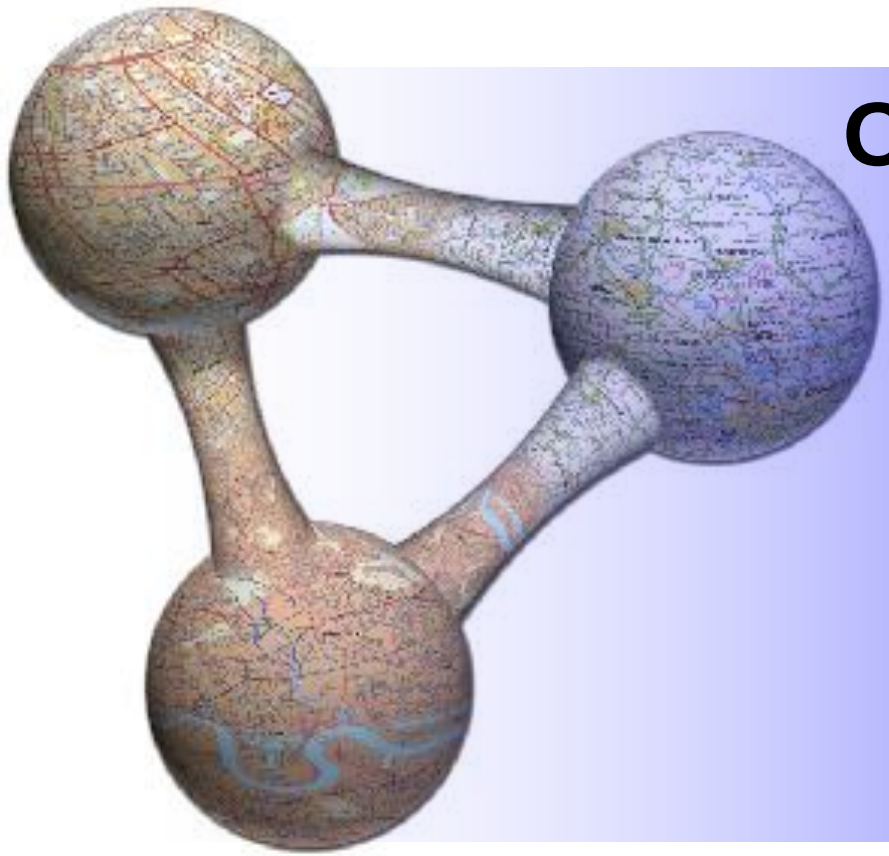


Open Data, Big Data and Smart Cities

Rosario Uceda-Sosa
rosariou@us.ibm.com
Cognitive Computing Department
IBM T.J. Watson Research Labs

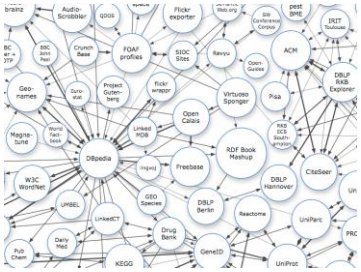
The questions

1. Storing data. How varied is web data, really?
2. Information ecosystems. How are people likely to use data? What ecosystems (app developers, domain experts, publishers) will be built to leverage this data?
3. Usable, scalable semantics. What's the role of semantic technologies in these ecosystems?

The answers will vary by domain. We'll use Smart Cities as our main threads, but we'll also refer

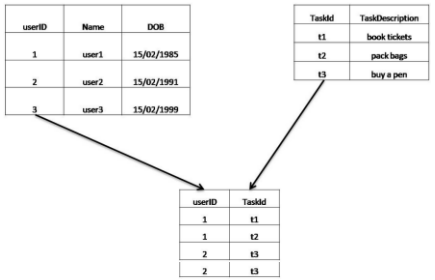
1. Storing data. A historic shift in data paradigms

2000's, (Open) LINKED DATA is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." *(From linkeddata.org)*



Abstraction: Network →
Data (any format) connected through
(inferred) navigational relations

1972, RDBMS. Instead of records being stored in some sort of linked list of free-form records as in Codasyl, Codd's idea was to use a "table" of fixed-length records, with each table used for a different type of entity. A linked-list system would be very inefficient when storing "sparse" databases where some of the data for any one record could be left empty. The relational model solved this by splitting the data into a series of normalized tables (or *relations*), with optional elements being moved out of the main table to where they would take up room only if needed. Data may be freely inserted, deleted and edited in these tables, with the DBMS doing whatever maintenance needed to present a table view to the application/user. **(*)**



Abstraction: Entitites (tables) in *fixed* records →
Efficient, scalable and easy to use

1962, CODASYL. The **Codasyl** approach was based on the "manual" navigation of a ***linked data set*** which was formed into a large network. Records could be found either by use of a primary key (known as a CALC key, typically implemented by hashing), by navigating relationships (called sets) from one record to another, or by scanning all the records in sequential order. Later systems added B-Trees that to provide alternate access paths. Many Codasyl databases also added a query language that was very straightforward. However, in the final tally, **CODASYL was very complex and required significant training and effort to produce useful applications.** (*)

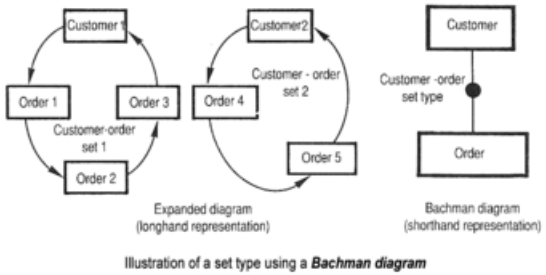


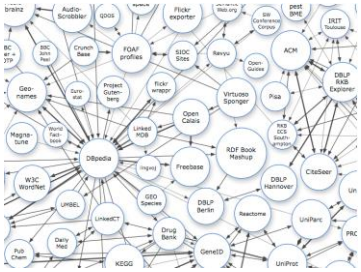
Illustration of a set type using a ***Bachman diagram***

Abstraction: Network →
Variable size, schemaless data
connected through navigational relations

(*) From http://en.wikipedia.org/wiki/Database#1960s_navigational_DBMS

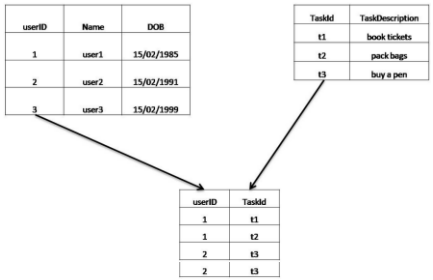
1. Storing data.

*PROs and CONs of data storage formats do not apply to linked data, because Linked Data has to do with the **semantic relations** between data, not with their formats.*



Abstraction: Network →
Data (any format) connected through (inferred) navigational relations

PRO: Easy-to-understand paradigm, scalable, robust …
CON: … Only for models with strong, stable schemata and not many relations



Abstraction: Entitites (tables) in *fixed* records →
Efficient, scalable and easy to use

PRO: Natural, flexible representation of entities and relations
CON: Too fine grained model. Hard for users to understand and systems to process

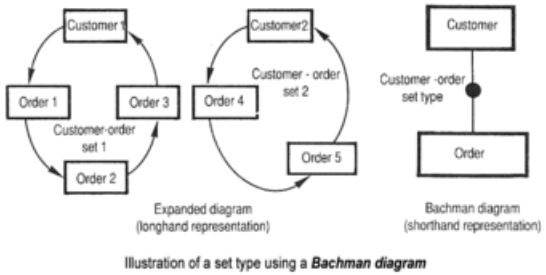


Illustration of a set type using a ***Bachman diagram***

Abstraction: Network →
Variable size, weak schema data connected through navigational relations

(*) From http://en.wikipedia.org/wiki/Database#1960s_navigational_DBMS

1. Storing data. The future of city data catalogs in the web

We've looked into 900 data sources, coming from Washington D.C. and the London Datastore, as well as governments like Kenya and India and categorized the information into one of these categories.

Instances (events, business licenses)

Entity types	Payment, Service Request, Transportation project, Landlot, Agency
Identifiers	Landlot_ID, BusinessLicense#, DriverLicense, Career Center ID, District#
Geospatial	GPS Point, Municipal area, Landlot area, bus route, land thoroughfare coordinates
Time	Timestamps, recursive schedules, ranges
Domain-specific (finite) categories	District names, service types, business licenses types, status, business use
Text	Descriptions, comments

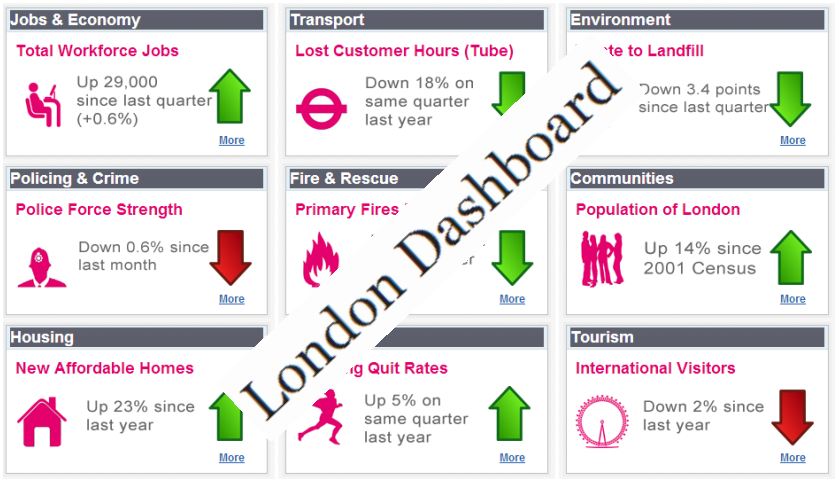
Stats

Measurements	Building dimensions, donation amounts, bus fleet size Air quality index, vehicles per hour
Indicators	Demographics, percentage buses on time, Avg income

Measurement and indicator information is quite complex (as discussed in Prof. Mark Fox's talk)

1. Storing (big) data. Open Data is already here… Is Linked Data coming?

Not clear whether the (instance) data itself will be *linked*… Most data is being published in tabular, XML or JSON formats … BUT there will be value in linking and analyzing this data. To do this, linked schemas are necessary.



Map		Map: Crime Incide	Incidents are deri	0	P
Tabular		Case Data from Sa	Cases created sin	0	S
Tabular		Data Catalog		0	
Tabular		Film Locations in S	If you love movie		
Tabular		Businesses Registe	Use this link		
Chart		Third-Party Spenc	San Fran		E
External		Building Footprints	This	0	F
External		Planning Neighbor		0	C
External		SFPD Reported Inc		0	P
Filter		Data Catalog for 3		0	C
Tabular		SFPD Incidents - P		0	P
External		City Lots (Zipped S		0	C
Chart		Campaign Financ		0	E
Filter		Graffiti SF311 R	es created sin	0	S
Filter		Campaign Finance		0	E
External		Streets of San Fran	View of Street Cen	0	Geography	
External		Zoning Districts	The Zoning Distric	0	Geography	planning, zoning,
Map		San Francisco Pipe	Snapshot of San Fr	0	Construction and I	development, hou

San Francisco Data

Washington D.C.	City of Huntsville, AL
Town of Hillsborough, CA	City of Dunwoody, GA
Toronto, ON	City of DeLeon, TX
San Francisco, CA	City of Corona, CA
Roosevelt Island, NY	City of Alpharetta, GA
Raleigh City Hall, NC	Chicago, IL
Quebec, QC	Brookline, MA
Olathe, KS	Boston, MA
Newnan, GA	Bonn, Germany
Newark, OH - Service Department	Bloomington, IN
Manor, TX	Baltimore, MD
Howard County Department of Public Works, MD	Bainbridge Island, WA
Grand Rapids, MI	
Fontana City Hall	
Darwin, Australia	
City of Tucson, AZ	
City of Russell Springs, KY	
City of Richmond	
City of Newberg, OR	
City of New Haven, CT	

1. Storing (big) data. The future of city data catalogs in the web

•RDBMS's are -and will continue to be- the most efficient way to store and retrieve large volumes of data, especially transactional data.
•City data is stored in many formats (XML, ESRI, metadata documents), but almost all of them can be easily recorded in databases, because most data is already (semi-)structured. (Example below from Washington D.C. open data)
•Even social network data can be (semi-)structured by standards like FOAF, SKOS and the current DCAT (W3C Draft)
•All this data is semantically connected: by time, location, event, department, subject, etc.

→ The issue is not RDBMS vs Linked Data, but how to capture ***variable, dynamic, linked schemas***, regardless of where or how the instance data is stored.

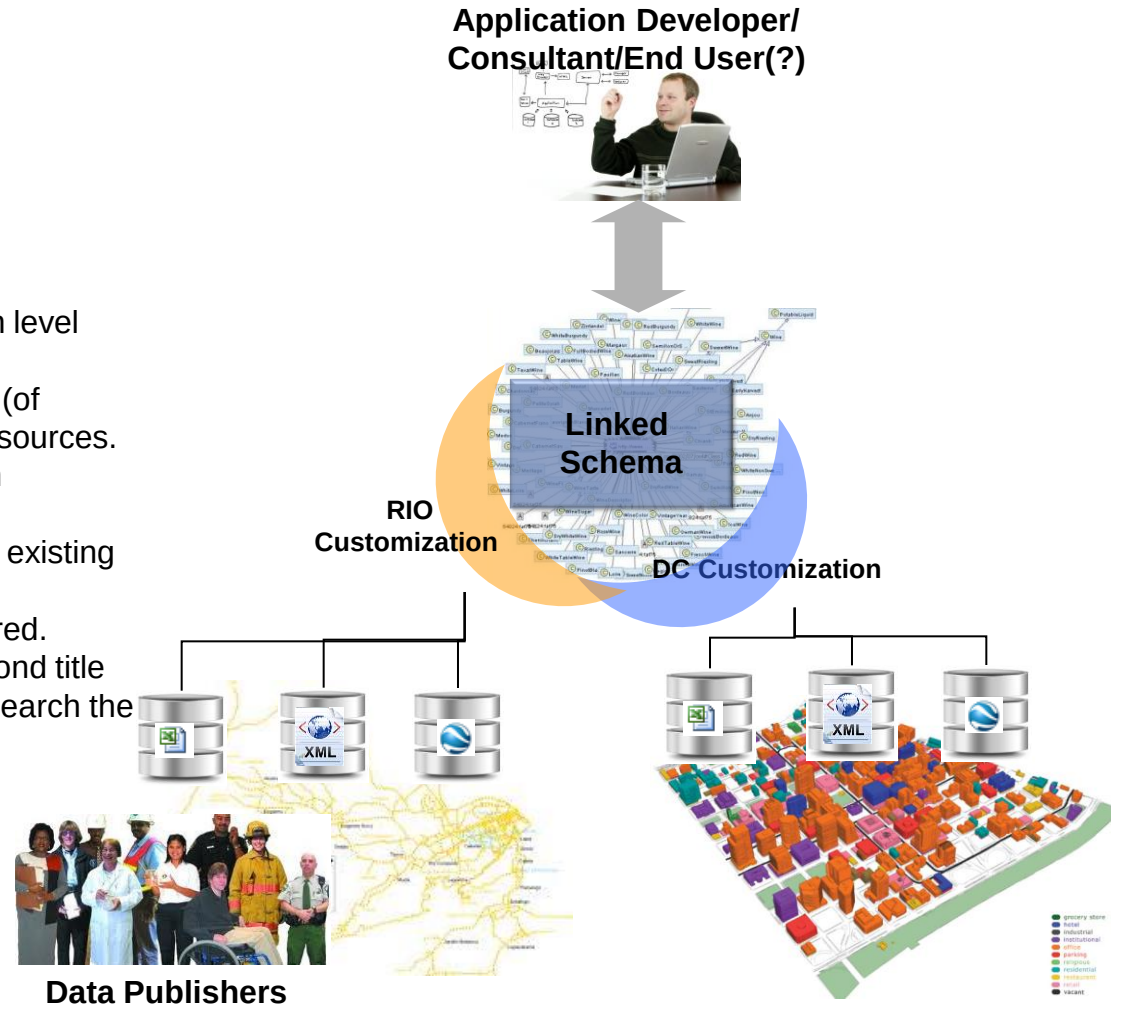
Title	Source	XML	Text/CSV	Atom (GeoRSS support)	KML/ESRI Shapefile	Maps	Download
ITSA Invoice and Vendor Payment Schedule	Optimal Solutions and Technologies	04/06/2013	04/06/2013	04/06/2013			
Provides information about the ITSA Invoice and Vendor Payment Schedule.							
ITSA Program Managers - Time to Fill	Optimal Solutions and Technologies	04/06/2013	04/06/2013	04/06/2013			
Provides information about ITSA Program Managers' Time-to-fill for positions competed in the CBE community.							
311 Service Requests	Citywide Call Center	04/06/2013	04/06/2013	04/06/2013	04/06/2013	See it on Google Maps	Custom download
Provides data on service requests submitted to the Mayor's 311 Call Center and the online Service Request Center. XML Schemas							
Basic Business Licenses	DCRA	04/05/2013	04/05/2013	04/05/2013	04/05/2013	See it on Google Maps	
Provides information on Basic Business Licenses issued by DCRA							
Building Permits	DCRA	04/06/2013	04/06/2013	04/06/2013	04/06/2013	See it on Google Maps	Custom download
Provides information on building permits granted by the Department of Consumer and Regulatory Affairs (DCRA). XML Schemas							
CBEs active in ITSA Program	Optimal Solutions and Technologies	04/01/2013	04/01/2013	04/01/2013	04/01/2013	See it on Google Maps	
Provides information about the Certified Business Enterprises (CBEs) currently registered in the ITSA program and eligible to submit candidates for open requirements.							
CJIS Juvenile Arrests and Charges	MPD	09/14/2012	09/14/2012	09/14/2012			Custom download
Provides juvenile arrests and charges reported by the Metropolitan Police Department (MPDC), aggregated to block level. XML Schemas							
Completed Construction Projects 2003	DDOT	01/01/2004	01/01/2004		01/01/2004	See it on Google Maps	
Provides information on completed construction projects reported by DDOT in 2003.							

2. Information ecosystems. It's all about metadata

Suppose that we can keep a flexible schema that is (1) Authoritative, (2) Complete and (3) Extensible

We could use it to

- Integrate many data sources at the application level WITHOUT altering them.
- Create views, categorizations and annotations (of relations) that do not affect the underlying data sources. In a word, enhance the data schemas based on applications' needs.
- Port solutions to other cities, using 80% of the existing model and customizing 20% for the new city.
- Shield applications from *where* the data is stored.
- Link data sources in (open) data catalogs beyond title and metadata of the data source. Users could search the instance data inside the catalogs



2. Information Ecosystems. Metadata management

Suppose that we can keep a flexible schema that is (1) Authoritative, (2) Complete and (3) Extensible

For this we'd need:

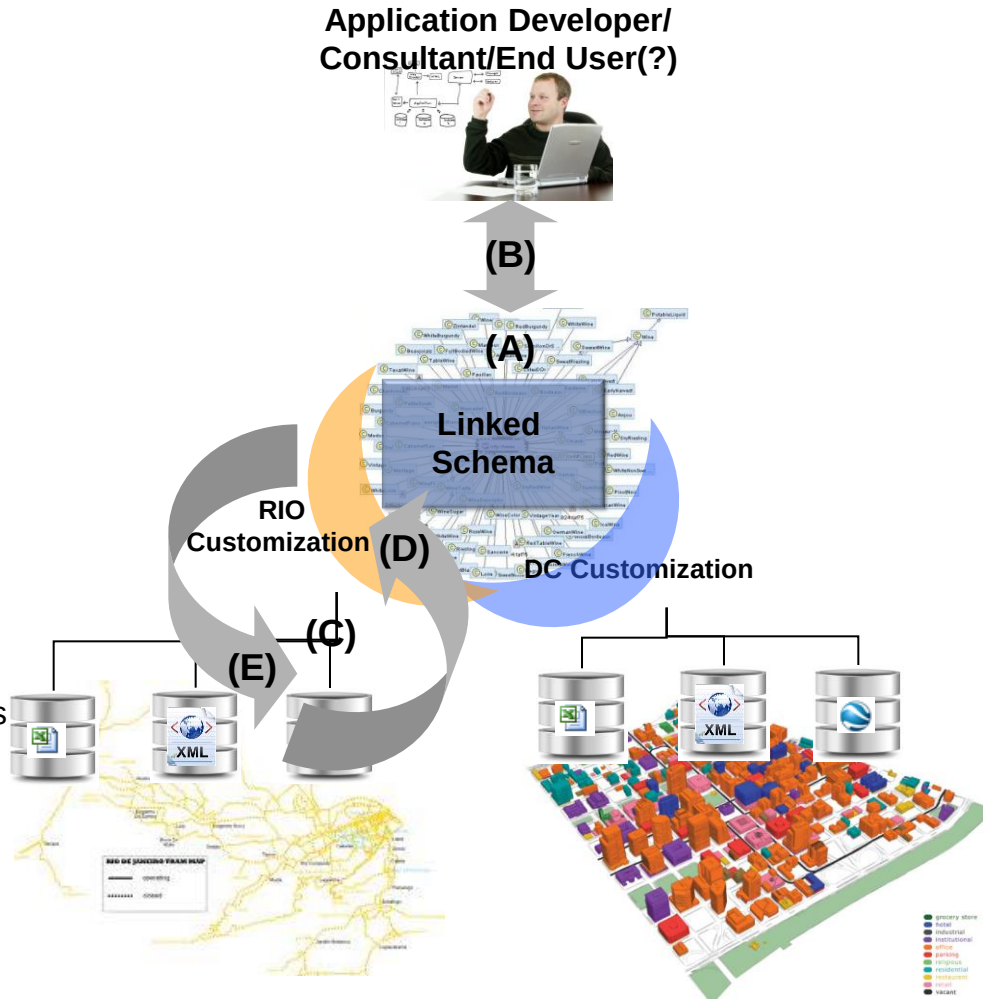
(A) A flexible, simple language for the schema.
- ✓ RDF/OWL, specifically OWL-QL with simple inferencing (Subsumption, Entity)

(B) APIs and code for the management, storage and query of the schema
- ✓ Jena library, Triplestores (DB2RDF)

(C) A language/system for mapping the data sources to/from the schema
- ❖ (Partial) Invention required: D2RQ and other mappings are inefficient.
- ❖ (Partial) Invention required: access control, billing for data use, etc.

(D) A way to retrieve *instance* data from the data sources
- ❖ (Partial) Invention required. Caching and storage need to be worked out.

(E) A way to write instance data to the data sources *(not all applications will require this)*
- ❖ (Partial) Invention required. Access control, transaction support, etc.



Scenario. Events in Washington D.C.

Suppose a Smarter City application managing city operations wants to provide a GUI for city administrators to monitor and query events in a city (311, 911, service requests, licenses requests, infrastructure work, etc.) A similar application is a GUI that allows end users or app developers to query city data.

Suppose we have two initial data sources, which contain semantically similar information: Both describe events with identifiers, events happen in time/space, have a subject, refer to services (competencies) provided by a department, etc.

911 Requests (2010)

CCN	REPORTDATETIME	OFFENSE	BLOCKSITEADDRESS	LATITUDE	LONGITUDE	WARD	DISTRICT	PSA
9185124	1/1/2010 0:00	THEFT F/AUTO	1900 B/O FENWICK ST N	38.91425947	-76.98444392	5	FIFTH	50
9185134	1/1/2010 0:00	ADW	800 B/O 21ST ST NE	38.90149128	-76.97417631	5	FIFTH	50
9185104	1/1/2010 0:00	ROBBERY	700 B/O MORTON ST N	38.93190228	-77.02516974	1	THIRD	30
10000041	1/1/2010 0:00	ADW	2000 B/O GEORGIA AVE	38.91701078	-77.02189474	1	THIRD	30
10000086	1/1/2010 0:00	ROBBERY	600 B/O IRVING ST NW	38.92903929	-77.02213507	1	THIRD	30
10000115	1/1/2010 0:00	BURGLARY	3500 B/O STANTON RD S	38.84388456	-76.97891927	8	SEVENTH	70
10000087	1/1/2010 0:00	ADW	1300 B/O OKIE ST NE	38.91421407	-76.98602973	5	FIFTH	50
10000096	1/1/2010 0:00	ADW	700 B/O 7TH ST NW	38.89913006	-77.02191663	2	FIRST	10
10000147	1/1/2010 0:00	ADW	1500 B/O MONROE ST N	38.93226905	-77.03560479	1	THIRD	30
10000141	1/1/2010 0:00	THEFT	2700 B/O O ST SE	38.87074185	-76.96837205	7	SIXTH	60

311 Requests (2010)

SERVICEREQUESTID	SERVICEPRIO	SERVICETYPECOD	SERVICETYPECODEDESCRIP	SERVICEORDERD/	SERVICEORD	AGENCYABBREVIATION	RESOLUT
10-00000005	UNKNOWN	ABAVEHOP	Abandoned Vehicle Operations	1/1/2010 9:53	CLOSED	DPW	Nothing Fo
10-00000078	UNKNOWN	DEPAHEAL	DOH	1/2/2010 7:51	OVERDUE OF	DOH	Baited - fu
10-00000019	UNKNOWN	PARKENFO	Parking Enforcement	1/1/2010 11:47	CLOSED	DPW	Not Enforc
10-00000067	UNKNOWN	PARKENFO	Parking Enforcement	1/1/2010 21:01	CLOSED	DPW	Repeat Off
10-00000073	UNKNOWN	PRSVAVOP	Abandoned Vehicle Operations	1/1/2010 21:53	OVERDUE OF	DPW	UNKNOW
10-00000049	UNKNOWN	SIGNS	Signs	1/1/2010 14:54	CLOSED	DDOT	UNKNOW
10-00000039	UNKNOWN	SISYINOD	SIOD	1/1/2010 13:41	OVERDUE OF	DDOT	UNKNOW
10-00000098	UNKNOWN	SISYINOD	SIOD	1/2/2010 10:19	OVERDUE OF	DDOT	UNKNOW
10-00000082	UNKNOWN	SNOW	Snow	1/2/2010 8:37	OVERDUE OF	DPW	UNKNOW

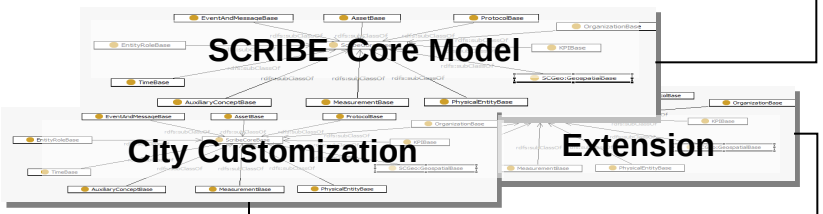
3. (An example of) Usable, scalable semantics. SCRIBE.

SCRIBE is a non-normative, authoritative, modular, extensible semantic model for Smarter Cities.

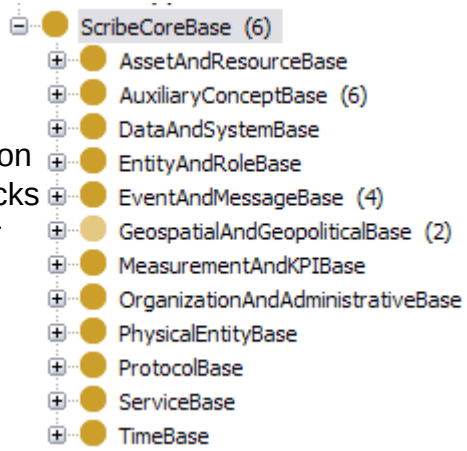
It consists of a Core model that includes common classes (events and messages, stakeholders, departments, services, city landmarks and resources, KPIs, etc.), *extensions* by domain and *customizations* by city.

Simple language
- Classes + Inheritance + Relations + Inferencing
- Based on standards (OWL-QL, SPARQL)
- Mappable to UML
- Metadata annotations and Tagging

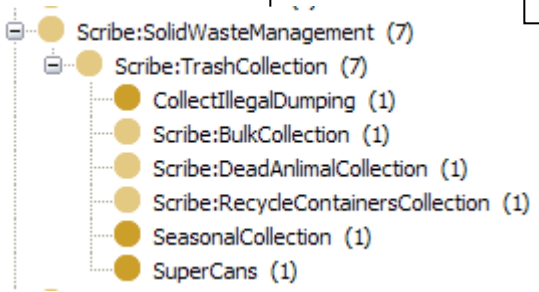
Authoritative
- Aligned with standards (CAP, NIEM, MISA/MRM, UCore)
- Validate with customer scenarios
- Validated with open city data



Common building blocks



Organization/Operation profile

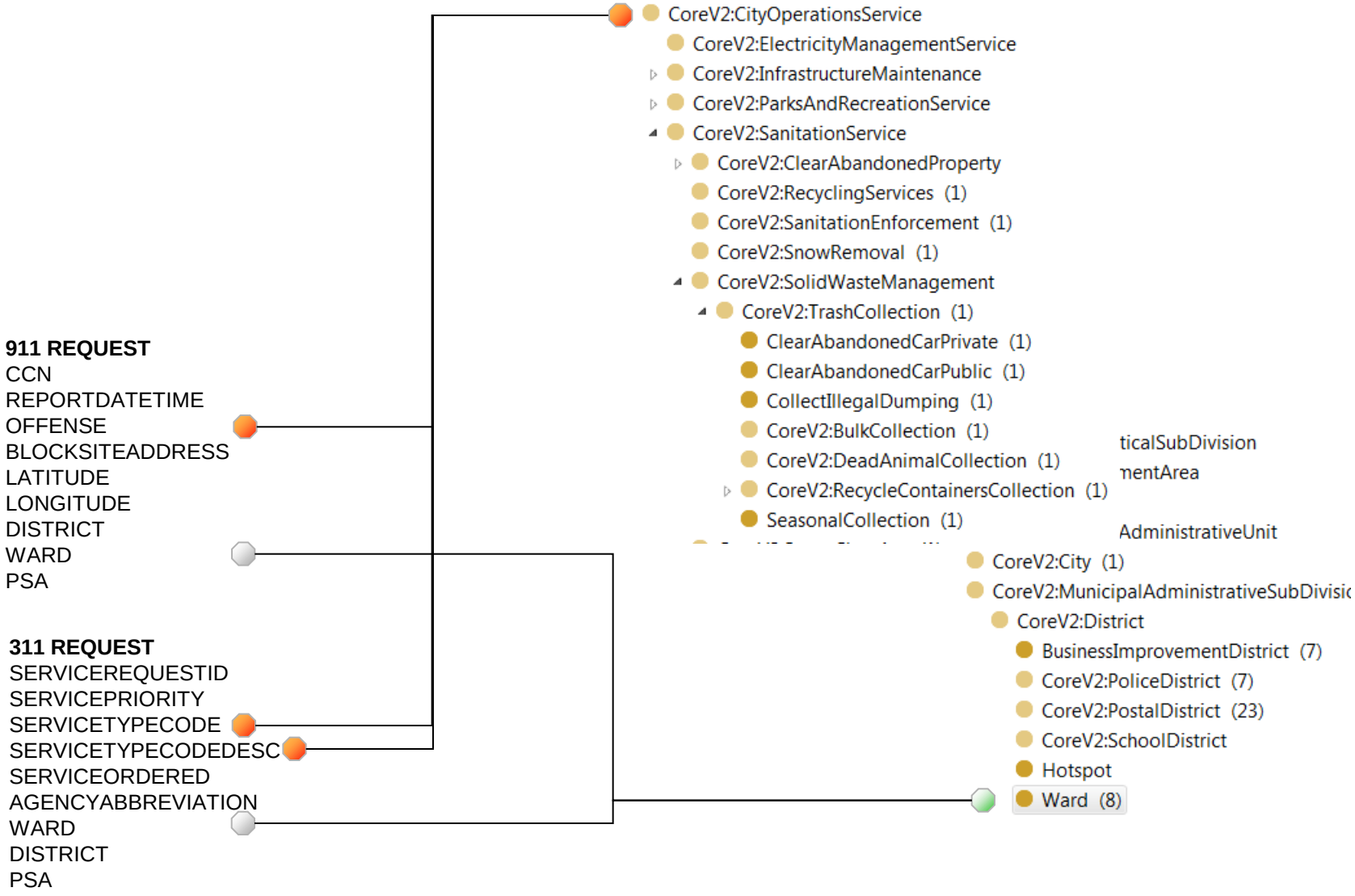


Features

AssetManagement
BuildingAndParcel
Transportation
Water
Weather

3. (An example of) Usable, scalable semantics. SCRIBE.

The SCRIBE solution is to ‘link’ the original data sources at the schema level to a semantic model with inferencing.



The SCRIBE Solution. Integrating data through the model

The data from DC Service Requests and Crime incidents can now be queried together as events, not just as service requests or criminal incidents.

Row	Event	descriptorLabel	District	Ward
1	10000716	STOLEN AUTO	SEVENTH	8
2	10000328	THEFT F/AUTO	FIFTH	5
3	10000315	THEFT F/AUTO	THIRD	1
4	10000672	THEFT	THIRD	1
5	10000713	STOLEN AUTO	THIRD	1
6	10000665	ROBBERY	SIXTH	7
7	10000200	STOLEN AUTO	SIXTH	7
8	10000237	THEFT	SECOND	2
9	10000353	THEFT	SECOND	2
10	10000147	ADW	THIRD	1
11	10000580	THEFT F/AUTO	THIRD	2
12	10000655	THEFT	SECOND	2
13	10000250	BURGLARY	SECOND	2
14	10000058	ROBBERY	SEVENTH	8
		...		
96	10000606	ROBBERY	THIRD	1
97	10000449	THEFT	SECOND	2
98	10000028	THEFT	THIRD	1
99	10000252	THEFT F/AUTO	FIFTH	5
100	9000412	THEFT F/AUTO	THIRD	2
101	10-00000012	BulkCollection	SEVENTH	8
102	10-00000046	SanitationEnforcement	SEVENTH	8
103	10-00000033	DCSeasonalCollection	FIRST	6
104	10-00000054	BulkCollection	THIRD	1
105	10-00000050	DCSuperCans	SECOND	3
106	10-00000060	BulkCollection	THIRD	1
107	10-00000038	DeadAnimal	SIXTH	7
108	10-00000039	DCSIOD	FIRST	6
109	10-00000028	DCTransportationOperationService		
110	10-00000062	DeadAnimal	SIXTH	7
111	10-00000044	BulkCollection	FOURTH	5
112	10-00000016	DCTrashCollection	SEVENTH	8

Query: All Events in DC, with type, District and Ward

```
SELECT DISTINCT ?Event ?descriptorLabel ?District ?Ward
WHERE {
  ?Event a Scribe:Event .
  ?Event Scribe:hasEventDescriptor ?x .
  OPTIONAL {
    ?Event Scribe:hasEventDescriptor ?descriptor .
    ?descriptor a Scribe:IncidentType .
    ?descriptor rdfs:label ?descriptorLabel .
  } .
  OPTIONAL {
    ?Event Scribe:eventLocatedIn ?district .
    ?district a Scribe:PoliceDistrict .
    ?district rdfs:label ?District .
  } .
  OPTIONAL {
    ?Event Scribe:eventLocatedIn ?ward .
    ?ward a WDC:Ward .
    ?ward rdfs:label ?Ward .
  } .
```

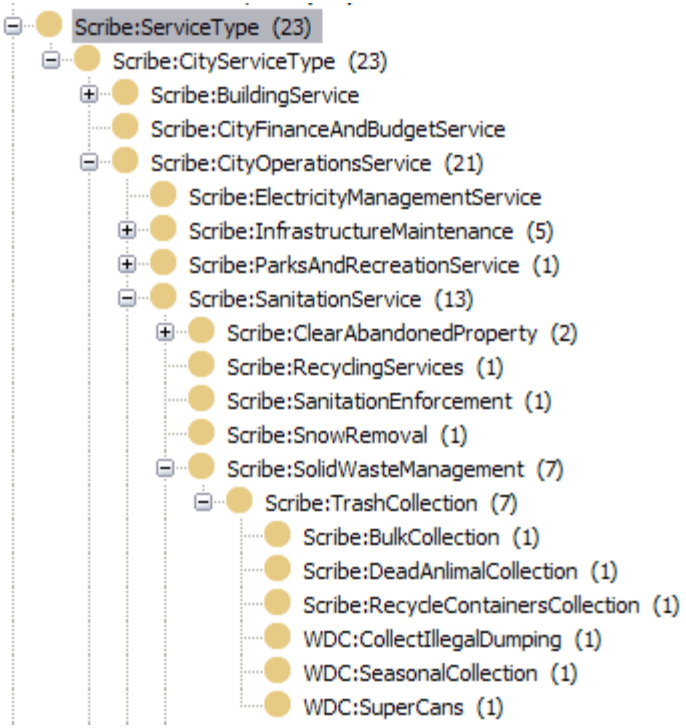
Notice that some of the data is missing in the original table… That's still ok

The SCRIBE Solution. Annotating instance data through the model

As shown previously. The inferencing in the ontology can be leveraged in a query.

Query: Public Sanitation Service Requests

```
SELECT *
WHERE {
   ?subject Scribe:requestAssociatedToServiceType ?serviceType .
   ?serviceType a Scribe:SanitationService .
}
```



Go back

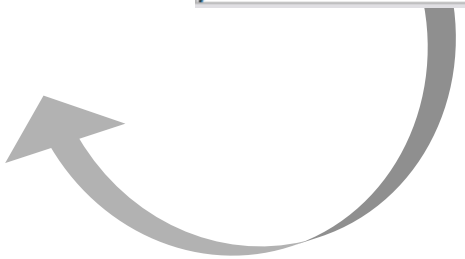
Row	ServiceRequestID	serviceType	dateOrdered	Status
1	10-00000099	DCBulkCollection	Date1/2/201010:19	CLOSI
2	10-00000094	DCRecycleContainerCollection	Date1/2/201010:13	CLOSI
3	10-00000096	DCBulkCollection	Date1/2/201010:16	CLOSI
4	10-00000084	DCTrashCollection	Date1/2/20108:50	CLOSI
5	10-00000088	DCSanitationEnforcement	Date1/2/20109:46	CLOSI
6	10-00000085	DCRecycling	Date1/2/20108:51	CLOSI
7	10-00000079	DCSuperCans	Date1/2/20108:11	CLOSI
8	10-00000082	DCSnowRemoval	Date1/2/20108:37	OPEN
9	10-00000090	DCBulkCollection	Date1/2/20109:59	CLOSI
10	10-00000087	DCBulkCollection	Date1/2/20109:37	CLOSI
11	10-00000081	DCTrashCollection	Date1/2/20108:36	CLOSI
12	10-00000089	DCBulkCollection	Date1/2/20109:50	CLOSI

The SCRIBE Solution. Linking instance data through the model

Everything in a semantic model is connected. The service request can be linked to the name of the dispatcher of the department.

Query: Select events associated to dept of Public Works and his dispatcher

```
SELECT DISTINCT ?Event ?department ?dispatcherName
WHERE {
   ?Event a Scribe:ServiceRequest .
   ?Event Scribe:requestHandledBy WDC:DepartmentOfPublicWorks .
   ?Event Scribe:requestHandledBy ?department .
   ?department Scribe:associatedTo ?dispatcher .
   ?dispatcher a Scribe:Dispatcher .
   ?dispatcher Scribe:roleOfPerson ?dispatcherName .
}
```



Event	department	dispatcherName
10-00000099	DepartmentOfPublicWorks	CharlesCrammer
10-00000055	DepartmentOfPublicWorks	CharlesCrammer
10-00000016	DepartmentOfPublicWorks	CharlesCrammer
10-00000038	DepartmentOfPublicWorks	CharlesCrammer
10-00000012	DepartmentOfPublicWorks	CharlesCrammer
10-00000035	DepartmentOfPublicWorks	CharlesCrammer
10-00000058	DepartmentOfPublicWorks	CharlesCrammer
10-00000005	DepartmentOfPublicWorks	CharlesCrammer
10-00000015	DepartmentOfPublicWorks	CharlesCrammer
10-00000096	DepartmentOfPublicWorks	CharlesCrammer
10-00000084	DepartmentOfPublicWorks	CharlesCrammer
10-00000045	DepartmentOfPublicWorks	CharlesCrammer
10-00000088	DepartmentOfPublicWorks	CharlesCrammer
10-00000002	DepartmentOfPublicWorks	CharlesCrammer
10-00000054	DepartmentOfPublicWorks	CharlesCrammer
10-00000046	DepartmentOfPublicWorks	CharlesCrammer
10-00000020	DepartmentOfPublicWorks	CharlesCrammer
10-00000085	DepartmentOfPublicWorks	CharlesCrammer
10-00000007	DepartmentOfPublicWorks	CharlesCrammer
10-00000029	DepartmentOfPublicWorks	CharlesCrammer
10-00000079	DepartmentOfPublicWorks	CharlesCrammer
10-00000011	DepartmentOfPublicWorks	CharlesCrammer
10-00000056	DepartmentOfPublicWorks	CharlesCrammer
10-00000004	DepartmentOfPublicWorks	CharlesCrammer
10-00000021	DepartmentOfPublicWorks	CharlesCrammer
10-00000044	DepartmentOfPublicWorks	CharlesCrammer

References

- **A direct map of relational data to RDF**, W3C Recommendation, 09-2012, http://www.w3.org/TR/rdb-direct-mapping/
- **R2RML: RDB to RDF Mapping Language**, W3C Recommendation, 09-2012, http://www.w3.org/TR/2012/REC-r2rml-20120927/
- **The D2RQ Platform v0.8 - Treating Non-RDF Relational Databases as Virtual RDF Graphs**, 2012**, http://d2rq.org/**
- Hannes Bohring and Soren Auer, **Mapping XML to Ontologies,** citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.8897
- Rodrigues, P. Rosa, J. Cardoso, **Mapping XML to existing OWL ontologies**, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.292
- DB2OWL, **A tool for automatic Database-To-Ontology mapping**, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.5970
- **Municipal Information Systems Association/Municipal Reference Model (MISA/MRM),** http://www.misa.on.ca/en/
- **National Information Exchange Model**, http://*www.niem.gov/*
- D. Gonzales, C. Ohlandt, E. Landree, C. Wong, R. Bitar and J. Hollywood. **The Universal Core Information Exchange Framework, Assessing its Implications for Acquisition Programs**, RAND report, 2011, http://www.rand.org/pubs/technical_reports/TR885.html
- D. Allemang, J. Hendler, **Semantic Web for the Working Ontologist, Effective Modeling in RDF and OWL**, Morgan Kaufman, 2008.
- R. Uceda-Sosa, B. Srivastava, B. Schloss, **Building a highly consumable semantic model for smarter cities,** in Proceedings of the AI for an Intelligent Planet, 2011.
- Noy, McGuinness, **Ontology Development 101: A Guide to Creating Your First Ontology.** http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html

Some good introductions to Linked Open Data

- Christian Bizer, Tom Heath, Tim Berners-Lee. **Linked Data, the story so far**, *http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf*
- *Tom Heath, Christian Bizer.* **Linked Data, Evolving the Web into a Global Data Space**. *http://linkeddatabook.com/editions/1.0/*

For more information

http://researcher.ibm.com/view_project.php?id=2505

OR

email rosariou@us.ibm.com