Data Warehousing & Infrastructure

Peter Gough
HISbi Head of Service

Kent Surrey Sussex Academic Health Science Network

BEAUTIFUL INFORMATION

Presentation Aims

Overview of some of the requirements for setting up a linked dataset:

- Initial planning
- Technical issues
- Pseudonymisation of data
- Key datasets
- Building the dataset

Initial Planning

Project Planning

- Get buy in from all key stakeholders
- Have a plan
- Agree who is the Data Controller
- IG/GDPR
- Project governance

Initial Planning

Data Planning

- Purpose of the system
- Users of the system
- Will it be re-identifiable?
- What data is required
- NHS data – Provider or Commissioner based (or both) ?
- Patient Master Index (PMI) is key
- Order of data

Pseudonymisation Methodology

Probably the most important decision of the project – generally needs to be made at the start as it will define how the IG and warehousing processes will be built

- First decision - Requirements:
    - Will the data be re-identifiable or not?
    - KID is non re-identifiable due to original IG setup, however this has caused problems
    - Current preferences tend towards re-identification

      Why?

      Who?
    - What is/are the pseudonym(s) – NHS Number? Something else?
    - Must have consistent pseudonym(s) across the system to allow analysis

Pseudonymisation Methodology

- Second decision - Methodology:
    - Various options for this

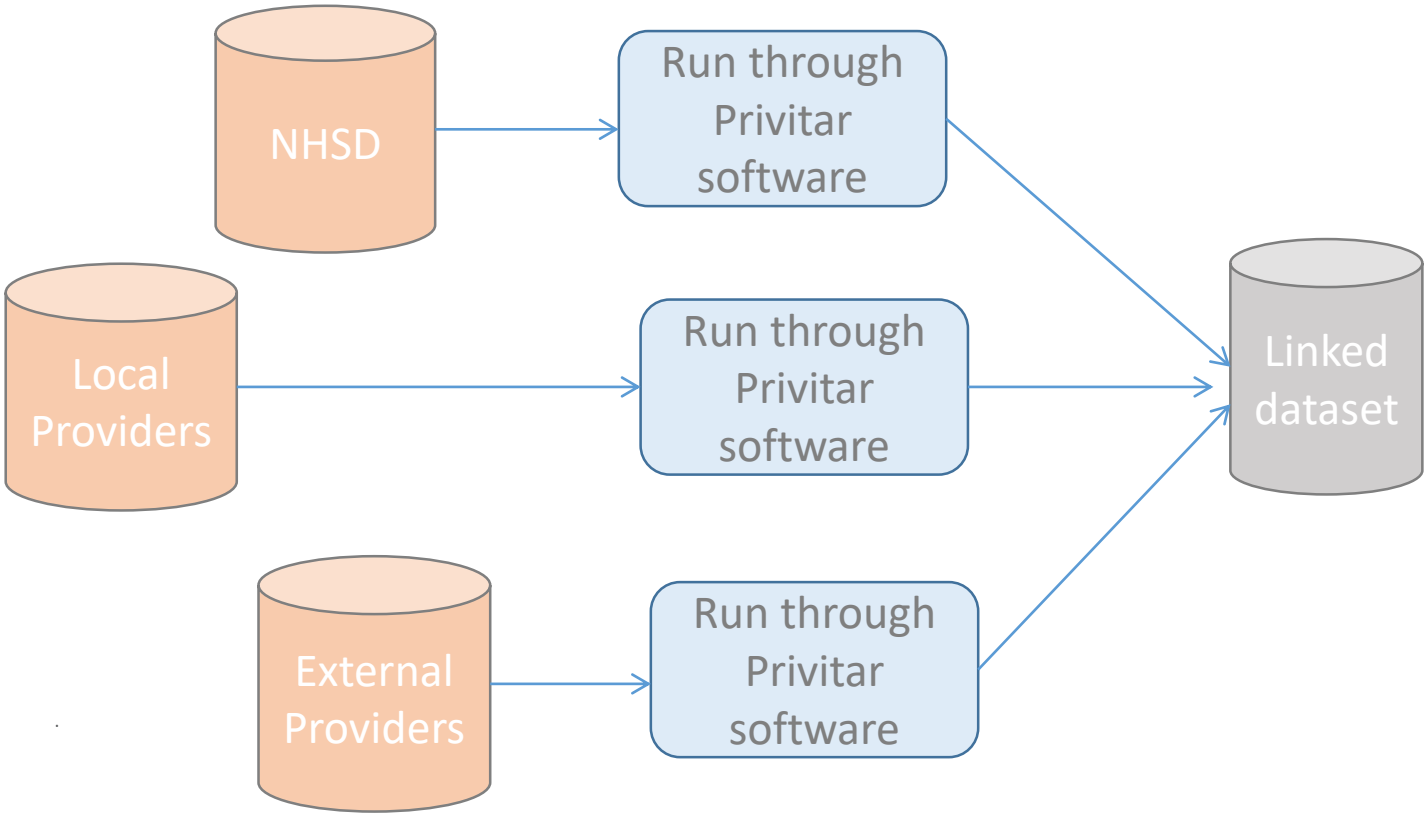
National Pseudonymiser Tool

In House design

Third Party tool

Pseudonymisation Methodology

Option 1 – National Pseudonymiser Tool from Privitar/NHSD



Pseudonymisation Methodology

Option 1 – National Pseudonymiser Tool

Pros:

- Nationally signed off methodology
- Likely to make NHSD DARS applications easier
- Can link data across care settings
- Can be used to link provider and commissioner datasets
- Likely to be cheap option

Pseudonymisation Methodology

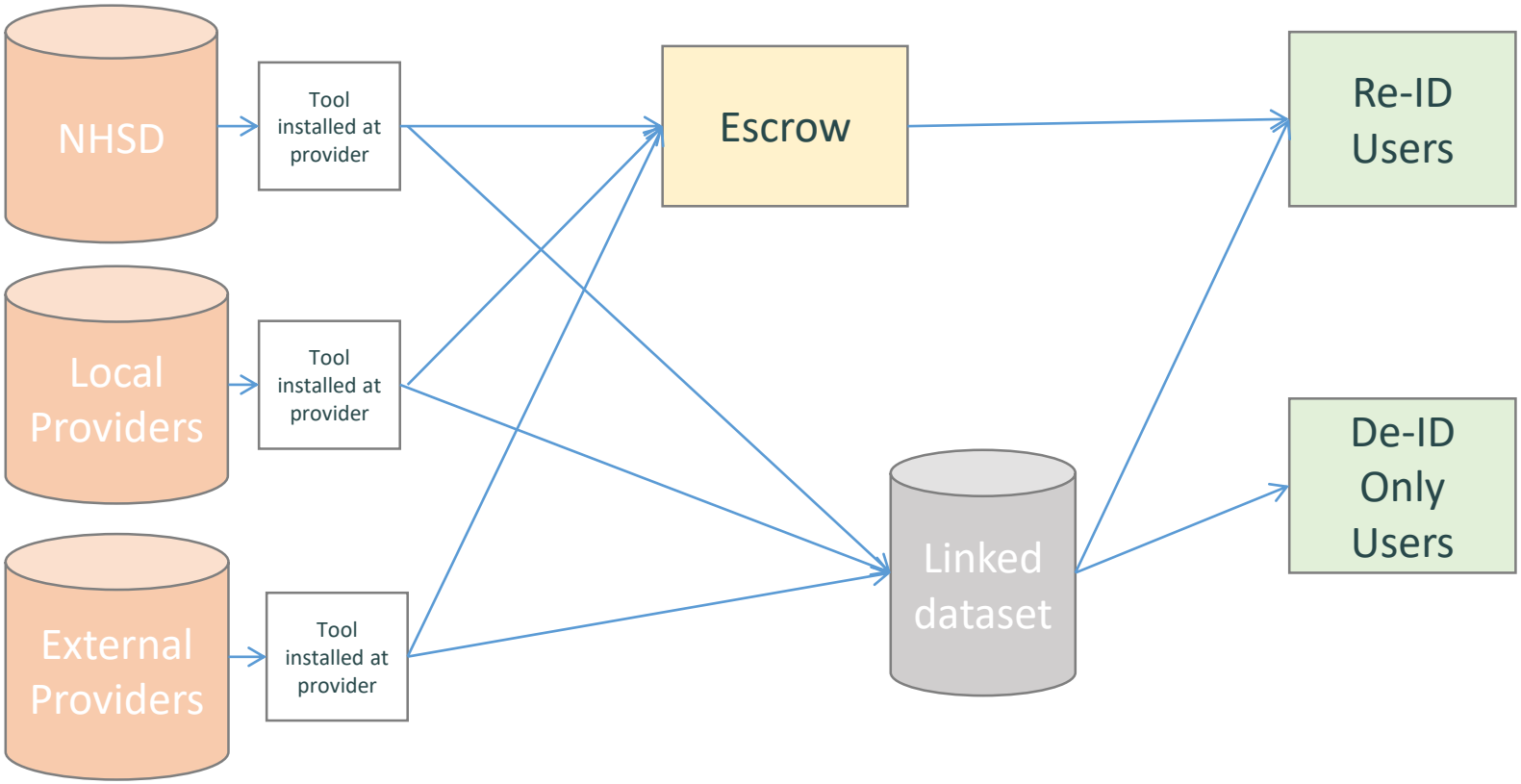
Option 1 – National Pseudonymiser Tool

Cons:

- Little information about final methodology available
- Appears to be De-ID only at present
- Timescales for implementation uncertain
- Controlled by NHSD
- Cannot make alterations (eg for probabilistic matching)

Pseudonymisation Methodology

Option 2 – In House Design



Pseudonymisation Methodology

Option 2 – In House Design

Pros:

- Control of methodology
- Can give out to providers with no external permissions
- Can link data across care settings
- Allows further development if required
- Should be cheap to implement
- Commercialisation could be an option

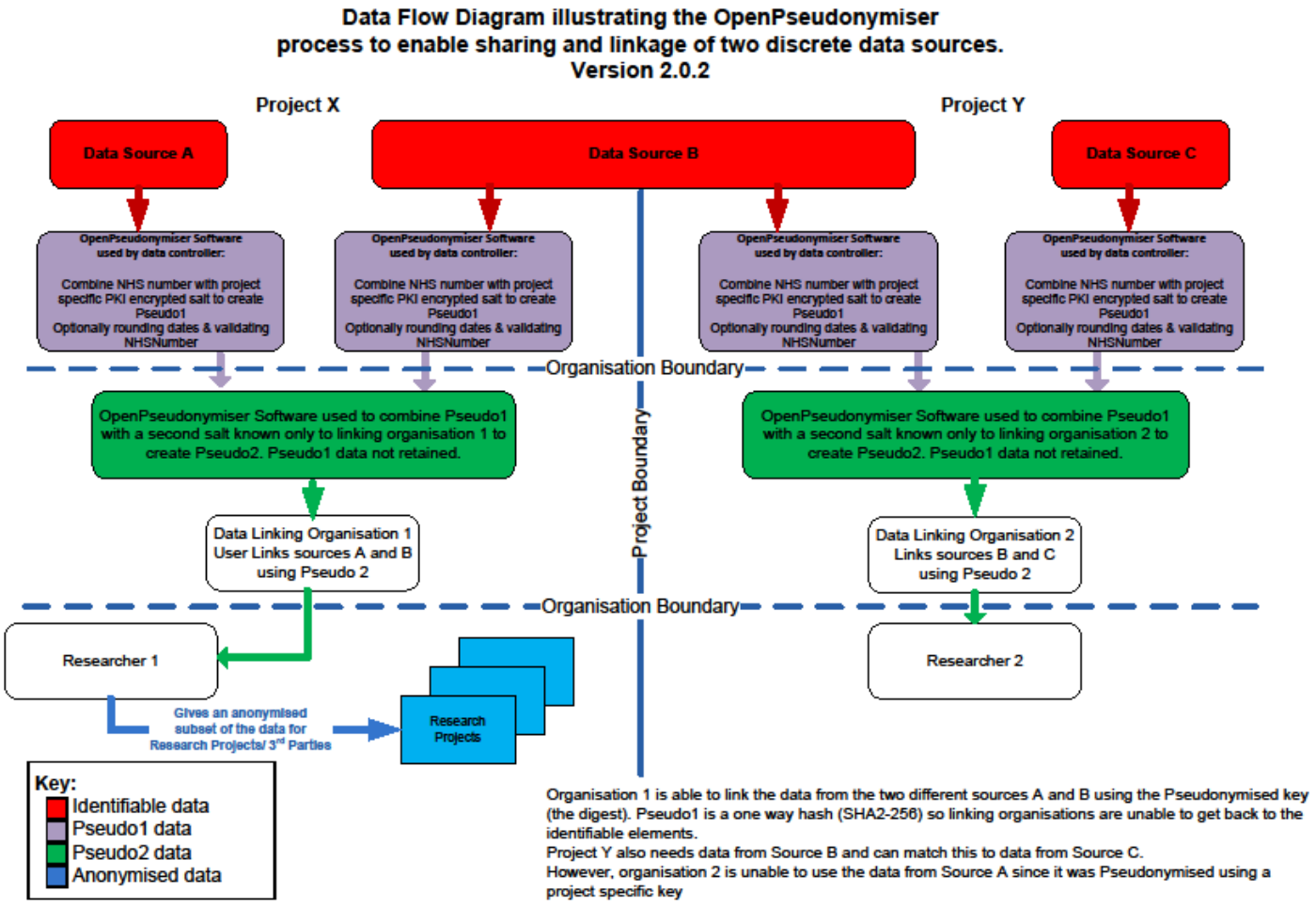
Pseudonymisation Methodology

Option 2 – In House Design

Cons:

- NHSD may need to sign off process before allowing their data to flow
- Would require customer confidence
- Need good documentation
- Would need an Escrow for re-identification if required
- Development time could be high

Pseudonymisation Methodology



Kent Surrey S
Academic Health Science Network

www.beautifulinformation.org

Pseudonymisation Methodology

Option 3 – Third Party Tool Design

eg. Open Pseudonymiser, CSU tools, commercial offerings

Pros:

- Signed off by NHSD
- Tested solutions
- Should be quick to get up and running
- Do not have to manage the software
- Easier for DARS applications

Pseudonymisation Methodology

Option 3 – Third Party Tool Design

eg. Open Pseudonymiser, CSU tools, commercial offerings

Cons:

- Unless Open Source the third party retains control of software
- Generally hard to edit for other uses
- Could be costly
- Methodology could cause problems

Pseudonymisation Methodology

Final thoughts:

- Best option will be project dependent
- National solution could become mandated
- Need to make decision at start of project to avoid full reload

Technical Build

Key questions:

- What Hardware
- What Software
- Hosting arrangements
- Who will do the work
- Space required
- How to keep the data secure from datasets containing PID

Data Build

- Write Design Spec
  - But expect this to change as the project progresses
  - KID started as just SUS and Social Care – now has 15 data sets including 90% GP practices in Kent
- Choose first datasets – proof of concept
  - PMI (key denominator for reporting)
  - Secondary Care data (probably easiest)
  - GP data (start this as it takes longest to get IG sign up)
- Then add next phase
  - Social Care
  - Community
  - Mental Health
- Then the 'would likes' – education, OOH, police data etc

Data Build – Data Sources

Commissioner or Provider based dataset (or both)

- Commissioner
  - A lot of data would have to come from NHSD via CSU
  - Are limited with what data can be linked based on DARS application
  - Likely to need mixed data from NHSD and local providers
  - Could be limited to monthly flows
  - Can get single extracts of (eg) SUS data
  - Data flows should be simplistic

Data Build – Data Sources

Commissioner or Provider based dataset (or both)

- Provider
  - Majority of data (bar PMI) comes from providers directly
  - Can look to expand into NHSD datasets when the system is up and running
  - Certain data would be missed for registered patients seen out of area
  - Have control over data flows/table structures etc
  - Have to manage a large number of data flows
  - Need a good working relationship with data providers
  - Quicker to get something up and running

Data Build - Datasets

- **Patient Master Index** (PMI) – arguably the most important dataset
  - Gives the denominator for research and analysis
  - Gives overall demographics of the area
  - Best way to get it is from NHSE via NHSD – although need a DARS
  - Pseudonymised at NHSD
  - Additional requirements
    - Segmentation tools
    - UPRN
  - Could create from GP data or provider data but these are problematic
  - Can use for opt out process

Data Build - Datasets

- **Secondary Care Data** – possibly the easiest dataset
  - SUS is easiest data source
  - Could use local data files but harder to manage
  - Either download via NHSD or local provider upload
  - Need to honour opt outs

Data Build - Datasets

- **GP Data** – probably the most rewarding dataset
  - Concentrate on GP sign up
  - Each practice is a data controller
  - Decide how to get data:
    - Direct from GP systems
    - Via third party such as Apollo
  - Understanding data is key
  - Need to ensure pseudonymisation at source

Data Build - Datasets

- Other Data
  - Agree process for each additional dataset
    - IG filled in
    - Source agreed
    - Spec created
    - Extract run and tested
    - Made live
  - Many health datasets to choose from – Mental Health, Community, Out of Hours, Ambulance, Continuing HealthCare etc
  - Many non health datasets – Social Care, police data, education data, fire service etc

Data Build - Datasets

- Other Data
  - Non health data with NHS Number – need to ensure opt outs are honoured and permissions to use data for linked datasets have been agreed
  - Becomes more interesting when the datasets do not have NHS number:
    - Use a different ID to link on (eg UPRN)
    - Assign an NHS Number

Data Build - Datasets

- Creating a pseudonym

  - Usually happens with non health data
  - Probabilistic matching – ie best guess of pseudonym
  - Used adapted UCL solution
  - Algorithm takes person details and compares them with known list
  - NHS Number assigned based on most probable match
  - Need to decide at what level a match is accepted
  - Needs to be built via a black box to avoid clear data being seen

Conclusions

- Plan, Plan, Plan
- Get buy in
- Start on IG as soon as possible
- Decide pseudonymisation process
- Get some data flowing

Peter Gough

Peter.gough@nhs.net

Kent Surrey Sussex
Academic Health Science
Network

BEAUTIFUL
INFORMATION