Blind Data Aggregation from Distributed and Protected Sources

Oluwafemi Ajayi, Richard O. Sinnott, Anthony Stell, National e-Science Centre, University of Glasgow
o.ajayi@nesc.gla.ac.uk

Alan Young
Clinical Trials Service Unit & Epidemiological Studies Unit
University of Oxford
Virtual Organisations

...collection of distributed resources shared by collection of users from one or more organizations...

- Provides conceptual framework for rules and regulations for resources that are offered/shared between VO institutions/members
  - Clinical domain much greater emphasis on expression and enforcement of rules and regulations (policies)
  - Less emphasis on dynamic (you don’t/shouldn’t find resources on the fly, you care very much which users are in the VO and what their role is...)

```
{Resources}   {Users}   {Resources}   {Users}
    \\
  VO
    \\
{VO specific agreements}
```
Virtual Organisations for Trials and Epidemiological Studies

- 3 year (£2.8M) MRC funded project started October 2005
- Plans to develop *framework for producing Grid infrastructures* to address key components of clinical trial/observational study
  - Recruitment of potentially eligible participants
  - Data collection during the study
  - Study administration and coordination
    - Involves Glasgow, Oxford, Leicester/Nottingham, Manchester, Imperial
      » Strong links with UK Biobank
VOTES Scottish Experiences

Scottish Data Space... up to now

- Scottish Care Information (SCI) Store
  - Hospital batch system rolled out across Scotland (lab data, patient records...)

- Scottish Morbidity Records (SMR)
  - Aggregated clinical records from last 40 years across Scotland
  - We have been given pseudo-anonymised
    - SMR01A General acute inpatient and day case discharges (3,719,206 records)
    - SMR04A Psychiatric and mental handicap hospitals and units: admissions, residents and discharges (241,599 records)
    - SMR06A Scottish cancer registrations (171,167 records)
    - SMR99A Deaths (173,615 records)

- General Practitioners Administration System for Scotland (GPASS)
  - Used by 85% of GPs across Scotland

- Consent
  - Opt-in/opt-out trial, study, disease area, ...

- Applied in range of areas/projects:
  - UK Biobank, Congenital anomaly, Brain trauma, Diabetes, Knee pain/obesity, Prostate cance....

- Community Health Index (CHI) number key to this!
Select your home organisation

The service you are trying to reach requires that you authenticate with your home organization. Please select an organisation using one of the methods below.

Choose from list

- Aberystwyth University
- JISC project SDSS (Fountainhall)
- JISC project SDSS (Thirlestane)
- JISC project SDSS (TypeKey Bridge)
- Kensington and Chelsea College
- Kidderminster College
- Kingston University
- Leeds Learning Network
- London School of Economics and Political Science
- Manchester Metropolitan University
- National e-Science Centre (Glasgow)
- National Science Learning Centre
- Newcastle University
- North Trafford College
- Nottingham Trent University
- ProctoAccess
- Reid Kerr College
- RSC South West
- Salford Software
- Science and Technology Facilities Council
- Sheffield Hallam University
Select your home organisation

Choose from list
- National e-Science Centre
- [Other options]

Search by keyword
- [Enter keyword]

Authentication Required

Enter username and password for "NeSC Glasgow Identity Provider" at https://mangelian.nesc.gla.ac.uk

User Name:
rosim

Password:
******

- [checkbox] Use Password Manager to remember this password.

[OK] [Cancel]

Need assistance? Visit the UK federation web site.
### Clinical Trial Query Portlet

**Role:** investigator  
**Trial name:** brainIT  
**Databases used:** store14, gndglass

#### Your SQL query

```sql
SELECT DISTINCT metaData.CHINum, metaData.DOB, metaData.firstName, metaData.lastName, metaData.Patient_Image_Details, metaData.PNSH_GCS_Motor, metaData.PNSH_Physical_Glucose  
FROM metaData INNER JOIN PatientMaster ON metaData.CHINum = PatientMaster.chi  
WHERE metaData.PNSH_GCS_Motor = '0' AND metaData.PNSH_Physical_Glucose >= '0'  
```

**Submit another query...**

### Your query results

<table>
<thead>
<tr>
<th>metaData.CHINum</th>
<th>metaData.DOB</th>
<th>metaData.firstName</th>
<th>metaData.lastName</th>
<th>metaData.PNSH_GCS_Motor</th>
<th>metaData.PNSH_Physical_Glucose</th>
</tr>
</thead>
<tbody>
<tr>
<td>020319597535</td>
<td>02/03/1959</td>
<td>MEADOW</td>
<td>INDEX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>020919797535</td>
<td>02/09/1979</td>
<td>SAIF ALI</td>
<td>MCELLIGOTT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
But...

- **The real world**
  - NHS and firewalls
    - Nope!
  - NHS and trust
    - software, people, policy
  - NHS and Grid
    - Globus, etc.
  - NHS and Grid security
    - RBAC, etc.
  - NHS and data disclosure/data linkage
    - Access, etc.

- NHS (and clinical data providers more generally) won’t deploy complex middleware! **Fullstop!**
  - The buck stops with them!
  - Their jobs!
  - Their paymasters jobs!
Vanguard Design Principles

- Privacy by design rather than by contract
  - Eliminate human error

- Autonomy
  - Not just a buzzword

- End-end encryption of data
  - Data obfuscation and anonymisation that stands up to data providers, ethics committees, policy advisor audit!

Resulted in design of VANGUARD system

- This is planned to be not yet another prototype proof of concept, but “we hope” a long term strategy for secure access to clinical data

- To be exploited by UK Biobank
  - Recruiting 500,000 people for wide array of disease areas
Primary Vanguard Components

- **Viewer**
  - allows end-users to access data
- **Guardian**
  - protects and manages source data repositories
- **Agent**
  - communicates between components, constructs queries and aggregates results
- **Banker**
  - handles resource allocation
Guardian Data Protection Levels

- Open - Guardian is willing to supply actual value
- Hash - Guardian is willing to supply anonymised one-way hash encoding of value
- Closed - Guardian will not supply value, but will perform queries that involve it as a selector

Protection can be applied globally or per user/role
Typical Example

How many days did mothers with HIV stay in hospital?
### Vanguard Linkage

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>hospID</td>
<td>Integer</td>
</tr>
<tr>
<td>mother</td>
<td>Integer</td>
</tr>
<tr>
<td>days</td>
<td>Integer</td>
</tr>
<tr>
<td>status</td>
<td>Integer</td>
</tr>
<tr>
<td>nhs</td>
<td>String</td>
</tr>
<tr>
<td>mother</td>
<td>Integer</td>
</tr>
<tr>
<td>dob</td>
<td>Date</td>
</tr>
<tr>
<td>weight</td>
<td>Real</td>
</tr>
<tr>
<td>sex</td>
<td>Int</td>
</tr>
<tr>
<td>chi</td>
<td>Int</td>
</tr>
<tr>
<td>hiv</td>
<td>Bool</td>
</tr>
<tr>
<td>hepatitis</td>
<td>Bool</td>
</tr>
</tbody>
</table>

• SELECT alpha.stay.days,H(alpha.birth.nhs)
  WHERE alpha.stay.mother = alpha.birth.mother

• SELECT H(gamma.linkage.nhs),H(gamma.linkage.chi)

• SELECT H(delta.disease.chi) WHERE delta.disease.hiv = true;

• Join on H(*.nhs) AND H(*.chi), then remove H(*.nhs) and H(*.chi)
Architecture and Typical Usage Scenario

Schematic Diagram of Vanguard Communications

PKA\(^{-1}(\text{Join}(\text{HA}_{x\text{Ares}}, \ldots), \text{HA}_{x\text{Bres}}, \ldots))\)

PKA\(^{-1}(\text{Join}(\text{HA}_{x\text{Cres}}, \ldots), \text{HA}_{x\text{Dres}}, \ldots))\)

PKV(Q_x, PKU_x)

All Hands Meeting, 2008
Anonymisation and Data Linkage

- Viewer submits distributed query to an Agent.
- Agent receives the query and decomposes the query.
- Guardians (e.g. Alpha, Beta, Gamma) pull queries that are meant for them from Agents.
- Guardians encrypt data and hash fields used for joining data. Pushes anonymised data to the Agent.
- Agent joins data received from multiple guardians using hashed key fields.
- Agent drops hashed fields from joined data.
- Viewer pulls data from the Agent and decrypts data using user’s private key.
Supporting Standards

- Agent as a web service.
- SHA1-HMAC for key-based hashing.
- AES for symmetrical encryption, offers performance and speed. Symmetrical keys encrypted using public-private key (RSA) encryption.
- Data exchange using webrowset xml format.
Screen Shots

UNIVERSITY of GLASGOW

All Hands Meeting, 2008
Questions ...?
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Ajayi, Oluwafemi; Sinnott, Richard O.; STELL, ANTHONY; Young, Alan

Title:
Blind data aggregation from distributed, protected sources

Date:
2008

Citation:

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/28849

File Description:
Blind data aggregation from distributed, protected sources

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.