MODELLING DATA ENVIRONMENTS WITHIN PROV TO ASSIST ANONYMISATION DECISION-MAKING

Muhammad Aslam Jarwar*, Mark Elliot*, Age Chapamn**, Fatemeh Raji**

*Centre for Digital Trust and Society, University of Manchester
**School of Electronics and Computer Science, University of Southampton
• Introduce the ADF
• The criticality of Provenance
• PROV and how to represent Data Environments
THE ADF

- Output of the UK Anonymisation Network
  - Collaboration of:
    - University of Southampton
    - University of Manchester
    - Office for National Statistics
    - Open Data Institute
    - UK Information Commissioner's Office

- Born out of work in the early 2000’s which led to the Data Environment Analysis Service with the Office for National Statistics.
HOW THE ADF WAS CREATED

- Four year process 2012-2016
- Series of multi stakeholder workshops
  - Answer the question what is anonymisation?
- Collation of thinking drawn into a book
- Drafts of the book reviewed by an international cross-sectoral panel of reviewers.
  - Hundreds of comments
  - Book published Summer 2016
  - Australian version published Autumn 2017
THE SECOND EDITION

• Review of the ADF took place in 2018/19
  • Input from our three advisory groups
    • User community
    • Legal group
    • Scientific expert
  • Changes for GDPR and National Legislation and further conceptual development.
• New materials
  • Compact version (practitioner facing)
  • Short summary documents
  • Posters and leaflets
THE NEW ADF: 10 STEP PROCESS

1. Describe/capture the presenting problem
2. Sketch the data flow and Determine Your Responsibilities
3. Map the properties of the data environment(s)
4. Describe and map the data
5. Engage with stakeholders
6. Evaluate the data situation
7. Select and implement the processes you will use to assess and control disclosure risk
8. Maintain stakeholders’ trust
9. Plan what to do if things go wrong
10. Monitor the evolving Data Situation
Data situations are often dynamic in that data move between environments for both processing and use. Thus, understanding contextual risk, and how to manage that risk through anonymisation, requires an awareness of, and capacity to map, the data flows between environments.
SKETCH THE DATA FLOW: A SIMPLE EXAMPLE

Organisation 1: Collects personal data and stores it
Organisation 2: Wants to use data for secondary purpose

Shares
Aspiration
- To (semi-)automate the capture and processing of provenance information
  - to enable anonymisation decision making
  - And other purposes – e.g. ethics applications
- Need a machine-interpretable way to express both anonymisation and relevant provenance concepts.

RP4 Provenance model
- Retrospective
- Prospective
- Prescriptive
- Proscriptive
- Permitted
· **PROV** is an interoperability standard
  · which defines amongst other things a data model to represent provenance information.
  · Can be used for may things from food to art to.. data..
  · It’s has an associated coding language
AN EXAMPLE PROV GRAPH

The core concepts and relationships supported by PROV
Taken from the PROV primer
https://www.w3.org/TR/2013/
NOTE-prov-primer-20130430/

Used to create graphs describing what happened to the data and processes
WHAT PROPERTIES DO WE NEED?

- A specification of those requirements is as follows:
  - R1: The data environment construct
  - R2: Nesting data environments within data environments
  - R3: Attaching attributes to data environments
  - R4: Relationships between data environments
  - R5: Annotation of relational constructs
  - R6: Representation of agents, data and processes in data environments
  - R7: Data governance instruments: contracts
  - R8: Access and control (direct and indirect)
WHAT'S MISSING FROM PROV

• Data Environments!
• Namespaces
  • inspired by the WWW architecture and was designed to make objects interoperable across technologies and platforms.
  • In PROV-DM, Namespaces are a Uniform Resource Identifier (URI)
  • A provenance graph can contain multiple - possibly many - Namespaces.

• Bundles
  • Are entities which provide provenance information regarding the creation and modification of a group of entities.
  • can contain entities, activities, agents, and the relationships between them.
  • can also support entities with attributes.
# EVALUATION

Requirements analysis for representing data environments

<table>
<thead>
<tr>
<th>Representation requirements</th>
<th>Bundle</th>
<th>Namespace</th>
<th>Namespace+</th>
<th>Bundles+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Environment Construct</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data Environments within Data Environments</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Attaching Attributes to Data Environments</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Relationships between Data Environments</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Annotation of relational constructs</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Representation of agents, data and processes within Data Environments</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data governance instruments: contracts</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Access and control</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
BUNDLES + REPRESENTATION
By extending PROV we are able to represent data environments in machine readable way.

This provides us with a language for reasoning about anonymisation decision making.

In related work we are building an approach to processing data governance instruments (policies, contracts, agreements, regulations, privacy statements, laws etc..) in a systematic way as an additional piece of the jigsaw...