

# MASTER YOUR DATA

Seth Anderson, AudioVisual Preservation Solutions  
AMIA 2013 | Richmond, VA

**get it**

**“CLEAN”**

**keep it**

**for**

**EXAMPLE**

*CA ERwin* Data Modeler

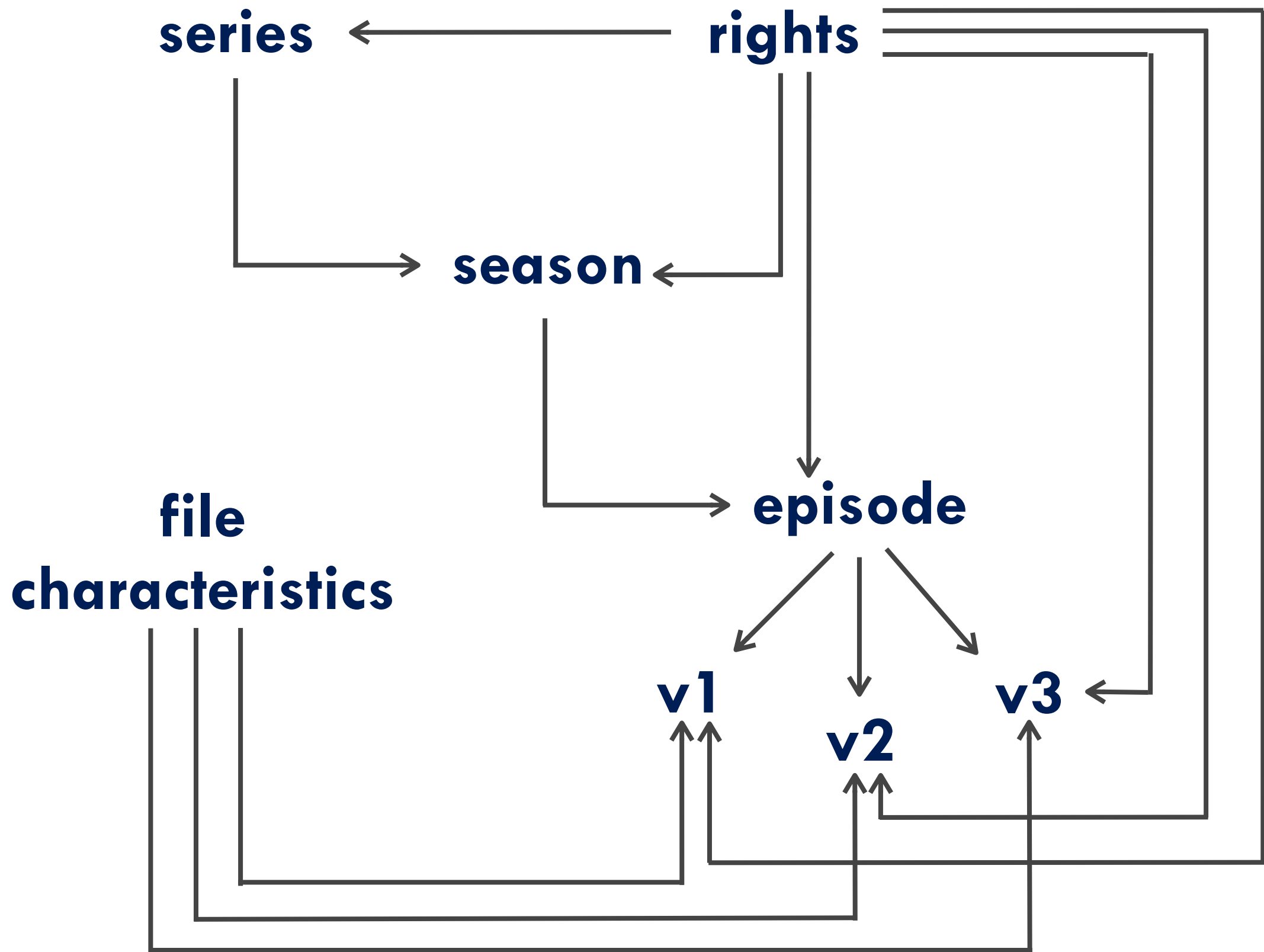
*Embarcadero ER/Studio*

*Sybase PowerDesigner*

*Dell Toad* Data Modeler

*IBM* Data Architect

**ENGINEERING**



Toad Data Modeler [C:\Program Files (x86)\Quest Software\Toad Data Modeler 4.0\Samples\Employee.tbl]

File Edit Objects View Notation Model Settings Tools Macros Help

Search Toad World (Oracle)

Application View, Gallery Explorer

Application View Gallery Explorer

Videorental Employee\*

All Items

Physical Model Explorer, Object Viewer

Physical Model Explorer Object Viewer

Videorental

Message Explorer, Loupe

Message Explorer Loupe

Customer Rating

Title NN (PFK)

Director NN (PFK)

Rating

Id	Date	Time	Message
2	6/17/2011	3:44:50 PM	---
3	6/17/2011	3:44:50 PM	---
4	6/17/2011	3:44:50 PM	Number of Errors: 0
5	6/17/2011	3:44:50 PM	Number of Warnings: 0

*Dell* Toad Data Modeler

**VOCABULARY**

**THESAURUS**

**ONTOLOGY**

**TOOLS**

*Synaptica* Knowledge  
Management System

*SmartLogic* Semaphore

*Mondeca* Smart Content Factory

*Access Innovations* Data  
Harmony

**VOCABULARY**



IPSV APR55 - Semaphore Ontology Manager

File Terms Tools Publish Configure Windows Help

- Business and industry
  - Business people
  - Business practice and regulation
    - Business advice services
    - Business development
    - Business management
      - Business planning
        - Corporate policy
        - Feasibility studies
        - Prioritising
      - Cooperation
      - Decision making
      - Financial management
      - Human resource management
      - Information management
      - Knowledge management
      - Management control
      - Organisational development
      - Performance management
      - Programme management
      - Project management
      - Public relations
      - Resource management
      - Risk management
    - Business studies
    - Business travel
  - e-Commerce
  - Ethical business practices
    - Fair trading
  - Health and safety at work
    - Insolvency
  - Marketing
  - Regulation and deregulation
- Business sectors
- Companies
- Consumer affairs
- Energy and fuel
- International trade
- Economics and finance
- Education and skills
- Employment, jobs and careers
- Environment
- Government, politics and public administration
- Health, well-being and care
- Housing
- Information and communication
- International affairs and defence
- Leisure and culture
- Life in the community
- People and organisations
- Public order, justice and rights
- Science, technology and innovation
- Transport and infrastructure

Business planning

Hierarchical

Type	Term
BT	Business management
NT	Corporate policy
NT	Feasibility studies
NT	Prioritising

Associative

Type	Term
RT	Business continuity planning
RT	Council policies and plans
RT	Organisational development

Equivalence

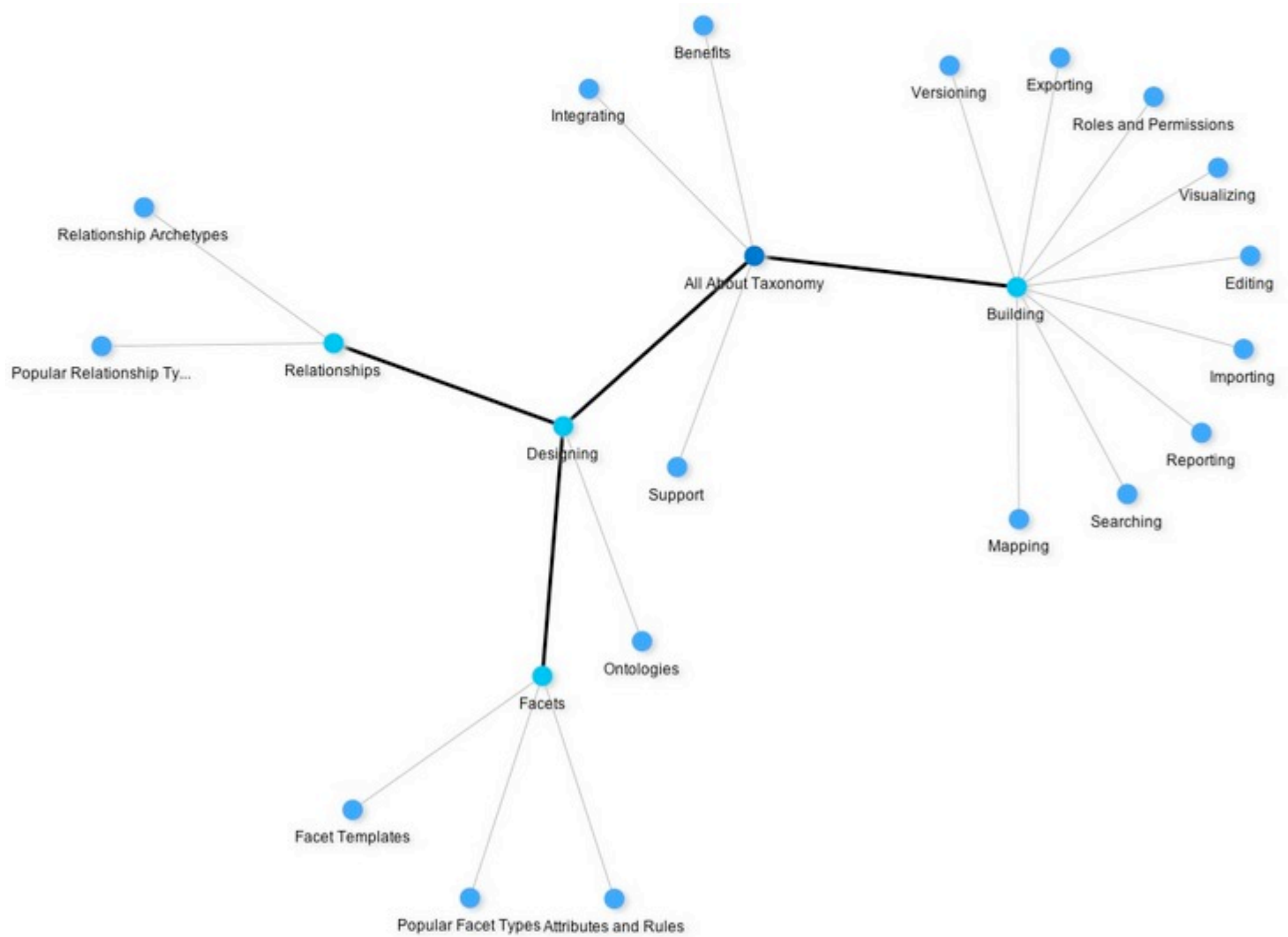
Type	Term
LF	Action plans
LF	Business Building Regulation approval
LF	Business Planning advice
LF	Business Planning environment
LF	Business Planning support
LF	Business Planning Zones
LF	Business plans
LF	Business Premises extensions
LF	Corporate Planning
LF	Forward Plans
LF	planning (business)
LF	Planning Applications business
LF	Planning Permission businesses
LF	Planning Permission work home
LF	Strategic planning

Hierarchical | Alphabetical | Hierarchy Select | Removed

Relationships | Family | Term Information | Term Attributes | Properties

IPSV APR55 | View: Edit | admin @ localhost:8001 | Server version: 5.1+35436

# SmartLogic Semaphore



# *Synaptica* Knowledge Management System

**the**  
**BIG**  
**ones**

*Adaptive Metadata Manager*

*ASG Rochade*

*IBM Infosphere*

*Informatica*

*Talend (open source!)*

**END-TO-END**

**Informatica PowerCenter Designer - [Mapping Designer - DataIntegrationMappings - [PC\_Repository]]**

Repository Edit View Tools Layout Versioning Mappings Transformation Window Help

DataIntegrationMappings - [PC\_Reposit] 100%

Repositories

- PC\_Repository
  - DataIntegrationMappings
    - Business Components
    - Sources
    - Targets
    - all\_cust\_data
    - DQ\_output
    - full\_cust\_data
    - Cubes
    - Dimensions
    - Transformations
    - Mapplets
    - Mappings
      - m\_customer\_data
      - m\_dq\_customer\_data
    - User-Defined Functions
  - DataQuality\_Rules
    - Business Components
    - Sources
    - Targets
    - Cubes
    - Dimensions
    - Transformations
    - Mapplets
      - mplt\_Company\_Name\_...
      - mplt\_Company\_Name\_F...
      - mplt\_Familyname\_and...
      - mplt\_Firstname\_and\_S...
      - mplt\_Individual\_Name\_...
      - rule\_Color\_Parse
      - rule\_Company\_Name\_S...
      - rule\_Completeness
      - rule\_Completeness\_Mul
      - rule\_Country\_Identifica
      - rule\_Country\_Name\_St...
      - rule\_CreditCard\_Numbe
      - rule\_Credit\_Card\_Secu
      - rule\_Currency\_Code\_C
      - rule\_Date\_Parse
      - rule\_Email\_Parse
      - rule\_Email\_Parse\_Into

**Mapping Designer**

**USAdditions (Flat File) Source Definition**

K	Name	Datatype
	ID	numb
	SYSTEM	string
	COMPANY	string
	ADDR1	string
	ADDR2	string
	ADDR3	string

**SQ\_USAdditions Source Qualifier**

Name	Datatype
ID	decimal
SYSTEM	string
COMPANY	string
ADDR1	string
ADDR2	string
ADDR3	string

**UStesting (Flat File) Source Definition**

K	Name	Datatype
	ID	numb
	SYSTEM	string
	COMPANY	string
	ADDR1	numb
	ADDR2	string
	ADDR3	string

**SQ\_UStesting Source Qualifier**

Name	Datatype
ID	decimal
SYSTEM	string
COMPANY	string
ADDR1	decimal
ADDR2	string
ADDR3	string

**Union Union Transformation**

Name	Datatype
ID	decimal
SYSTEM	string
COMPANY	string
ADDR1	string
ADDR2	string

**Shortcut\_to\_rule\_USA\_P... Mapplet Shortcut**

Name	Datatype
Input_NEWGROUP	
IN_NAME_NEW...	string
Output_OUTPUT	
Out_FullName_O...	string
Out_NameDesign...	string
Out_NamePrefix...	string

**Shortcut\_to\_rule\_USA\_P... Mapplet Shortcut**

Name	Datatype
Input_NEWGROUP	
In_Phone_NEWG...	string
Output_OUTPUT	
Out_Phone_Std...	string
Out_Phone_Dash...	string
Out_Phone_No_S...	string

**DQ\_output (Flat File) Target Definition**

K	Name	Datatype
	ID	numb
	SYSTEM	string
	COMPANY	string
	FirstName	string
	Firstname_Std	string
	Surname	string

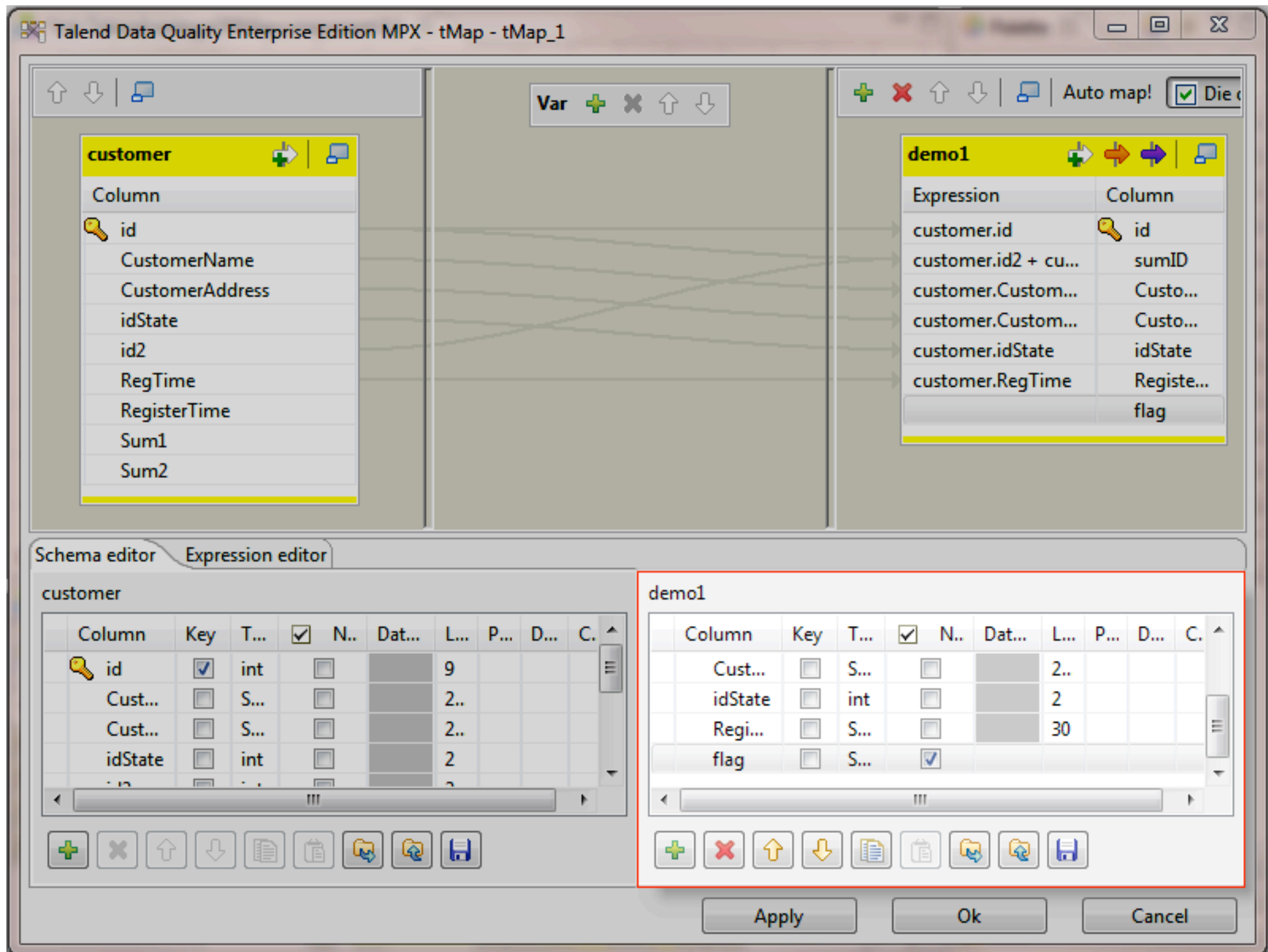
Parsing mapping m\_dq\_customer\_data...  
...parsing completed with no errors.

\*\*\*\*\* Mapping m\_dq\_customer\_data is VALID \*\*\*\*\*  
mapping m\_dq\_customer\_data updated.

Save Fetch Log Generate Validate Debugger Session Log Notifications

Ready NUM

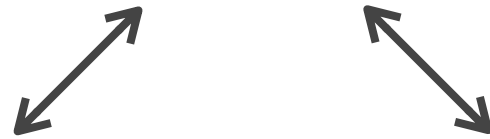
# Informatica Power Center



*Talend*

**dirty data**

**VOCAB**



**MODEL** ↔ **RULES**

**clean data**

**OPEN**

**REFINE**



**3/4" U-matic**

OR

**Umatic**

OR

**3/4" video**

**3/4 U-matic  
videocassette**

OR

**U-matic video**

OR

**Three quarter inch  
video tape cassette  
(U-matic)**

Create Project « Start Over Configure Parsing Options Project name  Create Project »

	INSTID	UKPRN	Region of institution	Institution	HE students Postgraduate	Mode of study Full-time	Part-time	Column	Domicile United Kingdom	Other European Union	Non- European-U
1.	0047	10000291	EAST	Anglia Ruskin University		1720	1465		1975	325	
2.	0108	10007759	WMID	Aston University		2270	920		1340	285	1
3.	0048	10000571	SWES	Bath Spa University		680	2595		3215	25	
4.	0109	10007850	SWES	The University of Bath		1950	3295		2950	575	1
5.	0026	10007152	EAST	University of Bedfordshire		3600	1780		1955	200	3
6.	0127	10007760	LOND	Birkbeck College(#9)		1235	3840		4325	325	
7.	0052	10007140	WMID	Birmingham City University		1975	2140		2680	150	1
8.	0110	10006840	WMID	The University of		6740	4910		7260	715	3

Parse data as Update Preview

- Excel files**
- JSON files
- Line-based text files
- CSV / TSV / separator-based files
- Fixed-width field text files
- PC-Axis text files
- RDF/N3 files
- XML files
- Open Document Format spreadsheets (.ods)

Worksheets to Import

- Table\_1 211 rows

- Ignore first  line(s) at beginning of file
- Parse next  line(s) as column headers
- Discard initial  row(s) of data
- Load at most  row(s) of data

- Store blank rows
- Store blank cells as nulls
- Store file source (file names, URLs) in each row

## Cluster & Edit column "AcademicDept"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

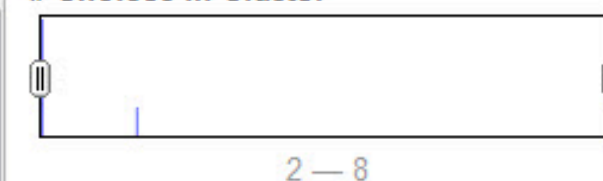
Method key collision

Keying Function fingerprint

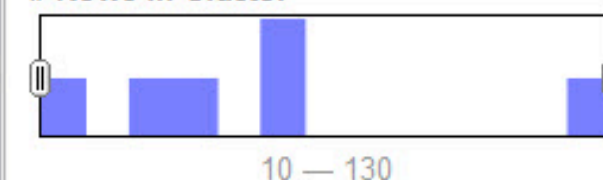
6 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	62	<ul style="list-style-type: none"> <li>Institute of Social and Industrial Relations. (35 rows)</li> <li>Institute of Social and Industrial Relations. (16 rows)</li> <li>Institute of Social and Industrial Relations. (4 rows)</li> <li>Institute of Social and Industrial Relations. (2 rows)</li> <li>Institute of Social and Industrial relations. (2 rows)</li> <li>Institute of Social and Industrial Relations. (1 rows)</li> <li>Institute of Social and Industrial Relations. (1 rows)</li> <li>Institute of Social and Industrial Relations (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Institute of Social and Indust
3	130	<ul style="list-style-type: none"> <li>School of Social Work. (98 rows)</li> <li>School of Social Work. (31 rows)</li> <li>School of Social Work. (1 rows)</li> </ul>	<input type="checkbox"/>	School of Social Work.
2	36	<ul style="list-style-type: none"> <li>Dept. of Philosophy. (35 rows)</li> <li>Dept. of Philosophy. (1 rows)</li> </ul>	<input type="checkbox"/>	Dept. of Philosophy.

# Choices in Cluster



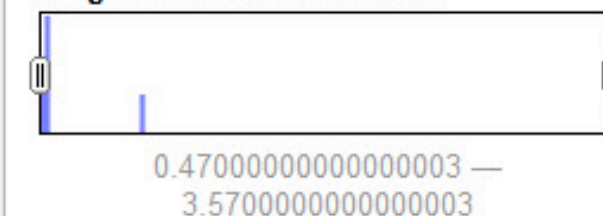
# Rows in Cluster



Average Length of Choices



Length Variance of Choices



Select All

Deselect All

Merge Selected & Re-Cluster

Merge Selected & Close

Close

**try them out**

**or**

**develop new tools**



[www.avpreserve.com](http://www.avpreserve.com)