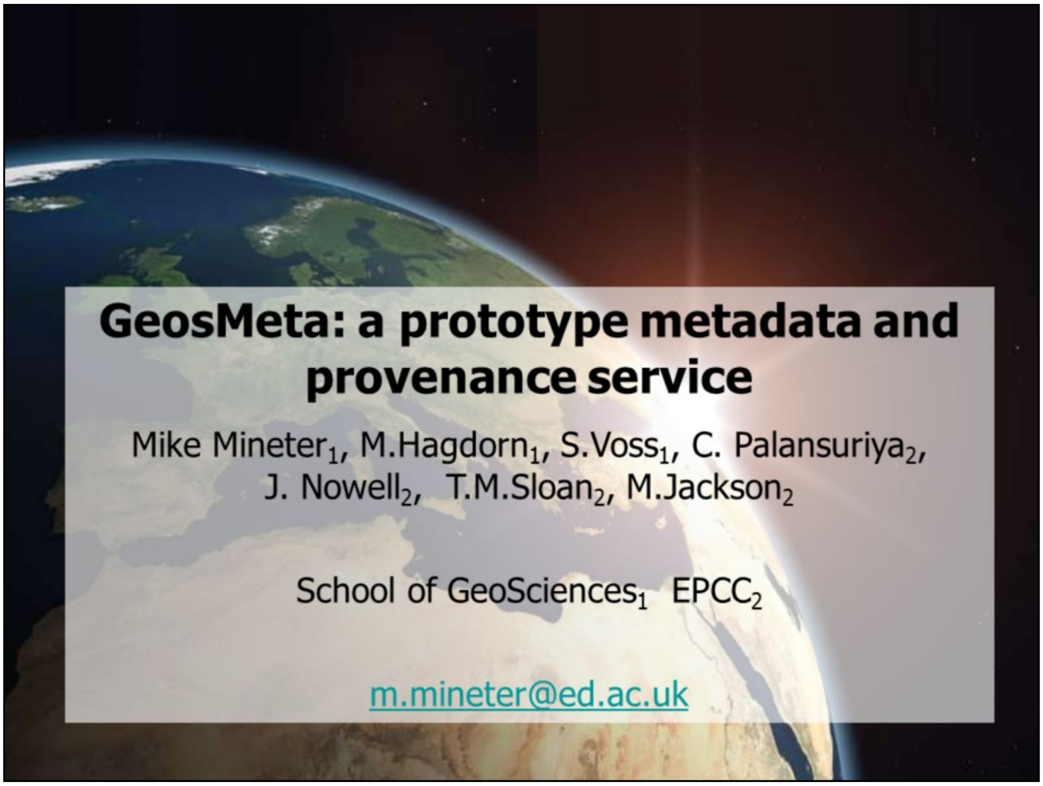


GeosMeta: a prototype metadata and provenance service

Mike Mineter[1], M.Hagdorn[1], S.Voss[1], C. Palansuriya[2], J. Nowell[2], T.M.Sloan[2], M.Jackson[2]

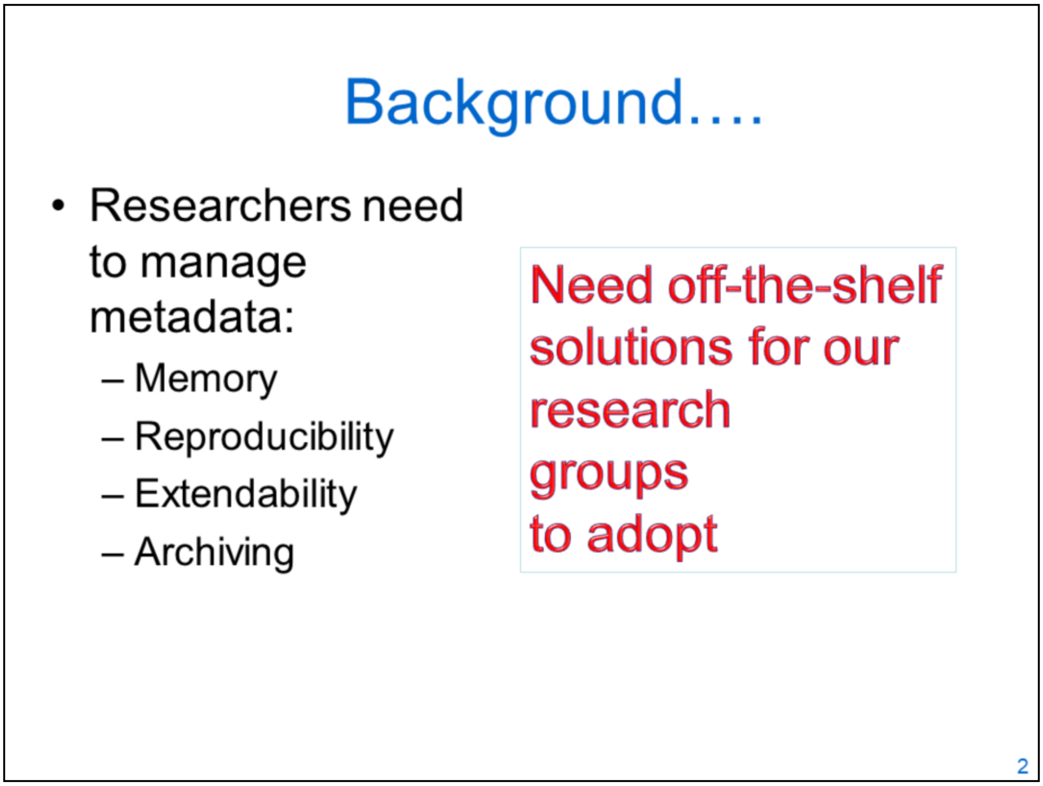
School of GeoSciences[1] EPCC[2]

m.mineter@ed.ac.uk

I work in climate science in the school of geosciences , with time bought by a number of projects.

One was keen for me to develop this further – its something that had been on a back burner for some time

since initial work done with EPCC



Background….

- Researchers need to manage metadata:
  - – Memory
  - – Reproducibility
  - – Extendability
  - – Archiving

Need off-the-shelf solutions for our research groups to adopt

A few years ago the need for data management plans, and reproducibility etc were being imposed, without tools to help.

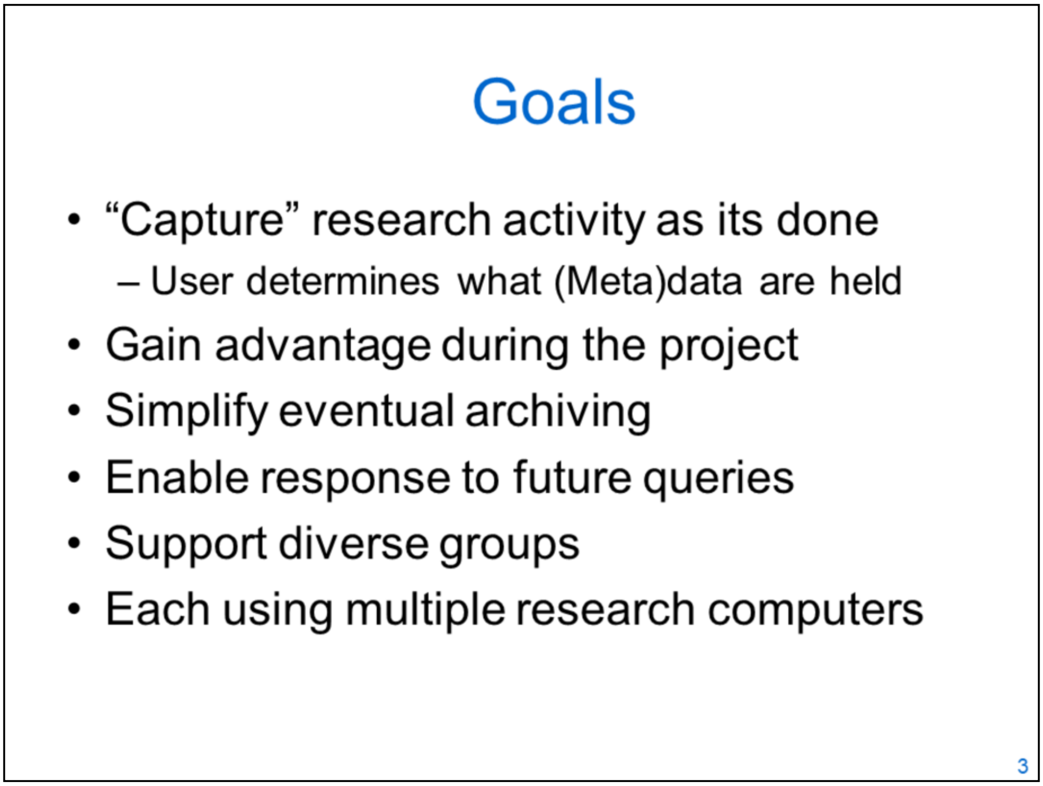
We wanted to give a tool that did help at archiving time, preventing the big “what did we do and we don’t have time to write it down” phase of archiving and data sharing.

We realised research is complicated, and tools to help us remember what we did are needed – enotebooks for example; developments with R came along in time.

Something that worked as an extension of the researchers memory, to give advantage during the research was the main goal and a major incentive for hoped-for adoption.

Any such tool offering a general way to do something requires a discipline from the researcher to hold to it.

We hope we have a balance in which the types of data in GeosMeta can be tuned and selected.



Goals

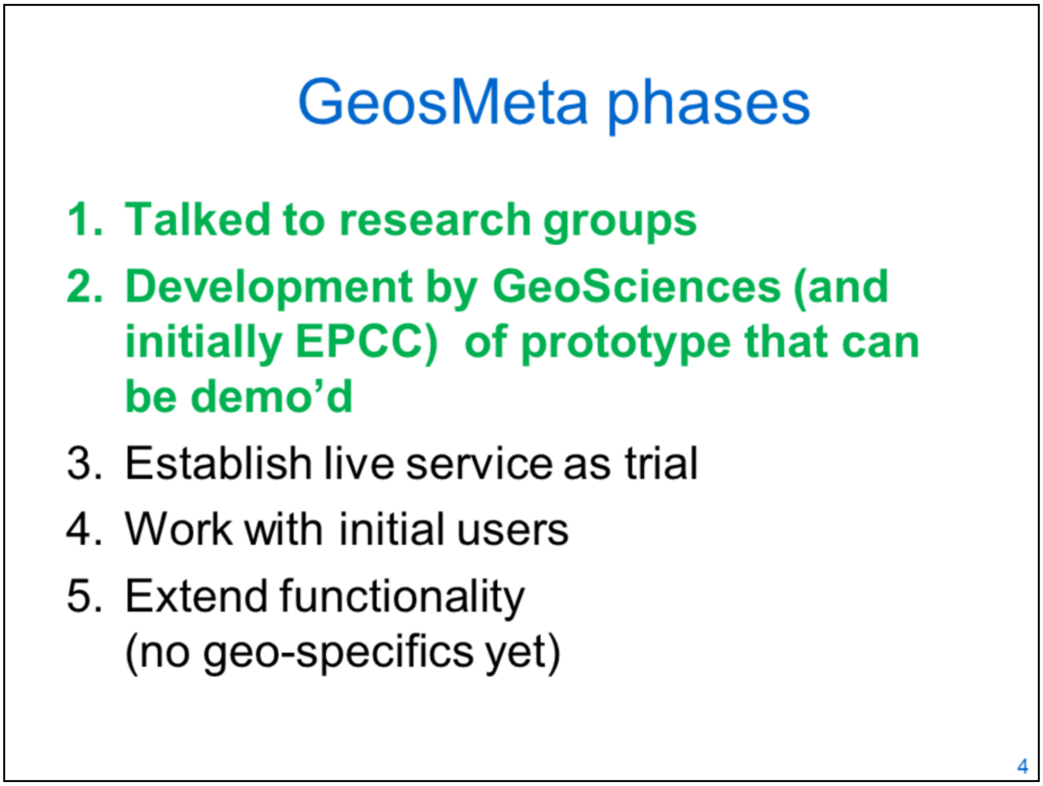
- "Capture" research activity as its done
  - User determines what (Meta)data are held
- Gain advantage during the project
- Simplify eventual archiving
- Enable response to future queries
- Support diverse groups
- Each using multiple research computers

Great diversity in group size, IT experience and complexity, kinds of data, metadata, ways of working

e.,g. climate modelling on ARCHER, on Eddie, analysis with big data on JASMIN, final analyses on desktop,

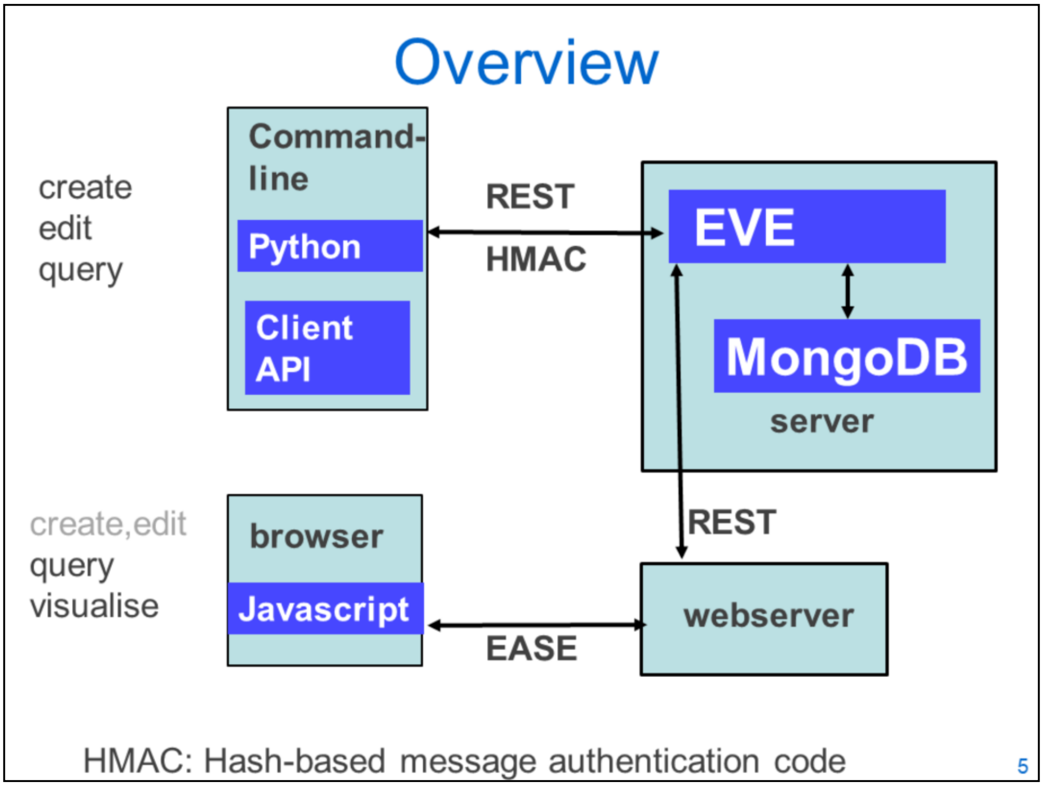
And research fieldwork.

Nothing is yet geo-specific.



GeosMeta phases

1. **Talked to research groups**
2. **Development by GeoSciences (and initially EPCC) of prototype that can be demo'd**
3. Establish live service as trial
4. Work with initial users
5. Extend functionality
   (no geo-specifics yet)



Overview

create
edit
query

Command-line

Python

Client API

REST

HMAC

EVE

MongoDB

server

create,edit
query
visualise

browser

Javascript

EASE

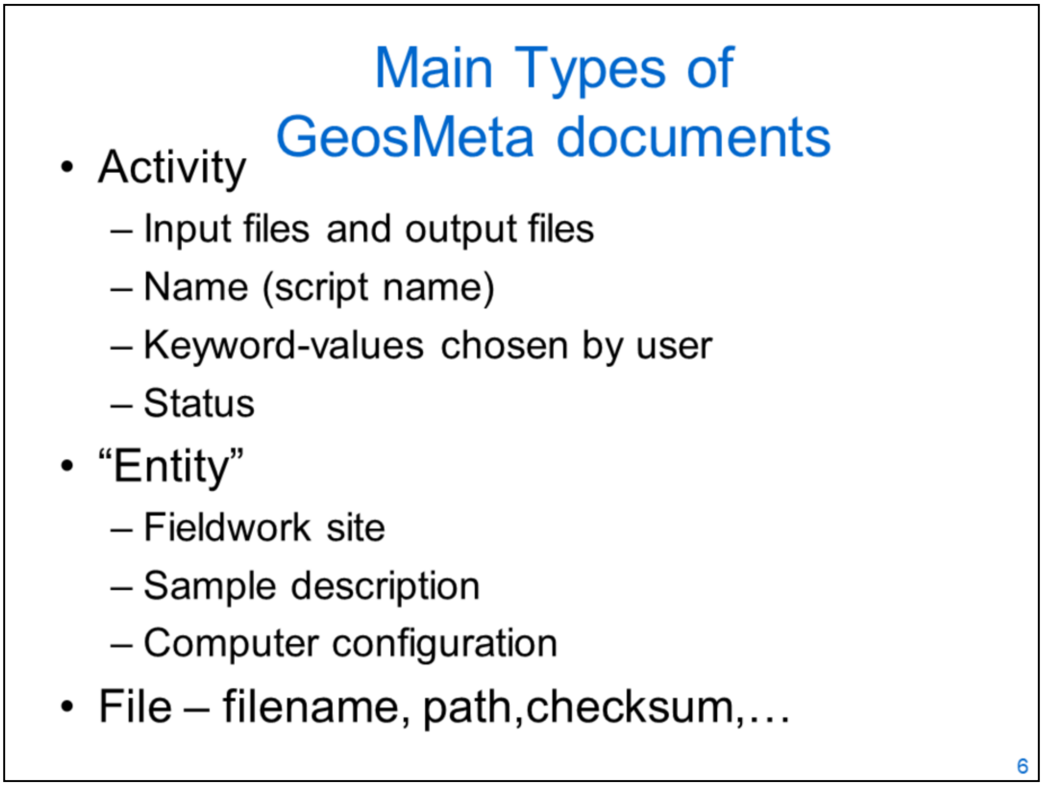
REST

webserver

HMAC: Hash-based message authentication code

Python client: uses HMAC, so can work from batch jobs also. After initial setup of a user, they have a secret managed like a private key – used for authentication.

Browser: uses EASE proxy server. EVE is extended to work with EASE proxy or HMAC



Main Types of GeosMeta documents

- Activity
  - Input files and output files
  - Name (script name)
  - Keyword-values chosen by user
  - Status
- “Entity”
  - Fieldwork site
  - Sample description
  - Computer configuration
- File – filename, path,checksum,…

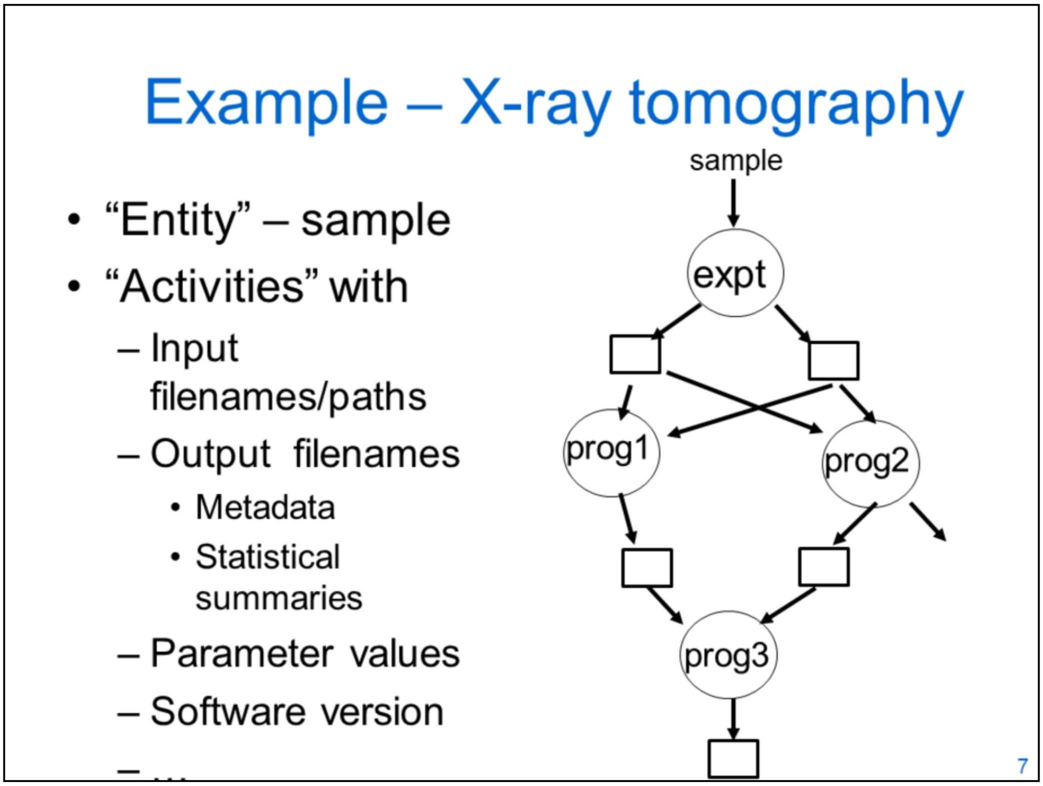
6

In practice entity is held in the “activity” document…. But the user might think in terms of activities that use/make files; files and associated metadata ; entities that hold useful stuff

File documents do not replicate the data in general: but as an example if a file has a global grid of temperatures the use might choose to hold the global mean, along with other attributes of the file.

The goal is that users can choose what they hold in the documents without needing to re-engineer the server.

They can easily test what works for them.



Example – X-ray tomography

- “Entity” – sample
- “Activities” with
  - Input filenames/paths
  - Output filenames
    - Metadata
    - Statistical summaries
  - Parameter values
  - Software version
  - …

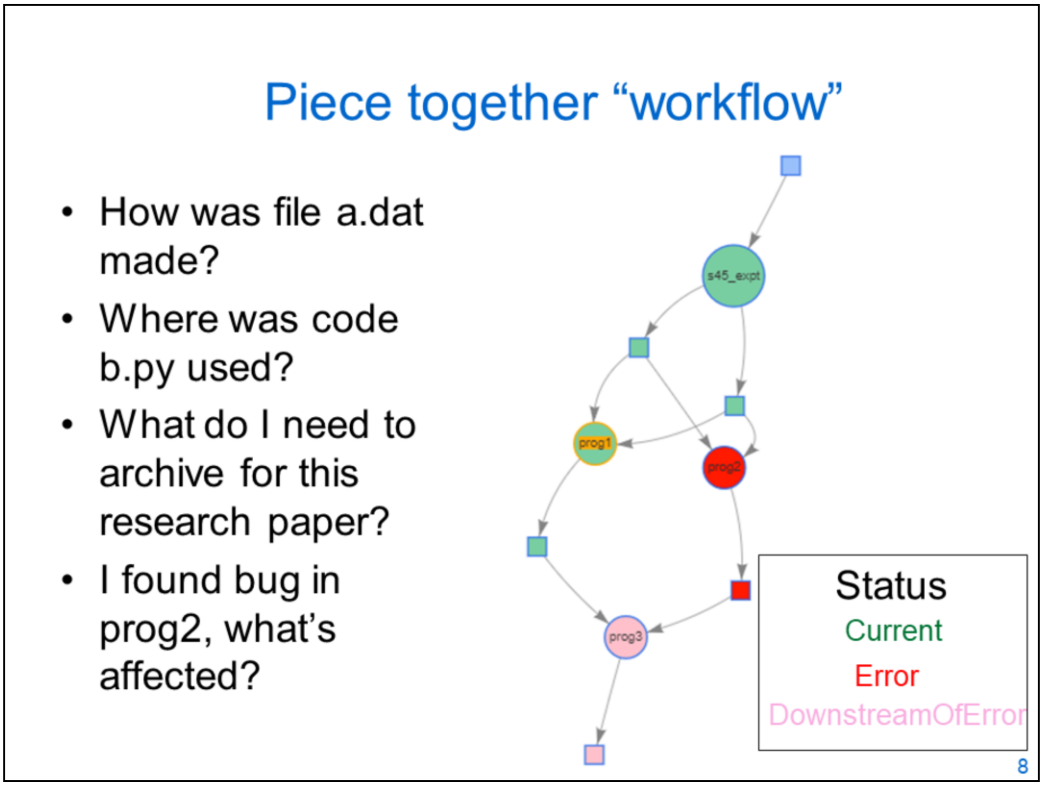
Squares represent files; circles activities,

Its not a “classic” workflow - but prog1 and prog2 might be run weeks apart, as the researcher refines their intentions.

The use of the filenames and paths permits GeosMeta to reassemble what was done with connections between activities (in effect scripts or programs)

Prog1, prog2 etc have to include a tiny bit of python either in the python or called from a bash script.

How this looks in the browser is on the next slide.



Piece together "workflow"

- How was file a.dat made?
- Where was code b.py used?
- What do I need to archive for this research paper?
- I found bug in prog2, what's affected?

Status

Current

Error

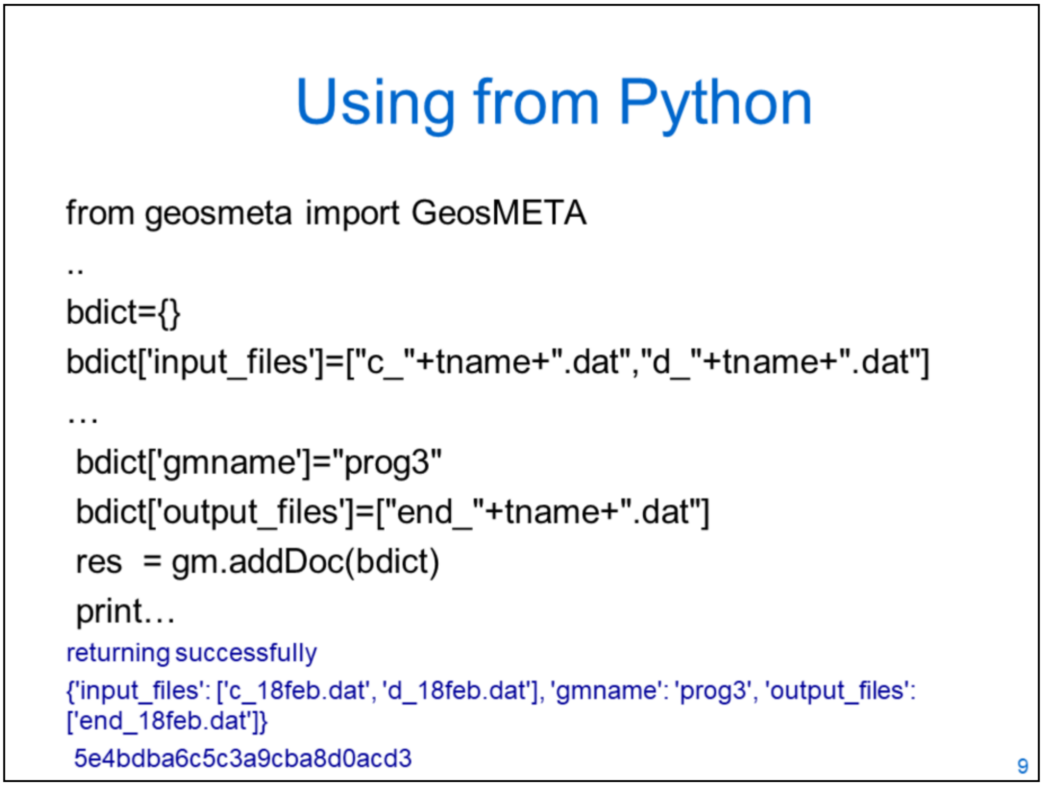
DownstreamOfError

8

So here, after prog2 was initially run, a bug was found.

The user reset the prog2 activity's status to "Error" and geosmeta then recognised that puts in doubt the prog3 step in the processing.

At archiving time, the user can ask how was this end-product made, gather the metadata for each upstream activity.



Using from Python

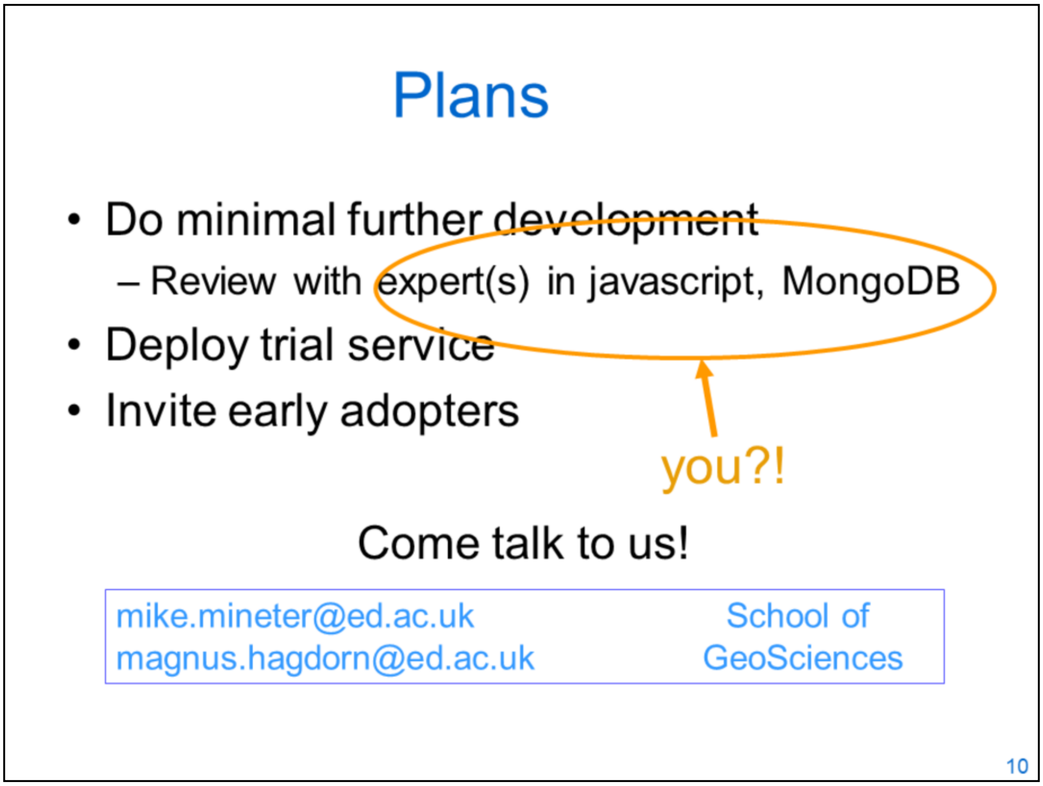
```
from geosmeta import GeosMETA
..
bdict={}
bdict['input_files']=["c_"+tname+".dat","d_"+tname+".dat"]
…
 bdict['gmname']="prog3"
 bdict['output_files']=["end_"+tname+".dat"]
 res  = gm.addDoc(bdict)
 print…
```

```
returning successfully
{'input_files': ['c_18feb.dat', 'd_18feb.dat'], 'gmname': 'prog3', 'output_files':
['end_18feb.dat']}
 5e4bdba6c5c3a9cba8d0acd3
```

9

An example I use in testing is shown here,

Tname is just a name given the test script so documents for that test are distinct.



Plans

- Do minimal further development
  - Review with expert(s) in javascript, MongoDB
- Deploy trial service
- Invite early adopters

you?!

Come talk to us!

mike.mineter@ed.ac.uk
magnus.hagdorn@ed.ac.uk

School of GeoSciences