# A Model for Semantic Equivalence Discovery for Harmonizing Master Data

**Baba Piprani**
**MetaGlobal Systems,**
**Canada**
**ORM '09 Workshop**
**Vilamoura - 2009**

# Baba Piprani
# MetaGlobal Systems, Canada

- Over 30 years experience teaching/implementing Object Role Modelling and implementing 100%  rules in SQL DBMSs
- Canadian Delegate in ISO SQL and Metadata Standardization in ISO/IEC JTC1 SC32 since 1986, and in the ISO WG on Conceptual Schemas(TR9007)
- Past Chair - ISO Conceptual Schema Modelling Facilities WG
- Past Chair - Govt. of Canada Treasury Board IT Standards (TBITS SQL)
- Author of several publications
- Highly successful DW and metadata based web implementations using Model Driven Architecture

# Abstract

IT projects often face the challenge of harmonizing metadata and data so as to have a “single” version of the truth. Determining equivalency of multiple data instances against the given type, or set of types, is mandatory in establishing master data legitimacy in a data set that contains multiple incarnations of instances belonging to the same semantic data record

The results of a real-life application define how measuring criteria and equivalence path determination were established via a set of “probes” in conjunction with a score-card approach.

There is a need for a suite of supporting models to help determine master data equivalency towards entity resolution---including mapping models, transform models, selection models, match models, an audit and control model, a scorecard model, a rating model.

An ORM schema defines the set of supporting models along with their incarnation into an attribute based model as implemented in an RDBMS.

# Agenda

1. Data Redundancy and Integration Issues
2. Use Case Scenario
3. Establishing Semantic Mapping – type level
4. Establishing Semantic Mapping – instance level
5. Determining Equivalency – Probes
6. ORM Mapping  Implementation

# Data Redundancy and Integration Issues

Major Issues….

- Data Duplication – not uncommon to come across multiple sets of redundant data on customers, products……registry prone items!
- Average figure 30%-60% redundant data
- No single version of the fact
- (Sometimes we don’t want a single version…)
- How then are we to integrate data ?

# Data Redundancy and Integration Issues

Master Data…2 kinds….

- Reference Data: e.g. types and categories pertaining to the properties or characteristics of data…reference tables like address type code employee type code…
- Master Registry (a.k.a. master file): e.g. data for customers, clients, vendors, products, etc…

Mergers/acquisitions/takeovers etc…necessitate rapid harmonization without opportunity to rebuild….

# Use Case....(not a useless case!)

Sept 11 2001…US terrorist attacks…

- Critical air traffic situation in Canada
- Hundreds of US bound planes re-routed to land in Canada
- Urgent need to determine / match
  - Runway properties
  - Airport emergency service facilities
  - Airside service facilities
  - Passenger airport facilities
  - Aircraft type, aircraft size, number of passengers
  - Fuel situation…..etc….

# Transport Canada situation….9/11

## Transport Canada had a serious situation…

- Regulatory agency responsible for transportation…
- Needed to access context related data across multiple systems..ran into….
  - 1:1 Airport Mappings were not always available…namely not all airport locations had IATA or ICAO international codes…(Canada and other countries have local codes (CFS Canada Flight Supplement )
  - Inconsistent terminology or identification referencing same concept
  - Rounding errors for latitude/longitude produced multiple airports at exact same locaion (3’ vs 29’ were rounded to 0’)
  - Cities within Canada were inconsistently referenced
- Not all applications presented consistent navigability for the involved concepts of aircraft type, runway properties , airport facilities, emergency facilities etc…

# Transport Canada situation….9/11

Transport Canada TOD (Transport Object Dictionary) project was born

- Harmonize data across various applications for a consistent common transparent access
- Establish common mappings across various applications
- On –demand data warehouse/application access
- Concepts:
  - Location
  - Carrier Org
  - Carrier Individual
  - Aircraft
  - Make . Model….etc

# Transport Canada situation….9/11

- 1st step towards Transport Object Dictionary (TOD)….
  - Establish semantic  equivalency
  - Match like instances
  - Map to common global identifier
- Problems encountered:
  - Lack of data quality
  - Typographical misspellings
  - Typo errors
  - Non-standard notations (e.g. organization suffix)
  - Abbreviations
  - Whitespace…etc

# Transport Canada situation….9/11

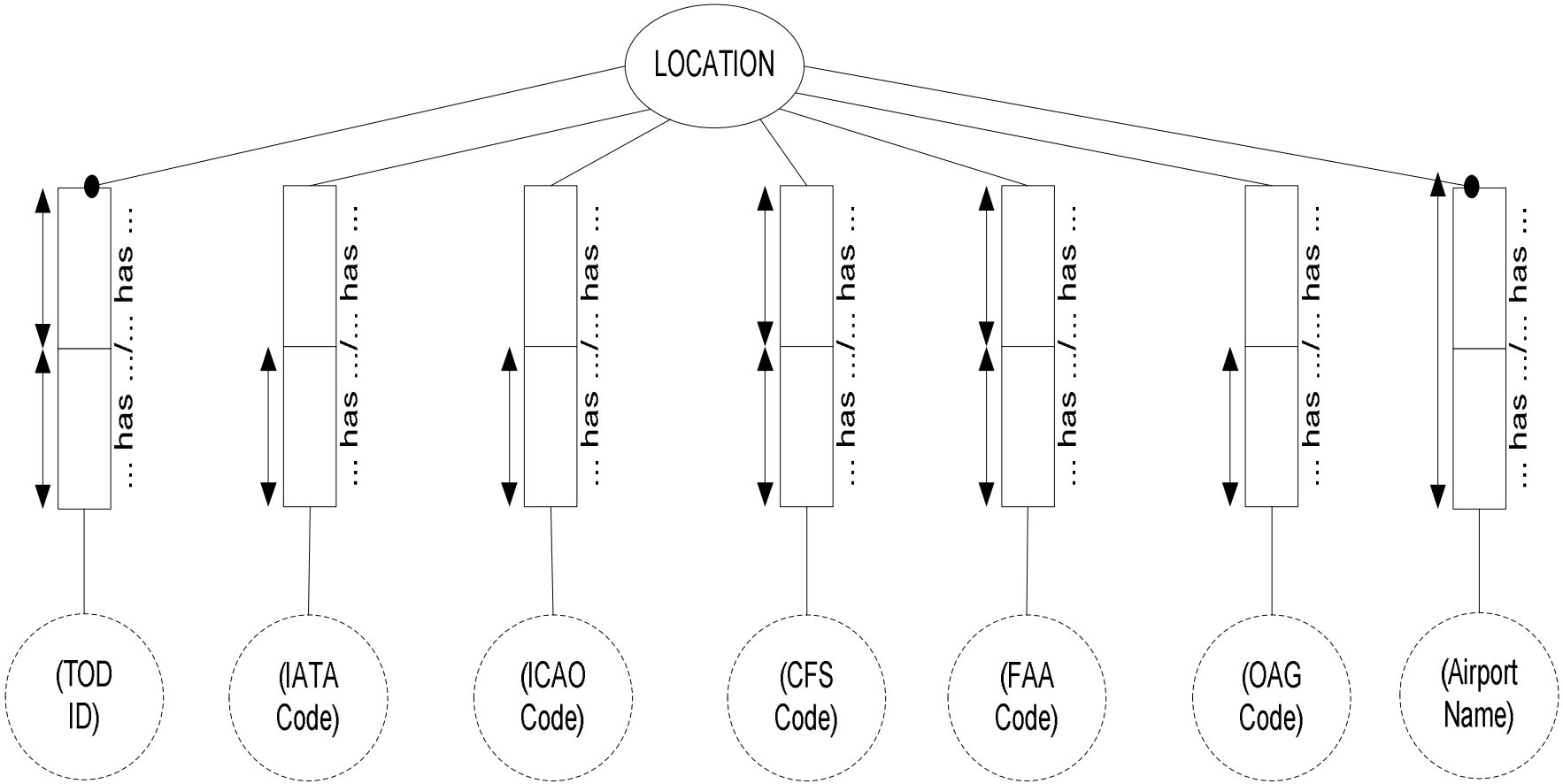
- Mappings needed to be harmonized at 2 levels
  - Metadata level – type
  - Value level – instance

- i.e. develop cross-walk to enable common mapping from participating applications to a global identifier based on
  - Type – i.e. qualified columns
  - Instance – local referencing schemes

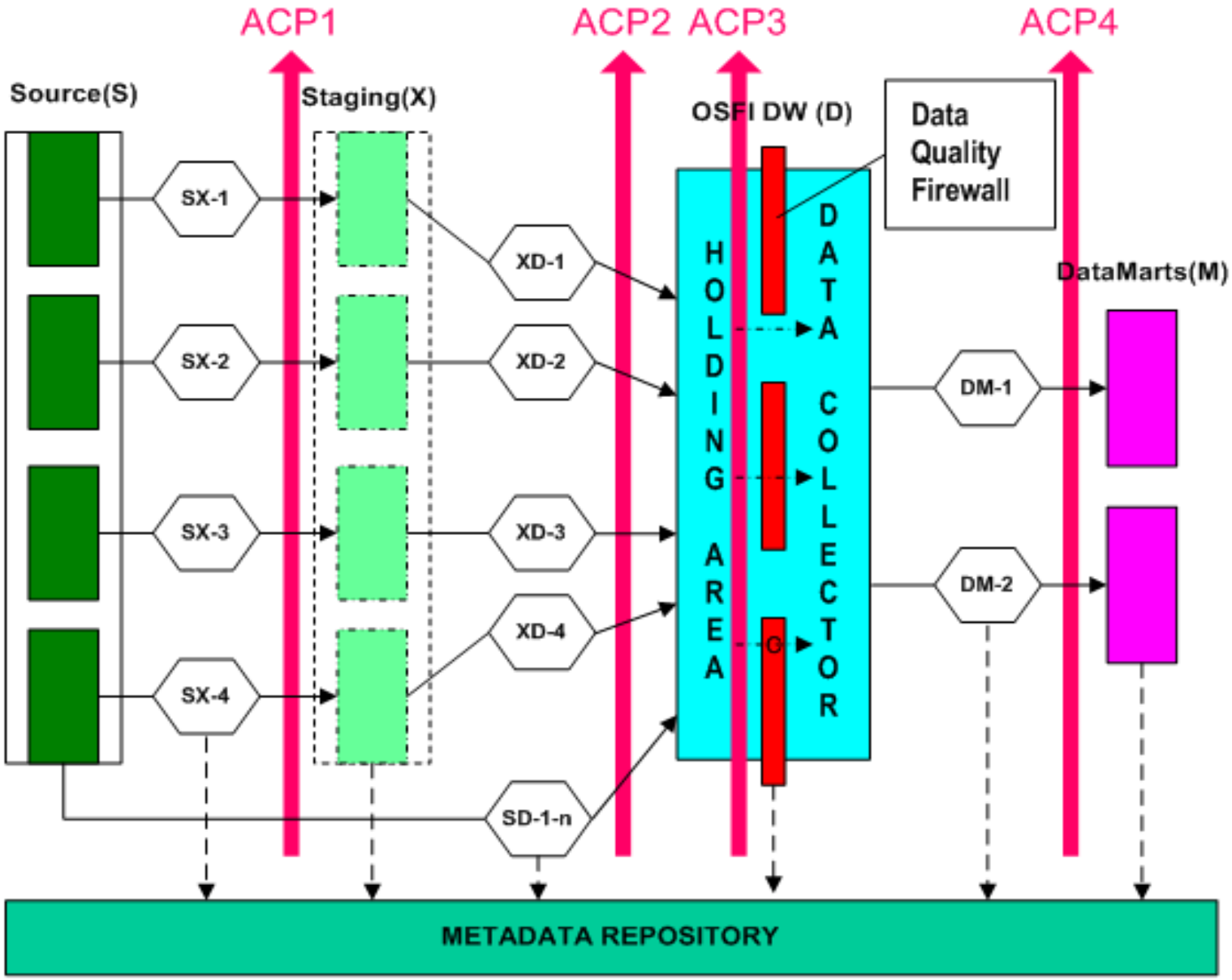
# Agenda

1. Data Redundancy and Integration Issues
2. Use Case Scenario
3. **Establishing Semantic Mapping – type level**
4. **Establishing Semantic Mapping – instance level**
5. Determining Equivalency – Probes
6. ORM Mapping  Implementation

# Mapping to Global Identifier



Audit control points in the data warehouse framework



# Apply Knowledge based concepts

**TOD Two pronged Approach:**
**to address: Mappings and,**
**Semantics**

**1) Mapping correlation of**
**a) Metadata**
**b) Values**
**2) Master Data Set**
**(for identification and usage)**



Legend:
**CMK → Common Mapping Key (for addressing Mapping problem)**
**TOD → Transport Object Dictionary (for addressing Semantics problem)**

# ORM Mapping Model - Type

- Between source data element to target qualified column list
- Involve other criteria
  - Boundary conditions
  - CAST descriptor
  - Mapping Criteria
  - Transform Criteria
  - Filter Criteria

# Establishing Semantic Mapping : Additional Criteria…

- *Boundary Conditions:* denotes range, or value or other integrity constraint on the involved column(s) for mapping purposes to be included in the data mapping, usually denoted and exercised by an SQL CHECK clause or User Defined Function. *e.g. The carrier UID in CPF is CHECK (carrierUID BETWEEN 1 AND 999999).* Contains the actual CHECK clause to be used that can be automatically included in a dynamic SQL statement.

- *CAST descriptor:* Any data type conversions on a global scale *(*see later section for value based CASTing), usually exercised via an SQL CAST predicate. Contains the actual CAST predicate to be used that can be automatically included in a dynamic SQL statement.

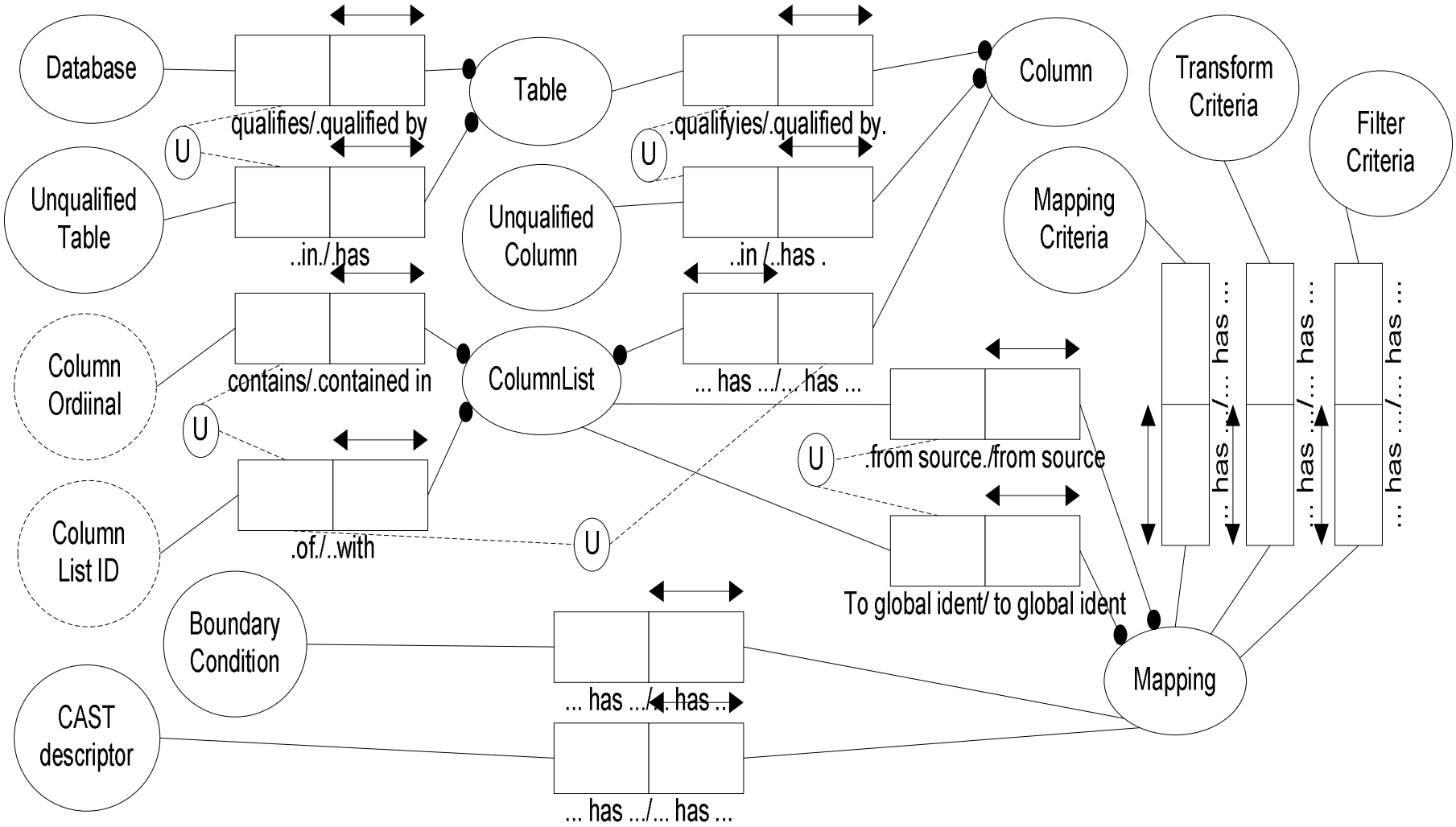
# Establishing Semantic Mapping : Additional Criteria…

- *Mapping Criteria:* Any matching criteria beyond the involved columns that affect the mapping *e.g. where type=xx, colour = yy etc*.

- *Transform Criteria:* Any conversion or transform that includes a change of variables or coordinates in which a function of new variables or coordinates is substituted for each original variable, *e.g. old Address Type = 2, becomes new Address Type = Business.* Contains the actual User Defined Function or code snippet to be used that can be automatically included in a dynamic SQL statement.

- *Filter Criteria:* Any applicable controlling criteria to limit the values of the selected set. Contains the actual User Defined Function or code snippet to be used that can be automatically included in a dynamic SQL statement.

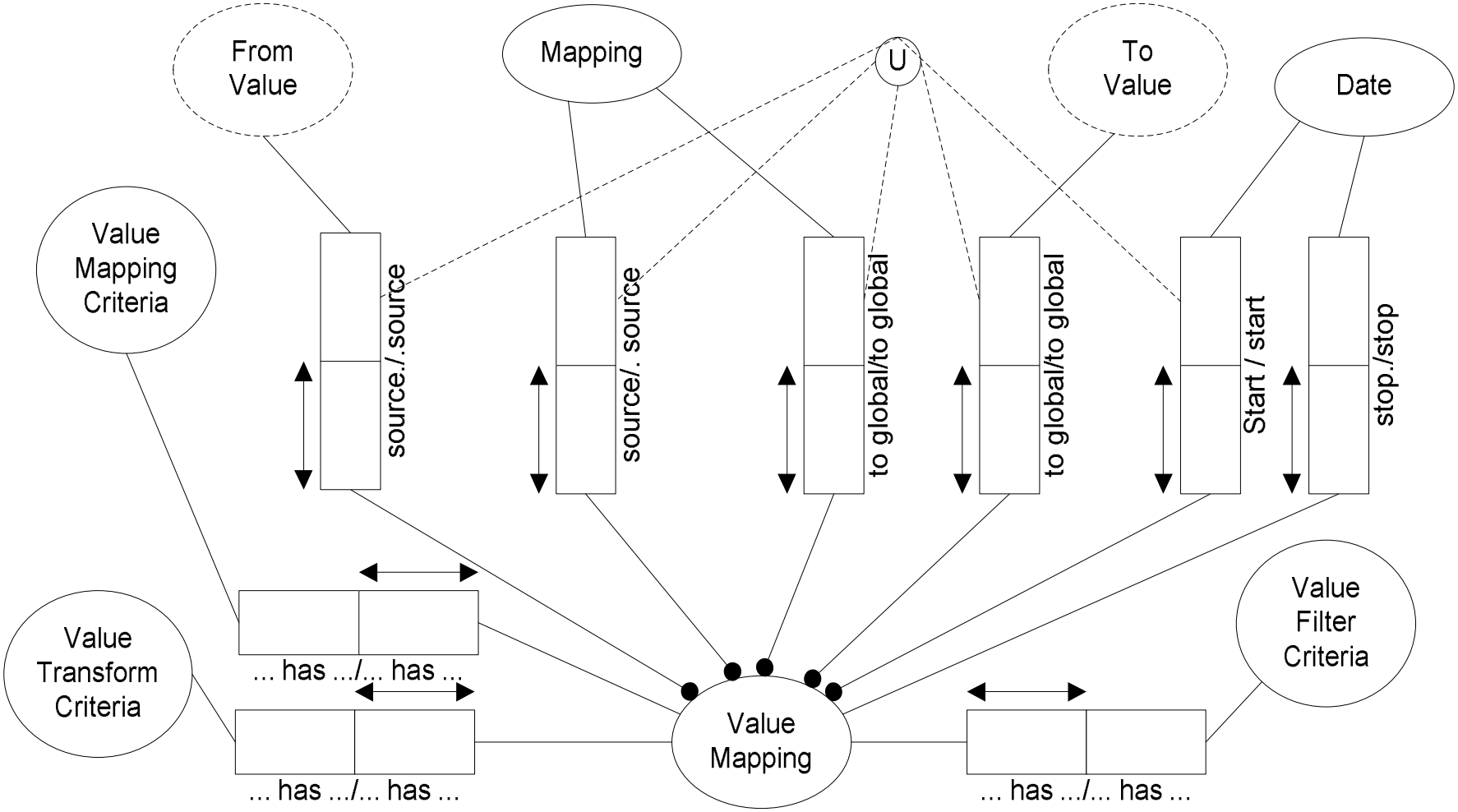
# ORM Mapping Model – Type Level



# ORM Mapping Model - instance

- Establish precise cross-walk navigation between applications
- Cross-correlate enable concordance between values from one application system to another via common global identifier
- Involve specific mapping criteria (if different from type)
  - Specific Mapping Value Criteria
  - Specific Transform Value Criteria

# ORM Mapping Model – Instance Level



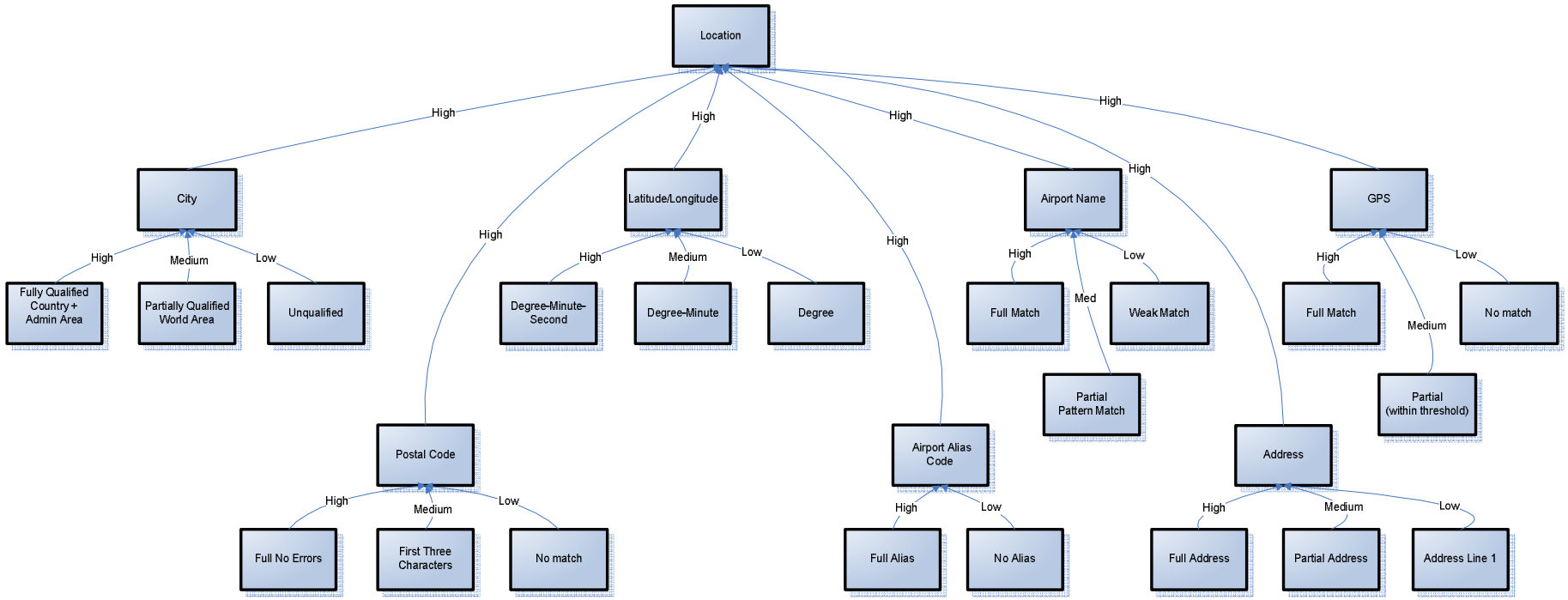
# Agenda

1. Data Redundancy and Integration Issues
2. Use Case Scenario
3. Establishing Semantic Mapping – type level
4. Establishing Semantic Mapping – instance level
5. **Determining Equivalency – Probes**
6. ORM Mapping  Implementation

# Determining Equivalency - probes

- At the value level
- Not easy….
  - Data quality errors
  - Most literature do not get this far
  - Type level mapping is all fine and dandy….the rubber hits the road at the value level!!!
  - Need additional matching properties (marriage broker)
    - decision tree + scorecard to determine go/nogo
    - e.g. location- known aliases (IATA/ICAO/CFS/other)
    - City
    - Latitude / longitude etc….

# Probe Based Model - Determining Location Equivalency




| Probe | City | Addr | AirptNm | PostCd | ALIAS | COORD | Score |
|---|---|---|---|---|---|---|---|
| LP1 | M | L | M | L | L | L | M+ |
| LP2 | L | H | L | - | L | H | H+ |
| LP3 | M | L | H | - | H | L | H+ |
| LP4 | L | L | L | - | L | L | L |
| LP5 | L | L | H | - | L | L | H- |
| LP6 | | | | | | | |
| LP7 | | | | | | | |
| LP8 | | | | | | | |
| LP9 | | | | | | | |

H = High [3 points]
M = Meidum [2 points]
L = Low [1 point ]

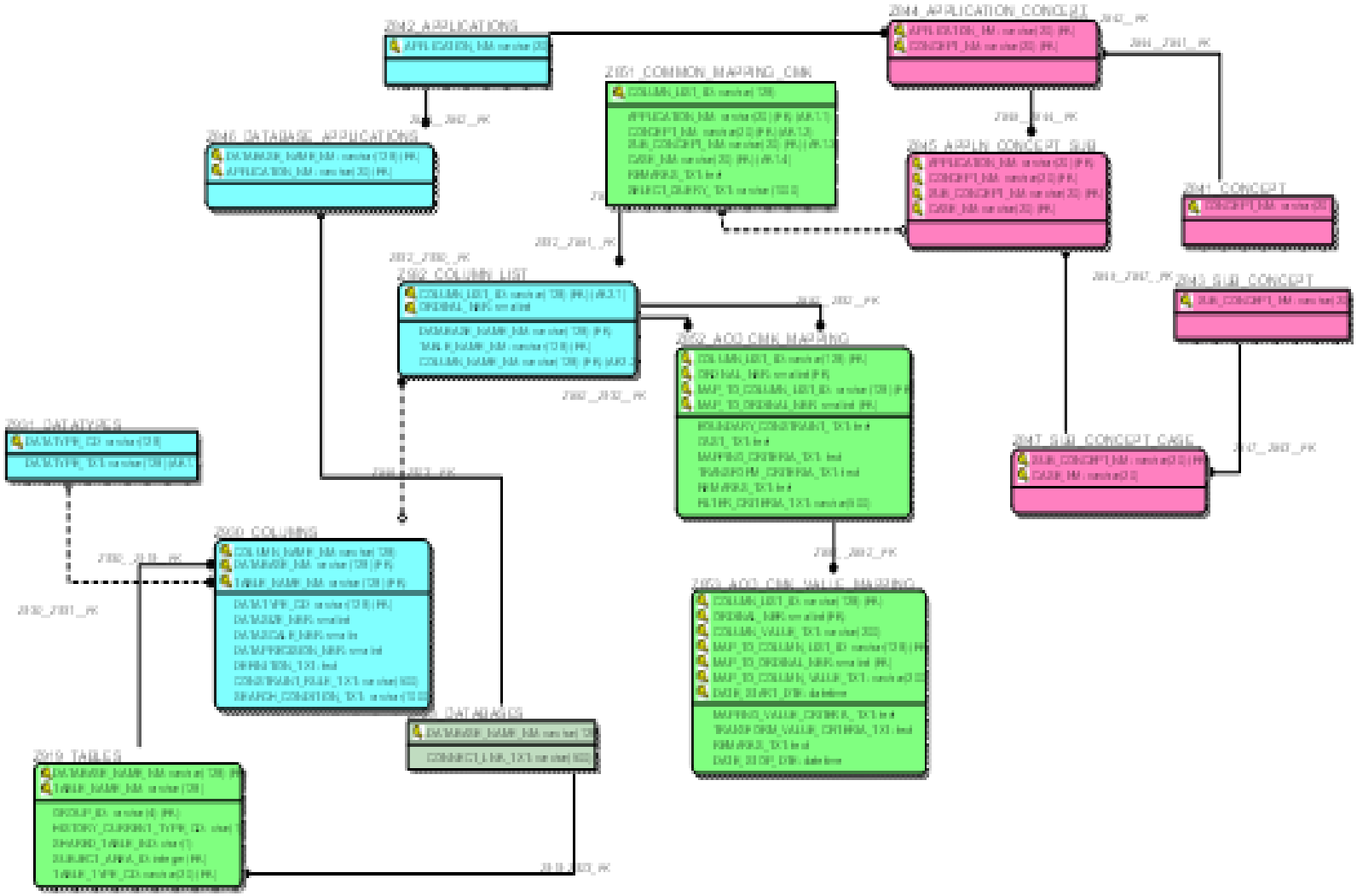
# Determine Score after probing…

- Get a scorecard report of matches
- Determine if equivalent location already exists,
- If score is high within established parameters: then assign previously defined global identifier, else, assign new
- probe versions….continuous improvement, self learning….

# Probe Scorecard

| Probe | City | Addr | AirptNm | PostCd | ALIAS | COORD | Score |
|---|---|---|---|---|---|---|---|
| LP1 | M | L | M | L | L | L | M+ |
| LP2 | L | H | L | - | L | H | H+ |
| LP3 | M | L | H | - | H | L | H+ |
| LP4 | L | L | L | - | L | L | L |
| LP5 | L | L | H | - | L | L | H- |
| LP6 | | | | | | | |
| LP7 | | | | | | | |
| LP8 | | | | | | | |
| LP9 | | | | | | | |

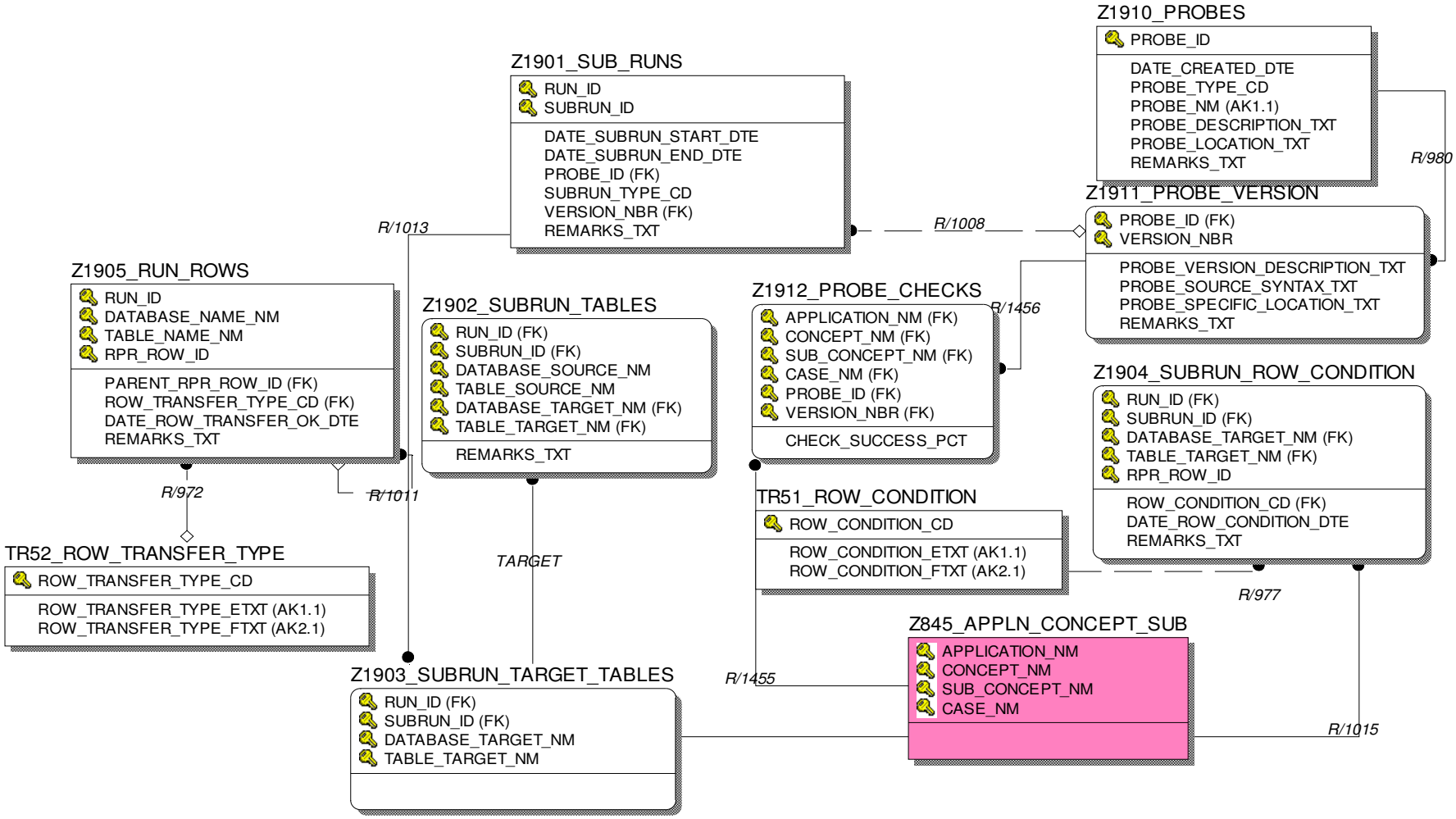
# Implementation Model of ORM Mapping Models (CMK)



# Audit and Control on Probe runs…

- Track which probe was run with what results..
- Assign score to instance from each probe run
- Probe sequence determined dynamically based on
  - prior score
  - value ranges/characteristics
  - combination of matched pairs
  - best probability of success…+ others (still exploring…)

# Probe Tracking Model (partial)

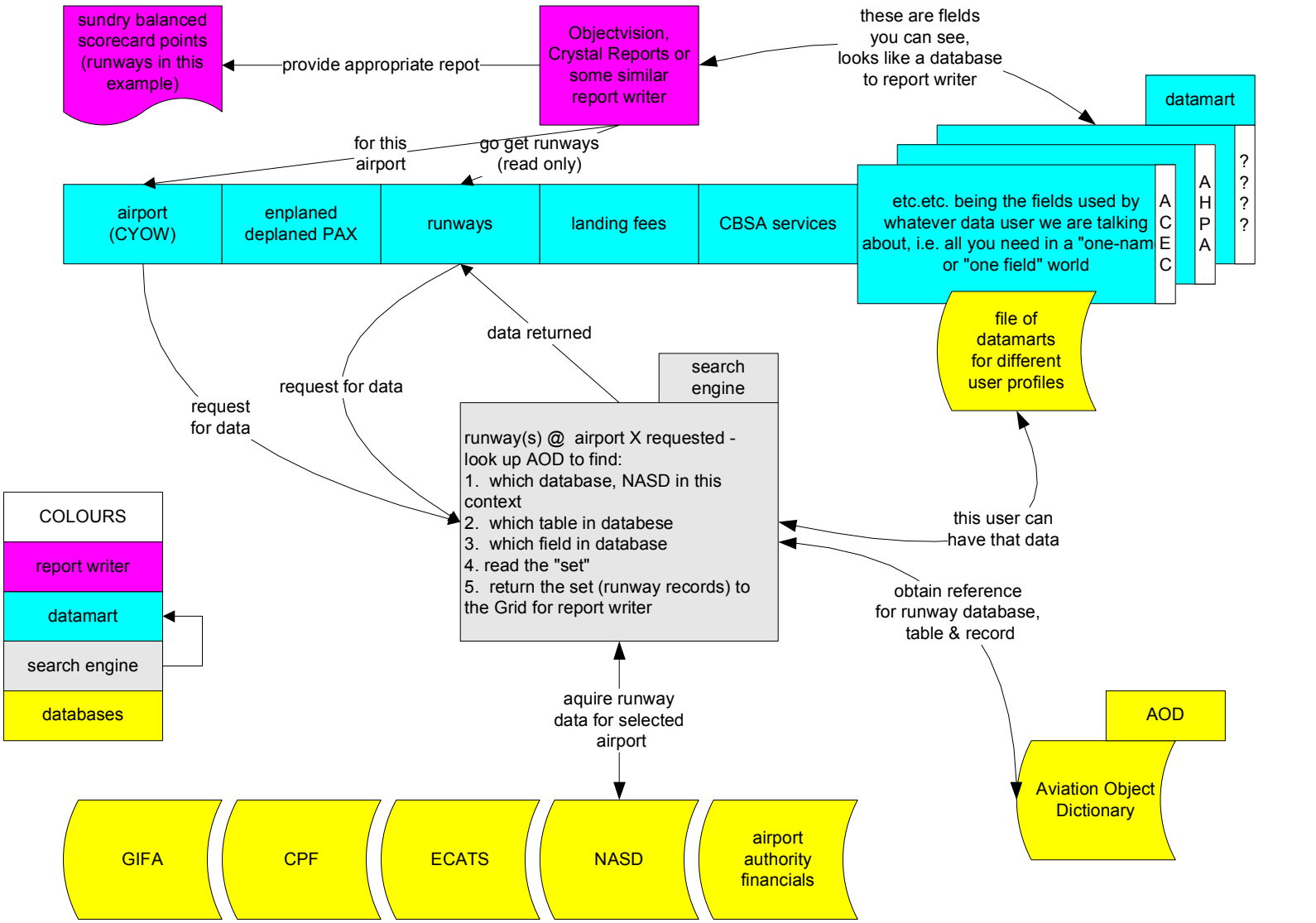


# Agenda

1. Data Redundancy and Integration Issues
2. Use Case Scenario
3. Establishing Semantic Mapping – type level
4. Establishing Semantic Mapping – instance level
5. Determining Equivalency – Probes
6. **ORM Mapping  Implementation**

# Successful Use Case… (definitely not useless case….)

- Shopping cart window of client requirements
- Source databases/ resources hidden or unknown to user
- Developed User Defined Functions (UDF) to fetch data from multiple sources after consult with the TOD master data metadata reference
  - Which involves loading of type and instance mappings
  - Source availability mappings + associated criteria etc)
- Currently refresh every week, eventually dynamic generation of UDFs for new shopping cart elements not in library, since all metadata is in TOD mapping library
- Totally designed using ORM….




Courtesy: Iain Henderson, Airport Policy, Transport Canada

# Acknowledgement

I would like to acknowledge and thank the Transport Canada TOD (Transport Object Dictionary) team, staff and users, Michel Villeneuve, Madhu Hundal, Tom Nash, Vasko Miovski, Dave McCutcheon, Rob Henkel, Wayne Shimoon, Jean-Pierre Sabbagh El Rami, Jean Yves Cadieux, Iain Henderson, and Dave Dawson for their support and advice in establishing the necessary mapping algorithms and developing probes in support of this successful application.

# And onward…

# Thank you….