"/>
                  </entity>
               </datapackage>
            </left>
            <operator>=</operator>
            <right>
               <value>11/12/2008</value>
            </right>
         </condition>
      </where>
   </query>
</dataquery>
```

- Alternative to Java API
  - ✓ Attribute selection
  - ✓ Joins
  - ✓ Conditions
  - ✓ Subqueries
  - ✓ Metadata promotion

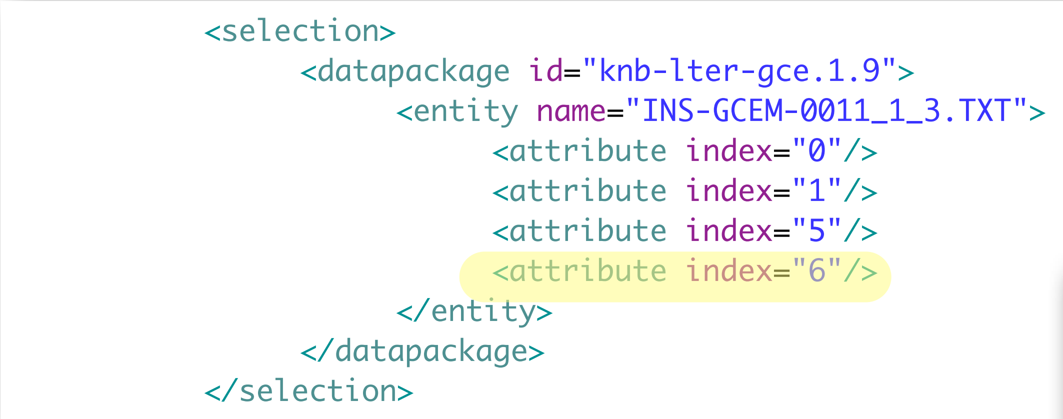
```sql
SELECT  Datos_Meteorologicos.DATE,
        Datos_Meteorologicos.TIME
FROM Datos_Meteorologicos
WHERE Datos_Meteorologicos.DATE = '11/12/2008';
```

Data Query Specification

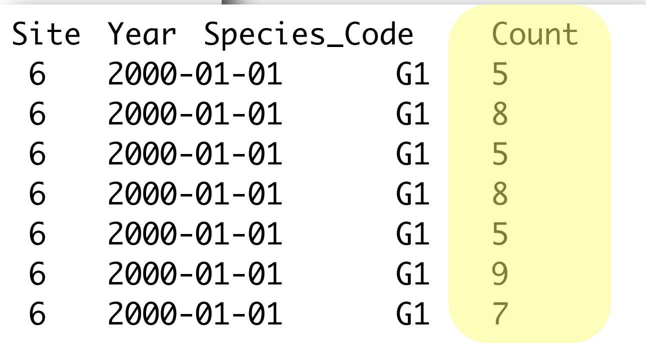


Attribute(s) Info:

Name	Site	Year	Month	Day	Transect	Species_Code	Count
Column Label							
Definition	GCE Sampling Site	Calendar year of the observation	Calendar month of the observation	Calendar day of the observation	Transect number (randomly placed)	Coded species name	Number of individuals observed
Type of Value	integer	integer	integer	integer	integer	string	integer
Measurement Type	ordinal	datetime	datetime	datetime	nominal	nominal	ratio
Measurement Domain	Domain Info	Format YYYY Precision 1	Format MM Precision 1	Format DD Precision 1	Def Transect number (randomly placed)	Domain Info	Unit number Precision 1 Type whole Min 0 Max
Missing Value Code	Code NaN Expl value not recorded or invalid	Code NaN Expl value not recorded or invalid	Code NaN Expl value not recorded or invalid	Code NaN Expl value not recorded or invalid	Code NaN Expl value not recorded or invalid		Code NaN Expl value not recorded or invalid



```
<selection>
     <datapackage id="knb-lter-gce.1.9">
          <entity name="INS-GCEM-0011_1_3.TXT">
               <attribute index="0"/>
               <attribute index="1"/>
               <attribute index="5"/>
               <attribute index="6"/>
          </entity>
     </datapackage>
</selection>
```

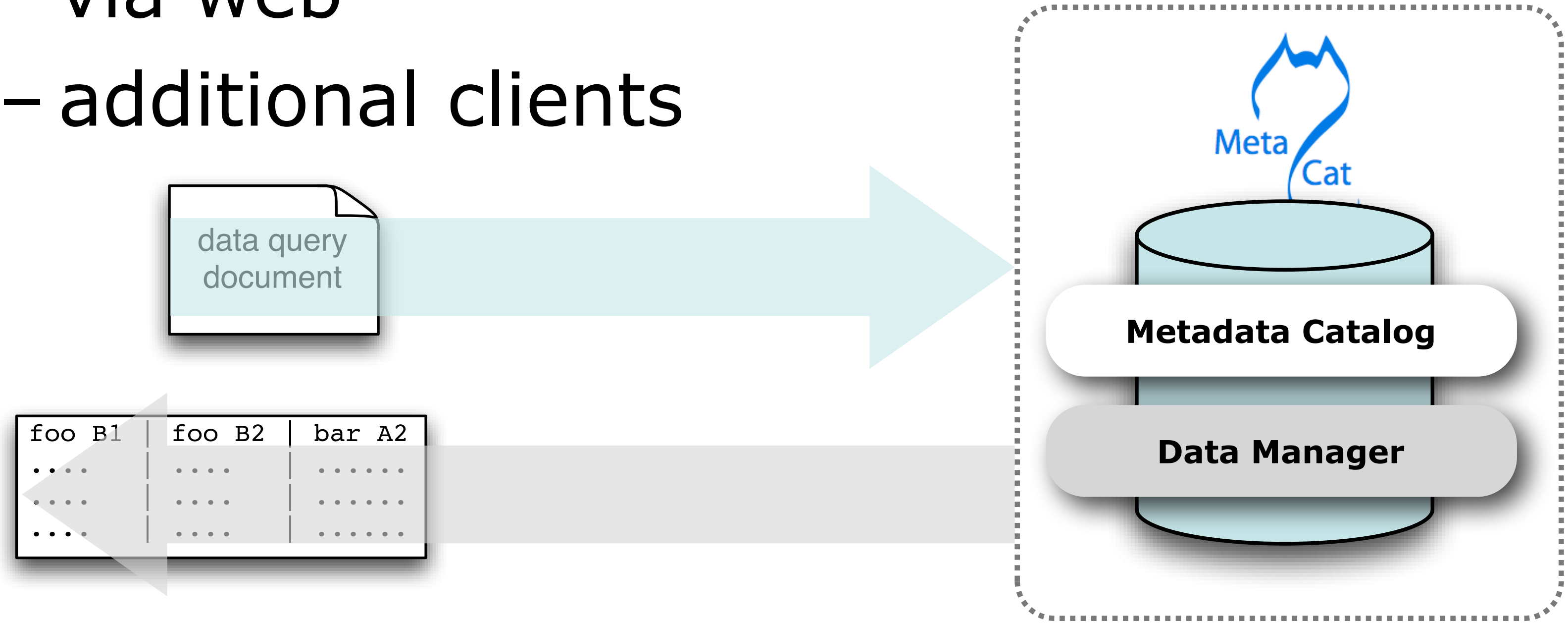


```
Site  Year  Species_Code    Count
 6    2000-01-01      G1    5
 6    2000-01-01      G1    8
 6    2000-01-01      G1    5
 6    2000-01-01      G1    8
 6    2000-01-01      G1    5
 6    2000-01-01      G1    9
 6    2000-01-01      G1    7
```

Metacat Integration

- Embed Data Manager
  - Eliminate individual client deployments

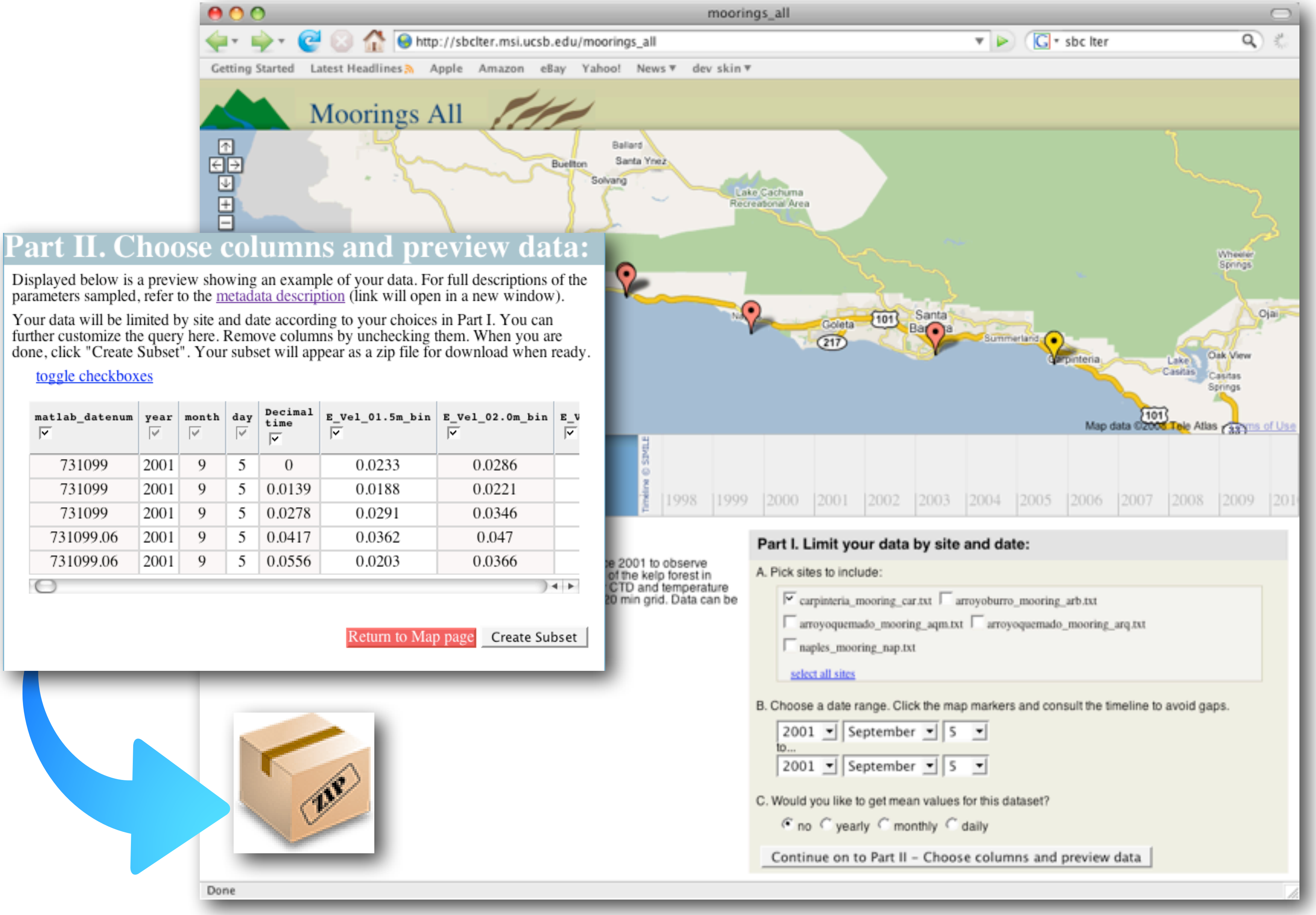
- Expose data query capabilities
  - via web
  - additional clients



foo B1	foo B2	bar A2
....	....	......
....	....	......
....	....	......

Santa Barbara Coastal LTER

Temporal and location queries



Part II. Choose columns and preview data:

Displayed below is a preview showing an example of your data. For full descriptions of the parameters sampled, refer to the metadata description (link will open in a new window).

Your data will be limited by site and date according to your choices in Part I. You can further customize the query here. Remove columns by unchecking them. When you are done, click "Create Subset". Your subset will appear as a zip file for download when ready.

toggle checkboxes

matlab_datenum ☑	year ☑	month ☑	day ☑	Decimal time ☑	E_Vel_01.5m_bin ☑	E_Vel_02.0m_bin ☑	E_V ☑
731099	2001	9	5	0	0.0233	0.0286	
731099	2001	9	5	0.0139	0.0188	0.0221	
731099	2001	9	5	0.0278	0.0291	0.0346	
731099.06	2001	9	5	0.0417	0.0362	0.047	
731099.06	2001	9	5	0.0556	0.0203	0.0366	

Return to Map page | Create Subset

e 2001 to observe of the kelp forest in CTD and temperature 20 min grid. Data can be

Part I. Limit your data by site and date:

A. Pick sites to include:

☑ carpinteria_mooring_car.txt ☐ arroyoburro_mooring_arb.txt
☐ arroyoquemado_mooring_aqm.txt ☐ arroyoquemado_mooring_arq.txt
☐ naples_mooring_nap.txt

select all sites

B. Choose a date range. Click the map markers and consult the timeline to avoid gaps.

2001 September 5
to...
2001 September 5

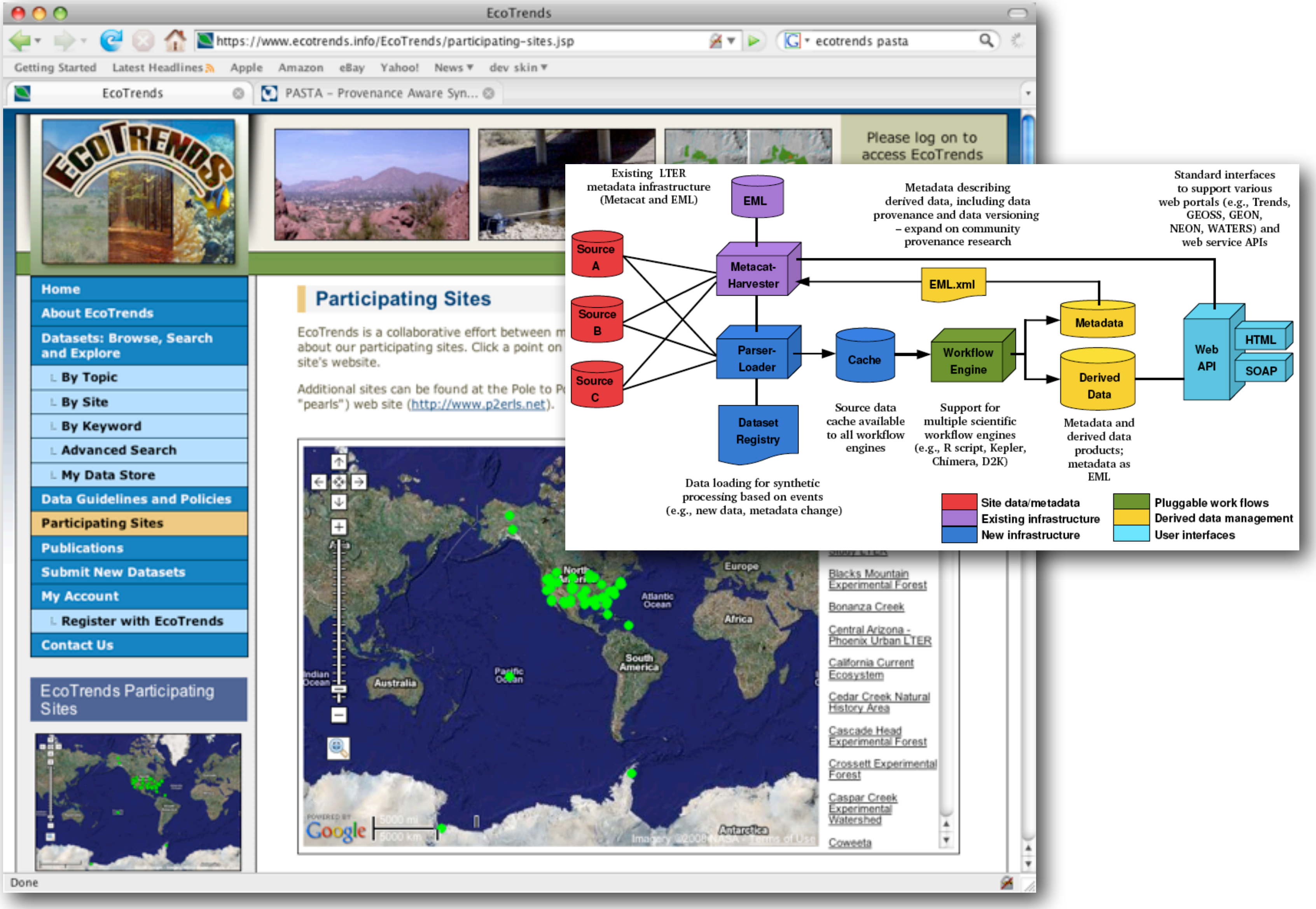
C. Would you like to get mean values for this dataset?

● no ○ yearly ○ monthly ○ daily

Continue on to Part II – Choose columns and preview data

LTER “PASTA”

Provenance Aware Synthesis Tracking Architecture



Participating Sites

EcoTrends is a collaborative effort between m
about our participating sites. Click a point on
site's website.

Additional sites can be found at the Pole to P
"pearls") web site (http://www.p2erls.net).

Study LTER
Blacks Mountain Experimental Forest
Bonanza Creek
Central Arizona - Phoenix Urban LTER
California Current Ecosystem
Cedar Creek Natural History Area
Cascade Head Experimental Forest
Crossett Experimental Forest
Caspar Creek Experimental Watershed
Coweeta

Existing LTER metadata infrastructure (Metacat and EML)

Metadata describing derived data, including data provenance and data versioning – expand on community provenance research

Standard interfaces to support various web portals (e.g., Trends, GEOSS, GEON, NEON, WATERS) and web service APIs

Source data cache available to all workflow engines

Support for multiple scientific workflow engines (e.g., R script, Kepler, Chimera, D2K)

Metadata and derived data products; metadata as EML

Data loading for synthetic processing based on events (e.g., new data, metadata change)

Wednesday, August 27, 14

QUALITY CHECKS

EML Quality Reporting

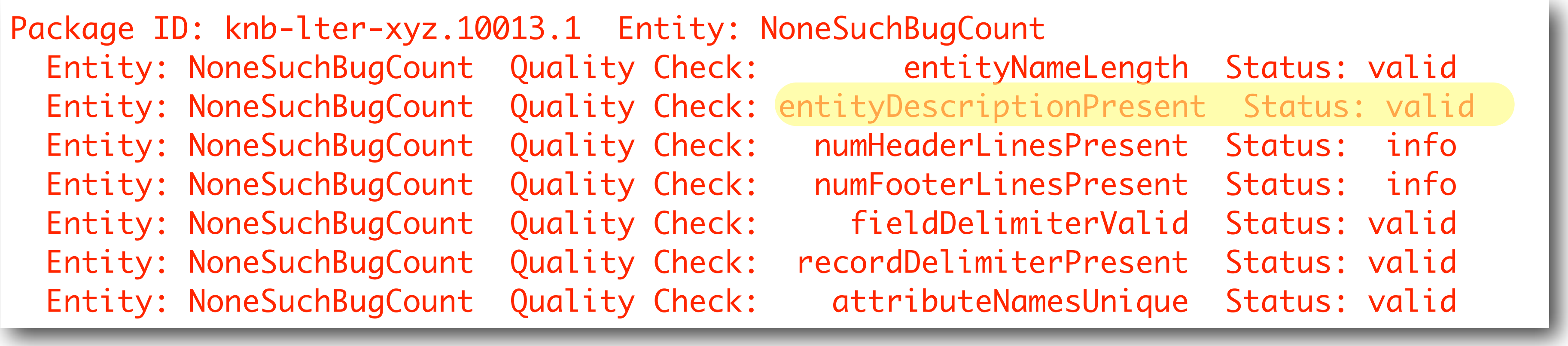
- Verify metadata fields
  - Title (length)
  - Publication date
  - Keywords present
  - Coverage elements present

- Data files
  - Accessible/downloadable
  - Attribute metadata match data (types)

EML Quality Reporting

- Check for optional fields in metadata

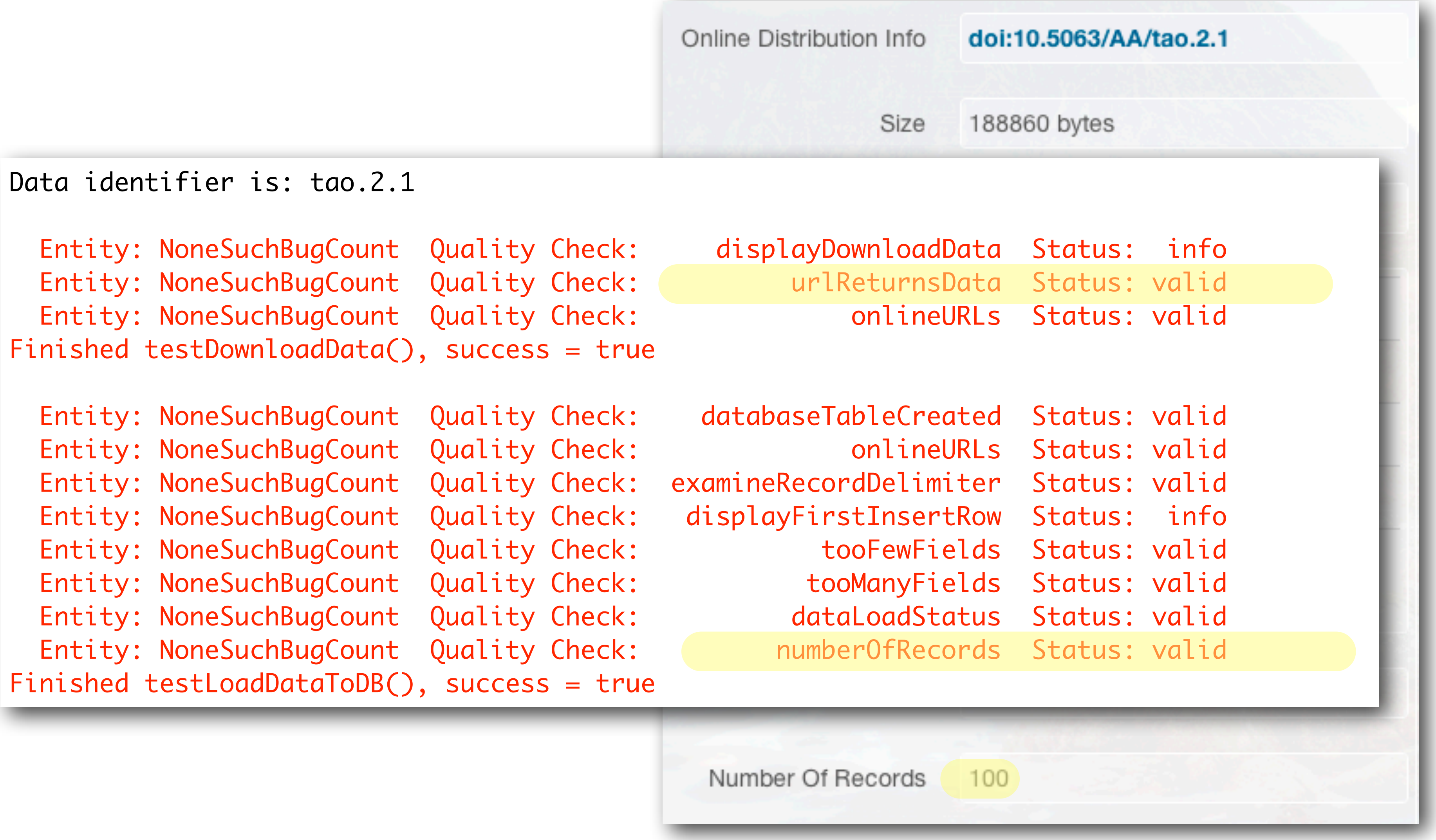
```
<qualityCheck qualityType="metadata" system="lter" statusType="warn" >
      <identifier>entityDescriptionPresent</identifier>
      <name>An entity description is present</name>
      <description>Check for presence of an entity description.</description>
      <expected>Field should have a data file description</expected>
      ...
    </qualityCheck>
```



```
Package ID: knb-lter-xyz.10013.1  Entity: NoneSuchBugCount
  Entity: NoneSuchBugCount  Quality Check:        entityNameLength  Status: valid
  Entity: NoneSuchBugCount  Quality Check: entityDescriptionPresent  Status: valid
  Entity: NoneSuchBugCount  Quality Check:    numHeaderLinesPresent  Status:  info
  Entity: NoneSuchBugCount  Quality Check:    numFooterLinesPresent  Status:  info
  Entity: NoneSuchBugCount  Quality Check:      fieldDelimiterValid  Status: valid
  Entity: NoneSuchBugCount  Quality Check:  recordDelimiterPresent  Status: valid
  Entity: NoneSuchBugCount  Quality Check:     attributeNamesUnique  Status: valid
```

EML Quality Reporting

- Verify data matches metadata



Online Distribution Info doi:10.5063/AA/tao.2.1

Size 188860 bytes

```
Data identifier is: tao.2.1

  Entity: NoneSuchBugCount  Quality Check:    displayDownloadData  Status:  info
  Entity: NoneSuchBugCount  Quality Check:         urlReturnsData  Status: valid
  Entity: NoneSuchBugCount  Quality Check:             onlineURLs  Status: valid
Finished testDownloadData(), success = true

  Entity: NoneSuchBugCount  Quality Check:   databaseTableCreated  Status: valid
  Entity: NoneSuchBugCount  Quality Check:             onlineURLs  Status: valid
  Entity: NoneSuchBugCount  Quality Check: examineRecordDelimiter  Status: valid
  Entity: NoneSuchBugCount  Quality Check:    displayFirstInsertRow  Status:  info
  Entity: NoneSuchBugCount  Quality Check:           tooFewFields  Status: valid
  Entity: NoneSuchBugCount  Quality Check:          tooManyFields  Status: valid
  Entity: NoneSuchBugCount  Quality Check:         dataLoadStatus  Status: valid
  Entity: NoneSuchBugCount  Quality Check:        numberOfRecords  Status: valid
Finished testLoadDataToDB(), success = true
```

Number Of Records 100

INTERNATIONALIZATION

Crucial Metadata

- Important for
  - Discovery
  - Interpretation
  - Attribution

Title

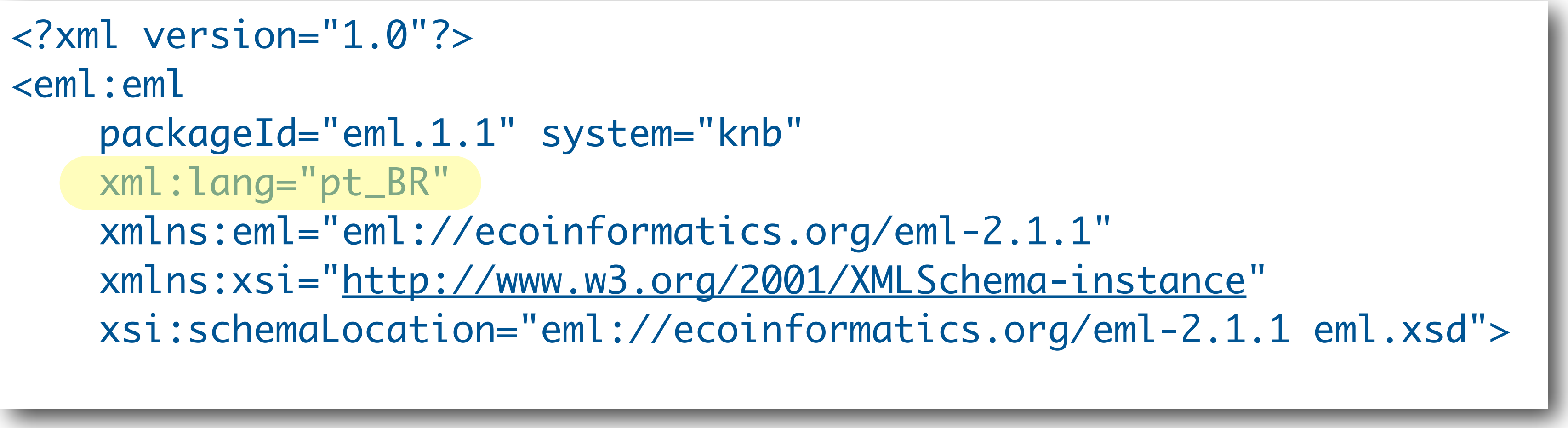
Keyword

Contact information

(names, organizations, addresses)

EML 2.1.1 documents

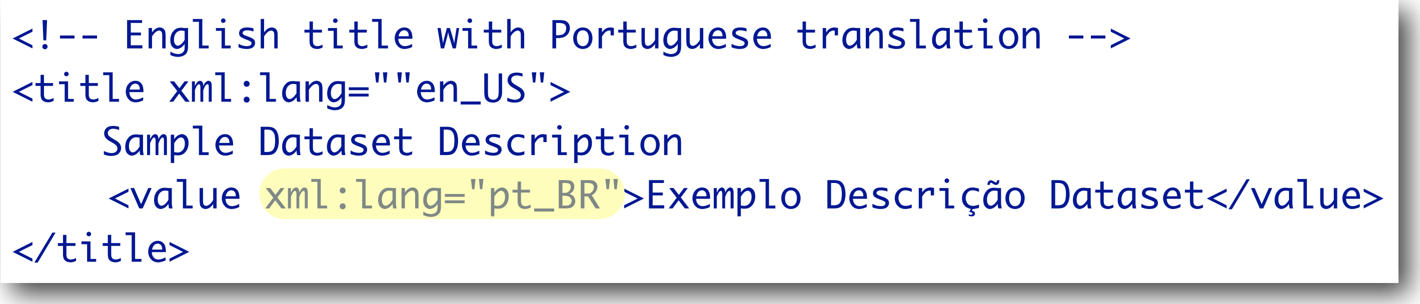
- Mark entire metadata record
  - all elements inherit Portuguese



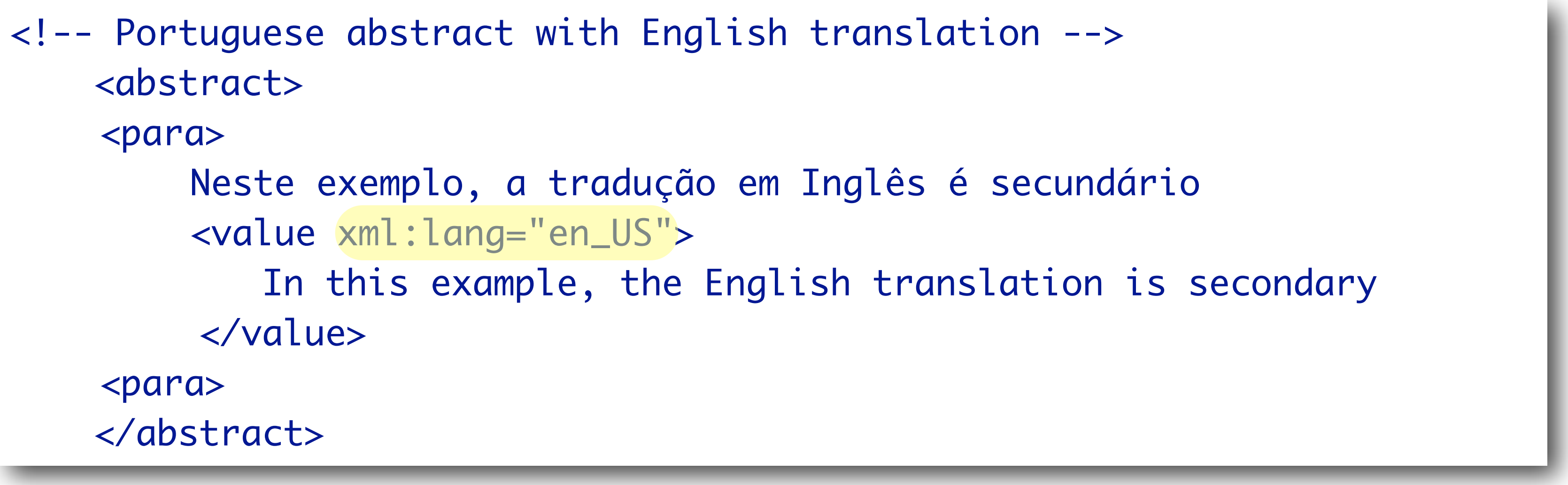
```
<?xml version="1.0"?>
<eml:eml
    packageId="eml.1.1" system="knb"
    xml:lang="pt_BR"
    xmlns:eml="eml://ecoinformatics.org/eml-2.1.1"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1 eml.xsd">
```

EML 2.1.1 elements

- Provide translations with original

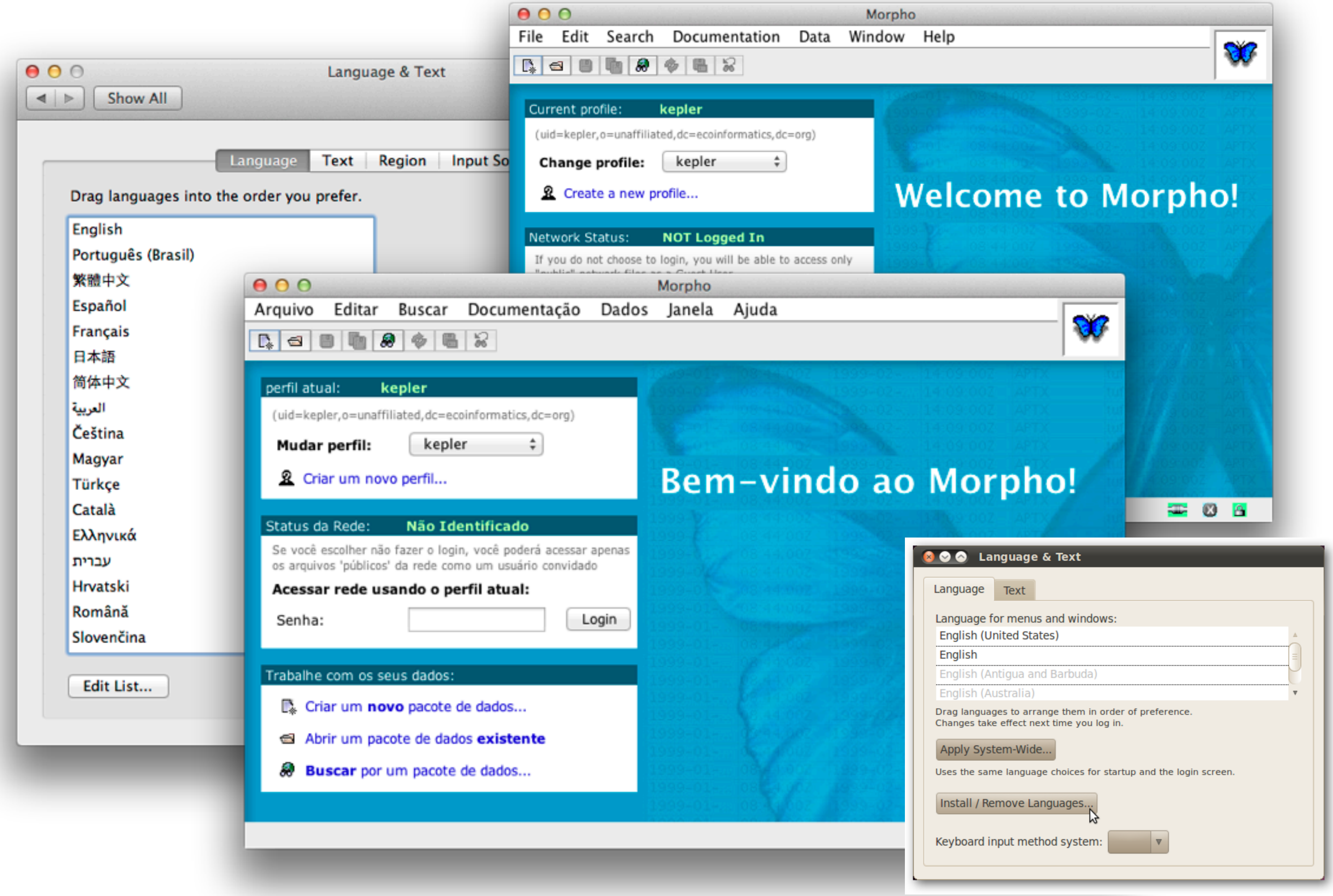


```
<!-- English title with Portuguese translation -->
<title xml:lang=""en_US">
    Sample Dataset Description
    <value xml:lang="pt_BR">Exemplo Descrição Dataset</value>
</title>
```



```
<!-- Portuguese abstract with English translation -->
    <abstract>
    <para>
        Neste exemplo, a tradução em Inglês é secundário
        <value xml:lang="en_US">
            In this example, the English translation is secondary
        </value>
    <para>
    </abstract>
```

Tool Support



Language & Text

Show All

Language | Text | Region | Input So

Drag languages into the order you prefer.

English
Português (Brasil)
繁體中文
Español
Français
日本語
简体中文
العربية
Čeština
Magyar
Türkçe
Català
Ελληνικά
עברית
Hrvatski
Română
Slovenčina

Edit List...

Morpho

File Edit Search Documentation Data Window Help

Current profile: kepler

(uid=kepler,o=unaffiliated,dc=ecoinformatics,dc=org)

Change profile: kepler

Create a new profile...

Welcome to Morpho!

Network Status: NOT Logged In

If you do not choose to login, you will be able to access only

Morpho

Arquivo Editar Buscar Documentação Dados Janela Ajuda

perfil atual: kepler

(uid=kepler,o=unaffiliated,dc=ecoinformatics,dc=org)

Mudar perfil: kepler

Criar um novo perfil...

Bem-vindo ao Morpho!

Status da Rede: Não Identificado

Se você escolher não fazer o login, você poderá acessar apenas os arquivos 'públicos' da rede como um usuário convidado

Acessar rede usando o perfil atual:

Senha: Login

Trabalhe com os seus dados:

Criar um **novo** pacote de dados...

Abrir um pacote de dados **existente**

Buscar por um pacote de dados...

Language & Text

Language | Text

Language for menus and windows:

English (United States)
English
English (Antigua and Barbuda)
English (Australia)

Drag languages to arrange them in order of preference.
Changes take effect next time you log in.

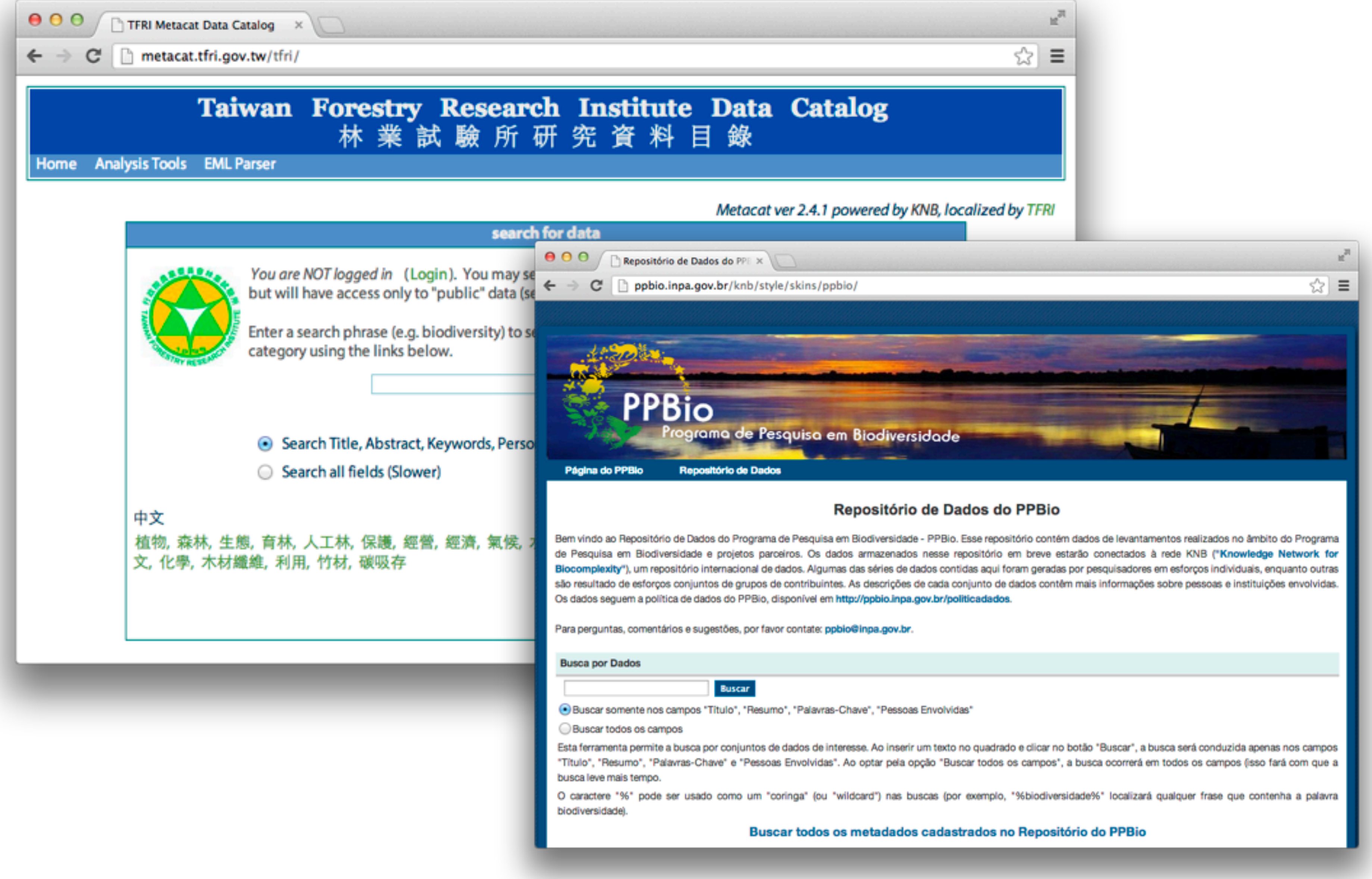
Apply System-Wide...

Uses the same language choices for startup and the login screen.

Install / Remove Languages...

Keyboard input method system:

Tool Support



Taiwan Forestry Research Institute Data Catalog
林業試驗所研究資料目錄

Metacat ver 2.4.1 powered by KNB, localized by TFRI

search for data

You are NOT logged in (Login). You may se
but will have access only to "public" data (se

Enter a search phrase (e.g. biodiversity) to se
category using the links below.

◉ Search Title, Abstract, Keywords, Perso
○ Search all fields (Slower)

中文
植物, 森林, 生態, 育林, 人工林, 保護, 經營, 經濟, 氣候,
文, 化學, 木材纖維, 利用, 竹材, 碳吸存

Repositório de Dados do PPE
ppbio.inpa.gov.br/knb/style/skins/ppbio/
PPBio
Programa de Pesquisa em Biodiversidade
Página do PPBio
Repositório de Dados

Repositório de Dados do PPBio

Bem vindo ao Repositório de Dados do Programa de Pesquisa em Biodiversidade - PPBio. Esse repositório contém dados de levantamentos realizados no âmbito do Programa de Pesquisa em Biodiversidade e projetos parceiros. Os dados armazenados nesse repositório em breve estarão conectados à rede KNB ("Knowledge Network for Biocomplexity"), um repositório internacional de dados. Algumas das séries de dados contidas aqui foram geradas por pesquisadores em esforços individuais, enquanto outras são resultado de esforços conjuntos de grupos de contribuintes. As descrições de cada conjunto de dados contêm mais informações sobre pessoas e instituições envolvidas. Os dados seguem a política de dados do PPBio, disponível em http://ppbio.inpa.gov.br/politicadados.

Para perguntas, comentários e sugestões, por favor contate: ppbio@inpa.gov.br.

Busca por Dados

◉ Buscar somente nos campos "Título", "Resumo", "Palavras-Chave", "Pessoas Envolvidas"
○ Buscar todos os campos

Esta ferramenta permite a busca por conjuntos de dados de interesse. Ao inserir um texto no quadrado e clicar no botão "Buscar", a busca será conduzida apenas nos campos "Título", "Resumo", "Palavras-Chave" e "Pessoas Envolvidas". Ao optar pela opção "Buscar todos os campos", a busca ocorrerá em todos os campos (isso fará com que a busca leve mais tempo.

O caractere "%" pode ser usado como um "coringa" (ou "wildcard") nas buscas (por exemplo, "%biodiversidade%" localizará qualquer frase que contenha a palavra biodiversidade).

Buscar todos os metadados cadastrados no Repositório do PPBio

Acknowledgements

- This material is based upon work supported by:
  - The National Science Foundation under Grant Numbers 9980154 (KDI), 0618501 (FIRST) and 0225676 (SEEK).
  - The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.
  - The Andrew W. Mellon Foundation.

- Resources
  - http://www.nceas.ucsb.edu/ecoinfo/
  - https://knb.ecoinformatics.org/
  - http://lno.lternet.edu/projects/pasta
  - http://sbc.lternet.edu/