The Sudamih Project

Supporting Data Management Infrastructure in the Humanities

Tuesday 11 May 2010

James A J Wilson
James.wilson@oucs.ox.ac.uk

SUPPORTING DATA MANAGEMENT
INFRASTRUCTURE IN THE HUMANITIES
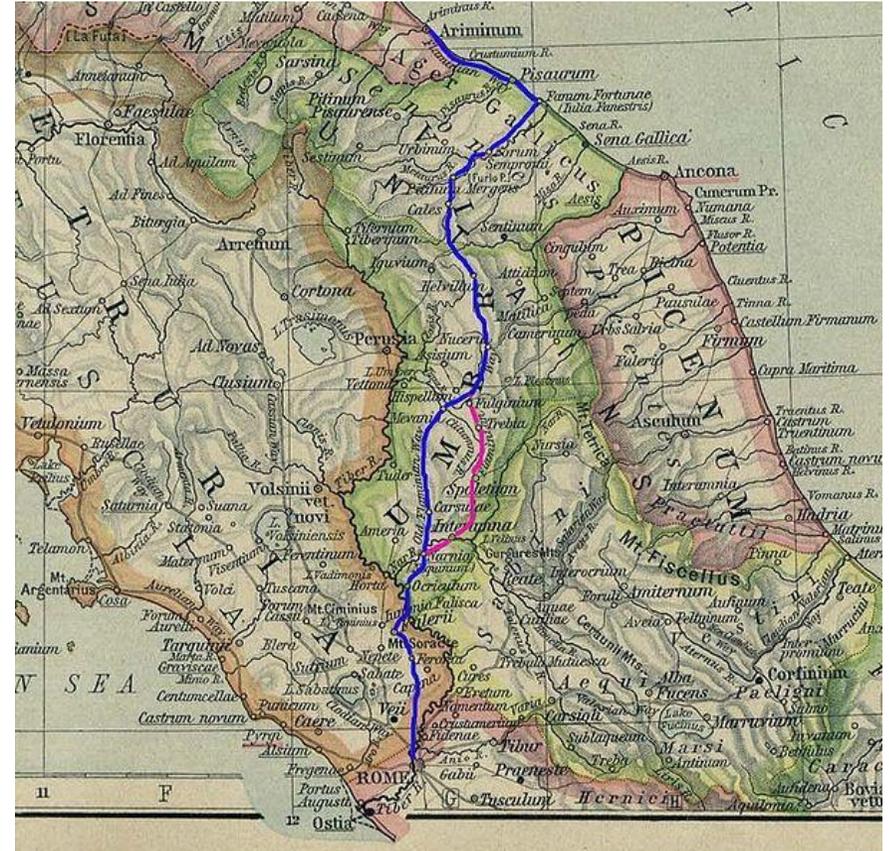(SUDAMIH)

JISC

UNIVERSITY OF OXFORD

Project Focus

- Understanding how scholars in the humanities manage the information they use in their research
    - Finding, storing, structuring, using, and re-using information

- Pilot data management training modules
    - Broad approach to data management
    - How can we improve existing practices?

- Pilot 'Database as a Service' (DaaS) system
    - What advantages can we bring to researchers through this?

- Cost models for data management services

Humanities Data – Example 1

Database of Ancient Cities

- Effectively a 'lone researcher' working for an Ancient History project that involves others

- Data stored in an Access database on his laptop

- Compiles information from Barrington Atlas, encyclopaedias, monographs, journal articles

- Records GIS references, names, dates, sources, evidence of economic activities, etc.

- Data not yet available for others to use
  - Wants to complete doctoral thesis first

- Doctoral study forming part of a

Humanities Data – Example 2

Media representations of Islamic security threats

- Multidisciplinary team of four researchers spanning humanities and social sciences

- Video recordings of television news broadcasts & transcriptions of these. Broadcasts from Britain, France, and Russia

- >1 TB, indexed in an XML directory

- Only *relevant* material indexed

- Four local copies of data & stored on University of Manchester servers

Humanities Data – Example 3

Organically evolved 'Database' of medieval songs

- Researcher began by using Endnote as simple bibliographical database. Over time has added new custom fields in order to describe medieval songs, such as

  - Composer, lyricist, rhyme scheme, number of lines, number of syllables, versification, and so forth

- Can now search for songs which share particular features

- Necessitated development of a standardised orthography for Middle French, personal to her system

  - i.e. Not familiar to other potential users

- Not familiar with database software

Characteristics of Humanities Data

- Long life-span

- Part of 'Life's Work'

- Compiled, not generated
    - May be from poetry, music, art, material objects, recordings of speech, news broadcasts, academic books and journal articles

- Unbounded / incomplete / inconsistent / interpreted

- May be very narrowly relevant to particular researchers

- May be intended for public; may be intended for personal use

Humanities data - Accessibility

- Where a public web interface is not envisaged as an output from the outset, there are problems sharing data:
  - It's messy
  - It's employs personal, idiosyncratic standards
  - It's partial and specific
  - It's existence is not widely known
  - Needs to milked for publications first
- However, humanities researchers are rarely opposed to sharing their data *in principle*.

Humanities Data - Storage

- Favourite storage medium – Laptop hard drive

- Favourite backing-up mechanism

    – External hard drive, every once in a while

- Frequently use more than one computer, with files transferred via memory stick

- Relatively little use of institutionally-provided storage

- Ignorant of, or confused by automatic back-up systems

- Don't overestimate researcher's awareness of centrally provided infrastructure

Data Management Concerns

- Occasional sense of foreboding regarding data accumulation
  - On top of things now, but problems in future?

- Concerns about speed of technological change, especially amongst those senior enough to have experienced it
  - Obsolescence of data formats

- Uncertainty about databases
  - Researchers often don't understand how databases work, when they are appropriate, and the kinds of output one could expect

- Enter, Sudamih!

Data Management Training Suggestions

"Training in ways to organize material would be useful – computer file structures, organizing paper notes, that sort of thing"

"Case studies and examples of what people have done in the past [to organize all their information]"

"Finding out how to connect pictures to searchable notes would be really useful."

"It might be useful to learn about specific bits of technology, such as scanner pens"

"A review of different software packages – an overview which covers their advantages and disadvantages and shows what they might be used for"

- Suggestions for Graduates included:

  – Good backing-up practices; recording your sources and what you've read; versioning; and just getting them to think about how they need to structure their information in advance

Data Management Training Content

"If a training course titled 'Data Management' were offered, most humanities students would consider it irrelevant to them" [Trainng Officer]

- Broad Data:
  - Organising computer files; backing up; versioning; managing email; linking notes to content; long-term curation issues; keeping track of sources

- Narrow Data:
  - Which type of software is best fits your needs?; Structuring data in relational databases; querying and retreiving information; long-term curation – data formats and migration issues; using the DaaS

- Surgery service for technical project funding bids

Data Management Training Approach

"Most people are so inundated with opportunities to attend training and conferences and workshops that they don't have time to take up many of them. People tend not to worry about data management until it becomes an issue and there's something specific they need to do, but even then the usual attitude these days is to try to work it out for yourself on the basis of what you already know" [Music Faculty Lecturer]
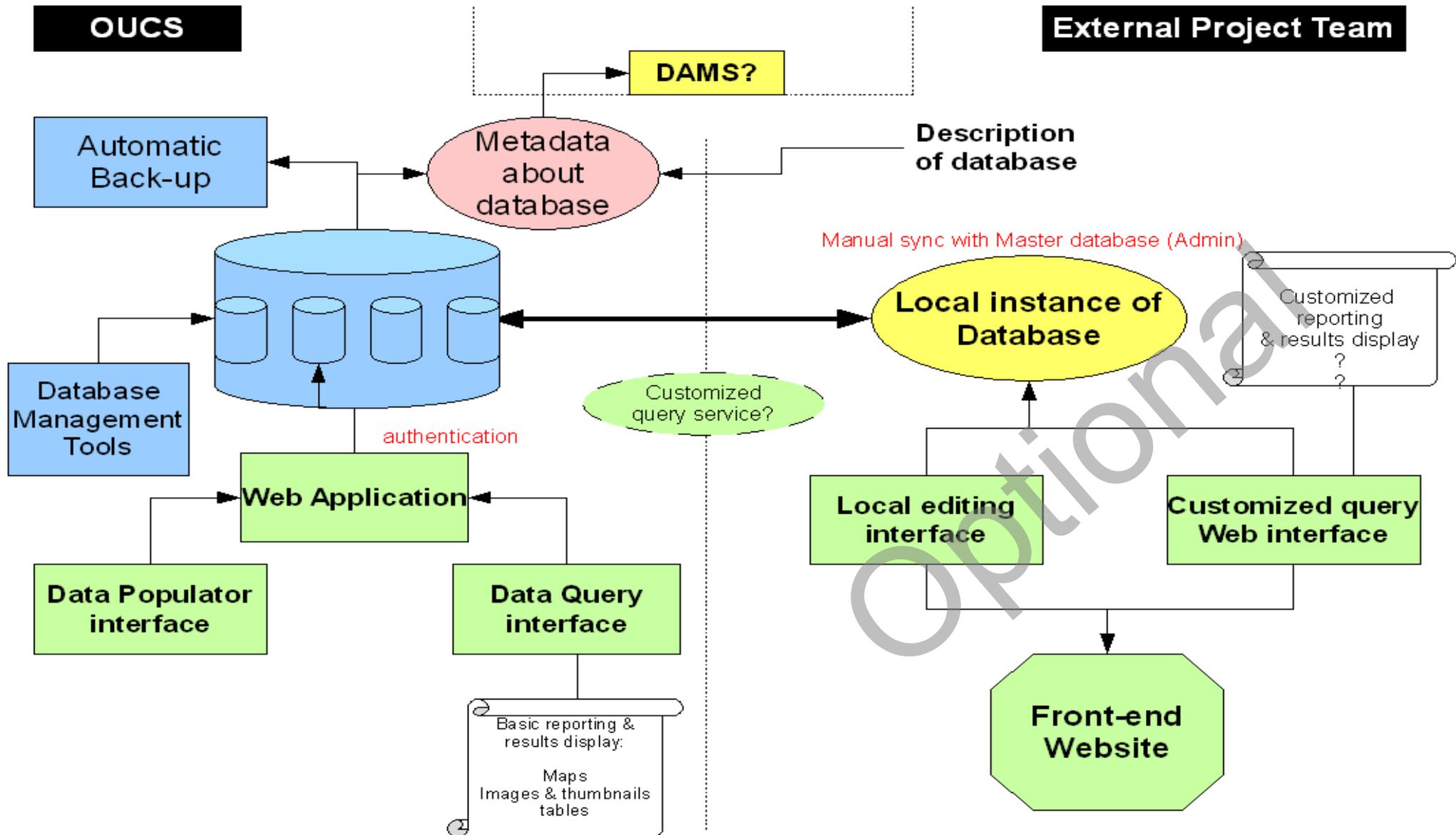
- Recognized need, but may be a 'hard sell'?
    - Identify actual research problems faced, don't sell it as generic skills training
    - Employ a mixture of face-to-face courses with online content to supplement
    - Get data management training into existing sessions if possible
    - Make it compulsory if possible
    - Get graduate students early, but not before they have some sense of the need – after 6 months or a year
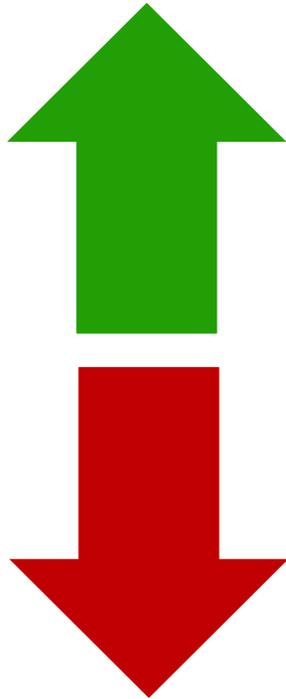
Database as a Service

- Benefits of central database service

    - Regular back up

    - Managed metadata

    - Integration into rediscovery services

- User requirements identified:

    - Ability to input and search text in non-Roman alphabets

    - Multiple media types [pilot will cover text, image, and geospatial data]

    - Fine-grained access and editing controls

    - (customisable) Web interface

    - Searchable in many different ways

    - Linking data to research outcomes

Database as a Service - Architecture

Sudamih DaaS Proposed Architecture

OUCS

External Project Team

DAMS?

Metadata about database

Description of database

Automatic Back-up

Manual sync with Master database (Admin)

Local instance of Database

Customized reporting & results display ? ?

Database Management Tools

Customized query service?

authentication

Web Application

Local editing interface

Customized query Web interface

Data Populator interface

Data Query interface

Basic reporting & results display:

Maps
Images & thumbnails tables

Front-end Website

Optional

Trends in Humanities Research

Collaborative Projects

Short-term Projects

Database Projects

Specific Doctoral Projects / Postdocs

'Lone Researcher'

- Changes driven partly by funding opportunities
- Be wary, however. Trends can change & backlashes begin

Thanks!

Any Questions?

Contact me at james.wilson@oucs.ox.ac.uk