DMPonline / DaMaRO Workshop

Rewley House, Oxford
28 June 2013

DataStage

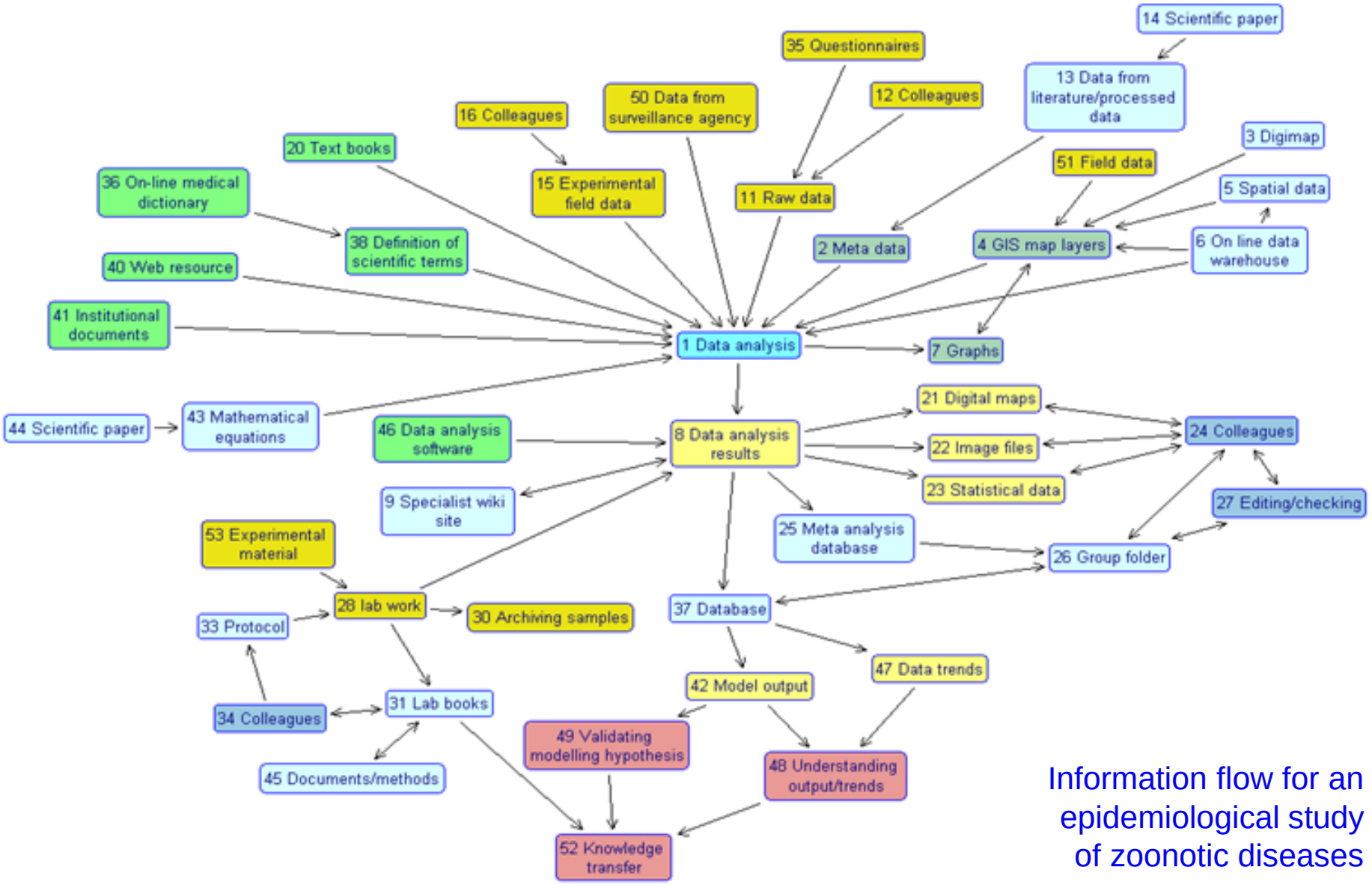
- a simple file management system

David Shotton

Oxford e-Research Centre and
Department of Zoology
University of Oxford, UK

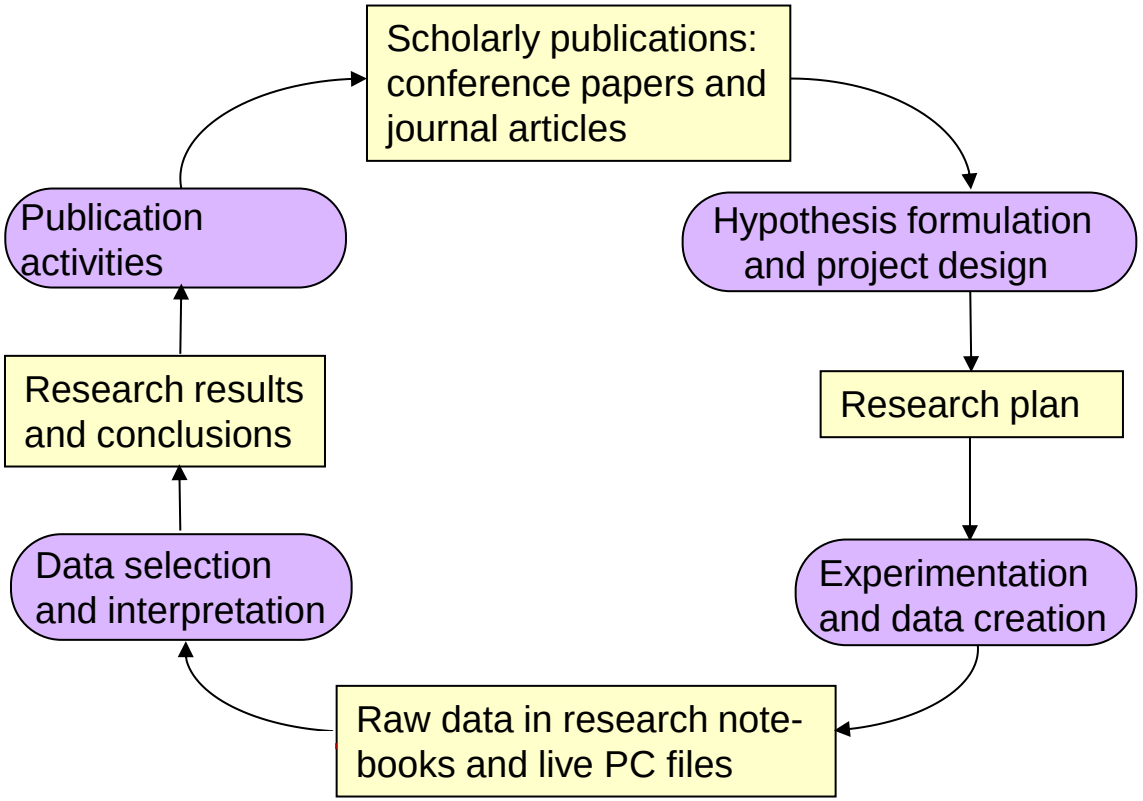
david.shotton@zoo.ox.ac.uk

Information management in biological research can be quite complex

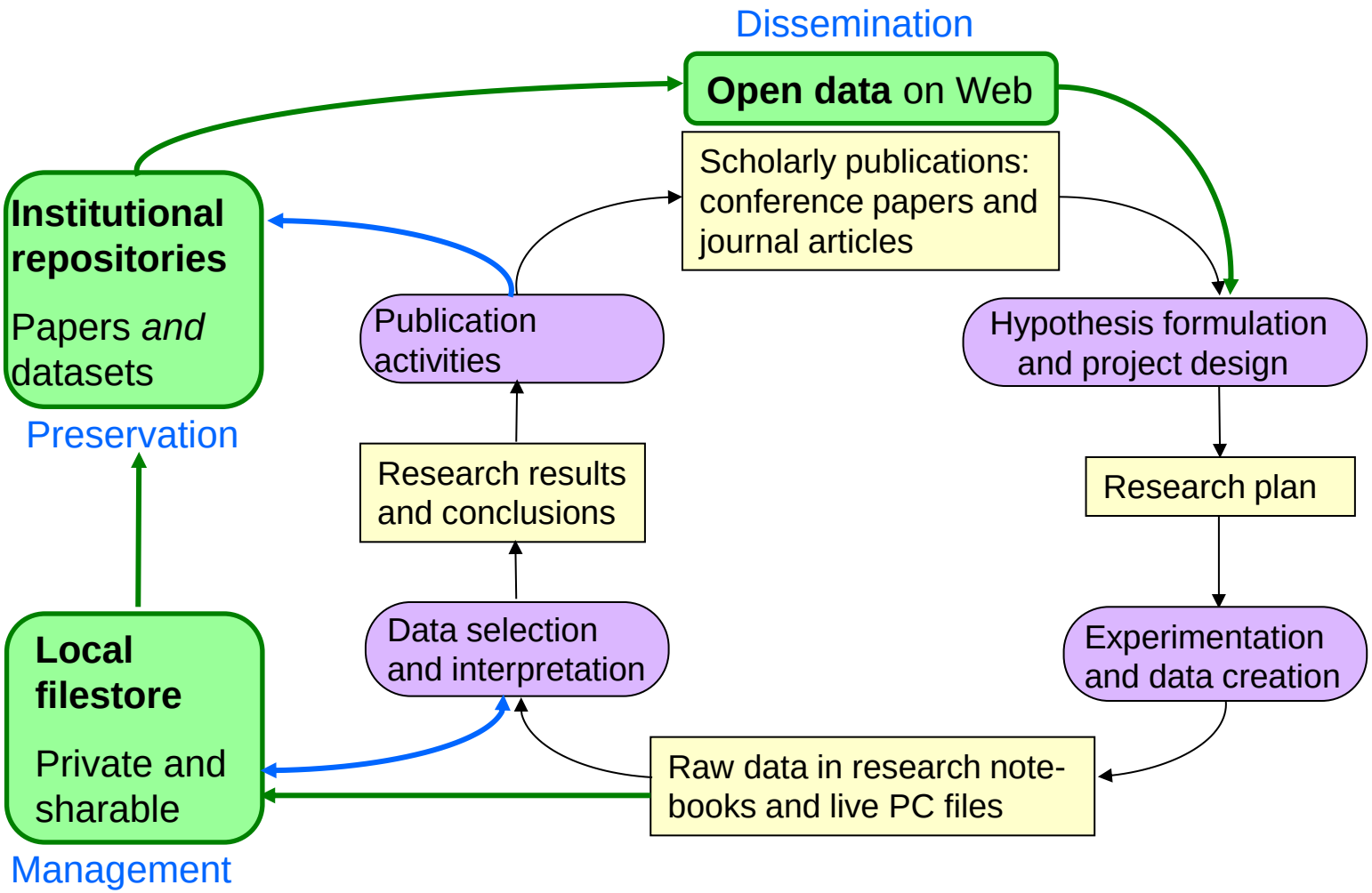


Information flow for an epidemiological study of zoonotic diseases

The conventional research data lifecycle



The enhanced research data lifecycle



The need for easy-to-use data management tools

- To facilitate uptake by busy researchers, we need easy-to-use data management tools that provide immediate benefit
- We practice 'sheer curation' (http://en.wikipedia.org/wiki/Sheer_curation):
  - working with users rather than against them
  - enable use of data file formats familiar to them
  - make contribution attribution activities sufficiently lightweight and transparent that they do not impose significant cognitive overhead
- The first task in data management is to ensure 'live' data files are properly stored and reliably backed up
- For this we have developed DataStage

What is DataStage?

- DataStage is a simple file management system for use by research groups to permit organization of 'live' data files, including those that will not get archived
- One DataStage instance is used per research group
- Each group member has
  - a Private Folder: only (s)he can write, and only (she) and PI can read
  - a Shared Folder: only (s)he can write, and all group members can read
  - a Collaborative Folder: All group members can write and read
- The DataStage server can be deployed locally or in the cloud
- The DataStage file store appears to each user as a mapped drive
- It can also be accessed via a Web browser
- It is tailored to permit creation of data packages from selected data files, and their submission to the Oxford DataBank, our institutional data repository, where they can be archived and published with DataCite DOIs
- To encourage submissions, metadata requirements are minimal

DataStage and DataBank

DataFlow

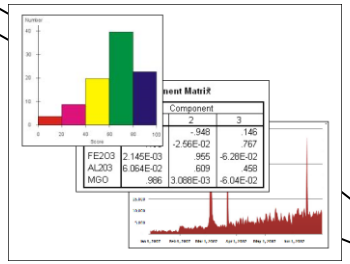
Researchers

DataStage file system

DATASTAGE

DataStage: http://datastage-demo.bodleian.ox.ac.uk

DataBank: http://databank-demo.bodleian.ox.ac.uk



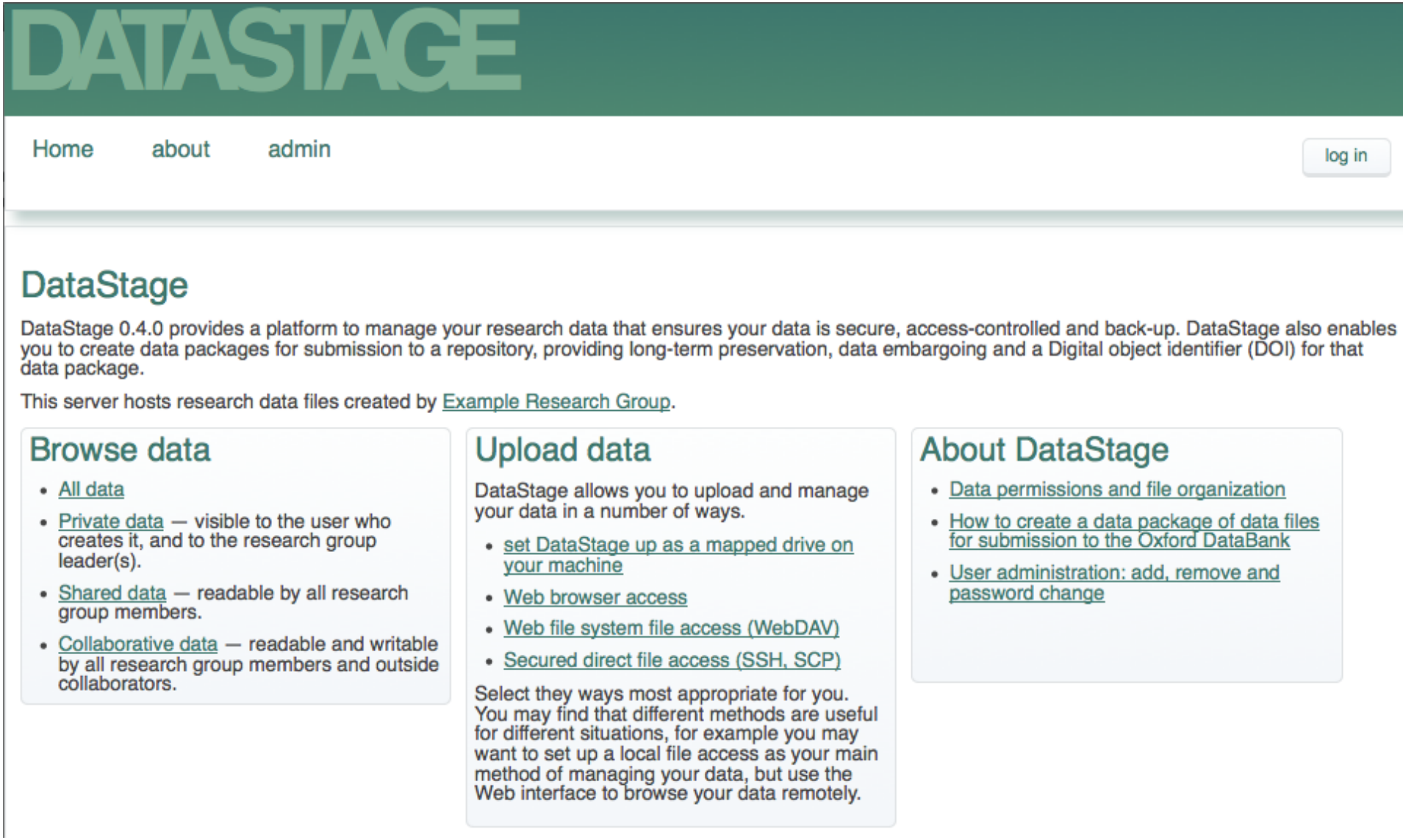
SWORD deposit

Researchers, other users

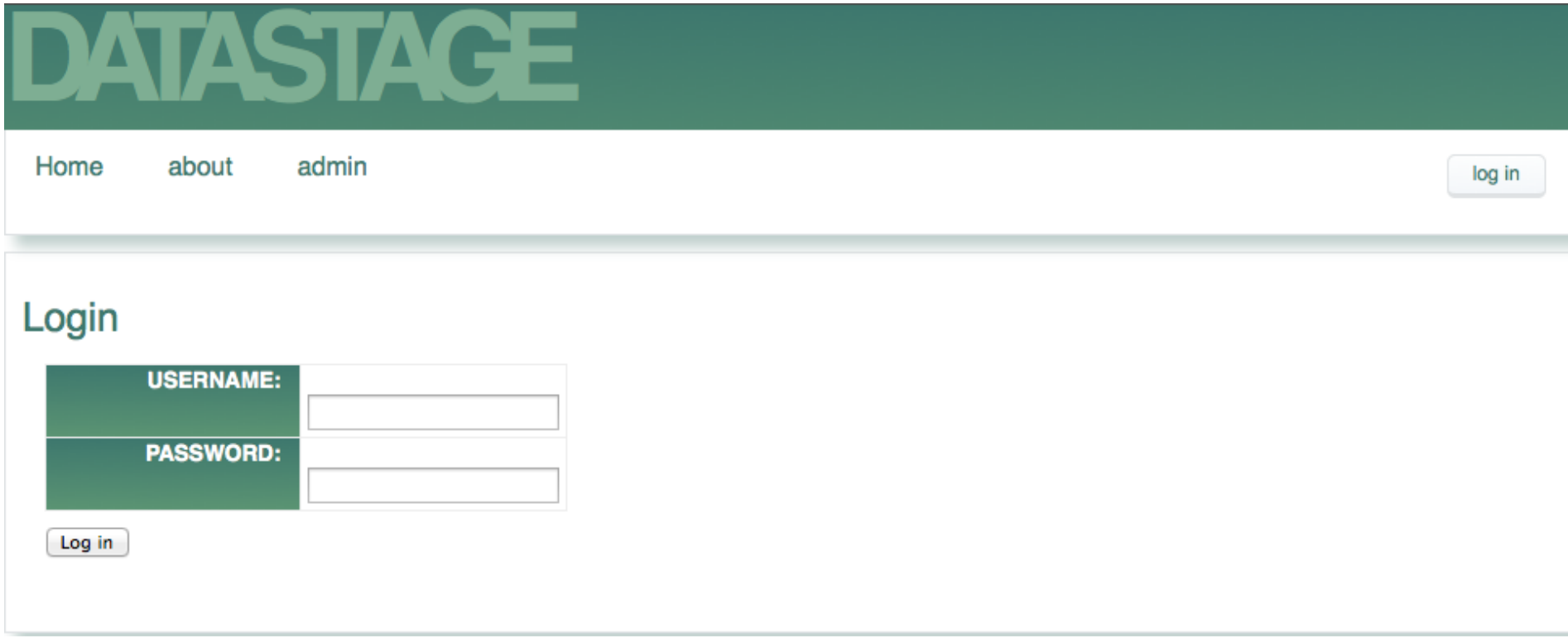
DATABANK

DataBank repository

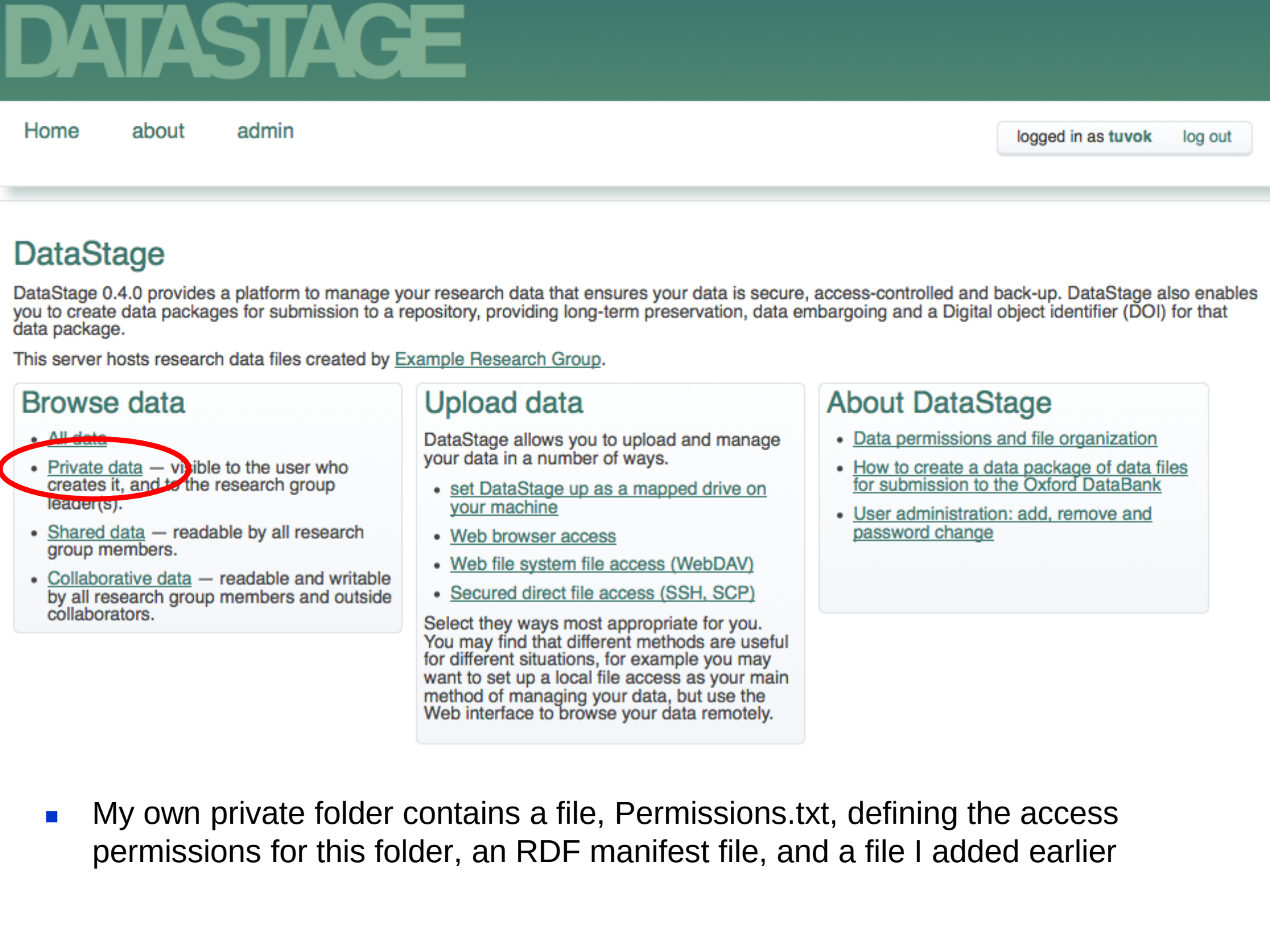
The DataStage web interface front page



The DataStage log-in page

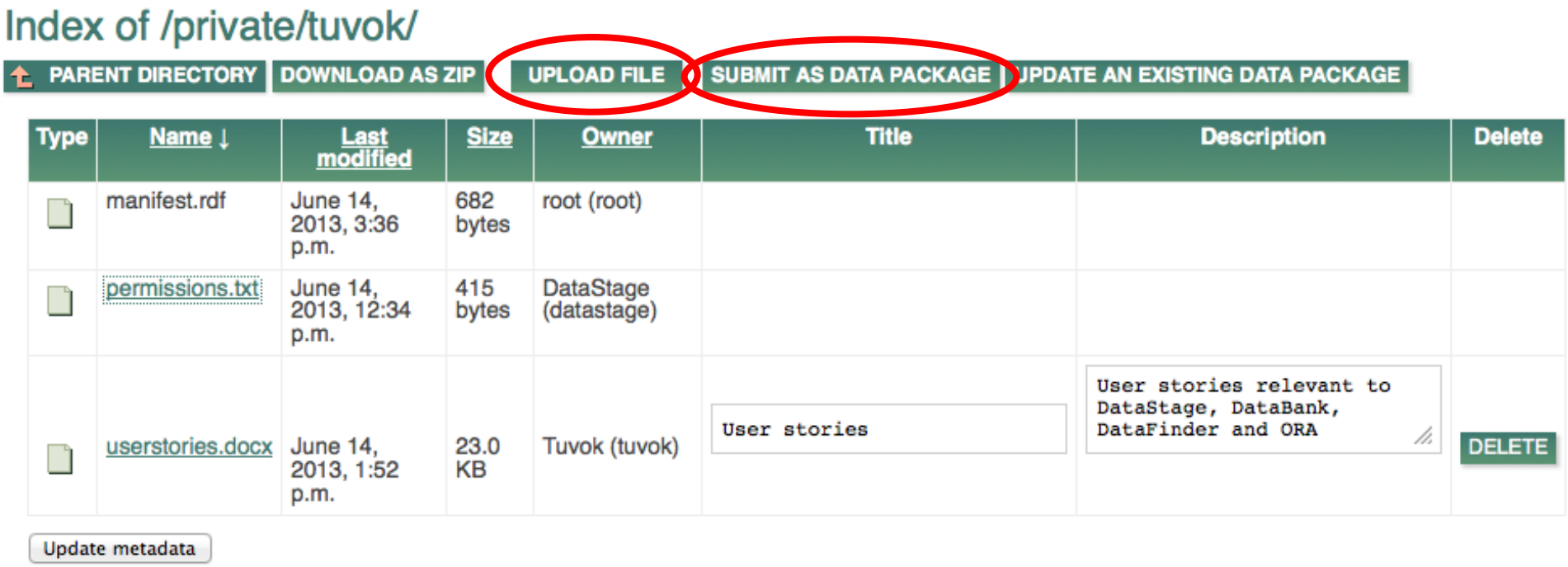


- Log in with a locally assigned username and password
- Because the same username and password are used both to create the mapped drive and for web access, that use quite different protocols, we have been unable to integrate with the University Single Sign-On System



- My own private folder contains a file, Permissions.txt, defining the access permissions for this folder, an RDF manifest file, and a file I added earlier

Upload files to this folder, then submit them as a data package



- I previously added the file userstories.docx to my private folder
- Additional files can be uploaded to the DataStage server one by one
- All the files in this folder (or any other selected folder or sub-folder) can be submitted as a data package to DataBank, using the Submit button

Benefits of DataStage

- Simple and easy to use
- Both mapped drive and browser access
- The DataStage Server can be hosted locally or on the cloud
- Can be used both for local file management and backup on a day-to-day basis, and also for data submission to a repository for archiving and publication
- Repository submission uses the SWORD2 protocol, and so can be directed to any SWORD2-compliant repository (e.g. Dspace), not just to DataBank
- Minimal metadata requirements to encourage usage
- Open source software
- Packaged as a Debian package for one-click installation on an Open Source Ubuntu Linux system
- Separate administrative interface for configuring group members, permissions, etc.

DataCite metadata entry tool for more complete metadata

DataCite

Helping you to find, access, and reuse data

This form itself has been created by Tanya Gray and David Shotton of the University of Oxford, and is not part of the official DataCite metadata standard. It is an open source tool that can be used by anyone . Please notify problems or send comments both to tanya.gray@zoo.ox.ac.uk and to david.shotton@zoo.ox.ac.uk with the subject line 'DataCite Metadata Input Form'.

DataCite Metadata Input Form

Load an existing DataCite Metadata Report into this form, or just start typing to create a new DataCite Metadata Report. SAVE your report after completing each subsection, for data security. This form creates a single DataCite Metadata Report as an XML document containing the metadata entered into this Web form, which can be saved to your local hard drive as a .xml file by using the "Save" tab. Partially completed reports can be saved and later re-loaded into a blank form using the 'Load' tab. This form is based on the DataCite Metadata Kernel v2.2 (July 2011). In the DataCite specification, Properties 1-5 are mandatory while Properties 6-17 are optional.

DataCite Metadata Input Form | Load | Save | Validate | Help

Create a separate DataCite metadata report for each entity that has (or will have) a unique DOI. If data files within a data package have, or will be given, individual DOIs, first create and save a DataCite metadata report for the data package as a whole, then modify it as appropriate for each data file and save each modified DataCite metadata report with a separate file name.

Click + to add another item below the first with empty metadata input fields, click Duplicate to copy an existing item or section including the metadata that has been entered into the input fields, or ✖ to delete one. Click ? for guidance. Unused fields removed by clicking "X" can be reinstated by clicking the button that appears below this sentence labelled "View optional fields that can be added to the input form" and then selecting the deleted fields to be restored.

DataCite metadata report — Reset form — Clicking on this link will delete all entered data!

Date of last update to this metadata record (Format: yyyy-mm-dd)

2013-03-04

Version number for this version of the metadata record (Please use an integer, incrementing from 1)

1

[1] DOI of dataset, if known

10.1234/123456

Creators

1. [2] Creator of Data Collection + Duplicate *Name the creator(s) of the dataset being annotated, in priority order, or the authors of the publication. The name used may be a corporate/institutional name or a personal name. Format for personal names: FamilyName, GivenName(s). Use + to add additional names if there are multiple authors.*

[2.1] Creator name

Shotton, David

[2.2] Personal identifier (if available)

0000-0001-5506-523X

[2.2.1] Personal identifier scheme (if personal identifier given)

ORCID

Titles of dataset

1. [3] Title + Duplicate

DatasetTitle

[3.1] Title type (Leave blank if main title)

Metadata can be viewed as HTML or output as RDF

DataCite

Helping you to find, access, and reuse data

Report

Date of last update to this metadata record (Format: yyyy-mm-dd)
2013-03-04
Version number for this version of the metadata record (Please use an integer, incrementing from 1)
1
[1] DOI of dataset, if known
10.1234/123456
[1.1] Identifier type
DOI
Creators
[2] Creator of Data Collection
[2.1] Creator name
Shotton, David
[2.2] Personal identifier (if available)
0000-0001-5506-523X
[2.2.1] Personal identifier scheme (if personal identifier given)
ORCID
Titles of dataset
[3] Title
DatasetTitle
[3.1] Title type
[4] Publisher
Oxford Databank
[5] Publication year
2013
Subjects
[6] Subject
TestData
[6.1] Subject schema
MySubjectScheme
Contributors
[7] Contributor
[7.1] Contributor type
Funder
[7.2] Contributor name
JISC
[7.3] Contributor identifier

TestDataCiteMetadata.xml.ttl — Edited

```
@prefix cito: <http://purl.org/spar/cito/> .
@prefix datacite: <http://purl.org/spar/datacite/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcmitype: <http://purl.org/dc/dcmitype/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix fabio: <http://purl.org/spar/fabio/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix frapo: <http://purl.org/cerif/frapo/> .
@prefix frbr:  <http://purl.org/vocab/frbr/core#> .
@prefix literal: <http://www.essepuntato.it/2010/06/literalreification/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix prism: <http://prismstandard.org/namespaces/basic/2.0/> .
@prefix pro: <http://purl.org/spar/pro/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/TR/2004/REC-rdf-schema-20040210/> .
@prefix scoro: <http://purl.org/spar/scoro/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix tisit: <http://www.ontologydesignpatterns.org/cp/owl/timeinterval.owl> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

:thisDataCiteMetadataDocument a fabio:EntityMetadata ;
dcterms:created "2013-03-04"^^xsd:date ;
        prism:versionIdentifier "1"^^xsd:integer;
cito:citesAsSourceDocument :thisDataset .

        :datasetCreator  a foaf:Person ;
        foaf:name "Shotton, David" ;

            datacite:hasIdentifier [ a datacite:PersonalIdentifier ;
            literal:hasLiteralValue "0000-0001-5506-523X" ;
            datacite:usesIdentifierScheme datacite:orcid ] ;

dcterms:title "DatasetTitle" ;

dc:publisher "Oxford Databank" ;
        fabio:hasPublicationYear "2013"^^xsd:gYear ;

    dcterms:subject [ rdf:label "TestData" ;
    skos:inScheme [ rdf:type fabio:TermDictionary ;
    rdf:label "MySubjectScheme" ] ] ;

    dcterms:contributor  [ a foaf:Agent ;
    foaf:name "JISC"  ;
    pro:holdsRoleInTime [ pro:withRole scoro:funder ];
    datacite:hasIdentifier [ a datacite:FunderIdentifier ;
    literal:hasLiteralValue "9876" ;
    datacite:usesIdentifierScheme datacite:fundref ] ] ;
dcterms:created "2013-02-28"^^xsd:date ;
dcterms:dateSubmitted "2013-03-02"^^xsd:date ;

   dcterms:language [ rdf:value "eng"^^dcterms:ISO639-3 ] ;

    dcterms:type [ rdf:value "XSLT script" ;
    datacite:hasGeneralResourceType dcmitype:Software ] ;
```

Acknowledgements

- Graham Klyne and Bhavana Ananda have been primarily responsible for the development of DataStage, with Katherine Fletcher as our invaluable project manager
- Tanya Gray created the DataCite metadata input form
- Richard Jones of Cottage Labs, working with us and with the DataBank developer Anusha Ranganathan at the Bodleian Libraries, implemented the SWORD2 protocol to encode DataStage submissions to DataBank
- The JISC funded the DataFlow Project that lead to the development of DataStage and improvements to DataBank
- I am grateful to them all

end