Data Management in the Humanities

-The Sudamih Project

Friday 16 July 2010

Dr. James A J Wilson, Oxford University Computing Services
James.wilson@oucs.ox.ac.uk

SUPPORTING DATA MANAGEMENT
INFRASTRUCTURE IN THE HUMANITIES
(SUDAMIH)

Environmental research data

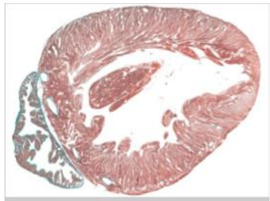
“CRU accepts with hindsight that they should have devoted more attention in the past to archiving data” ...

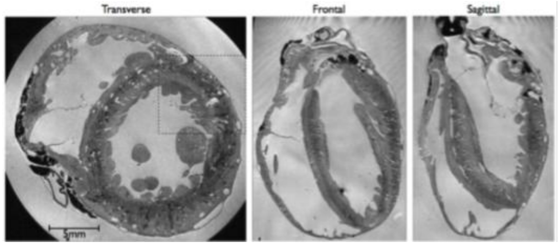
“We saw no evidence of any deliberate scientific malpractice in any of the work of the Climatic Research Unit ... Rather we found a small group of dedicated if slightly disorganised researchers who were ill prepared for being the focus of public attention. As with many small research groups their internal procedures were rather informal.”

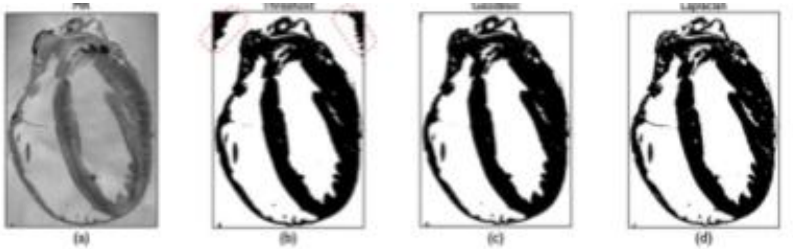
Report of the International Panel set up by the University of East Anglis to examine the research of the Climatic Research Unit

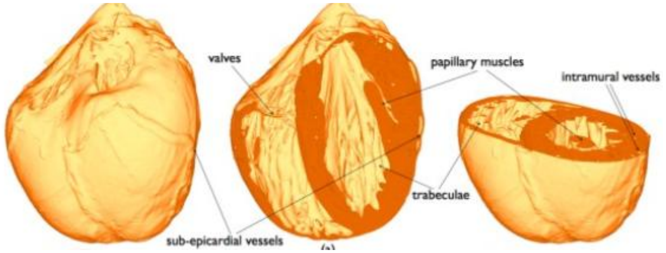
Scientific research data

- 3D Heart Project at Oxford
  - Very high resolution images of hearts are used to create computer models upon which in-silico experiments can be conducted
  - But imagining and staining techniques are not flawless
  - Cost estimated at £668,000
  - Data should be findable, reusable, and verifiable









Research Data and Technology

- Improvements in data-gathering technologies and processes are enabling new types of research
- ‘Data-driven research’ is now a possibility

> “Pioneering archives ... have demonstrated just how powerful such legacy data sets can be for generating new discoveries – especially when data are combined ... and analysed in ways that the original researchers could not have anticipated”
>
> Editorial from ‘Nature’ 461, 10 September 2009

- But technology does not just help research, it places new demands upon researchers and institutions

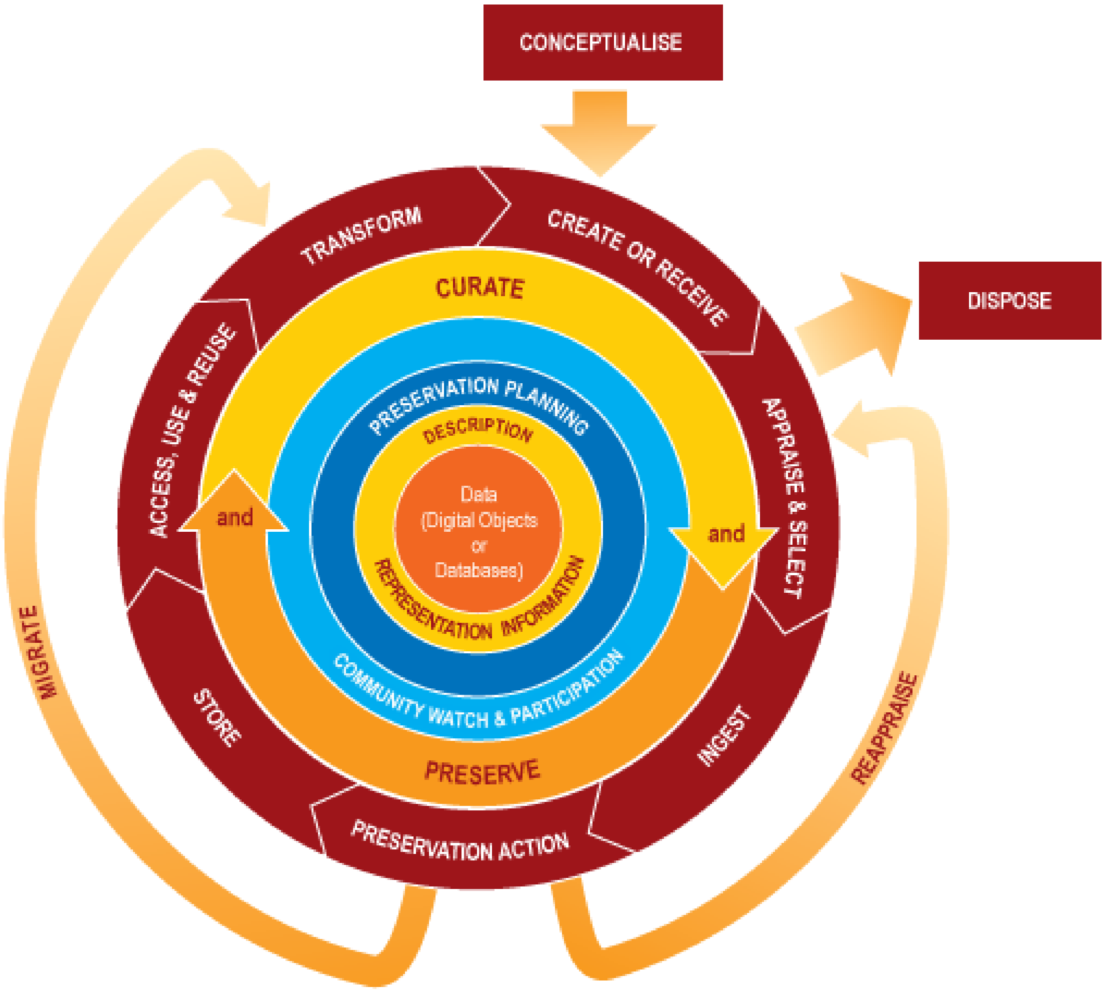
But data in the humanities is different, isn’t it?

- yes, and no.
- Hand-crafted, not captured
- Frequently lacks uniformity and consistency

BUT

- Shares many of the same issues
  - Accessibility
  - Re-usability
  - Must be understood to be properly used
  - Expensive, and value should be maximised
  - Needs to be preserved and curated

The data curation lifecycle



The Sudamih Project

Supporting Data Management Infrastructure in the Humanities

- 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
  - 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

What constitutes data?

“The term ‘data’ may be problematic, as lots of humanities students may react that they don’t really work with data, because this will make them think of big databases”

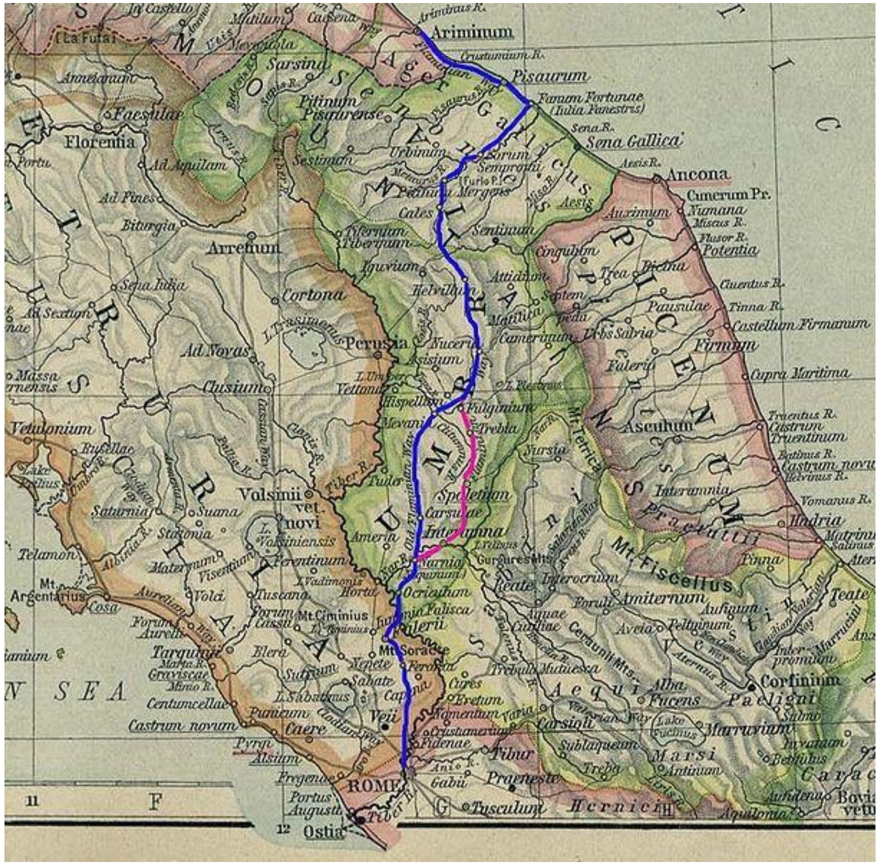
[Humanities Training Officer]

- ‘Broad’ data – all the information materials that go into the research process, leading to research outputs
- ‘Narrow’ data – information in a structured electronic format that can be processed by computer
- Boundaries not necessarily clear in humanities scholarship

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



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

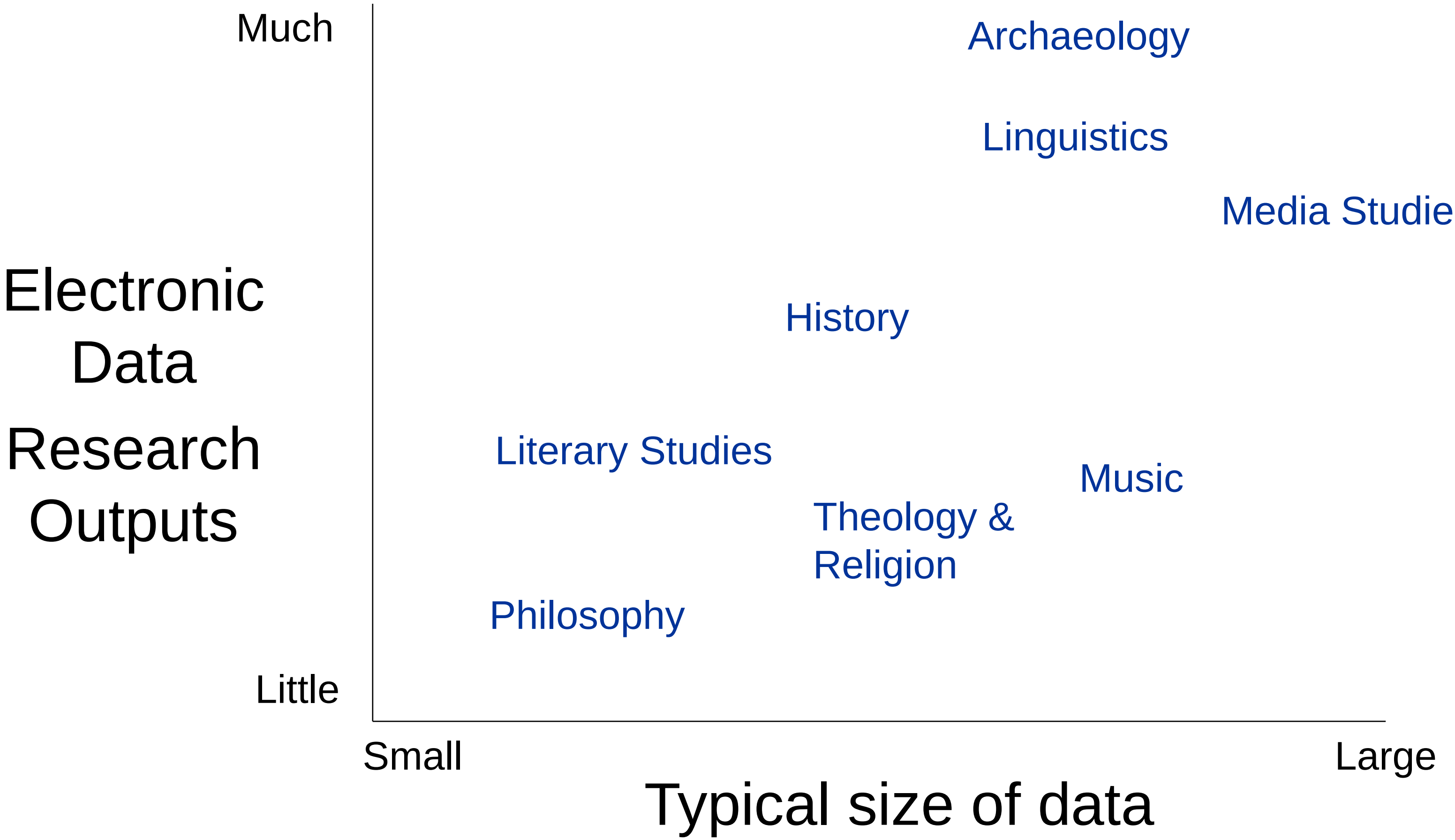
Characteristics of Humanities Data

- 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 seem/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*.

Indicative disciplinary differences



Much

Electronic Data Research Outputs

Little

Small

Large

Typical size of data

Archaeology

Linguistics

Media Studies

History

Literary Studies

Music

Theology & Religion

Philosophy

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

The DaaS Proposal

- A web-based system that will enable researchers to quickly and simply build a relational database which will be
  - centrally hosted and maintained
  - regularly backed-up
  - easily shared with collaborators and the public
  - capable of dealing with text, images, and geospatial data
  - managed metadata
  - integration into discovery services
- Open Source, so that other Institutions can modify and deploy the software

Response to the DaaS

- Positive
  - Several interviewees considering database projects
  - Researchers already involved with mature online databases not so inclined to move, however
  - Potential solution for Web resources lacking stable, long-term hosting arrangements
  - Good way to ‘open up’ inaccessible datasets

DaaS Requirements

“Flexibility is important – people need to have a variety of fields”

“It would have to be customisable”

“Many of our databases contain GIS data in one form or another, and need to be able to interface with mapping software”

“We’d like to be able to link data with research outputs”

“It would need to be usable with a range of languages – with full support for diacritics and other alphabets: Cyrillic and so on”

“Version control is important, because people make mistakes, and you also need the ability to set up editing permissions at different levels”

“If it’s an online system, it would be essential for people to be able to download the data and work with it using desktop applications”

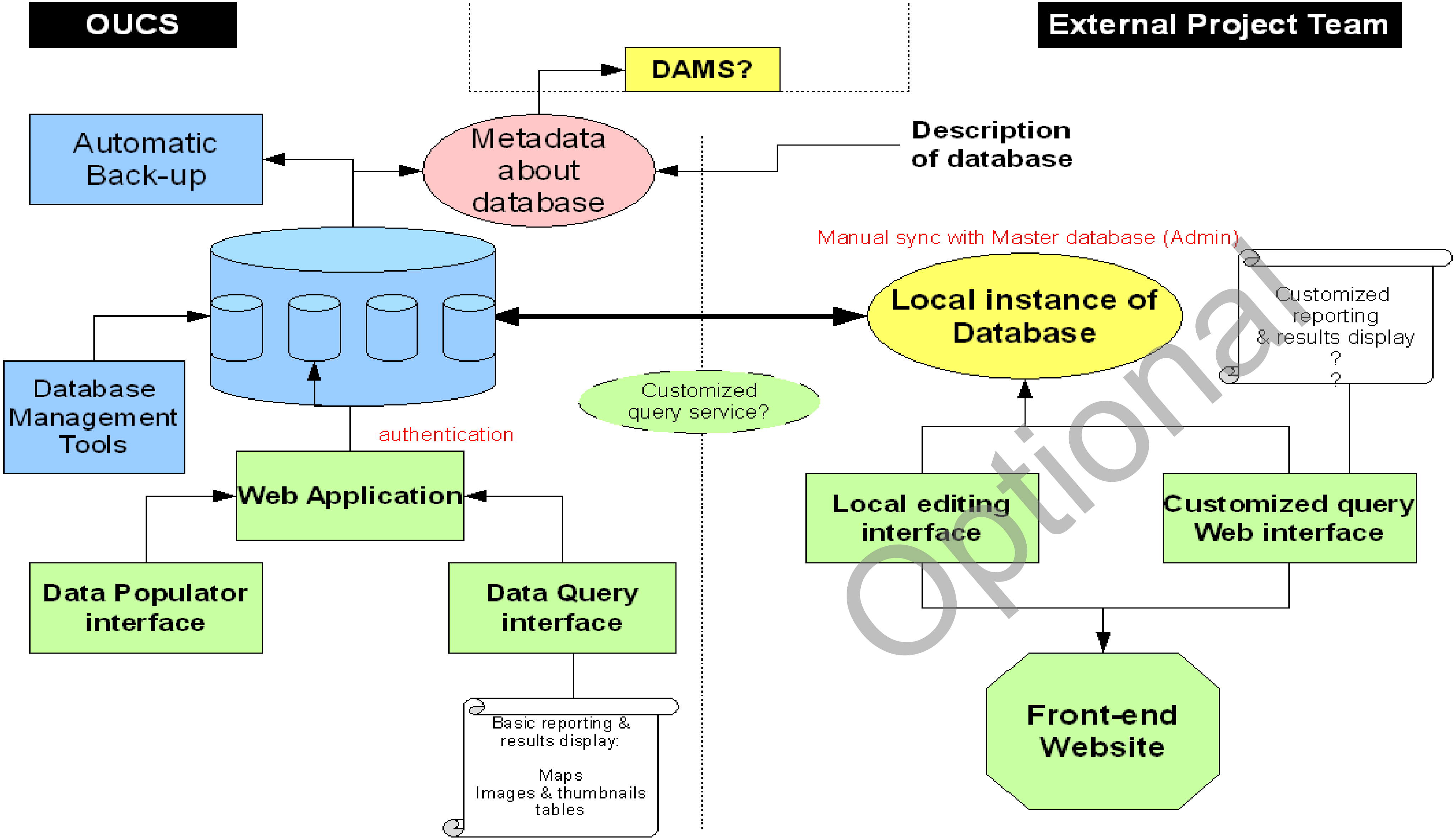
“It would be important to have the ability to import existing Access databases”

DaaS Reservations

- When you store your data online, if the system goes down you are dependent on somebody else to fix it
- Putting research on a publicly findable database would be an open invitation to be contacted and questioned by everybody
- Other concerns
  - Security
  - Speed (of accessing and editing)
  - Permanence
  - Cost (project funding does not last beyond the project)

Database as a Service - Architecture

Sudamih DaaS Proposed Architecture



OUCS

External Project Team

DAMS?

Automatic Back-up

Metadata about database

Description of database

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

Data Ownership

- Researchers think that they own their data
- Oxford, however, thinks otherwise

5. 1. The University claims ownership of all intellectual property specified in section 6 of this statute which is devised, made, or created:

(a) by persons employed by the University in the course of their employment;

(b) by student members in the course of or incidentally to their studies; ...

6. The intellectual property of which ownership is claimed under section 5 (1) of this statute comprises:

(1) works generated by computer hardware or software owned or operated by the University;

(2) works created with the aid of university facilities including (by way of example only) films, videos, photographs, multimedia works, typographic arrangements, and field and laboratory notebooks; ...

(6) databases, computer software, firmware, courseware, and related material not within (1), (2), (3), (4), or (5), but only if they may reasonably be considered to possess commercial potential;

- Potential problem for trust in DaaS

Technical Training in the Humanities

- Many humanities researchers uncertain about opportunities presented by technology

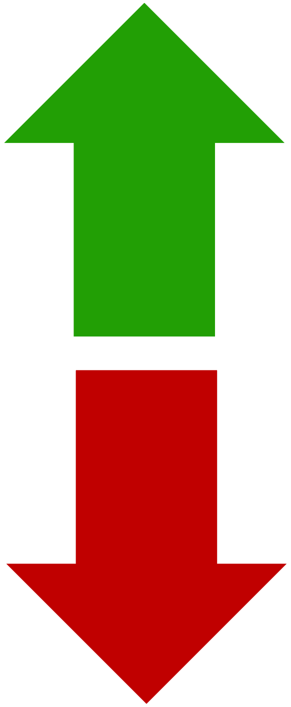
“People often start off by creating massive Word or Excel files, then two years down the line they have to change everything as they realize that the best way to manage the material is in a database”

  - Many wanted an overview of software and tools,

    “somewhere historians could get advice about which software package is best suited to their project”

  - Help organising notes and connecting related source material
  - Senior researchers wanted advice on making technical bids
  - Technical advisory service would be popular
  - Lack of knowledge of what University already provides

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

Future Challenges

- Improved understanding of technology amongst researchers
- Improved sense of responsibilities regarding creating and curating research data
- Improved infrastructure for research
- Better framework for collaboration
- Communication between humanities researchers and technologists

Thanks!

Any Questions?

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