Make things as simple as possible, but no simpler

Einstein

Essentially, all models are wrong, but some are useful

Box & Draper

If you can’t measure it, you can’t manage it…

Various known
Your Presenter

Role – Service/Project Manager and Information Analyst

**Projects**

- Data Migration and Cleansing
- Data Standards Implementations
- Sales and Account Development
- Business and Data Analysts
- Accruals Accounting

**Employment**

- National Audit Office (E&AD)
- Capgemini
- Freelance
- Atos Origin

**Recent Training**

- TOGAF9 Enterprise Data Architect
- Prince2 Practitioner
- Value Analysis.

» Ian.woodrow@atosorigin.com
Introduction

» Information Management Service
» Corporate Data Model
» Person Metadata Model
» Data Standards
» Data Profile Reporting
» Questions and Queries
» Contact Me.
Information Management Service

Corporate Data Model
- Interfaces Model
- Application Data Models
- Model Management
- Data Standards
- Data Profiling (Attribute Level)
- Business Information Data Browser
- Taxonomy and Domains
- Catalogue Management (Events, ..)
- Promotion of Re-Use
- Data Champions
- Centre for Data Excellence

Governance Process
- Data Profiling (CDM Level)
- Data Quality (De-Duplication)
- Cross Application Mappings
- Data Investigation
- Business Rule Management
- Business and Data Analysis
- Business Process Support
- Business Process Improvement
- Schema Management
- Data Migration

5 Defining a Person Metadata Model to Improve Data Quality
Corporate Data Model

- PEOPLE/ORGANISATION
- EVENTS
- LOCATIONS
- COMPANY
- DOCUMENTS
- SERVICES
Applications

- Registry
- Membership
- Benefit
- Biometrics
- Resources
- Overseas

Defining a Person Metadata Model to Improve Data Quality
Select Target Attribute

- Identify Local Application Alternatives
- Look for Commonality
- Review Government Standards
- Review UK National Standards
- Review Other Government Standards (e.g. NIST)
- Review International Standards
- Assess Applicability
- Decide Solution
- Confirm Solution
- Publish Solution

Target Attribute Metadata

- Datatype
- Length
- Values Lists
- Description
- Constraints
- Defaults
Tooling: Best of Breed Solution

PRODUCTS

Erwin Data Modeller
Model Manager
Process Modeller (BPWin)

CHANNELS

INTRANET
TECH DOC LIBRARY

Discovery
CDM Person Subject Area

10 Defining a Person Metadata Model to Improve Data Quality
Cross-reference CDM to Applications

<table>
<thead>
<tr>
<th>ATTRIBUTE</th>
<th>CDM</th>
<th>REGISTRY</th>
<th>MEMBERSHIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family Name</td>
<td>PERSON.FAMILY NAME</td>
<td>REGISTER.PRINCIPAL NAME</td>
<td>PEOPLE.SURNAME</td>
</tr>
<tr>
<td>Given Names</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date of Birth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nationality(s)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language (s)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternative Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organisation Start Date</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organisation Reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>International ID (Passport)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NI Number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Line Address</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postcode</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone Number(s)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Publishing the Results**

---

**Attribute: FAMILY NAME**

(CHAR(35))

<table>
<thead>
<tr>
<th>App Entity Name</th>
<th>App Attribute Name</th>
<th>Application Name</th>
<th>App Attribute Datatype</th>
<th>CDM Entity Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPLICANT</td>
<td>FamilyName</td>
<td>Y</td>
<td>VARCHAR(80)</td>
<td>PERSON</td>
</tr>
<tr>
<td>FILE COVERS</td>
<td>Surname</td>
<td>N</td>
<td>VARCHAR(45)</td>
<td>PERSON</td>
</tr>
<tr>
<td>LC PEOPLE</td>
<td>FAMILY_NAME</td>
<td>C</td>
<td>VARCHAR(75)</td>
<td>PERSON</td>
</tr>
<tr>
<td>NAMEDPERSON</td>
<td>Surname</td>
<td>P</td>
<td>VARCHAR(50)</td>
<td>PERSON</td>
</tr>
<tr>
<td>PEOPLE</td>
<td>FAMILY_NAME</td>
<td>C</td>
<td>VARCHAR(75)</td>
<td>PERSON</td>
</tr>
<tr>
<td>SUBJECT</td>
<td>TXT_CUSTOMER_NAME</td>
<td>C</td>
<td>VARCHAR(40)</td>
<td>PERSON</td>
</tr>
<tr>
<td>TRLAPPLICATION</td>
<td>AP_SURNAME</td>
<td>R</td>
<td>VARCHAR(50)</td>
<td>PERSON</td>
</tr>
<tr>
<td>TRLPARTICIPANTS</td>
<td>SURNAME</td>
<td>S</td>
<td>VARCHAR(50)</td>
<td>PERSON</td>
</tr>
<tr>
<td>TBLPERSON</td>
<td>PN_SURNAME</td>
<td>H</td>
<td>VARCHAR(150)</td>
<td>PERSON</td>
</tr>
<tr>
<td>TBLSuspects</td>
<td>FAMILY_NAME</td>
<td>N</td>
<td>VARCHAR(50)</td>
<td>PERSON</td>
</tr>
<tr>
<td>THIRD PARTY CORRESPONDENT</td>
<td>TXT_CUSTOMER_SURNAME</td>
<td>C</td>
<td>VARCHAR(40)</td>
<td>PERSON</td>
</tr>
<tr>
<td>WORKER</td>
<td>Surname</td>
<td>G</td>
<td>CHAR(30)</td>
<td>PERSON</td>
</tr>
</tbody>
</table>

---

12 Defining a Person Metadata Model to Improve Data Quality
Person Metadata Model – Vision

» To Create a Whole Customer View
  » Person Identity
  » Key Events
  » Sufficient for Identity Resolution

» Organisation Facts
  » Siloed Data Sources
  » Several Suppliers
  » Considering:
    » Master Data Management
    » Service Oriented Architecture

» Technical Facts
  » Different Databases
    » SQL Server
    » Oracle
    » Access
  » No ETL Tool on server
Benefits of Person Metadata Model Approach

» Address data shortfall in existing applications (enrichment)
» Form basis for new application data models
» Bench mark internal developments and suppliers offerings
» Baseline for Profiling
Person Metadata Model

Person Baseline

- Full Name
  - Family Name
  - Given Names
- Names
- Date of Birth
- Gender
- Nationality
- Language
- Alternative Details
- Organisation Start Date

Identifiers
- Organisation Reference
- International ID
- Nationality

- NI Number
- UK Contact Details
- UK Address
- Telephone Number
Person Baseline Analysis

Defining a Person Metadata Model to Improve Data Quality
## CDM Attribute Scoring

<table>
<thead>
<tr>
<th>MODEL COMPARISON ATTRIBUTE</th>
<th>CDM Maximum</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Name</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Family Name</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Given Names</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Date of Birth</td>
<td>6</td>
<td>2 for each part</td>
</tr>
<tr>
<td>Gender</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Nationality(s)</td>
<td>6</td>
<td>3 + 3 for dual Nat</td>
</tr>
<tr>
<td>Language(s)</td>
<td>4</td>
<td>2 + 2 for multiple Lang</td>
</tr>
<tr>
<td>Alternative Details</td>
<td></td>
<td>2 for each name, only 2 if held at a composite part (maximum of 2)</td>
</tr>
<tr>
<td>Organisation Start Date</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Internal Reference Available externally</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Home Office Reference</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>International ID (Passport)</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Country Issuing International ID</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>International ID Details Linked</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>NI Number</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>First Line Address</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Postcode</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Telephone Number(s)</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>6</td>
</tr>
</tbody>
</table>
### Defining a Person Metadata Model to Improve Data Quality

<table>
<thead>
<tr>
<th>MODEL COMPARISON ATTRIBUTE</th>
<th>CDM Maximum</th>
<th>Options Registry</th>
<th>Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Name</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Family Name</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Given Names</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Date of Birth</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Gender</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Nationality(s)</td>
<td>6</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Language(s)</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Alternative Details</td>
<td>4</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Organisation Start Date</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Internal Reference Available externally</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home Office Reference</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>International ID (Passport)</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Country Issuing International ID</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>International ID Details Linked</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NI Number</td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>First Line Address</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Postcode</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Telephone Number(s)</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>56</strong></td>
<td><strong>6</strong></td>
<td><strong>26</strong></td>
</tr>
</tbody>
</table>
Data Standards - Approach

Select Target Attribute

» Identify Local Application Alternatives
» Look for Commonality
» Review Government Standards
» Review UK National Standards
» Review Other Government Standards (e.g. NIST)
» Review International Standards
» Assess Applicability
» Decide Solution
» Confirm Solution
» Publish Solution
Data Standards – Observations and Issues

» Local Standards
  » Name
    - Two Fields (Given Names and Family Name)
    - Single Composite Fields
  » Problems
    - Which is the Family Name (single field option)
    - Cultures with no Family Name
    - Length, Which Character set to Use
    - Using National Standards to International Situation
  » Gender
    - Single Character pretty universally (but also text)
    - Which standard to apply (M/F or H/D, not known, not disclosed, not specified)
Data Standards: Given Names Resolution

» eGIF

» Originally CDM mandated to be compliant with eGIF
  » Too short

» Application Standard
  » Acknowledged as almost long enough!

» New Standard for length 100 characters prevails

» Promoted the publication on CDM website/Release Note
Data Standards: Gender Resolution

» Application examples
  » Values: M/F, M/F/U, M/F/D

» eGIF:
  » 4 values 0,1(Male), 2(Female), 9

» Use of ISO/IEC 5218:2004 standard, enables international interchange

» Standard for 4 UK values with mappings

» Promoted the publication on CDM website/Release Note
Data Profile Reporting

» Standard Data Profiling
  - Field Type Distribution
  - Field Uniqueness
  - Relationship Integrity

» CDM Profiling
  - Person Baseline
  - Event Baseline

» Issues
  - Root Cause Analysis
  - People/Process/Technology Approach
Field Type Distribution Report

Presents the percentage of rows within a particular field that fall into any of the following field type categories:

» Null
» Integer
» String
» Decimal
» Space
Field Uniqueness Report

Presents the percentage uniqueness of a column

- Null Percentage
- All Distinct e.g. ID fields
- Reference Data (few distinct values)
- Form a view of adequate uniqueness
Relationship Integrity Report

Presents the widows and orphans by number/percentage

<table>
<thead>
<tr>
<th>TABLE 1.FIELD 1</th>
<th>TABLE 2.FIELD 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows Loaded</td>
<td>Non-Matching</td>
</tr>
<tr>
<td>3016</td>
<td>1247</td>
</tr>
<tr>
<td>Rows</td>
<td>TABLE 1.Field1</td>
</tr>
<tr>
<td>Non-Matching</td>
<td>Values</td>
</tr>
<tr>
<td>1247</td>
<td>1769</td>
</tr>
<tr>
<td>Matching</td>
<td>TABLE 1.Field1</td>
</tr>
<tr>
<td>1312090</td>
<td>Values</td>
</tr>
<tr>
<td>Matching</td>
<td>TABLE 2.Field2</td>
</tr>
<tr>
<td>15742872</td>
<td>Non-Matching</td>
</tr>
<tr>
<td>Non-Matching</td>
<td>TABLE 2.Field2</td>
</tr>
<tr>
<td>2522782</td>
<td>Rows</td>
</tr>
<tr>
<td>Rows</td>
<td>TABLE 2.Field2</td>
</tr>
<tr>
<td>15742872</td>
<td>Rows Loaded</td>
</tr>
</tbody>
</table>

Join Analysis
valid as of xx/x/xxxx
Profiling Examples: Given Names

» Presence – 101 Null
» Unique Values – 99,999 N/A
» Patterns – 4,000
  » Leading Spaces – 100
  » Contains Digits – 50
  » Initials Only – 25
  » Trailing Spaces – 50
  » Leading Punctuation – 100
  » Non-printable Characters – 10
» Minimum Value – 0
» Maximum Value – Â…
Profiling Examples: Gender

- Presence – 25,000 Null
- Unique Values – 25 values
- Patterns – 8
  - Leading Spaces – 1
  - Contains Digits – N/A
  - Initials Only – 2
  - Trailing Spaces – 3
  - Leading Punctuation – 4
  - Non-printable Characters – 5
- Minimum Value – male
- Maximum Value – vnm
## Gender – Unique Values

<table>
<thead>
<tr>
<th>Value</th>
<th>Frequency</th>
<th>Dist %</th>
<th>Length</th>
<th>Soundex</th>
<th>Metaphone</th>
<th>Pattern</th>
<th>Mask</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>357227</td>
<td>75.375</td>
<td>4</td>
<td>M400</td>
<td>ML</td>
<td>a4</td>
<td>AAAA</td>
</tr>
<tr>
<td>Female</td>
<td>83865</td>
<td>17.696</td>
<td>6</td>
<td>F540</td>
<td>FML</td>
<td>a6</td>
<td>AAA...</td>
</tr>
<tr>
<td>Unknown</td>
<td>6523</td>
<td>1.376</td>
<td>7</td>
<td>U525</td>
<td>UNKNN</td>
<td>a7</td>
<td>AAA...</td>
</tr>
<tr>
<td>male</td>
<td>5</td>
<td>0.001</td>
<td>4</td>
<td>M400</td>
<td>ML</td>
<td>a4</td>
<td>AAAA</td>
</tr>
<tr>
<td>M</td>
<td>2</td>
<td>0.001</td>
<td>1</td>
<td>M000</td>
<td>M</td>
<td>a</td>
<td>A</td>
</tr>
<tr>
<td>male</td>
<td>1</td>
<td>0.001</td>
<td>5</td>
<td>M400</td>
<td>ML</td>
<td>_a4</td>
<td>AAAA</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>0.001</td>
<td>2</td>
<td>M000</td>
<td>M</td>
<td>pa</td>
<td>A</td>
</tr>
<tr>
<td>`m</td>
<td>1</td>
<td>0.001</td>
<td>2</td>
<td>M000</td>
<td>M</td>
<td>pa</td>
<td>`A</td>
</tr>
<tr>
<td>f</td>
<td>1</td>
<td>0.001</td>
<td>1</td>
<td>F000</td>
<td>F</td>
<td>a</td>
<td>A</td>
</tr>
<tr>
<td>FEMALE</td>
<td>1</td>
<td>0.001</td>
<td>6</td>
<td>F540</td>
<td>FML</td>
<td>a6</td>
<td>AAA...</td>
</tr>
<tr>
<td>female</td>
<td>1</td>
<td>0.001</td>
<td>6</td>
<td>F540</td>
<td>FML</td>
<td>a6</td>
<td>AAA...</td>
</tr>
<tr>
<td>GHA</td>
<td>1</td>
<td>0.001</td>
<td>3</td>
<td>G000</td>
<td>F</td>
<td>a3</td>
<td>AAA</td>
</tr>
<tr>
<td>ind</td>
<td>1</td>
<td>0.001</td>
<td>3</td>
<td>I530</td>
<td>INT</td>
<td>a3</td>
<td>AAA</td>
</tr>
<tr>
<td>ma</td>
<td>1</td>
<td>0.001</td>
<td>2</td>
<td>M000</td>
<td>M</td>
<td>a2</td>
<td>AA</td>
</tr>
<tr>
<td>make</td>
<td>1</td>
<td>0.001</td>
<td>4</td>
<td>M200</td>
<td>MK</td>
<td>a4</td>
<td>AAAA</td>
</tr>
<tr>
<td>MALE</td>
<td>1</td>
<td>0.001</td>
<td>4</td>
<td>M400</td>
<td>ML</td>
<td>a4</td>
<td>AAAA</td>
</tr>
<tr>
<td>vnm</td>
<td>1</td>
<td>0.001</td>
<td>3</td>
<td>V500</td>
<td>FNM</td>
<td>a3</td>
<td>AAA</td>
</tr>
</tbody>
</table>
CDM Profiling – Report and Presentation

» Report
  » Executive Summary
  » Observations
    - People
    - Process
    - Technology
  » Profile Scope
  » How to read the Document
  » Business Drivers
  » Profile Objective
  » Profile Approach
  » Assumptions
  » Observation Details
  » Next Steps

» Presentation
  » Introduction
  » Process Overview
  » Observations
  » Data Facts
    » Standard Reports
    » Person Baseline Analysis
  » Findings
  » Recommendations
  » Next Steps

30 Defining a Person Metadata Model to Improve Data Quality
Addressing the Issues

» Root Cause Analysis
  » Fact Based
  » To preclude problem recurring

» Examine
  Code Specifications
  UI and UI Standards Use of dropdowns
  Field Validation Talk with Users

» Reporting Approach
  » People
  » Process
  » Technology

» General BA Tools.
Thank You