

# e-Nabling Data

---

## Some observations

Samuelle Carlson\*

Dawn Nafus†

Ben Anderson\*

\*Chimera, University of Essex

†Intel Corp



# The Menu

---

- Background
- A tour of the results
  - All data are not created equal
  - Explicit and Implicit forms of knowledge
  - Disconnecting data from people
  - Fragmentation
- Some conclusions

# Background

---

- The data rush
  - The coming deluge (Hay & Trefethen 2003)
  - Integrating heterogeneous data sources
  - Economies of re-use
- The collaboration rush
  - Inter-disciplinary innovation
  - 'Follow that problem'
  - Communities of practice and 'competence on demand'
  - Economies of scale
- The gold rush
  - e-Science, e-Social Science
  - £98 million 01-04 (DTI, Science Budget 01-04) and growing

# Problems

---

- Dis-embedding knowledge?
  - To create data for re-users who are distributed
    - Geographically (here <-> there)
    - Temporally (now <-> then)
    - Intellectually (them <-> us)
- Exchange (and reciprocity)?
  - Incentives? Paybacks?
- The gap of rhetoric?
  - Between what is, what is imagined and what may be

# Our approach

---

- ESRC NCeSS Small Grant
  - 12 months: Feb 04 - March 05
- 4 case studies of practice
  - **SkyProject** - a national collaboration of scientists focused on the use of data from a few key observation sites,
  - **SurveyProject** - a large scale quantitative social science survey data collection project which has data re-use enshrined in its objectives,
  - **CurationProject** - a collaboration between an anthropology department and an anthropological museum
  - **AnthroProject** - a long term research project driven by two anthropologists and implemented through a series of research student projects

# Contrasts

---

- **SkyProject**

- ‘big science’, very high use of e-Tech,
- many developers, few users
- data as numbers, no data protection issues
- ‘hard-pure’ (Becher, 1987)

- **SurveyProject**

- Quantitative social science, lower use of e-Tech,
- few developers, many users
- data as numbers, data protection issues
- ‘soft-applied’

- **Curation/AnthroProjects**

- Qualitative social science, lower/experimental use of e-Tech,
- few developers, many users
- seriously heterogeneous data, data protection issues
- ‘soft-pure’

- **‘hard-applied’**

- See for example Hine (2005), Fry (2006)

# In the field

---

- Interviews with key staff
  - Snowball recruitment within case studies
  - Mapping/representations of relationships
- Participant observation of
  - Research cycle, meetings, conferences, seminars
  - online developer sessions/meetings (SkyProject)
- Analysis of
  - Web-based documentation (SurveyProject)
  - wiki and IM logs (SkyProject)
- Looking for:
  - Use of e-Tech
  - Adaptations and challenges

# A tour of the results

---

- All data are not created equal
- Explicit and implicit forms of knowledge
- Disconnecting data from people
- Fragmentation

# Born digital

---

- SkyProject, SurveyProject
  - Inherently digital data
- CurationProject, AnthroProject
  - Heterogeneous, historical multiple media
- When data is not inherently digital

QuickTime™ and a  
TIFF (LZW) decompressor  
are needed to see this picture.

# Born digital

---

- The ESRC *demand*s that all new data they fund is deposited at the UKDA
  - But in practice - qualitative deposit = interview transcripts/tapes
- Dangers:
  - Easily digitised or numerical and text-based raw material may be favored
  - Qualitative studies may (have to) adopt quantitative approaches and sacrifice specificity
  - Digitisation may lead to 'information loss' [CurationProject is keeping the cards]
  - Cost of maintenance does not decline - physical collections remain

# A tour of the results

---

- All data are not created equal
- **Explicit and implicit forms of knowledge**
- Disconnecting data from people
- Fragmentation

# Reusable materials must be...

---

- Easy to disseminate
  - Transportable
  - Intelligible
- The necessarily implies
  - Codification
  - Explicit documentation
  - Reduction & representation
- All our case studies attempted to do this in different ways

# Representation Practices

---

- SkyProject/StarProject:
  - Already externalise knowledge, established standards and good practice
  - Range of representation practices (including numbers, images, models)
  - Code, algebra and statistics as stable language shared by a research community
- CurationProject/AnthroProject:
  - Primary data and methods remain tacit knowledge (craft-like)
  - Researcher as instrument and gatekeepers
  - No incentives, indeed strong resistance to formalise and externalise methods
- A quant/qual divide?

# Representation Practices

---

- SkyProject/StarProject:
  - Already externalise knowledge, established standards and good practice
  - Range of representation practices (including numbers, images, models)
  - Code, algebra and statistics as stable language shared by a research community
- CurationProject/AnthroProject:
  - Primary data and methods remain tacit knowledge (craft-like)
  - Researcher as instrument and gatekeepers
  - No incentives, indeed strong resistance to formalise and externalise methods
- A quant/qual divide?
- In fact there is a continuum between the two
  - And there is tacit & craft knowledge EVERYWHERE as we will see

# A tour of the results

---

- All data are not created equal
- Explicit and implicit forms of knowledge
- **Disconnecting data from people**
- Fragmentation

# Data and claims

---

- Appropriate uses

- People have claims on the data (anthropology, medicine...) and some data is

- What if

"We have a photo... showing relationships among people that these same people deny having had..."

"the museum is a neutral place for some 'dangerous' objects"

- Dangers.

- eScience might favor a restricted type of materials: those that can be anonymised and on which there are few claims

# Context and cooking

---

- Trust

- How to trust the quality/relevance of data when disconnected from the original
- Systems of error

SkyProject scientists need to know:  
atmospheric conditions, which detector,  
quality of instrument - imperfections,  
calibration algorithms,  
filters used etc etc

Similar issues for SurveyProject users  
... and traceable

# Cooking and context trails

---

- For SurveyProject it is built in to the documentation processes
    - but still 'information' gets lost in translation
  - For SkyProject it is not built in to the documentation processes
  - For SkyProject it is not built in to the documentation processes
- "all you can rely on is the person to have ethics; it is the person who makes decisions and selections every day."*
- "There is a huge amount of trust"*
- ...specificities not 'true vs false' nor
- ...to fully specify context is to reproduce the world!
- Yet there is significant tacit knowledge and trust to be seen on the 'hard' side
    - full specifying context for SurveyProject would also reproduce the world

# A tour of the results

---

- All data are not created equal
- Explicit and implicit forms of knowledge
- Disconnecting data from people
- **Fragmentation**

# Micro-contributions

---

- Micro-modularisation of outputs
  - No longer 'papers', 'grants', 'datasets', 'monographs'
  - Code-fixes, transforms, coding schemes, derived data, interpretations, annotations, 'virtual' datasets
  - User-contributions (CurationProject), OSS model
- Traces
  - Respondents realised that their microcontributions can be tracked, summed and accounted
  - *Ambivalence*
- Effects
  - Short-cuts and speeds 'publication' - IPR issues
  - Foregrounds 'junior' staff, but do incentives/rewards follow? Reveals methods of 'senior' staff!
  - New reputations systems emerge - how to enable peer review in microcontributions?

# The Menu

---

- Background
- A tour of the results
  - All data are not created equal
  - Explicit and Implicit forms of knowledge
  - Disconnecting data from people
  - Fragmentation
- **Some conclusions**

# Digital Limits

---

- Issues of digitisation
  - Historical and heterogeneous data
- Issues of claims
  - Revelatory of collections - another Elgin Marbles?
- Issues of conservation
  - Who to ask?
  - How to predict future use purpose?
- Issues of anonymisation
  - Especially when identity is inscribed in the data
- Systematic constraints on what is e-enabled
  - and so what is available for current and future e-science

**Qualitative disciplines  
may  
hit these limits first**

# Provenance

---

- In all cases
  - No data is self-contained and self-explanatory
  - Complementary external information needed before understanding and trust
- Data is always cooked
  - For re-use every detail of the cookery needs to be revealed
  - And not all sciences are adept at doing this

# Costs

---

- CurationProject still has physical collections
  - And also now has to support databases
  - These are all costs
- How can these cost levels be sustained?

# The Selfish Scientist

---

- Why do scientists collaborate?

- Because...

- 'there is competition; people are often interested in the same thing – like Papparazzi who would like a better view. The tradition ... is very personal: if you have an interest, you apply for instrument time and you don't want somebody to pick it up from repositories. The repository keeps it for a time for the one who observed it.'

- 'they are solitary hunters; they don't hunt as a pack. Alliances will form and break-up. They won't even leave trails of what they're researching: they will go somewhere and then somewhere else just to erase the track (...) they don't want an audit trail until they publish'.

aspirations...? 'standing bodies'

# Thank you

---

- [benander@essex.ac.uk](mailto:benander@essex.ac.uk)
- <http://www.essex.ac.uk/chimera/projects/edkm/>