MASTER YOUR DATA

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

get it

“CLEAN”

keep it

for

EXAMPLE

CA ERwin Data Modeler

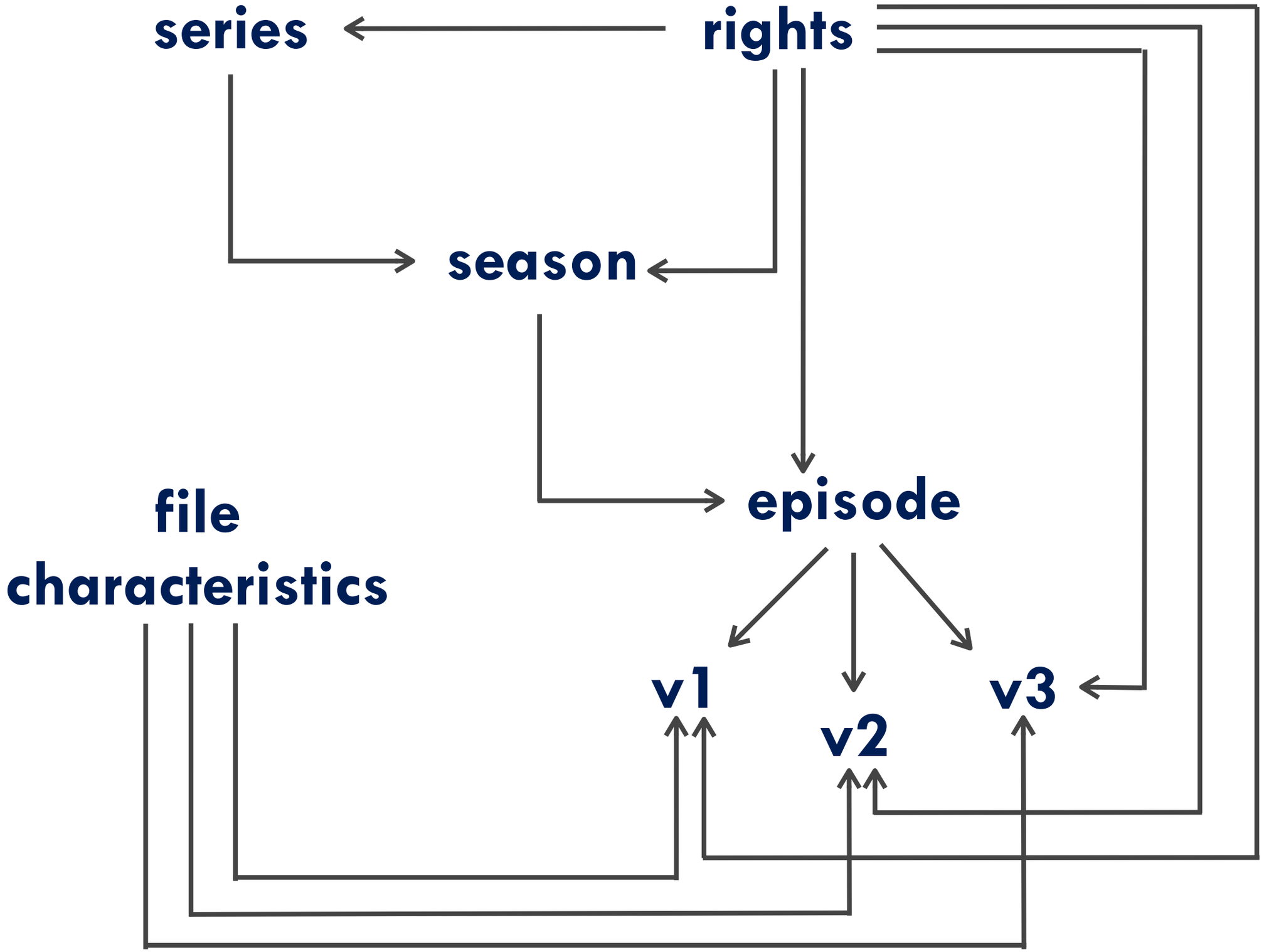
Embarcadero ER/Studio

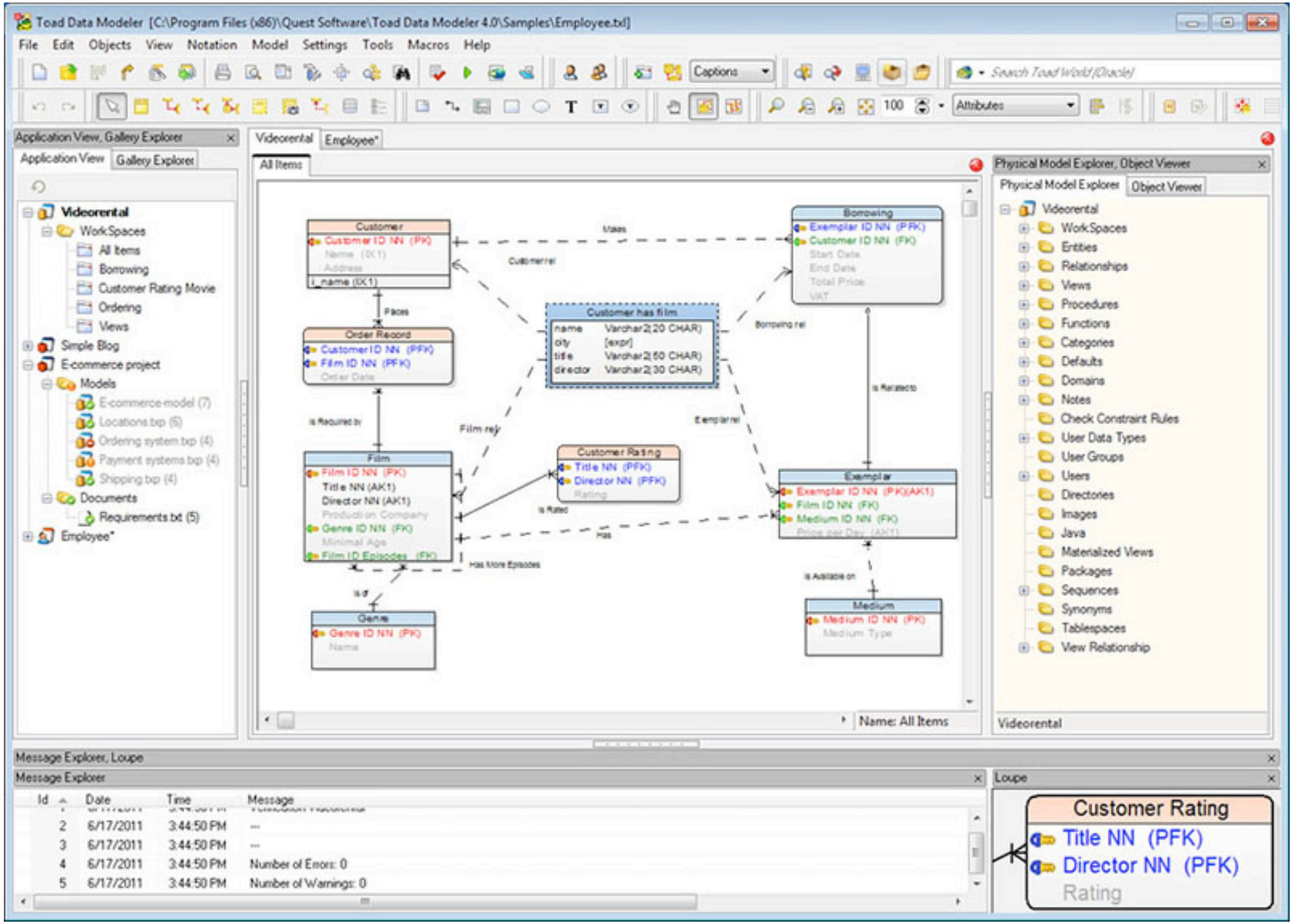
Sybase PowerDesigner

Dell Toad Data Modeler

IBM Data Architect

MODELING


series
rights
season
episode
file
characteristics
v1
v2
v3



Dell Toad Data Modeler

VOCABULARY THESAURUS

ONTOLOGY TOOLS

Synaptica Knowledge Management System

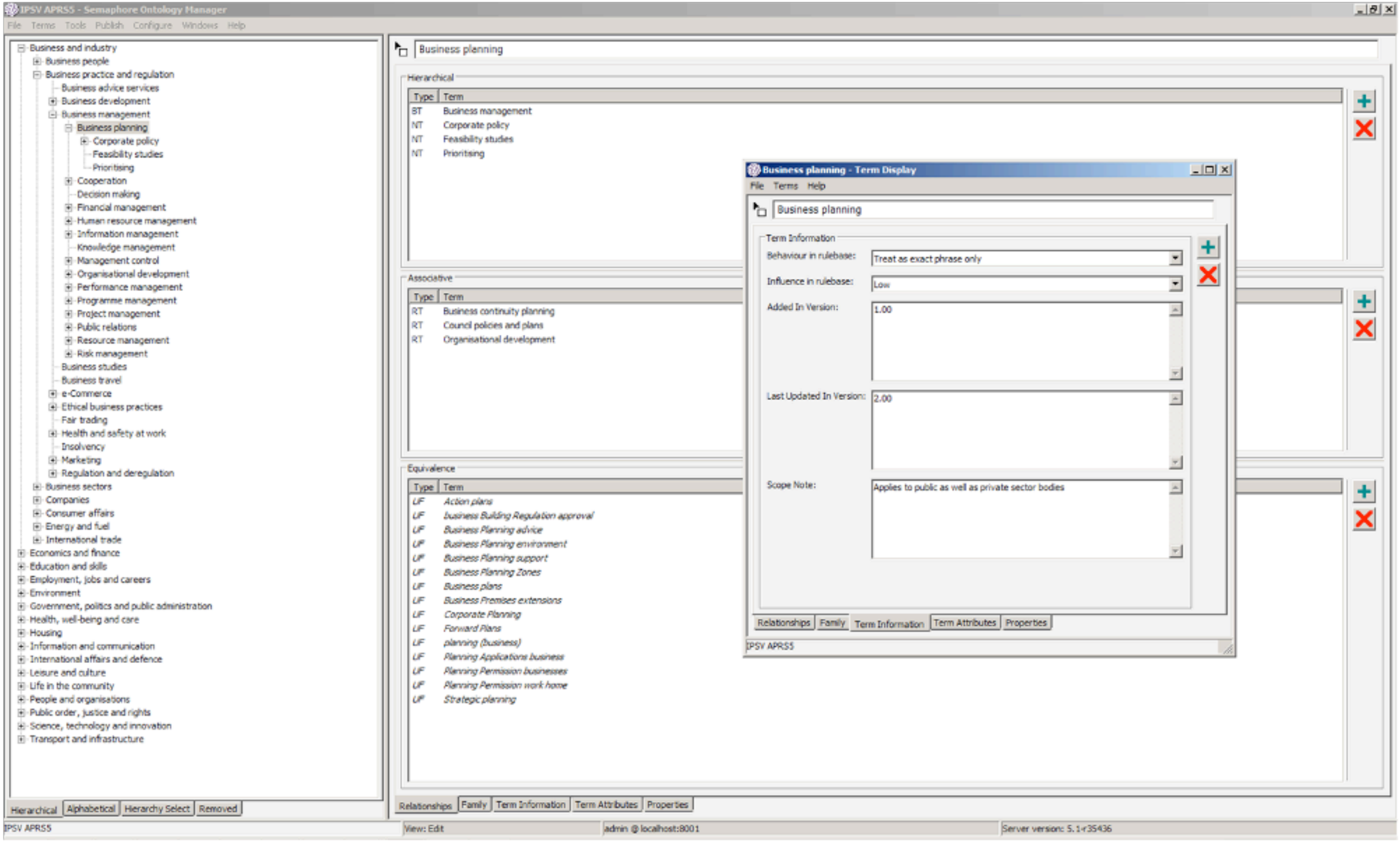
SmartLogic Semaphore

ULARY

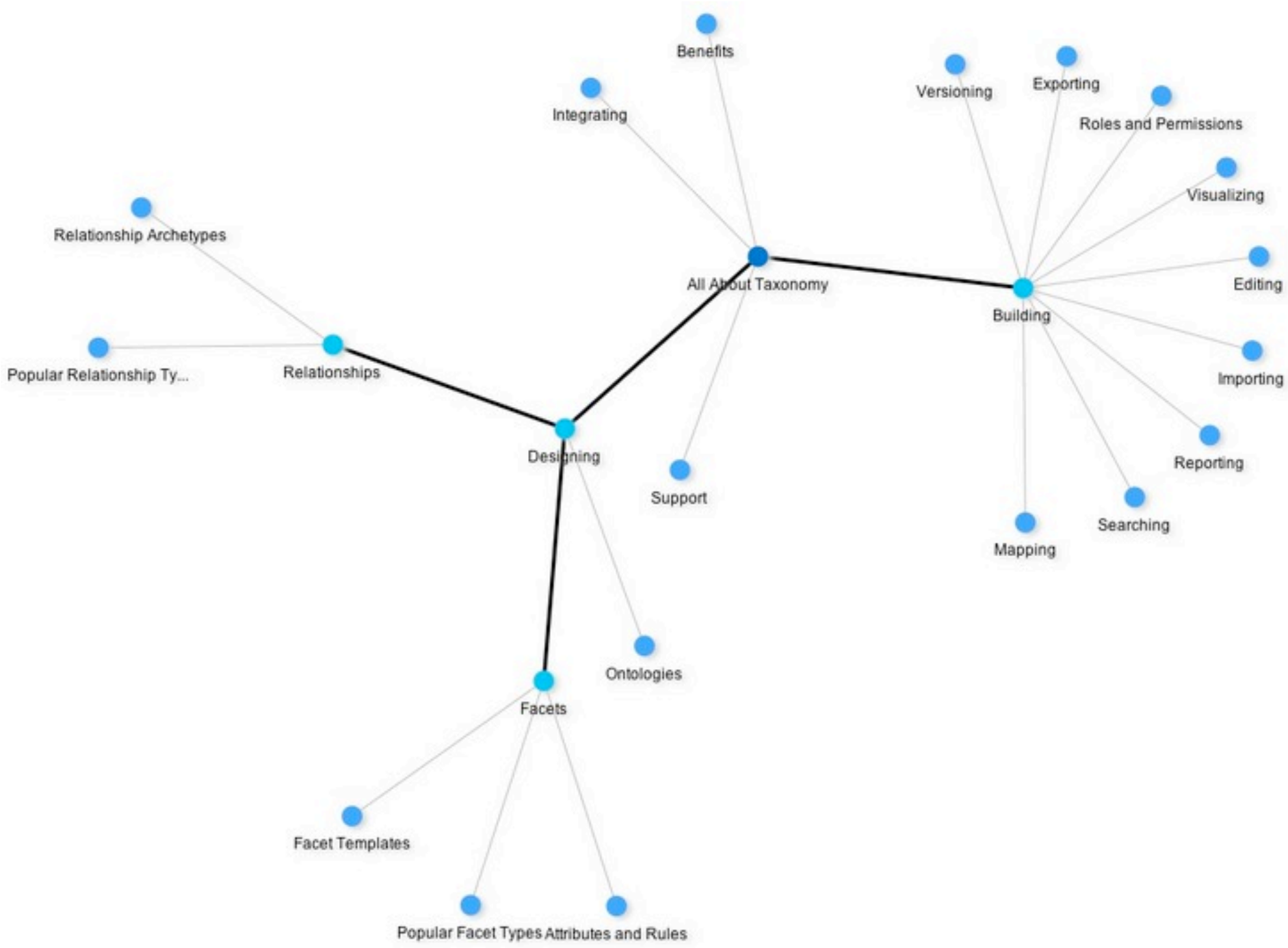
Mondeca Smart Content Factory

Access Innovations Data Harmony

VOCAB



SmartLogic Semaphore



Synaptica Knowledge Management System

the BIG ones

Adaptive Metadata Manager

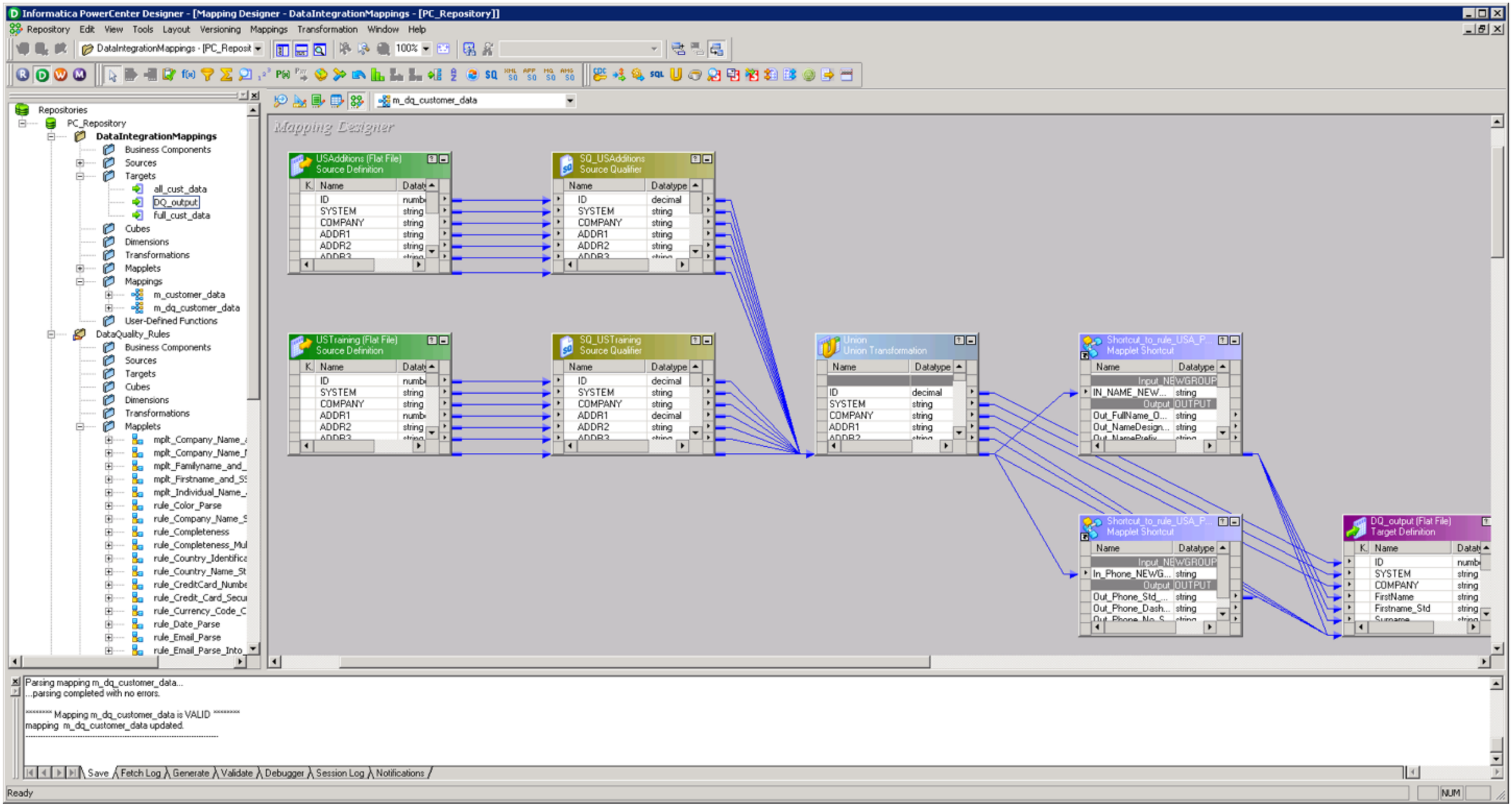
ASG Rochade

IBM Infosphere

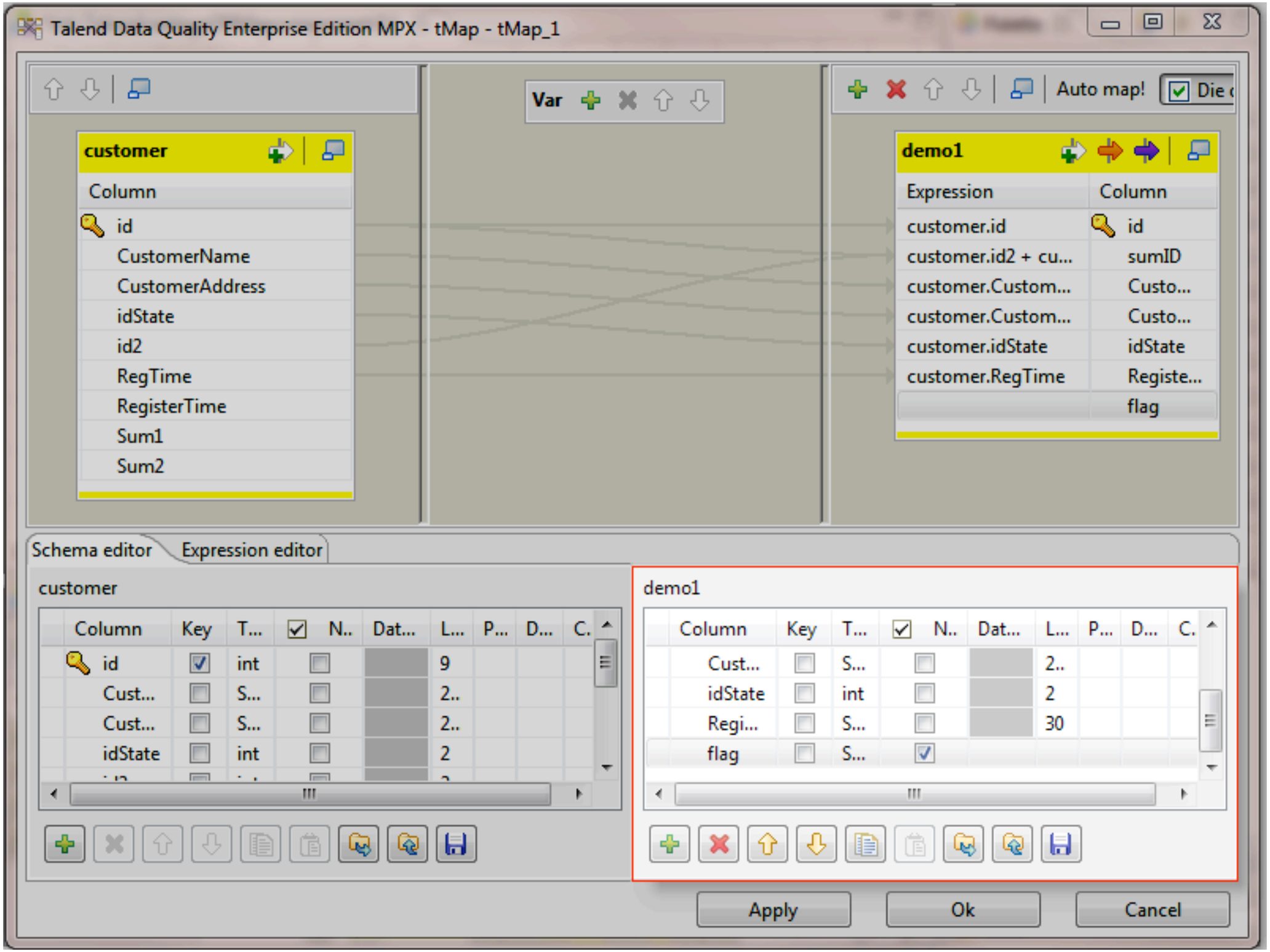
Informatica

Talend (open source!)

END-TO-END

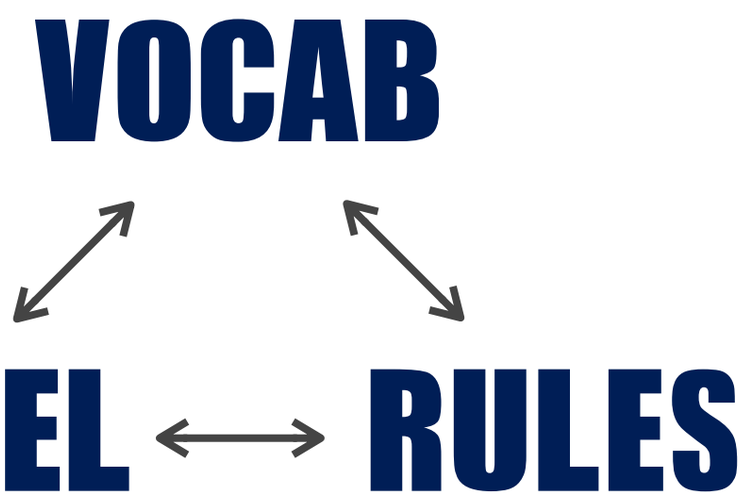


Informatica Power Center


Talend Data Quality Enterprise Edition MPX - tMap - tMap_1
Var
Auto map!
customer
Column
id
CustomerName
CustomerAddress
idState
id2
RegTime
RegisterTime
Sum1
Sum2
demo1
Expression
Column
customer.id
id
customer.id2 + cu...
sumID
customer.Custom...
Custo...
customer.Custom...
Custo...
customer.idState
idState
customer.RegTime
Registe...
flag
Schema editor
Expression editor
customer
Column
Key
T...
N..
Dat...
L...
P...
D...
C.
id
int
9
Cust...
S...
2..
Cust...
S...
2..
idState
int
2
demo1
Column
Key
T...
N..
Dat...
L...
P...
D...
C.
Cust...
S...
2..
idState
int
2
Regi...
S...
30
flag
S...
Apply
Ok
Cancel

Talend

dirty data



clean data

OPEN

REFINE

3/4” U-matic

OR

3/4 U-matic videocassette

OR

Umatic

OR

U-matic video

OR

3/4” video

OR

Three quarter inch video tape cassette (U-matic)

Google Refine

127.0.0.1:3333

Google refine *A power tool for working with messy data.*

Create Project
Open Project
Import Project

« Start Over | Configure Parsing Options | Project name HESA 2010/11 Students by Institution | 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

Version 2.5 [r2407]

Help
About

Parse data as

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)

Update Preview

Worksheets to Import
☑ Table_1 211 rows

☑ Ignore first 1 line(s) at beginning of file
☑ Parse next 5 line(s) as column headers
☑ Discard initial 1 row(s) of data
☐ Load at most 0 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 ...

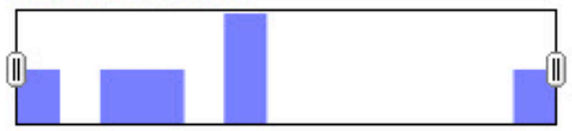
Method key collision Keying Function fingerprint 6 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	62	• Institute of Social and Industrial Relations. (35 rows) • Institute of Social and Industrial Relations. (16 rows) • Institute of Social and Industrial Relations. (4 rows) • Institute of Social and Industrial Relations. (2 rows) • Institute of Social and Industrial relations. (2 rows) • Institute of Social and Industrial Relations. (1 rows) • Institute of Social and Industrial Relations. (1 rows) • Institute of Social and Industrial Relations (1 rows)	☑	Institute of Social and Industi
3	130	• School of Social Work. (98 rows) • School of Social Work. (31 rows) • School of Social Work. (1 rows)	☐	School of Social Work.
2	36	• Dept. of Philosophy. (35 rows) • Dept. of Philosophy. (1 rows)	☐	Dept. of Philosophy.

 # Choices in Cluster

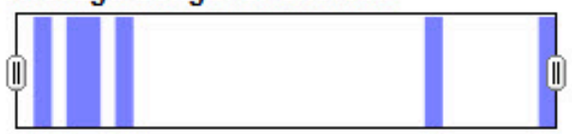
2 — 8

 # Rows in Cluster



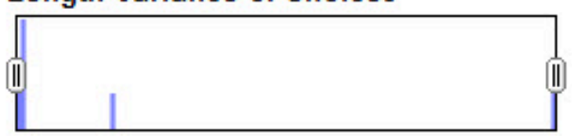
10 — 130

Average Length of Choices



17 — 49

Length Variance of Choices



0.47000000000000003 — 3.5700000000000003

Select All Deselect All **Merge Selected & Re-Cluster** Merge Selected & Close Close

try them out

or

develop new tools

www.avpreserve.com