



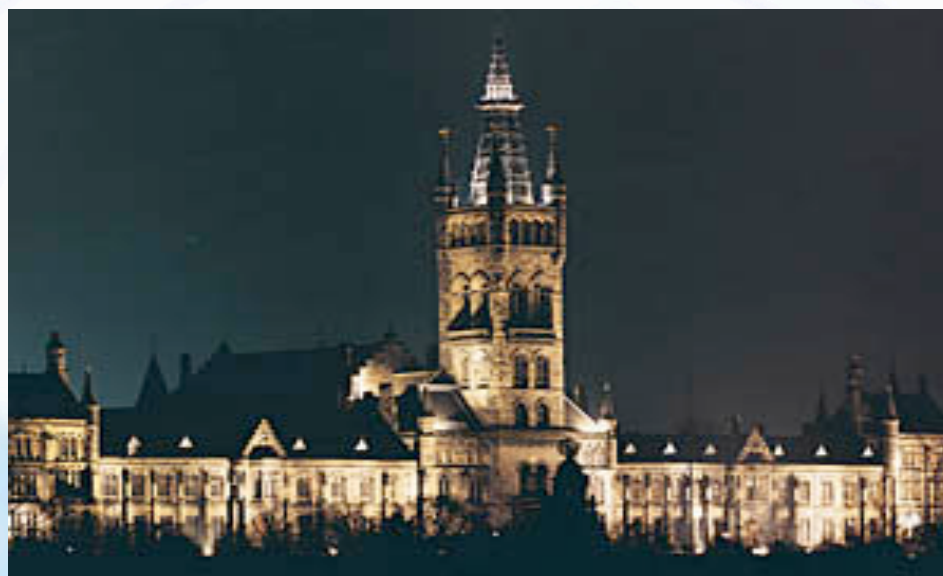
University
of Glasgow

Ajayi, O. and Sinnott, R.O. and Stell, A.J. (2008) *Blind data aggregation from distributed, protected sources: the future model for security-oriented collaborations*. In: UK e-Science All Hands Meeting, 8-11 Sept 2008, Edinburgh, UK.

<http://eprints.gla.ac.uk/7390/>

Deposited on: 8 September 2009

Blind Data Aggregation from Distributed and Protected Sources

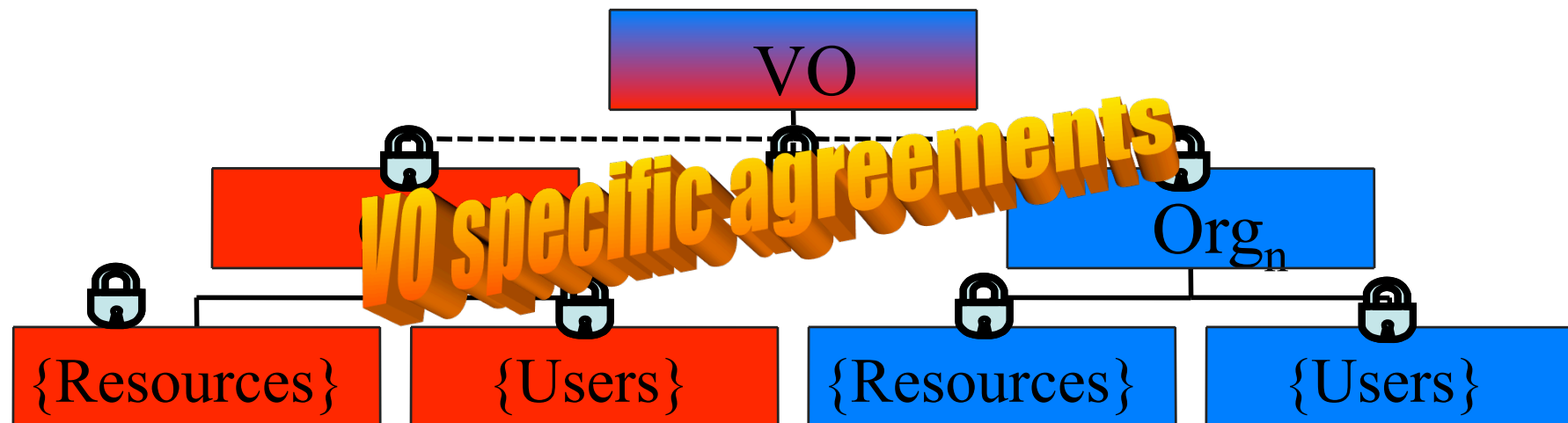


Oluwafemi Ajayi,
Richard O. Sinnott, Anthony Stell,
National e-Science Centre,
University of Glasgow
o.ajayi@nesc.gla.ac.uk

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

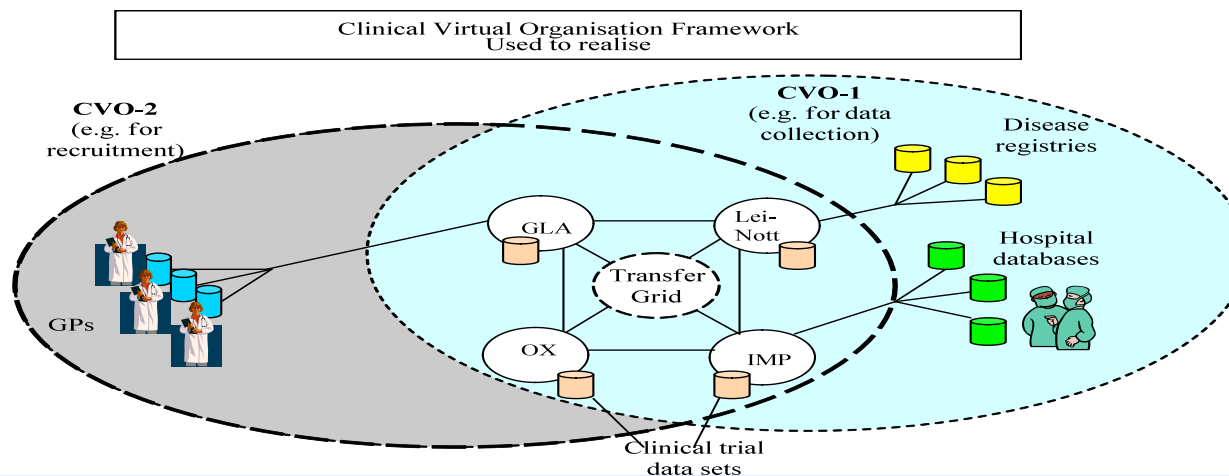
Virtual Organisations

- *...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...)



VOTES

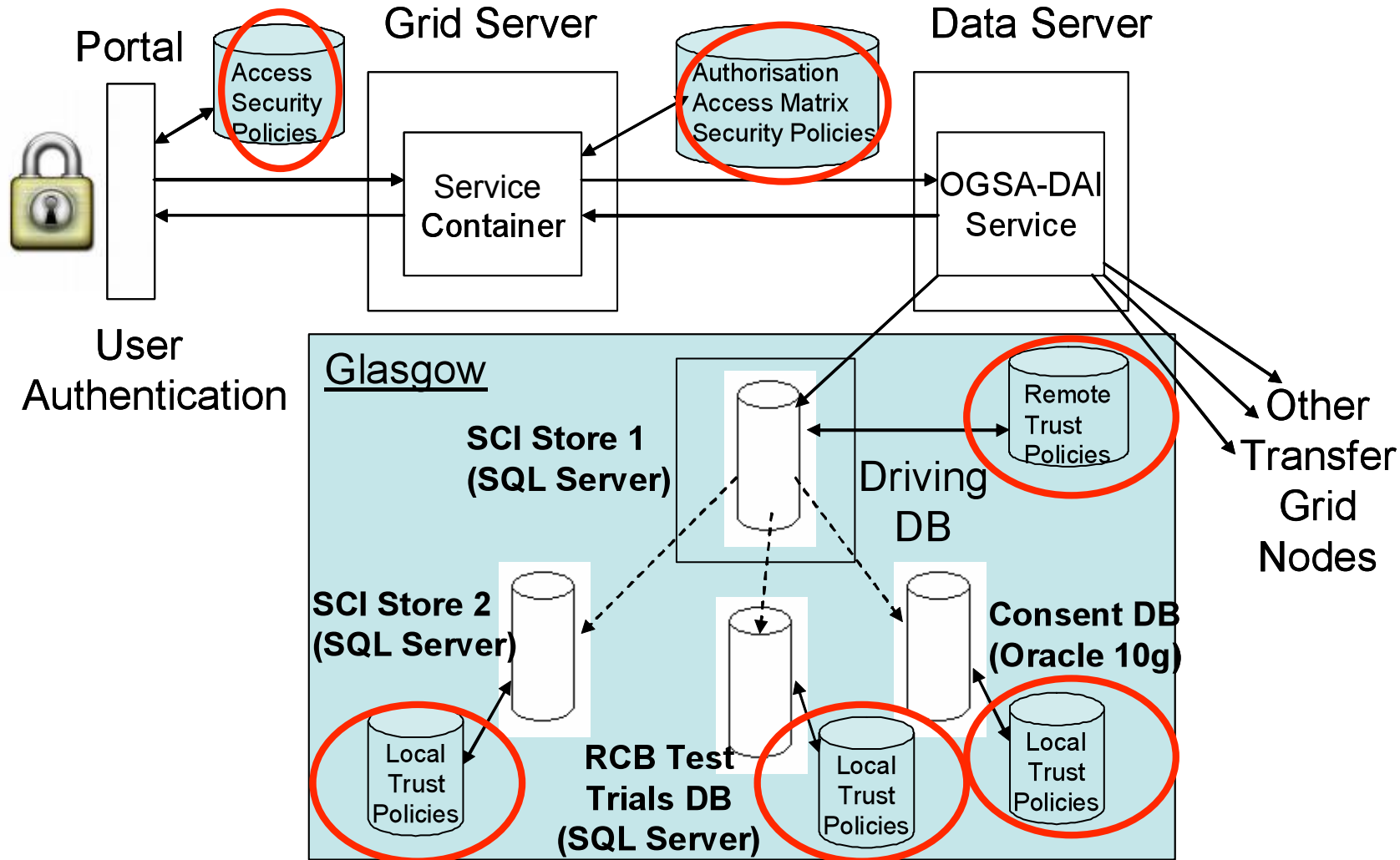
- **Virtual Organisations for Trials and Epidemiological Studies**
 - 3 year (£2.8M) MRC funded project started October 2005
 - Plans to develop framework for producing Grid infrastructures to address key components of clinical trial/observational study
 - ▶ Recruitment of potentially eligible participants
 - ▶ 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 cancer....
 - **Community Health Index (CHI) number key to this!**

VOTES Distributed Data Framework



Select your home organisation - Mozilla Firefox

File Edit View History Bookmarks Tools Help

 <https://wayf.ukfederation.org.uk/shibboleth-wayf/uk.wayf?shire=https%3A%2F%2Ftethys.nesc.gla..> Google

Select your home organisation



Selection options

The service you are trying to reach requires that you authenticate with your home organisation. 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
- ProtectNetwork
- Reid Kerr College
- RSC South West
- Salford Software
- Science and Technology Facilities Council
- Sheffield Hallam University

tion web site.

Select your home organisation



Selection options

The service you are trying to access requires authentication from your home organisation. Please select an organisation from the list below.

Choose from list

Search by keyword

Authentication Required



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

User Name:

Password:

Use Password Manager to remember this password.

{w} Need assistance? Visit the UK federation [web site](#).

Data Federation

Distributed Data Framework

Clinical Trial Query Portlet


Select a trial that you would like specific information on.

Select a specific clinical trial

votes1
votes1
votes2
rcb1
brainIT
gpass1

User Information

National e-Science Centre User ID card

Name	richard.cinnott	
Role	votes1_investigator votes2_investigator rcb1_investigator gpass1_investigator brainIT_investigator gemeps_rat_genome_researcher nanoCMOS_deviceModeller nanoCMOS_systemCircuit nanoCMOS_auroraLicense nanoCMOS_taurusLicense espe_paediatric_endocrinologist espe_paediatric_nurse	
Organization	University of Glasgow	
Unit	National e-science Centre	
Single-Sign-On Life Time	300	

Clinical Trial Query Portlet

Role: investigator

Select from the list below the parameters you would like to search on for this trial and apply the parametric conditions that will help refine your search.

Parameter selection for "brainIT" clinical trial

Select a different trial

- Metadata.CHInum
- Metadata.DOB
- Metadata.firstName
- Metadata.lastName
- Metadata.NSH_Initial_Gluco
- Metadata.Patient_Image_Det
- Metadata.PNSH_GCS_Motc
- PatientMaster.CHI
- PatientMaster.FamilyName

Clinical Trial Query Portlet

Role: nurse

Select from the list below the parameters you would like to search on for this trial and apply the parametric conditions that will help refine your search.

Parameter selection for "brainIT" clinical trial

Select a different trial

- Metadata.DOB
- PatientMaster.PostCode
- PatientMaster.Sex

Submit Query

National e-Science Centre User ID card

Name	john watt
Role	votes1_nurse votes2_nurse rch1_nurse brainIT_nurse gpass1_nurse gemeps_lymphnodes_genome_researcher nanoCMOS_taurusLicense nanoCMOS_systemCircuit
Organization	University of Glasgow
Unit	National e-Science Centre
Single-Sign-On Life Time	300

Data Federation

Distributed Data Framework

Clinical Trial Query Portlet

Role: investigator

Trial name: brainIT

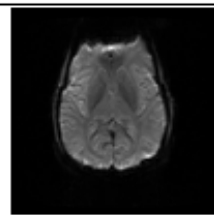
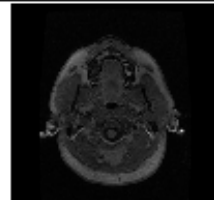
Databases used: store14 , gridglass

Your SQL query

```
SELECT DISTINCT MetaData.CHInum, MetaData.DOB, MetaData.firstName,
MetaData.lastName, MetaData.PNSH_GCS_Motor, MetaData.Patient_Image_Details,
MetaData.PNSH_GCS_Motor, PatientMaster As PatientMaster INNER JOIN
OPENDATASOURCE('Microsoft SQL Server', 'Server=tcp:store14.gridglass.ac.uk;
macross..JIANGJ.METADATA ON metadata.CHInum = PatientMaster.chi
WHERE MetaData.PNSH_GCS_Motor >= '0' AND MetaData.PNSH_GCS_Motor >= '0'
```

Submit another query...

Your query results

MetaData.CHInum	MetaData.DOB	MetaData.firstName	MetaData.lastName	MetaData.NSH_Initial_Glucose	MetaData.Patient_Image_De
020319597535	02/03/1959	MEADOW	INDEX	144.0	
020919797535	02/09/1979	SAIF ALI	MCELLIGOTT	138.6	

User Information

National e-Science Centre User ID card

Name richard sinnott

votes1_investigator
votes2_investigator
rcb1_investigator
gpass1_investigator
brainIT_investigator

Role gemeps_rat_genome_researcher
nanoCMOS_deviceModeller
nanoCMOS_systemCircuit
nanoCMOS_auroraLicense
nanoCMOS_taurusLicense
espe_paediatric_endocrinologist
espe_paediatric_nurse

Organization University of Glasgow

Unit National e-Science Centre

Single-Sign-On Life Time 300



Joining on CHI

But...

- The real world...

- NHS and firewalls

- ▶ Nope!

- NHS and trust

- ▶ software, people, pr

- NHS and

- ▶ Globu

- NHS and

- ▶ RBAC

- NHS and

- ▶ Acce

- NHS (and clinical data providers more generally) won't deploy complex middleware! Fullstop!

- ▶ The buck stops with them!
- ▶ Their jobs!
- ▶ Their paymasters jobs!

LIABILITY!

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

alpha.stay

Field	Type
hospID	Integer
mother	Integer
days	Integer
status	Integer

alpha.birth

Field	Type
nhs	String
mother	Integer
dob	Date
weight	Real
sex	Int

gamma.linkage

Field	Type
nhs	String
Chi	Int
Active	Bool

delta.disease

Field	Type
chi	Int
hiv	Bool
hepatitis	Bool

How many days did mothers with HIV stay in hospital ?

Vanguard Linkage

alpha.stay

Field	Type
hospID	Integer
mother	Integer
days	Integer
status	Integer

alpha.birth

Field	Type
nhs	String
mother	Integer
dob	Date
weight	Real
sex	Int

gamma.linkage

Field	Type
nhs	String
Chi	Int
Active	Bool

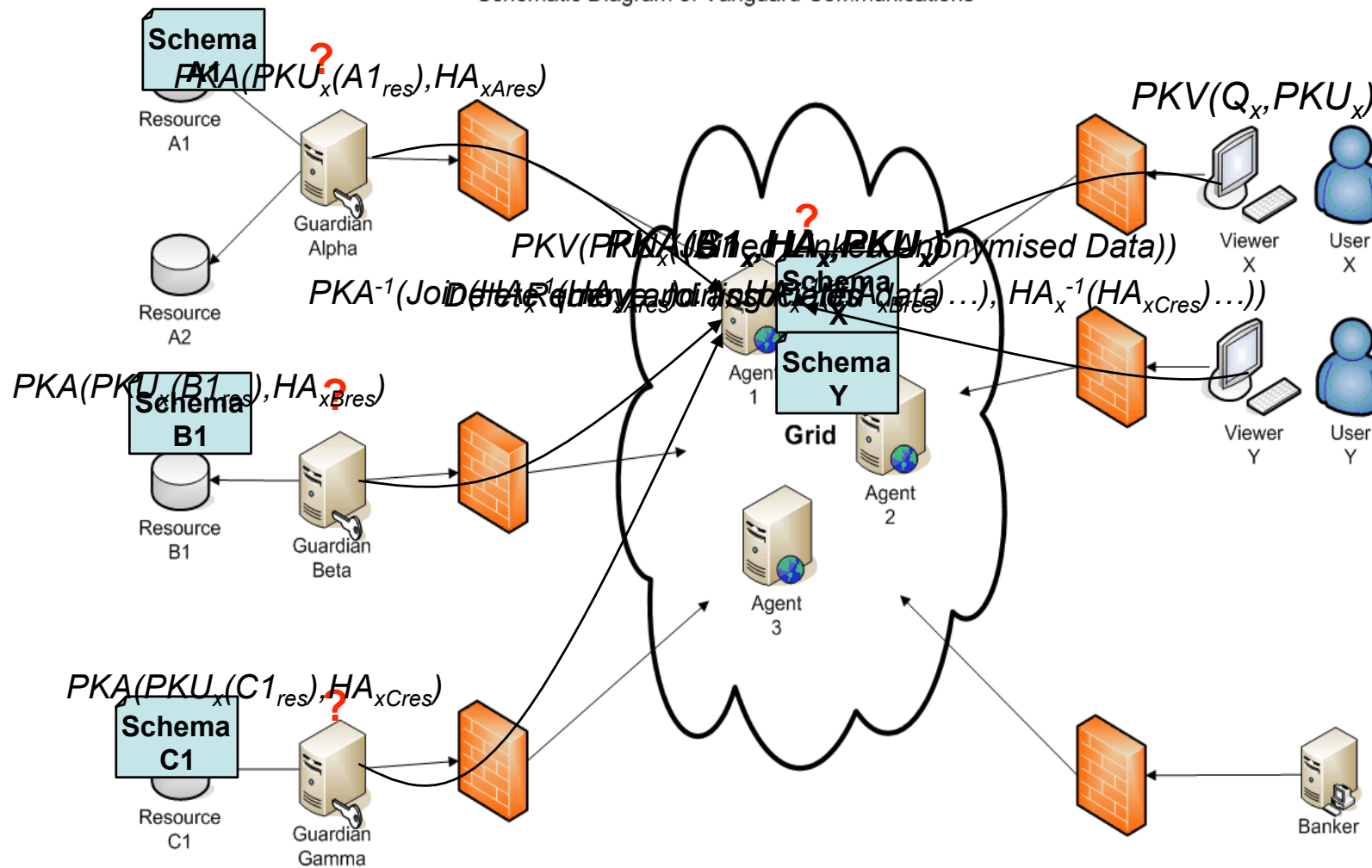
delta.disease

Field	Type
chi	Int
hiv	Bool
hepatitis	Bool

- 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



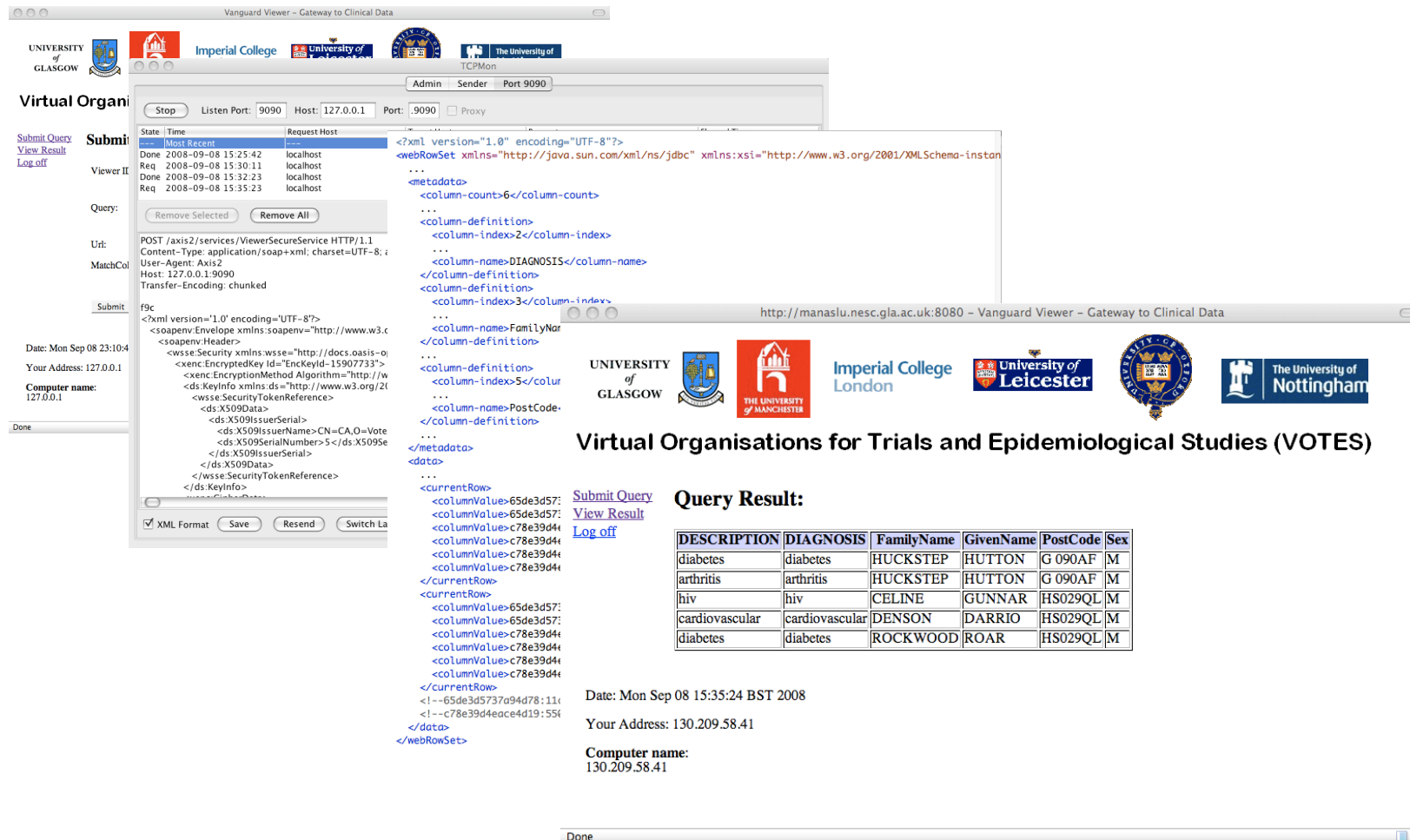
Anonymisation and Data Linkage

- 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.
- WS-Security for message level security - signing and encrypting soap messages. Uses digital certificates.
- Data exchange using webrowset xml format.

Screen Shots



Virtual Organisations for Trials and Epidemiological Studies (VOTES)

Query Result:

DESCRIPTION	DIAGNOSIS	FamilyName	GivenName	PostCode	Sex
diabetes	diabetes	HUCKSTEP	HUTTON	G 090AF	M
arthritis	arthritis	HUCKSTEP	HUTTON	G 090AF	M
hiv	hiv	CELINE	GUNNAR	HS029QL	M
cardiovascular	cardiovascular	DENSON	DARRIO	HS029QL	M
diabetes	diabetes	ROCKWOOD	ROAR	HS029QL	M

Date: Mon Sep 08 15:35:24 BST 2008
Your Address: 130.209.58.41
Computer name: 130.209.58.41

Questions ...?

