The Sudamih Project

Developing Data Infrastructure for the Humanities: Attitudes and Requirements

Tuesday 11 May 2010

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

JISC

SUPPORTING DATA MANAGEMENT INFRASTRUCTURE IN THE HUMANITIES (SUDAMIH)

UNIVERSITY OF OXFORD

The Sudamih Project

Supporting Data Management Infrastructure in the Humanities

- Premise: data is valuable!
- Understand current practices regarding humanities data
- Opportunities to improve current practices via development of university infrastructure
- Specific outputs:
  - Database as a Service (DaaS) system
  - Data Management training modules

Data

- Data? What 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]

- Two definitions of data
  - broad

'A thing given or granted; something known or assumed as fact, and made the basis of reasoning or calculation' – "Out of what Data arises the knowledge? (1691)" [OED]

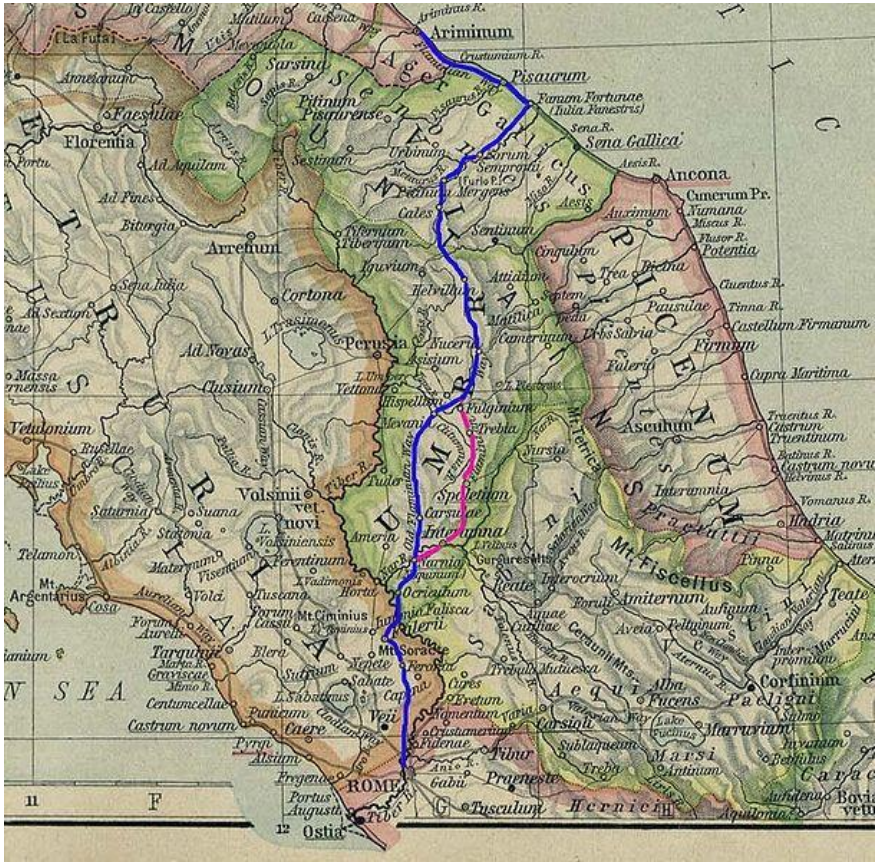
  - narrow

'The quantities, characters, or symbols on which operations are performed by computers and other automatic equipment' (since 1946) [OED]

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

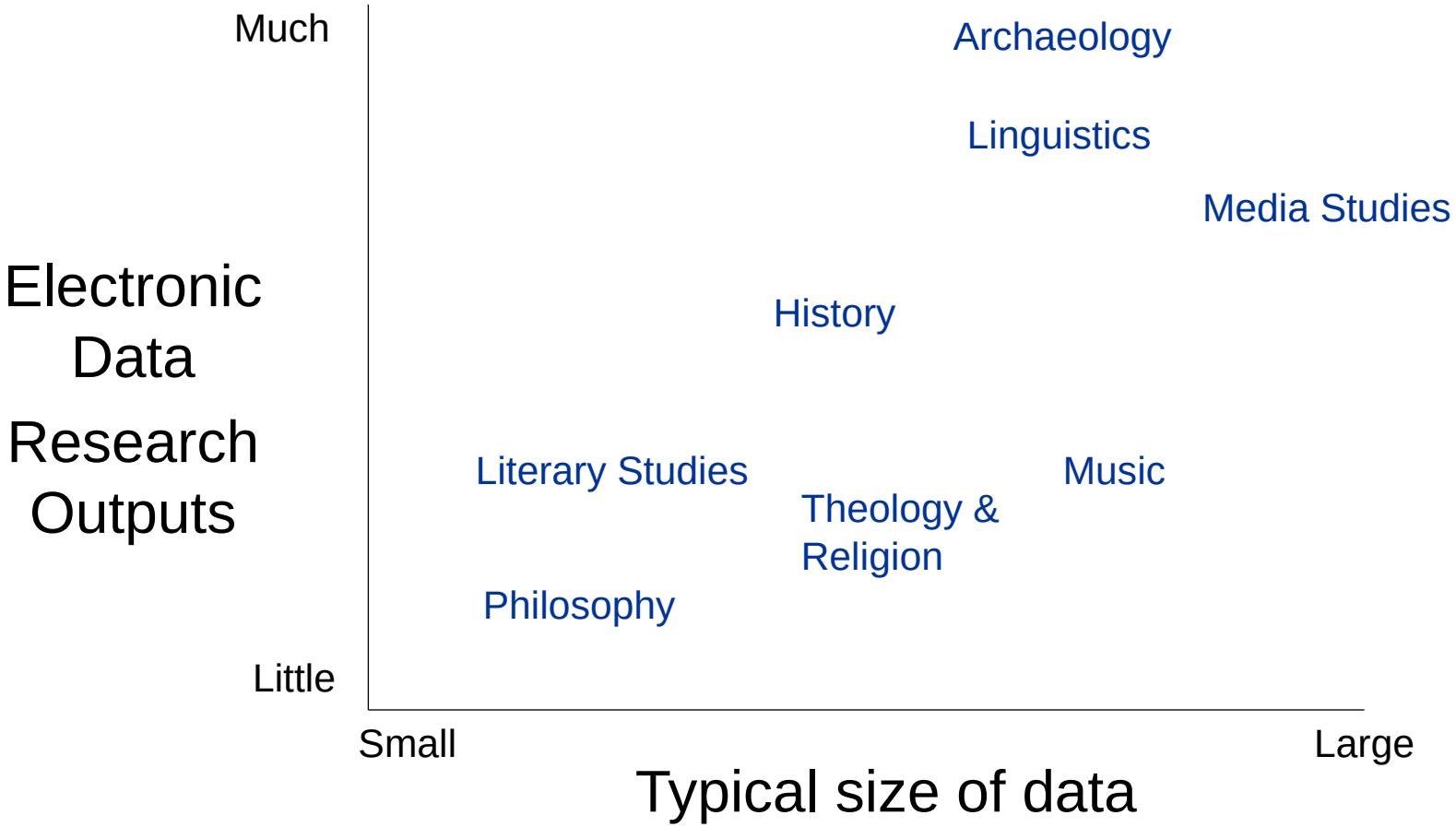
Organically evolved ‘Database’ of medieval songs

- 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

Data in the Humanities

- Long life-span (life's work nature of humanities)
  - Research tends to 'evolve' over time
- Compiled, not usually created
- May be found in 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
- Some data intended for public dissemination; some for private research and consequently hard to discover

Disciplinary differences



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*.
- Journals do not generally insist on data publication

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

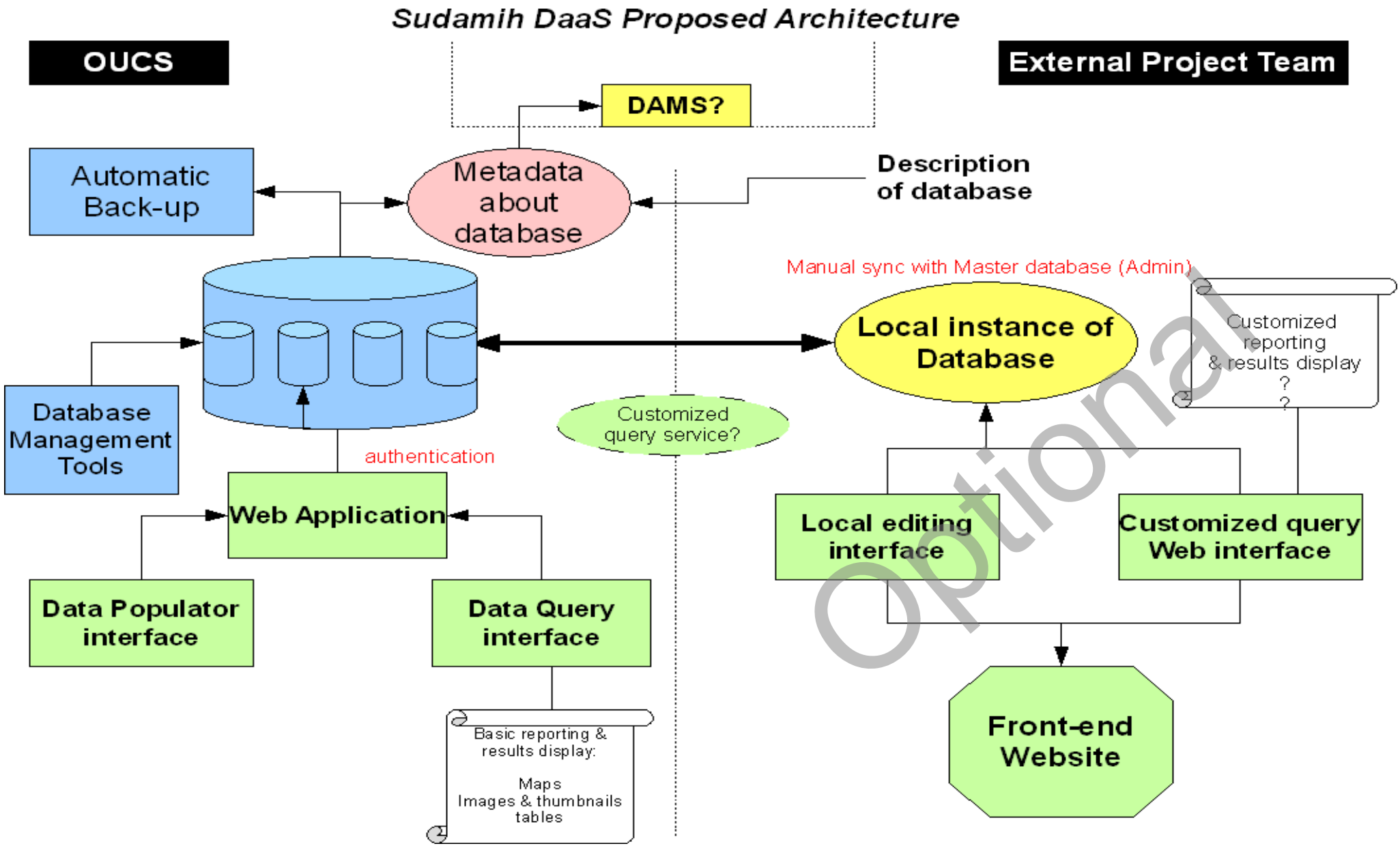
Data Management Concerns

- 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!

Database as a Service

- 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
  - Linking data to research outcomes
- Additional benefits of central database service
  - Regular back up
  - Managed metadata
  - Integration into rediscovery services

Database as a Service - Architecture



Data Management Training

- Or ‘Information Management Training’
  - Some interpreted this in the broad sense: how can I organise my information (electronic and hardcopy) so that I can find things quickly when I need them?
  - Some interpreted this in a narrow sense: what software tools do what? How should I structure a database?

Existing Data Management Training

- Current provision:
  - Some database software courses provided by computing services
  - Faculty-led academic practice training
  - Divisional skills and career training
  - Nothing specifically on data management
- The situation tended to be similar in other institutions that our interviewees had experience of.

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."

"Training would be better targeted at learning how to use specific tools rather than acquiring general skills"

"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 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

Data Management Training Content

- Broad Data:

"Learn how to organise your research information so that you can find things when you need them!"

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

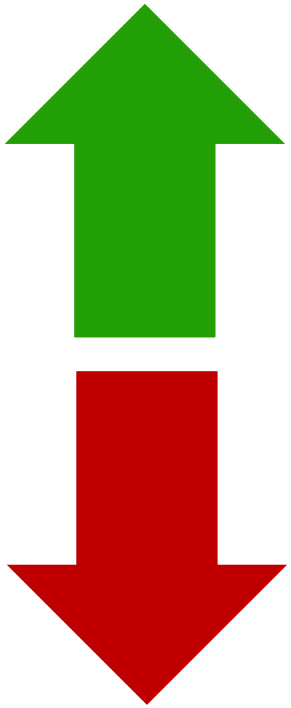
- Narrow Data:

"Learn how to structure your research information using spreadsheets, databases, bibliographic software and other tools!"

  - 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

Trends in Humanities Research



Collaborative Projects

Short-term Projects

Database Projects

Specific Doctoral Projects / Postdocs

‘Lone Researcher’

- Changes driven by mostly by funders – AHRC and JISC
- Be wary, however. Trends can change & backlashes begin

Thanks!

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