Why preserving databases matters and why it is harder than it sounds

Matthew Woollard
Director, UK Data Service

E-ARK / DPC / UK Data Service Workshop
University of Portsmouth
19 February 2015
UPDATE PROGRAMME
SET TITLE = "Why curating relational databases is different and why it’s more of a philosophical matter than one might immediately think”
WHERE TITLE = "Why preserving databases matters and why it is harder than it sounds”
<table>
<thead>
<tr>
<th>ID</th>
<th>COURSE_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>CS101</td>
</tr>
<tr>
<td>27</td>
<td>CS101</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COURSE_ID</th>
<th>FEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS101</td>
<td>£200</td>
</tr>
<tr>
<td>CS102</td>
<td>£150</td>
</tr>
</tbody>
</table>
ESRC Big Data Network

- Phase 1: Administrative Data Research Network

- Phase 2: Business and Local Government Research Centres
  - Urban Big Data Centre (University of Glasgow)
  - Business and Local Government Data Research Centre: (University of Essex)
  - Consumer Data Research Centre (University of Leeds and University College London)

- Phase 3: Third Sector and Social Media Data
UK Data Service: “Big Data Network Support”

- Pre-ingest / Collections Development
- Ingest
- Data Curation
- Access
- Training / Capacity Building
- Benefits Framework
- Public / Researcher Engagement
- Planning / Protocol Development
- Sustainability
Big Data Network Support: Preservation

• Capacity building within the UK Data Service
  • Research project --- archiving and providing access to administrative and transactional data
  • Research project --- archiving and providing access to social media
Operational databases --- states

- Operational
  - Current use version
- Reference
  - Non-operational snapshot
- Archive
  - Version maintained for compliance and business protection

(Source: Mullins (2007))
Operational databases --- states (2)

- Operational
  - Current use version
- Reference
  - Non-operational snapshot (contender for SIP)
- Compliance
  - Version constructed from non-operational records (i.e., not expected to be referenced again) (contender for SIP)
- Preservation
  - Bit-stream
  - Primary migration
Questions

• Does this model seem ring true?
• And, if yes, which is the best route to:
  • Reliability
  • Integrity
  • and Usability