Carga de Datos en la *Data Library* del IRI

John del Corral
jdcorral@iri.columbia.edu
traducido del inglés por :
Rémi Cousin
remic@iri.columbia.edu

Septiembre 30, 2011

Resumen: Muchos usuarios de la *Data Library* son profesionales del sector público (salud, agricultura y manejo de agua) que tienen datos relacionados con sus sectores para sus regiones o países y desean correlacionar estos datos con datos del clima y del medio ambiente que están disponibles en la *Data Library*. Dentro de la Data Library existe una herramienta para cargar datos de diferentes sectores, lo único que se requiere es , que estos datos tengan un formato a nivel espacial y temporal, que se ajuste a la plataforma de *Data Library*.

Objetivos : Guiar al usuario por los diferentes pasos que son necesarios para cargar un archivo local desde su computador hasta la *Data Library* del IRI usando una interface en internet.

Perspectiva general sobre la *Data Library* del IRI : La *Data Library* es un servicio en la red del internet que sirve para visualizar, analizar y bajar datos geofísicos de forma gratuita . La biblioteca contiene más de 400 bases de datos con aproximadamente 300 terabytes de datos. Hay varias opciones de visualización, incluyendo la superposición de diferentes capas incluyendo distribuciones geográficas o administrativas de cada sector. La Data Library tiene diferentes funciones de análisis estadísticos. La capacidad de cargar datos dentro de esta herramienta permite correlacionar datos de usuarios con datos climáticos . Visíte la pagina de la Data Library aquí : http://iridl.ldeo.columbia.edu.

La carga de datos en la *Data Library* del IRI es un proceso complejo. No se trata simplemente de abrir un archivo con su programa preferido. El archivo de datos esta integrado, almacenado, en una base de datos, y los metadatos están almacenados en otra base de datos. Ambas bases de datos están leídas por el programa de la *Data Library* para que sean accesibles en la misma . **Se sugiere empezar con una muestra pequeña de sus datos, hasta estar más familiarizado con el proceso de carga de datos y ahí cargar todas las series de datos que requieran.**

Etapas para Cargar Datos en la *Data Library* :

I. Formatos del Archivo del Usuario
II. Formatos de los Identificadores Espaciales y Temporales
III. Formulario de Internet para la Carga
IV. Adicionar Metadatos
V. Definir Dimensiones Espaciales y Temporales
VI. Control de Calidad
VII. Uso en la Data Library

I. Formatos del Archivo del Usuario

La *Data Library* puede cargar varios formatos generalizados de archivos de datos (tabular y SIG). El archivo que se cargue está almacenado en una tabla de base de datos. Pues, todas las herramientas de visualización y de análisis de la *Data Library* pueden ser utilizadas con esta nueva base de datos. La base de datos aparece en el directorio 'home' del usuario en la lista de bases de datos de la *Data Library*, y puede ser protegida por una contraseña.

Formatos de archivos tabularios que se aceptan para cargar en la Data Library

Excel (.xls, no .xlsx)
Tab separated (.tsv)
Comma separated (.csv)

Guías para incluir archivos tabularios

-una línea sola para los nombres de las columnas
-ninguna línea vacía desde la línea de los nombres de las columnas hasta el fin de los datos
-cada línea debe tener un referente espacial y temporal
-una columna o más con identificadores espaciales (código o nombre administrativo, o latitud y longitud) los cuales tienen que corresponder con los identificadores espaciales en los archivos SIG del usuario, o los archivos SIG de la *Data Library*
-una columna o más con identificadores temporales (año, mes, semana, o día) en formato texto, ISO 8601 estandar o Fecha Excel.
-identificadores espaciales deben ser iguales durante toda la serie temporal de datos que se cargue
- no se debe incluir ningún comentario en el archivo
-no se debe incluir ningúna suma o totales en el archivo
-no se debe duplicar ninguna línea
-No deben existir espacios en blanco , sihayun dato faltante debe ser llenados con identificadores espaciales y temporales, como -999 como valores de omision
-Archivos Excel tienen un límite de ~60000 líneas. Bases de datos más grandes deben ser archivadas en un formato .tsv o .csv
-La carga de datos en el sistema solo admite archivos Excel con una tabla unica

Formatos de archivos SIG

.dbf
.shp
.shx

Los archivos SIG están generalmente representados por 3-5 archivos. Los requisitos mínimos son una serie de 3 archivos (.shp, shx y .dbf). La proyección necesaria para cargar en la *Data Library* es la proyección geográfica, es decir que la coordenadas de longitud y de latitud esten en grados decimales, no en metros. Es posible convertir una proyección que no es geográfica, como *Universal Transverse Mercator (UTM),* en una proyección geográfica utilizando ArcView o ArcMap.

La serie de archivos SIG tiene una tabla de información adicional sobre cada rasgo geográfico. Esta información puede ser: nombre, población, superficie, suma de niños de edades de 5 o menos, etc.

Guías Generales para todos archivos

- no se debe incluir ningún carácter especial como ‘, `, @, $, ^, (,), -, el nombre del archivo o cualquiera entrada del archivo
-el uso del carácter guión bajo (‘_’) está permitido
-Nunca empiece un nombre de columna con una valor numérico, o con un nombre
si usa -números decimales ,tienen que seguir la convención del punto decimal, no la de la coma decimal.
-valores vacíos deben ser llenados con un indicador de valor vacío, como por ejemplo -999
-los nombres de los archivos no deben tener ningún blanco
-longitud y latitud tienen que ser en grados decimales

La fuente de los datos tiene que ser identificada y reconocida. Por favor indique cualquier restricción del uso de los datos, y cualquier proceso de autorización necesario para el acceso de los datos.

Un *username* de *Data Library* es necesario para cargar datos en la *Data Library*. Por favor contacte help@iri.columbia.edu o jefft@iri.columbia.edu, si no tiene un *username*. Si ya tiene un *username* del IRI, utilícelo.

II. Formatos de los Identificadores Espaciales y Temporales

Los identificadores espaciales en archivos tabulares tienen que corresponder con los identificadores en los archivos SIG del usuario o en las bases de datos SIG de la *Data Library* del IRI. Las bases de datos SIG de la *Data Library* se pueden encontrar en : http://iridl.ldeo.columbia.edu/SOURCES/.Features/. Una base de datos SIG de la *Data Library* tiene generalmente identificadores espaciales numéricos (código) y alfabéticos (nombres). Este identificador de la *Data Library* debe ser añadido al archivo tabular del usuario, antes de que se realice cualquier correlación con variables climáticas .

Como ejemplos de identificadores espaciales, se tiene la división administrativa de países, regiones climáticas, distritos o municipios o áreas especificas de ciertos cultivos los cuales tienen referentes espaciales distintos.

Ejemplos de identificador espacial por jerarquia

Nombres de Divisiones Administrativas de un País

-Primer Nivel Administrativo (estados, provincias, o regiones)
-Segundo Nivel Administrativo (distritos, zonas, departamentos, o municipios)
-Tercer Nivel Administrativo (sub-distritos, sub-zonas, o municipios)
-Cuarto Nivel Administrativo (poco utilizado, pero puede incluir ciudades, o pueblos o barrios o veredas)

Codigos de Identificación asociados con una entidad

-Código de Provincia (por ejemplo 01)
-Código de Distrito (por ejemplo 0101 -- incluyendo el código de provincia, o 0001-sin incluir el código de provincia)
-Código de Sub-Distrito (por ejemplo 010101 -- incluyendo los códigos de provincia y de distrito, o 000001 -- sin incluir los códigos de provincia o de distrito)

Los identificadores temporales que recomendamos son los que siguen el estándar Internacional ISO 8601 o el formato Fecha Excel.

Ejemplos de ISO 8601 (cuando se utiliza este formato, llame la columna 'time')

-Año YYYY (por ejemplo 1997)
-Mes YYYY-MM (por ejemplo 1997-07)
-Día YYYY-MM-DD (por ejemplo 1997-07-16)
-Semana YYYY-MM-DD/YYYY-MM-DD (por ejemplo 1997-07-16/1997-07-22)

Ejemplos Excel
-Año YYYY (por ejemplo 1997)
-Mes (por ejemplo 7/1997 o Jul-1997)
-Día (por ejemplo 7/16/1997 o 16-Jul-1997)

Ejemplos aceptados, pero que no son estándares:

dia	mes	ano		
3	Jan	2001		

mes	ano			
01	2001			

Nota : los nombres de las columnas no son impuestos, pero hay cueros caracteres que no se puede usar , por ejemplo no se usan tilde en 'dia' y no n en 'ano'. Las valores alfabéticas de los meses tienen que corresponder a la ortografía inglesa (por ejemplo para enero : 'Jan' y no 'Ene').

Por favor considere los ejemplos siguientes de identificadores temporales para que no tenga problemas cuando cargue su archivo en la Dta Library.

Archivo Excel (.xls)

fecha	distrito_id	distrito_nombre	casos	precipitacion
1/1997	504	Colonia	5	30
2/1997	504	Colonia	3	24

Archivos *Tab Separated* (.tsv)

time	distrito_id	distrito_nom	casos	precipitacion
1997-01	504	Colonia	5	30
1997-02	504	Colonia	3	24

Archivos *Comma Separated* (.csv)

time,distrito_id,distrito_nombre,casos,precipitacion
1997-01-01/1997-01-07,504,Colonia,2,5
1997-01-08/1997-01-14,504,Colonia,1,8

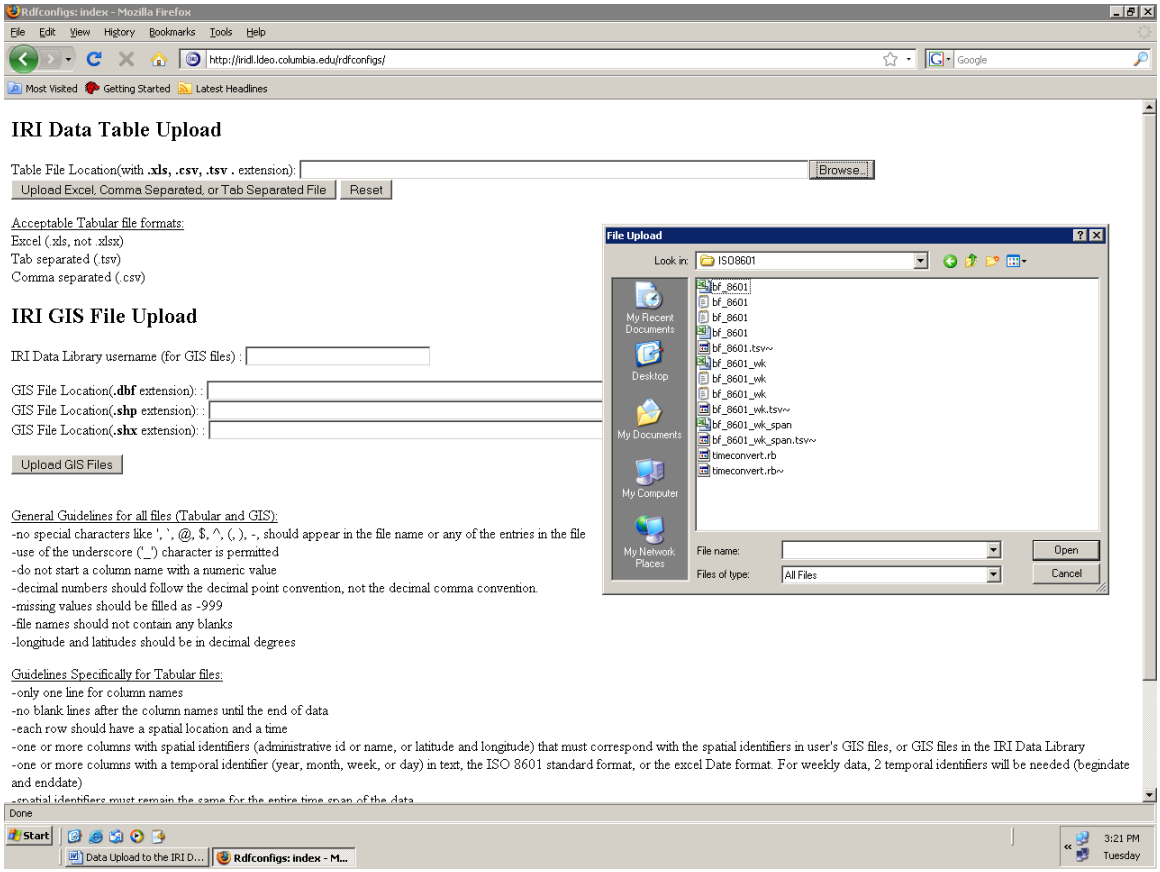
Archivos SIG

municipios.dbf
municipios.shp
municipios.shx

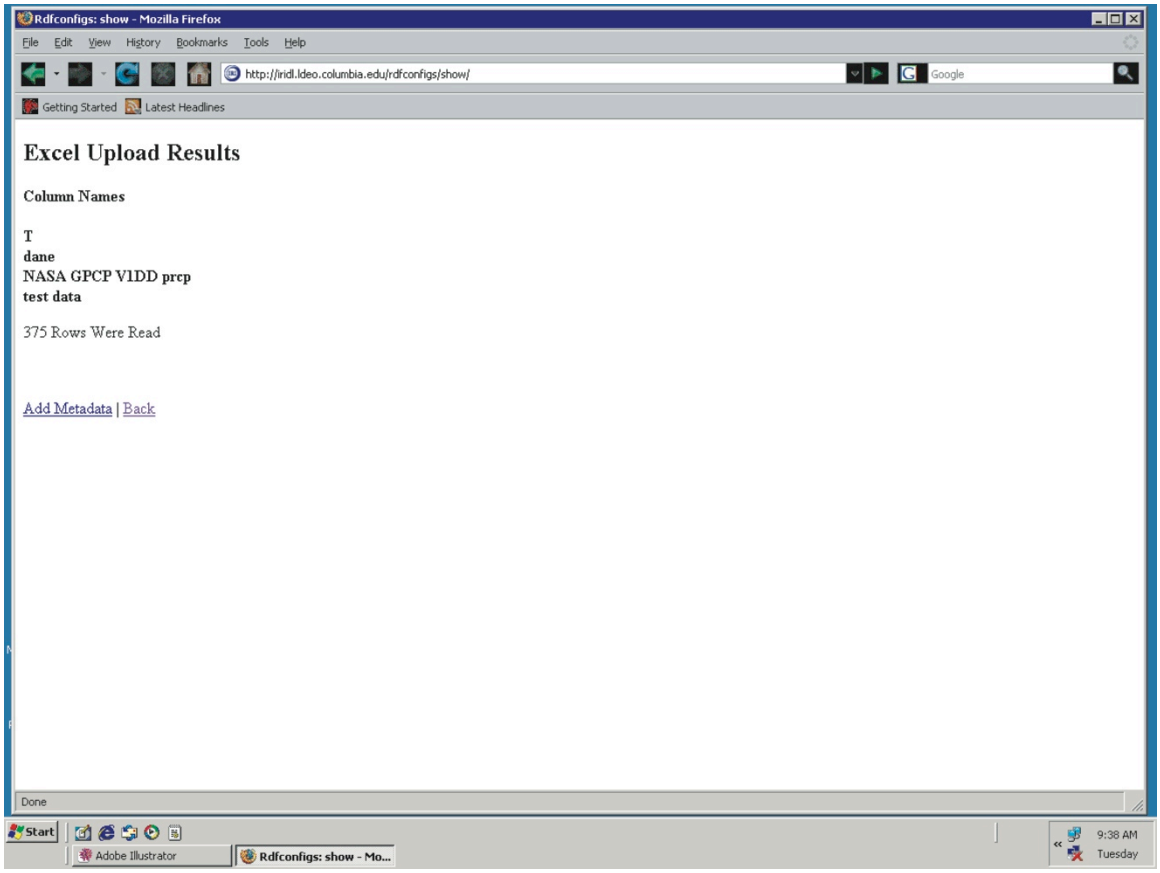
Ahora que ya ha verificado sus sus datos están , avance en la página internet para realizar la carga de datos.

III. Formulario de Internet para la carga de datos.

El formulario de internet está disponibel en la siguiente direccion URL, http://iridl.ldeo.columbia.edu/rdfconfigs/. No olvide incluir la última barra oblicua (‘/’). Utilice el botón ‘Browse’ para localizar su archivo (en su computador). Para archivos Excel y tabulares, indique un solo archivo . Para archivos SIG, indique tres archivos (.shp, .shx, and .dbf) y su *username* de *Data Library*. **Asegúrese de que el nombre del archivo no tiene valores en blanco**. Haga click sobre el botón ‘Upload’.



Si cargo exitosamente los datos , ahora encontrara una página ‘Upload Results’. Esta página muestra los nombres de las columnas, y la suma de líneas leídas. Si un mensaje de error aparece, haga click sobre la flecha ‘Atrás’ para volver al formulario de internet, y buscar errores, o intentar con otro archivo.

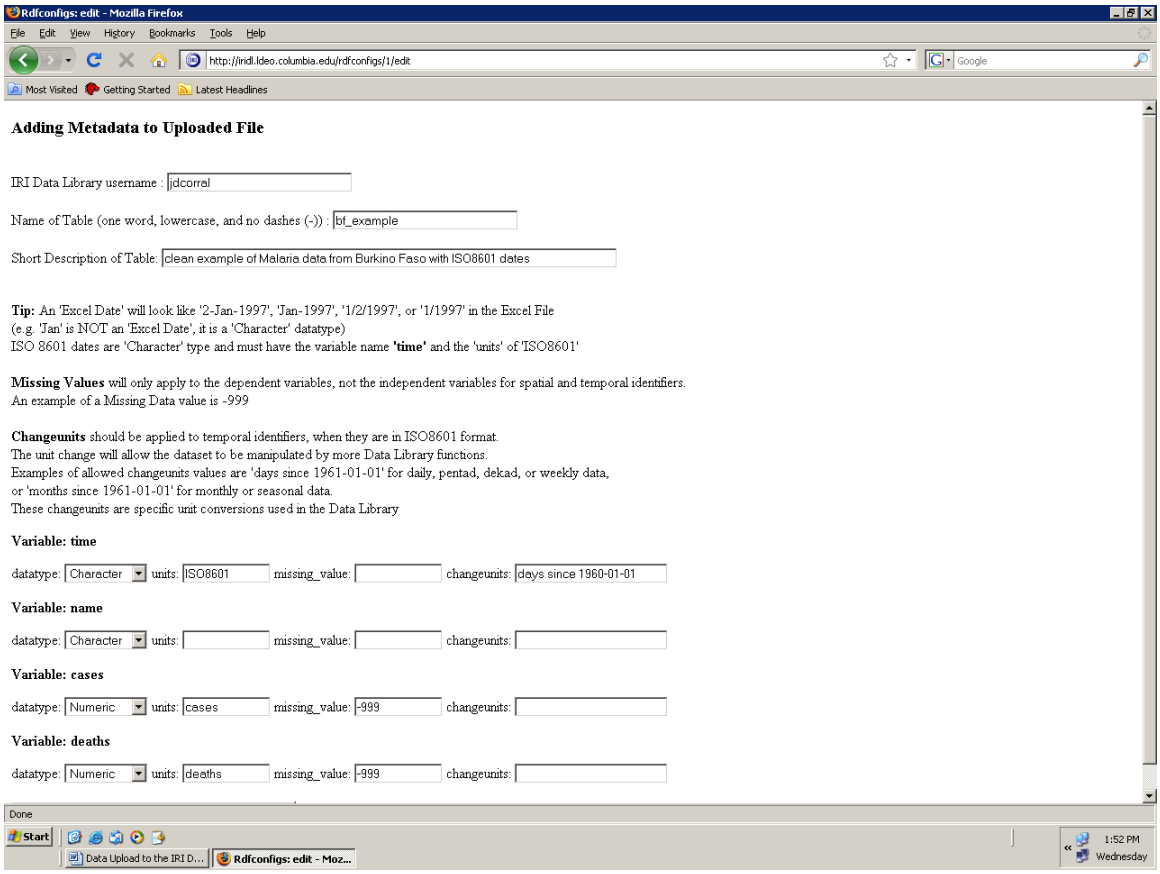


IV. Añadir Metadatos

Al final de la página ‘Upload Results’, hay un enlace que dice ‘Add Metadata’. Haga click en este enlace para ir al formulario ‘Adding Metadata’. Los tipos de metadatos necesarios están en *username*, un nombre corto para la base de datos, una descripción de la base de datos con la fuente de los datos, y el *datatype* para cada columna de datos. Indicar las unidades para cada columna de datos es algo opcional.

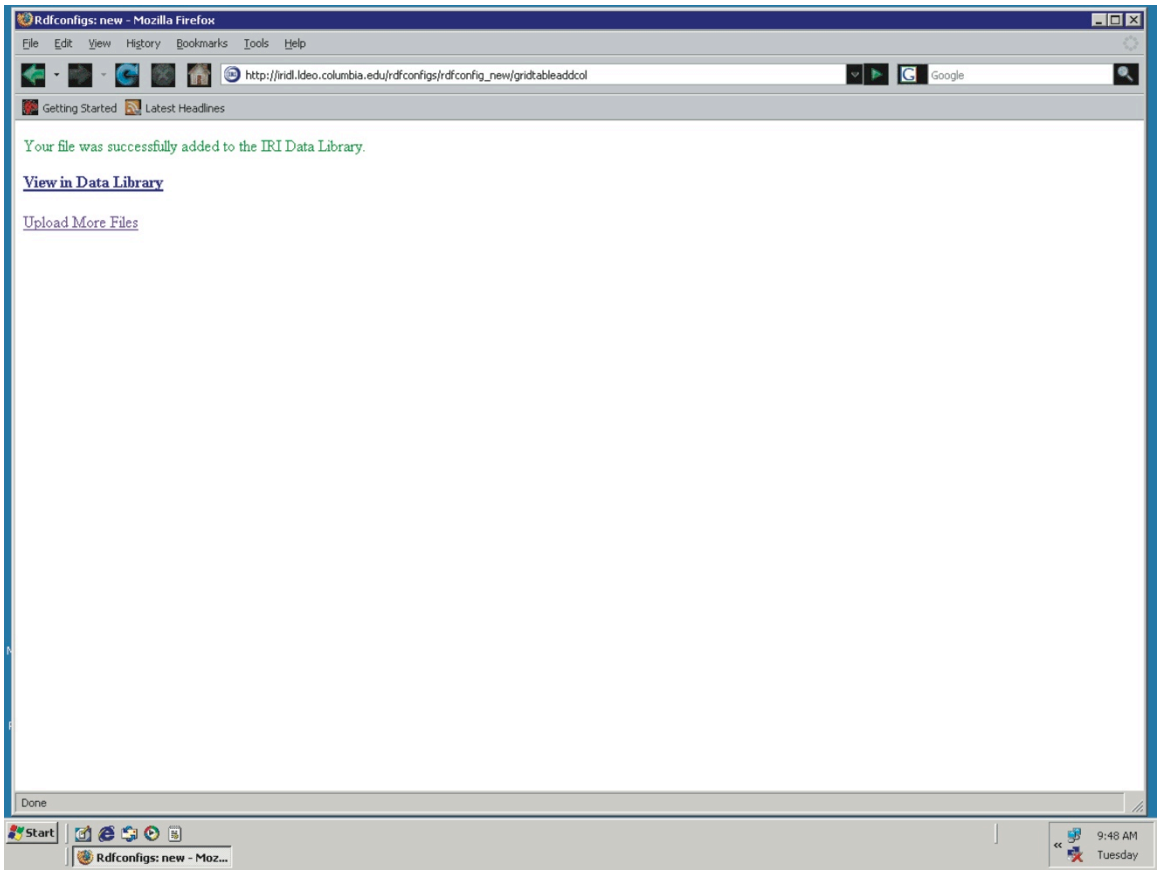
El *datatype* ‘Excel date’ es un *datatype* muy específico. Es un *datatype* utilizado por Excel para representar una fecha en un formato predefinido como ‘2-Jan-1997’, ‘Jan-1997’, ‘1/2/1997’, o ‘1/1997’. Un *datatype* ‘Numeric’ puede representar números decimales o enteros.

Un *datatype* ‘Character’ esta utilizado para nombres o texto, y para fechas ISO 8601. Las unidades de fechas ISO 8601 tienen que ser definidas ISO8601 también. Cuando se utiliza identificadores temporales que son estándares, utilice el *datatype* ‘Character’ para ‘Jan’ y ‘Numeric’ para ‘01’ en el caso del mes de enero.



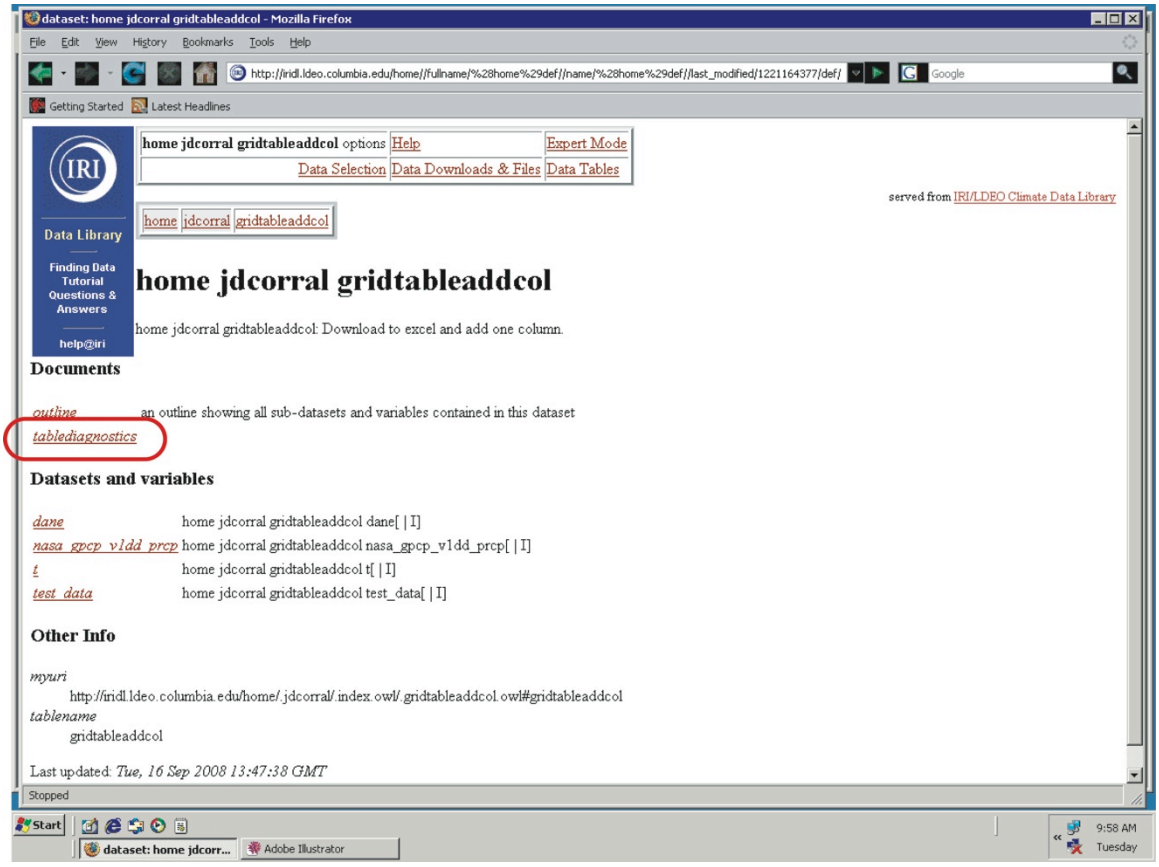
Envíe los metadatos haciendo click sobre el botón ‘Submit Metadata and add Dataset to Data Library’. Aparece en la página ‘Success’ si los metadatos fueron aceptados.

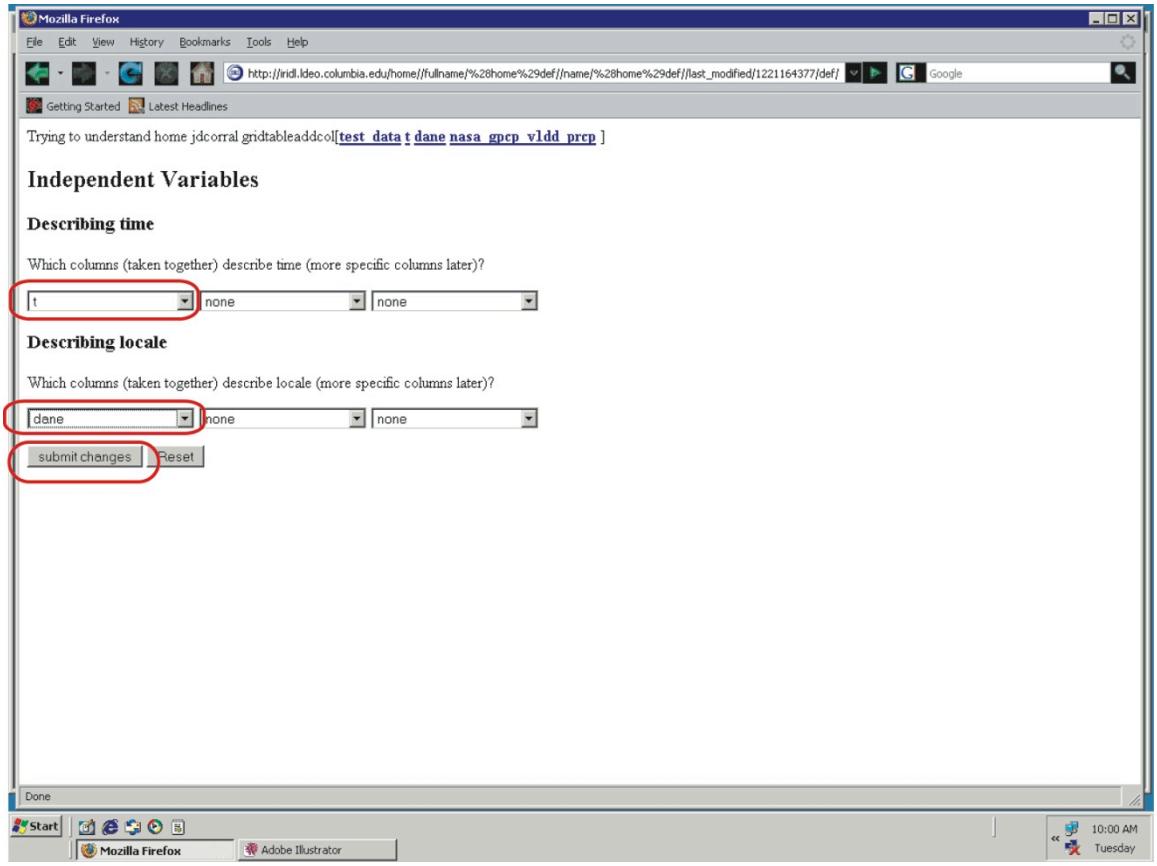
Esta página muestra un enlace hasta la nueva base de datos en la *Data Library*. La base de datos está localizada en la sección personal del usuario de la *Data Library* (home .username) o en la sección de los estudiantes *CIPH* en la *Data Library* (home .ciph .students .username). Si aparece un mensaje de error, haga click en la flecha ‘Atrás’ para volver al formulario internet, y buscar errores.



V. Definir las Dimensiones Espaciales y Temporales

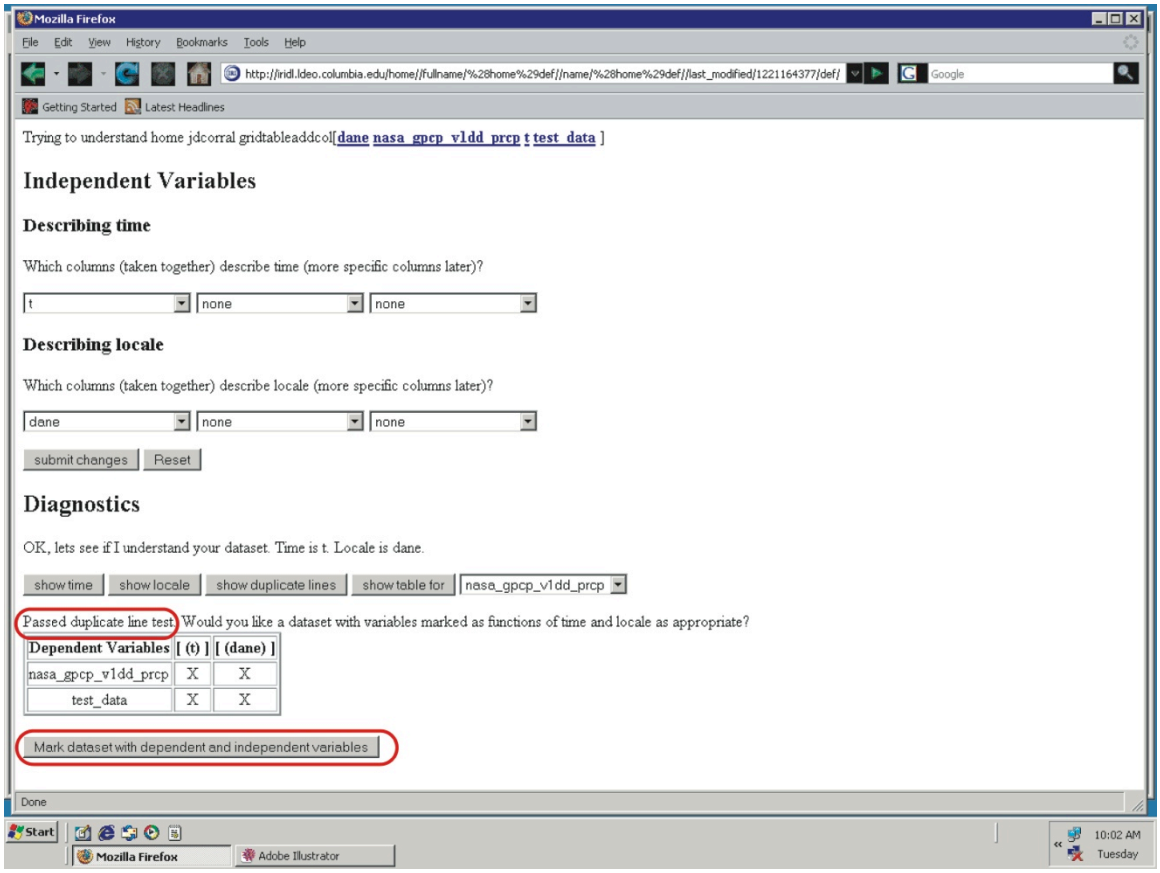
La 'table diagnostics' aparece en la página de la nueva base de datos. Haga click sobre este enlace para seleccionar cuales columnas de la base de datos deben ser utilizadas para definir las dimensiones espaciales y temporales de esta base de datos. Se llaman variables independientes 'Independent Variables' de la base de datos. Seleccione una columna o más para describir la fecha y una columna o más para definir la localidad. Haga click sobre el botón 'submit changes' para aplicar estas definiciones.



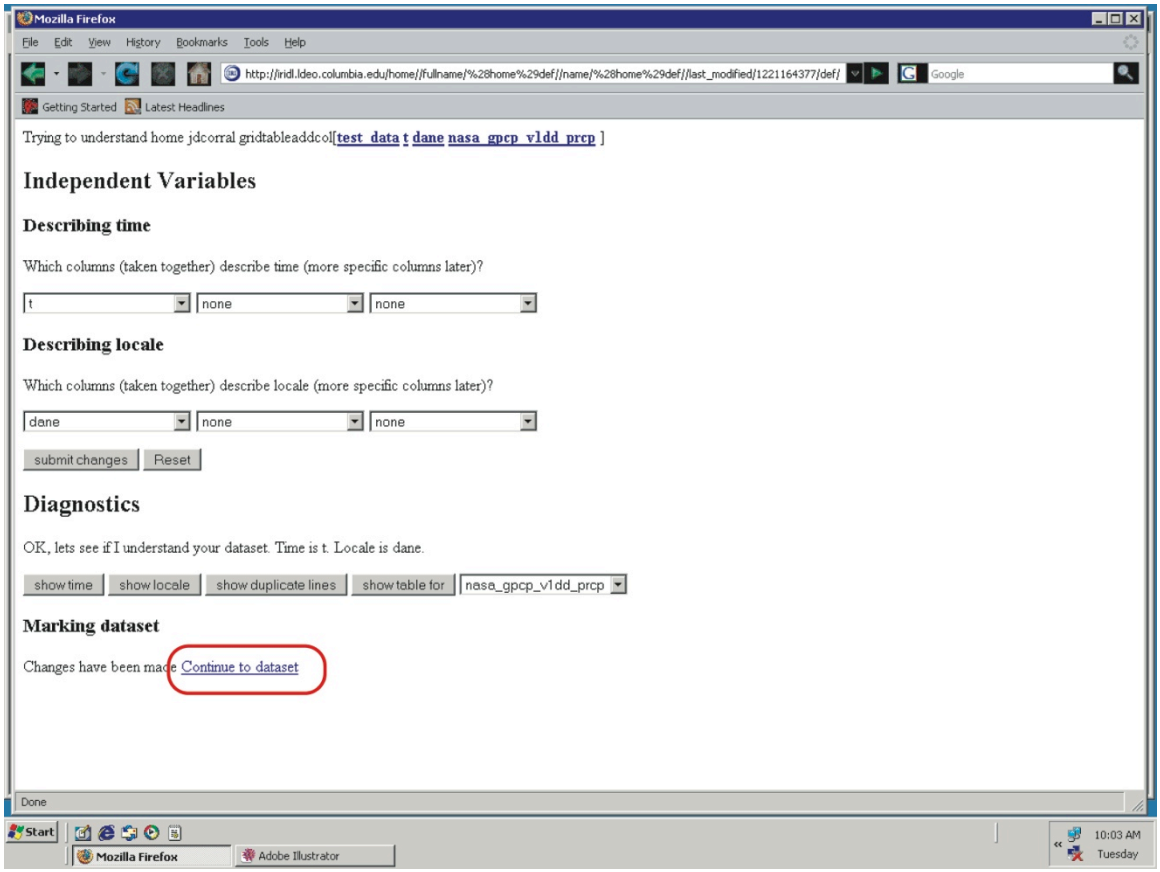

Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://iridl.ldeo.columbia.edu/home//fullname/%28home%29def//name/%28home%29def//last_modified/1221164377/def/
Google
Getting Started Latest Headlines
Trying to understand home jdcorral gridtableaddcol[test_data t dane nasa_gpcp_v1dd_prcp]
Independent Variables
Describing time
Which columns (taken together) describe time (more specific columns later)?
t
none
none
Describing locale
Which columns (taken together) describe locale (more specific columns later)?
dane
none
none
submit changes
Reset
Done
Start
Mozilla Firefox
Adobe Illustrator
10:00 AM
Tuesday

VI. Control de Calidad

Cuando las dimensiones espaciales y temporales están definidas, la *Data Library* evalúa si hay líneas duplicadas en la base de datos. El usuario puede ver dos resultados 1) un mensaje 'Passed duplicate line test', o 2) una lista de las líneas duplicadas. Cuando la *Data Library* hace la evaluación, el usuario ve una tabla de variables dependientes 'Dependent Variables' y sus dimensiones. La etapa siguiente es hacer click sobre el botón 'Mark dataset with dependent and independent variables'.

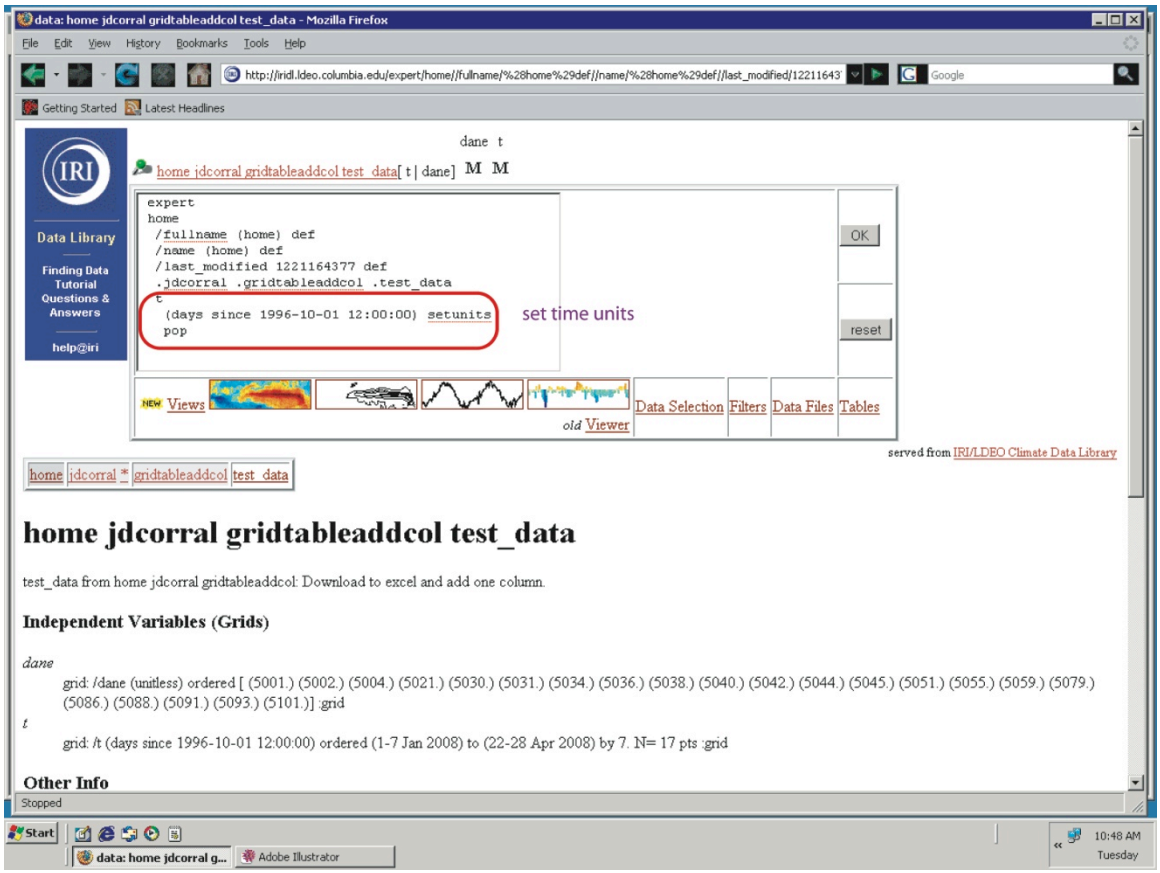


La página siguiente muestra un resumen de la base de datos. Si el resumen es satisfactorio, haga click sobre el botón ‘Continue to dataset’ para ver la base de datos en el directorio ‘home’ del usuario en la *Data Library*.



VII. Uso en la Data Library

Si no utilizó fechas ISO8601 con conversión de unidades, es posible que la unidad de tiempo de la base de datos tenga que ser modificada. La mayoría de las bases de datos de la *Data Library* siguen una convención para la unidad de tiempo 'days since 1996-10-01 12:00:00' o 'months since 1960-01-01', o 'julian_day'. Si el usuario quiere correlacionar su base de datos con bases de datos en la *Data Library*, y utilizar las herramientas de visualización de la *Data Library*, la unidad de la dimensión temporal tiene que corresponder a la convención de la *Data Library*. Hay una función de la *Data Library* que se llama 'setunits'. Esta se puede utilizar para convertir la unidad temporal de la base de datos en una unidad que concuerda con la *Data Library*. Esta función puede ser utilizada después de hacer click sobre el enlace 'Expert Mode' en el esquina superior derecha de la página de la base de datos (vea el ejemplo siguiente).



Si tiene cualquier problema con la carga de datos, por favor contacte a John del Corral (jdcorral@iri.columbia.edu) o Rémi Cousin (remic@iri.columbia.edu).