

Towards Web Semantics

Spreadsheets and the U.S. Government

Lee Feigenbaum, Cambridge Semantics
Brand Niemann, U.S. EPA
Semantic Technology Conference
May 21, 2008

Agenda

- Overview – *what is this project?*
- The Situation – *what's done today?*
- Challenges – *what are we trying to improve?*
- The Approach – *why do semantics help?*
- The Solution – *what does the future look like?*

Agenda

- **Overview** – *what is this project?*
- The Situation – *what's done today?*
- Challenges – *what are we trying to improve?*
- The Approach – *why do semantics help?*
- The Solution – *what does the future look like?*

Genesis: October Recommendations

- A New Enterprise Data Management Strategy for the US Government Based on:
 - The premise of reusing the data and information rather than changing the data systems themselves:
 - Putting the business and technical rules, logic, etc. into the data itself using markup languages.
 - The concepts and standards of the Semantic Web:
 - Also called the Data Web or Web 3.0.
 - The most important tenets of the reuse are:
 - Bring the data and the metadata back together.
 - Bring the structured and unstructured data and information back together.
 - Bring the data and information description and context back together.
- Looking for partners to work with Federal Government and US EPA data and metadata sources.

See Brand's presentation at 2:45 today

What is this project?

- Collaboration between Brand Niemann of the EPA and Cambridge Semantics to explore potential applications of semantics to spreadsheets within the U.S. Government
- An attempt to reuse government spreadsheet data on the Web as linked, open data
- The beginnings of a semantic repository of government statistical data

Agenda

- Overview – *what is this project?*
- **The Situation** – *what's done today?*
- Challenges – *what are we trying to improve?*
- The Approach – *why do semantics help?*
- The Solution – *what does the future look like?*

The U.S. *Statistical Abstract*

- Published yearly since 1878 by the U.S. Census Bureau
- A comprehensive collection of social, political, and economic statistics
- Compiled from information from over 250 agencies
 - Mostly federal agencies; some private organizations

Building the *Statistical Abstract*

- Source data is received in many forms
 - Excel spreadsheets; CSV files; accompanying notes
 - Raw data; aggregated data (by quarter, by year, ...)
- A Census Bureau team compiles the data into standard, multi-tier Excel data tables
- Spreadsheets include data along with head notes, foot notes, and free text defining the statistics' semantics

Publishing the *Statistical Abstract*

- The Census Bureau is chartered only with producing a high-quality hardcopy version of the annual *Abstract*
- Individual data tables are available for download as Excel spreadsheets
 - The 2008 *Abstract* contains almost 1400 Excel files
- PDFs are also available of the integrated text and tables
- A small subset of the data is displayed on the Web via the American FactFinder

Agenda

- Overview – *what is this project?*
- The Situation – *what's done today?*
- **Challenges** – *what are we trying to improve?*
- The Approach – *why do semantics help?*
- The Solution – *what does the future look like?*

Confusing Terms

- Some terms' meanings are unclear
 - Ex: city, micropolitan area, metropolitan area
- Some terms' definitions change over time
 - Ex: the ability to identify with multiple race groups
- Different sources use the same terms in different (but related!) ways
 - Ex: Does family include one-person units?

Goal: Tie data terms directly to their meanings

Leverage the Web

- The Census Bureau is funded only to produce a print version of the *Statistical Abstract*
- Putting the statistics on the Web must be simple and quick
 - Currently this means making Excel and PDF files downloadable from the Web
- Professional indexers: A proper index for this wealth of information would be longer than the *Abstract* itself!

Goal: Get the data on the Web in a useful way

Empower the Data

- *Statistical Abstract* data is vital for:
 - Legislative budget allocations
 - Citizen watchdog groups
 - Understanding the state of the U.S. society and economy
- We'd like to be able to query, sort, or filter the data on any of its dimensions and spanning multiple data tables

Goal: Model the data with expressive fidelity

Compiling Statistics

- To produce the *Abstract*, Census Bureau employees must convert, transform, and aggregate data from many different source formats
- The meaning of statistics in source data must be painstakingly researched from glossaries and personal research. Conclusions are documented for future use.

Future goal: Ease the process of going from source data to Statistical Abstract data tables

Challenge: Multi-tiered data

A3 [See notes]								
	A	B	C	D	E	F	G	H
1	Table 1210. Participation in Various Arts Activities: 2002							
2								
3	[See notes]							
4								
5								
6	Item	Adult population (millions)	Jazz	Classical music	Other Dance \1	Painting	Pottery \2	Sewing
9	Total	205.9	1.3	1.8	4.2	8.6	6.9	16.0
11	Sex:							
12	Male	98.7	1.8	1.5	3.3	6.4	4.9	2.4
13	Female	107.2	0.9	2.1	4.9	10.6	8.7	28.5
15	Race and Ethnicity:							
16	White alone	150.1	1.5	2.1	4.1	9.4	7.6	17.6
17	African American alone	23.7	1.2	0.4	3.5	5.6	4.1	9.4
18	Other alone	9.5	0.5	2.3	5.8	7.4	6.5	14.9
19	Hispanic	22.7	0.5	0.7	4.2	6.8	5.1	12.5
21	Age:							
22	18 to 24 years old	26.8	1.9	2.5	6.0	15.4	9.3	10.4
23	25 to 34 years old	36.9	1.2	1.4	4.5	10.2	7.8	13.0
24	35 to 44 years old	44.2	1.5	1.8	3.9	8.1	7.4	15.3
25	45 to 54 years old	39.0	2.0	2.5	4.2	8.2	7.5	18.6
26	55 to 64 years old	25.9	0.8	1.5	3.4	6.7	5.6	19.1
27	65 to 74 years old	17.6	0.5	1.4	3.7	4.8	4.6	20.5
28	75 years old and older	15.5	0.4	0.7	2.5	3.1	2.4	18.0
29								
30	Education:							
31	Grade school	11.6	0.1	0.4	0.7	1.7	1.6	12.0

Challenge: Human-friendly semantics

A3 [See notes]								
	A	B	C	D	E	F	G	H
1	Table 1210. Participation in Various Arts Activities: 2002							
2								
3	[See notes]							
4								
5								
6	Item	Adult population (millions)	Jazz	Classical music	Other Dance \1	Painting	Pottery \2	Sewing
9	Total	205.9	1.3	1.8	4.2	8.6	6.9	16.0
11	Sex:							
12	Male	98.7	1.8	1.5	3.3	6.4	4.9	2.4
13	Female	107.2	0.9	2.1	4.9	10.6	8.7	28.5
15	Race and Ethnicity:							
16	White alone	150.1	1.5	2.1	4.1	9.4	7.6	17.6
17	African American alone	23.7	1.2	0.4	3.5	5.6	4.1	9.4
18	Other alone	9.5	0.5	2.3	5.8	7.4	6.5	14.9
19	Hispanic	22.7	0.5	0.7	4.2	6.8	5.1	12.5
21	Age:							
22	18 to 24 years old	26.8	1.9	2.5	6.0	15.4	9.3	10.4
23	25 to 34 years old	36.9	1.2	1.4	4.5	10.2	7.8	13.0
24	35 to 44 years old	44.2	1.5	1.8	3.9	8.1	7.4	15.3
25	45 to 54 years old	39.0	2.0	2.5	4.2	8.2	7.5	18.6
26	55 to 64 years old	25.9	0.8	1.5	3.4	6.7	5.6	19.1
27	65 to 74 years old	17.6	0.5	1.4	3.7	4.8	4.6	20.5
28	75 years old and older	15.5	0.4	0.7	2.5	3.1	2.4	18.0
29								
30	Education:							
31	Grade school	11.6	0.1	0.4	0.7	1.7	1.6	12.0

Challenge: Human-friendly semantics

	A	B	C	D	E	F	G	H	I	J
1	Table 1210. Participation in Various Arts Activities: 2002									
2										
3	[Back to data]									
4										
5	HEADNOTE									
6	[In percent. For persons 18 years old and over. Covers activities									
7	engaged in at least once in the prior 12 months. See headnote in Table 1221]									
8										
9	FOOTNOTES									
10	\1 Dancing other than ballet (e.g. modern, folk and tap).									
11	\2 Includes ceramics, jewelry, leatherwork, and metalwork.									
12	\3 Includes making movies or video as an artistic activity.									
13										
14	Source: U.S. National Endowment for the Arts.									
15	Research Division Report #45.									
16	2002 Survey of Public Participation in the Arts.									
17										
18	For more information:									
19	http://www.nea.gov/research/ResearchReports_chrono.html									
20										
21										
22										

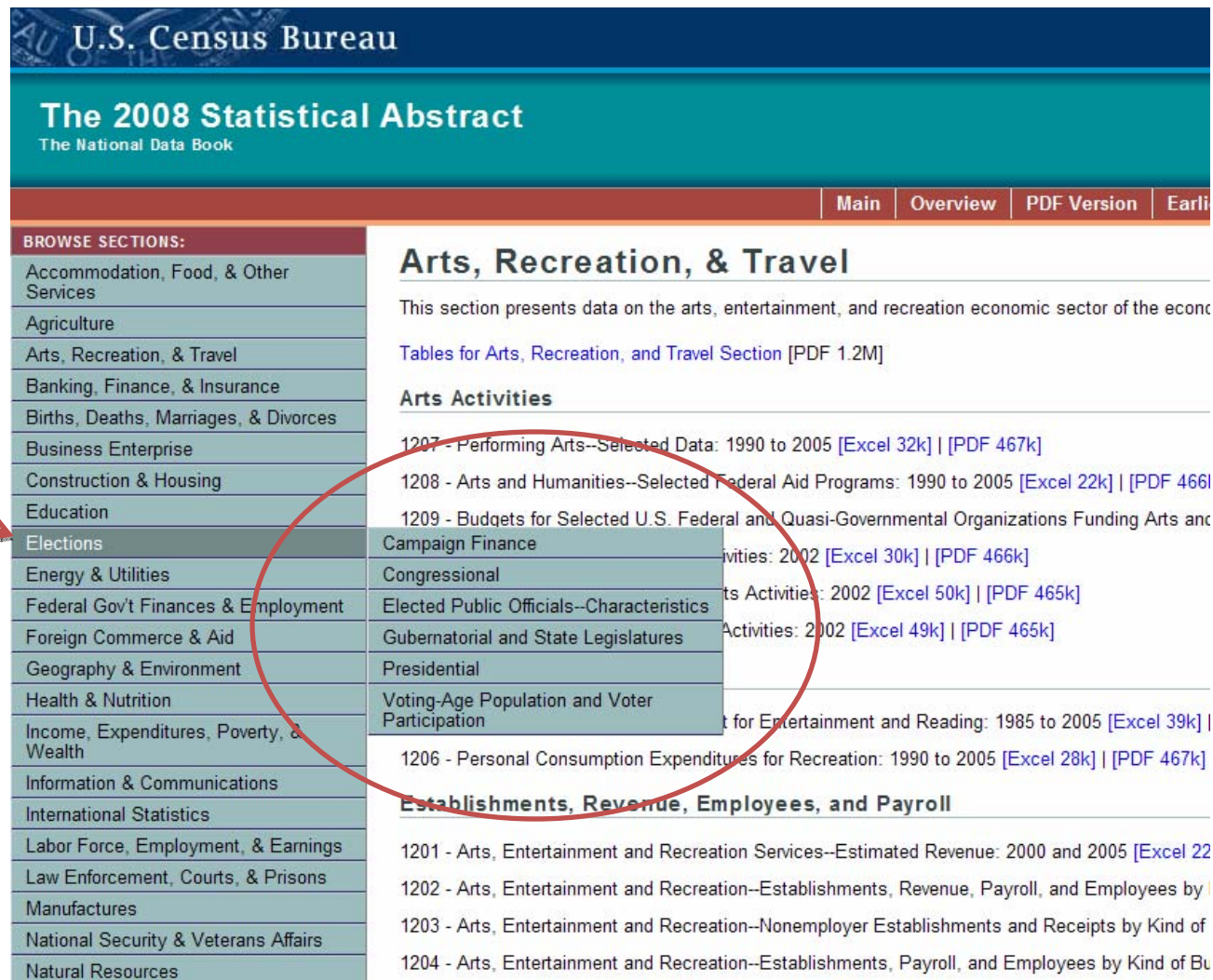
Challenge : Units and Multipliers

A3 [See notes]

	A	B	C	D	E	F	G	H
1	Table 1210. Participation in Various Arts Activities: 2002							
2								
3	[See notes]							
4								
5								
6	Item	Adult population (millions)	Jazz	Classical music	Other Dance \1	Painting	Pottery \2	Sewing
9	Total	205.9	1.3	1.8	4.2	8.6	6.9	16.0
11	Sex:							
12	Male	98.7	1.8	1.5	3.3	6.4	4.9	2.4
13	Female	107.2	0.9	2.1	4.9	10.6	8.7	28.5
15	Race and Ethnicity:							
16	White alone	150.1	1.5	2.1	4.1	9.4	7.6	17.6
17	African American alone	23.7	1.2	0.4	3.5	5.6	4.1	9.4
18	Other alone	9.5	0.5	2.3	5.8	7.4	6.5	14.9
19	Hispanic	22.7	0.5	0.7	4.2	6.8	5.1	12.5
21	Age:							
22	18 to 24 years old	26.8	1.9	2.5	6.0	15.4	9.3	10.4
23	25 to 34 years old	36.9	1.2	1.4	4.5	10.2	7.8	13.0
24	35 to 44 years old	44.2	1.5	1.8	3.9	8.1	7.4	15.3
25	45 to 54 years old	39.0	2.0	2.5	4.2	8.2	7.5	18.6
26	55 to 64 years old	25.9	0.8	1.5	3.4	6.7	5.6	19.1
27	65 to 74 years old	17.6	0.5	1.4	3.7	4.8	4.6	20.5
28	75 years old and older	15.5	0.4	0.7	2.5	3.1	2.4	18.0
29								
30	Education:							
31	Grade school	11.6	0.1	0.4	0.7	1.7	1.6	12.0

Data Notes

Challenge: Taxonomic Relations



U.S. Census Bureau

The 2008 Statistical Abstract

The National Data Book

[Main](#) | [Overview](#) | [PDF Version](#) | [Earli](#)

BROWSE SECTIONS:

- Accommodation, Food, & Other Services
- Agriculture
- Arts, Recreation, & Travel
- Banking, Finance, & Insurance
- Births, Deaths, Marriages, & Divorces
- Business Enterprise
- Construction & Housing
- Education
- Elections**
- Energy & Utilities
- Federal Gov't Finances & Employment
- Foreign Commerce & Aid
- Geography & Environment
- Health & Nutrition
- Income, Expenditures, Poverty, & Wealth
- Information & Communications
- International Statistics
- Labor Force, Employment, & Earnings
- Law Enforcement, Courts, & Prisons
- Manufactures
- National Security & Veterans Affairs
- Natural Resources

Arts, Recreation, & Travel

This section presents data on the arts, entertainment, and recreation economic sector of the econ

[Tables for Arts, Recreation, and Travel Section](#) [PDF 1.2M]

Arts Activities

- 1207 - Performing Arts--Selected Data: 1990 to 2005 [[Excel 32k](#)] | [[PDF 467k](#)]
- 1208 - Arts and Humanities--Selected Federal Aid Programs: 1990 to 2005 [[Excel 22k](#)] | [[PDF 466k](#)]
- 1209 - Budgets for Selected U.S. Federal and Quasi-Governmental Organizations Funding Arts and Activities: 2002 [[Excel 30k](#)] | [[PDF 466k](#)]
- 1210 - Arts and Humanities--Selected Federal Aid Programs: 2002 [[Excel 50k](#)] | [[PDF 465k](#)]
- 1211 - Arts and Humanities--Selected Federal Aid Programs: 2002 [[Excel 49k](#)] | [[PDF 465k](#)]
- 1212 - Arts, Entertainment and Recreation--Selected Federal Aid Programs: 1985 to 2005 [[Excel 39k](#)] | [[PDF 467k](#)]
- 1206 - Personal Consumption Expenditures for Recreation: 1990 to 2005 [[Excel 28k](#)] | [[PDF 467k](#)]

Establishments, Revenue, Employees, and Payroll

- 1201 - Arts, Entertainment and Recreation Services--Estimated Revenue: 2000 and 2005 [[Excel 22k](#)]
- 1202 - Arts, Entertainment and Recreation--Establishments, Revenue, Payroll, and Employees by Kind of Business: 2000 and 2005 [[Excel 22k](#)]
- 1203 - Arts, Entertainment and Recreation--Nonemployer Establishments and Receipts by Kind of Business: 2000 and 2005 [[Excel 22k](#)]
- 1204 - Arts, Entertainment and Recreation--Establishments, Payroll, and Employees by Kind of Business: 2000 and 2005 [[Excel 22k](#)]

Agenda

- Overview – *what is this project?*
- The Situation – *what's done today?*
- Challenges – *what are we trying to improve?*
- **The Approach** – *why do semantics help?*
- The Solution – *what does the future look like?*

Capture the data

- Decouple the data from its spreadsheet layout
- The RDF graph data model can explicitly capture the row/column relationships implicit within the *Statistical Abstract's* spreadsheets
- Once freed from the spreadsheet silos, the semantic data can be reused and remixed elsewhere
 - For now, on the Web
 - Later, statistical or mathematical packages, next-generation visualizations, ...

Capture the data's semantics

- Semantic Web technologies provide an incremental path for exposing the semantics within the *Statistical Abstract*:
 - A *glossary* of data tables' columns and their meanings
 - E.g. Differentiate different uses of the same term
 - A *taxonomy* of concepts depicted in the *Abstract*
 - E.g. Synonym- and hierarchy-aware search
 - An *ontology* of concepts and relationships spanning the different data sets in the Abstract
 - E.g. Compare statistics compiled under slightly different semantics

Reuse familiar tools & paradigms

- Layer semantics *in place* atop the data
 - Keep the information in Excel
 - Shy away from ETL approaches
 - Advocate familiar user interface approaches
 - Drag and drop, autocomplete, integrated activity panes

Agenda

- Overview – *what is this project?*
- The Situation – *what's done today?*
- Challenges – *what are we trying to improve?*
- The Approach – *why do semantics help?*
- **The Solution** – *what does the future look like?*

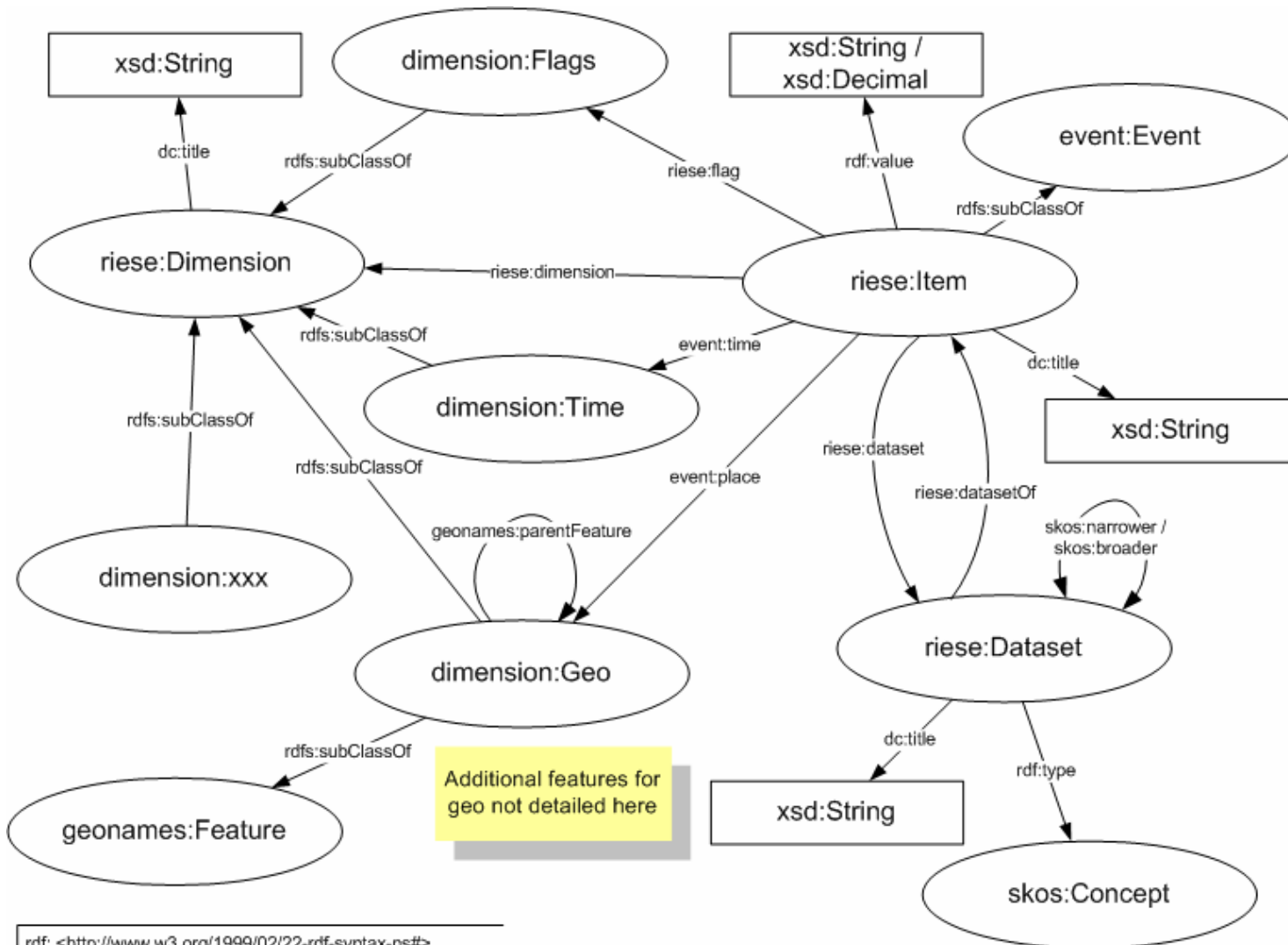
Modeling statistics in RDF

- Model statistics; don't try to model the underlying domains
- A data item is a value, possibly with units
- Data items are grouped into data sets
 - E.g. *number of Congress representatives*
- Data items are defined by their values for various dimensions (facets)

– E.g.

Time: 2002
State: California
Political Party: Republican

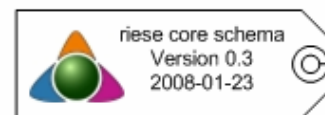
Modeling statistics in RDF



Source: The [Riese](#) team, Wolfgang Halb, Yves Raimond, and Michael Hausenblas.

```

rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
xsd: <http://www.w3.org/2001/XMLSchema#>
riese: <http://riese.joanneum.at/schema/core#>
dimension: <http://riese.joanneum.at/schema/dimension/>
event: <http://purl.org/NET/c4dm/event.owl#>
skos: <http://www.w3.org/2004/02/skos/core#>
geonames: <http://www.geonames.org/ontology#>
dc: <http://purl.org/dc/elements/1.1/>
    
```



Semantic Repository

- Tie definitions to rows and columns to build a glossary of terms
 - Census Bureau employees are already capturing these definitions in unstructured notes
- Search the glossary to reuse or extend existing definitions when appropriate
 - Can also search ISO11179 metadata repositories
- Expose this glossary on the Web
 - Terms linked to their meaning

“Paint” semantics onto the data

- Tight integration into Excel allows semantic concepts to be dragged and dropped from the semantic repository onto data tables
- The data table's implicit row/column relations are explicitly stored in an RDF semantic database
- Cells, columns, and regions are tagged with explicit semantics
- Publish the data tables on the Web

Enabling Technology

- Cambridge Semantics's SHAPE semantic middleware platform enables this via:
 - Open Anzo: An open-source RDF server supporting real-time updates, access control, data modularity, and more
 - Semantic Excel: Attach semantics to spreadsheet data and bind the data to an RDF database
 - Mojo: Use simple HTML templates to display RDF data on the Web

More benefits...

- A live connection between spreadsheets and the Web can eliminate the time usually needed for (re-)publishing cycles
- On the Web, the *Statistical Abstract* can be presented as linked, open data, woven into the emerging Semantic Web
- More precise and accurate search via guided query with autocomplete, informed by the semantic repository

Where do we go from here?

- Continue development of Semantic Excel to include features:
 - Richer ontology support
 - Spreadsheet aggregation
 - Cross-spreadsheet query
 - ...
- Explore other use cases that leverage semantics to make better use of spreadsheets in the U.S.
Government
- Integration with other sources of linked data (e.g. Riese or Joshua Tauberer's census data)

Questions?

Please contact Lee with any questions:

lee@cambridgesemantics.com

Thanks to Lars Johanson and Ian O'Brien of the U.S. Census Bureau for help in exploring this use case.