The Core User Directory

A *Pilot Project in Identity Management for the collegiate University*

ICT Forum Conference
July 2009

Lou Burnard, OUCS
Rob Hebron, independent consultant

Agenda

- Background drivers
- Overview of the CUD as delivered
  - Demonstration
  - Evaluators' experience
  - What makes it tick
- Where do we go from here
- Q&A

Background

- Personal Data is stored in hundreds of different places
  - University
  - College
  - Department/Division/Faculty
- Some data attributes are duplicated across different stores (sometimes inconsistently); some are unique to particular stores
- There is no way of reliably combining data about the same person held in different stores

Problems...

- People have multiple, changing, affiliations
- Data is often (re)created when an affiliation changes or is created
- Different parts of the collegiate university maintain their own lists of “authorised” users
- The Government requires counts of people, not their roles or affiliations
- Authorisation however may derive from a role
- And not all sources of information are equally authoritative...

Two specific examples

- To make its HESA return, the University needs to know how many students and staff it has.
  - But some students are also staff members, so they shouldn't count twice
- To pay graduate students operating the Graduate Supervision System, we would need to know the University Payroll identifier for their account (if they have one)
  - But there is no simple way of looking it up

The CUD Pilot Project

"Establish a University-wide Identity Management system which provides authentication and authorisation, and enables interoperability with national and international infrastructure." [ICT Strategic Plan, as endorsed by Council GPC, and PRAC last year]]

- What would an Identity Management system need?
  - A unique identifier for individuals associated with the University
  - A minimum set of attributes associated with that key
  - Links between the resulting database and other existing data sources within the University

CUD Goals

- CUD should provide three services:
  - *Data matching service* to link data linked to a person across multiple data stores
  - *Data provision service* to provide a “useful subset” of all data know about a person directly from the CUD
  - *Foreign key service* to enable unique keys in different data stores to be obtained and then used to obtain additional data directly from those systems
- CUD is person focussed. A person's identity is represented in the CUD by a single person record, and all data in other systems is linked to this representation

“Identity” Management

“*The quality or condition of being the same in substance, composition, nature, properties, or in particular qualities under consideration; absolute or essential*” - OED

“*sameness of essential or generic character in different instance*” - Webster

- A person only has one identity
- An identity may be represented in different data stores, but data contained in a data store does not constitute an identity
- Identity Management comprises the matching, linking and manipulation of data which represents people in data stores
- Functions such as account provisioning are ancillary and enabled by Identity Management

CUD Pilot Database

To be included in the CUD, a data attribute:

- Should be required by two or more data consumers...
- … or should be required by a central data service
- … or should be provided by two or more data providers

To be included in the CUD, a data source

- should have a persistent person-based key
- must contain sufficient data to identify people

CUD Pilot Functionality

- Provides interfaces to import data, either as batch or deltas
- Matches the data against data already in the CUD
- Consolidates a subset of data available into a central data store
- Provides interfaces to query data
- Provides interfaces to export data, either as batch or deltas
- … thus providing a central *data reconciliation* service

Initial User Community

- For the Pilot, we consolidated data from eight data sources
  - Card
  - OSS
  - OUCS
  - Staff Records
  - New College
  - DPAG
  - Earth Sciences
  - Telecomms
- We defined and evaluated a number of use cases for these

CUD Pilot

Demonstration of web interface to CUD running locally

Under the hood

- The software components of CUD are all open source
  - Mule ESB to provide the incoming and outgoing interfaces for data transfer
  - Drools (JBoss Rules) for rule-based data manipulation and transformation
  - Postgres for data storage
  - Data Rules and Routing Engine to manage data matching and data flows
  - Spring Framework to manage configuration, object lifecycle etc.

Under the hood – the DRRE

- The DRRE manages data flows:
  - Data is received on an interface and stored in a dynamically build class
  - Rule based data transformations made
  - Data matching
  - Rule based data transformations made
  - Data written to CUD database

Under the hood – interfaces

- Mule ESB is a well established project which provides full Enterprise Service Bus functionality
- CUD only uses it to provide incoming and outgoing data interfaces which use well defined protocols
  - File
  - SOAP
  - SMTP
  - POP3...
- Branching and conditional logic which is available in Mule is unused
- CUD does not have an inherent dependency on Mule – something like Apache ServiceMix could be used instead

Under the hood – rules

- Drools used to apply business rules which correspond to a condition → action format
- Drools uses the RETE algorithm to assess rules in parallel and deal with multiple rule matches, and is very fast (much faster than a traditional if {...} else if {...} else if {...} method of applying rules)
- Conditions always assess data passing through the CUD, actions always transform that data with add, modify and delete actions
- Helper classes enable dynamic lookups etc. to be done in a Drools ruleset

Under the hood – data matching

- Data matching is predicated on data existing in both the matching database and the data store that can be used to uniquely identify a person
- If a unique match cannot be made, but multiple possible matches exist then the matches can be stored for manual checking
- Multiple matching strategies can exist for a data sources, arranged in order of confidence:
  - DOB, Surname, First name – high confidence
  - DOB, Surname match – low confidence
  - Surname matches – very low confidence

Under the hood – data matching

- Matching strategies work by string comparison of attributes and are converted into SQL queries:
  - Exact matches
  - Case insensitive matches
  - Fuzzy matches (Lievenstien, Soundex, Metaphone, TRGM)
- Dates must be in common format to be matched (dd-MMM-yyyy)

Under the hood – the database

- The schema of the CUD database is designed to be very flexible for an ever changing range of data sources
  - Matching database
    - treats all attributes the same
    - stores them as strings to ease matching
  - CUD database
    - makes simple distinction between person attributes, unit attributes and units
    - stores attributes as strings with data type recorded

Results of the Pilot

- Technical issues of providing required functionality have been addressed
- What have we learned from the use cases?
  - In a production CUD data stores are likely to be both providers of data to the CUD and consumers of data from it
  - There are major issues relating to governance and policy

Policy issues to be decided

- Which data sources should be the CUD's primary data providers?
- Which data attributes should be included in the CUD?
- Data protection and privacy issues?
- Who decides and on what basis?

Which data?

- The Big Five data sources
  - Card, OSS, Payroll, Telecomms, OUCS
- Unit-role affiliations
- A given data attribute may be
  - Known to exist in a data source
  - Linkable from the CUD but maintained elsewhere
  - Copied into the CUD
- Pragmatic criteria need to be defined

Criteria

- Is this data attribute useful to more than one data consumer?
- Does access to this data contribute to the university's business?
- Is this data reliably available from an authoritative source, possibly on the basis of other data already present in the CUD?

Next steps

- Over the next year, the CUD might facilitate...
  - a unified email/phone directory service
  - linkage of student records and card image data
  - a data linking service
- We need to set up
  - High level authority to determine policies
  - Technical support for the CUD
  - Database administration function

Comments?