Mapping Existing Data Sources into VIVO

Pedro Szekely, Craig Knoblock, Maria Muslea and Shubham Gupta

University of Southern California/ISI

Outline

- Problem

- Current methods for importing data into VIVO

- Karma approach

- Demo

- Conclusions

Pedro Szekely

Problem: Data Ingest

VIVO Data Ingest Guide:

Data ingest refers to any process of loading existing data into VIVO other than by direct interaction with VIVO's content editing interfaces.

Typically this involves downloading or exporting data of interest from an online database or a local system of record.

Current Methods for Importing Data into VIVO

Pedro Szekely

VIVO Provided Ingest Methods

- Writing SPARQL Queries
  - Convert external data (e.g., CSV) into RDF
  - Map data onto VIVO ontology

<div style="background-color:orange; color:white; font-size:2em; padding:10px;">

= Programming

</div>

  - Option 1: Convert data into predefined CSV format
    - Supports limited set of data fields
  - Option 2: Edit existing XSL scripts for your data

Pedro Szekely

Example Data

People

VIVO-demo-data.xlsx

	A	B	C	D	E	F	G	H	I
1	person_ID	name	first	last	middle	email	phone	fax	title
2	3130	Burks, Rosella	Rosella	Burks		BurksR@univ.edu	963.555.1253	963.777.4065	Professor
3	3297	Avila, Damien	Damien	Avila		AvilaD@univ.edu	963.555.1352	963.777.7914	Professor
4	3547	Olsen, Robin	Robin	Olsen		OlsenR@univ.edu	963.555.1378	963.777.9262	Assistant Professor
5	1538	Moises, Edgar Estes	Edgar	Moises	Estes	MoisesE@univ.edu	963.555.2731x3565	963.777.8264	Professor
6	2941	Brian, Heath Pruitt	Heath	Brian	Pruitt	BrianH@univ.edu	963.555.2800	963.777.7249	Associate Curator
7	2401	Claude, Elvin Haney	Elvin	Claude	Haney	ClaudeE@univ.edu	963.555.2902	963.777.9730	Curator
8	2070	Mosley, Edmund	Edmund	Mosley		MosleyE@univ.edu	963.555.2945	963.777.9285	Assistant Professor
9	2561	Derek, Antoine Mccoy	Antoine	Derek	Mccoy	DerekA@univ.edu	963.555.2992	963.777.5454	Curator
10	1625	Hawkins, Callie	Callie	Hawkins		HawkinsC@univ.edu	963.555.3350x6480	963.777.4949	Professor

Organizations

VIVO-demo-data.xlsx

	A	B	C
1	org_ID	org_name	org_vivo_uri
2	101028	Energy Sciences Department	http://vivoweb.org/ontology/core#Department
3	102017	Geothermal Technology Department	http://vivoweb.org/ontology/core#Department
4	100000	University of VIVO	http://vivoweb.org/ontology/core#University
5	102000	College of Research	http://vivoweb.org/ontology/core#College
6	101000	College of Science	http://vivoweb.org/ontology/core#College

Positions

VIVO-demo-data.xlsx

	A	B	C	D	E	F
1	person_ID	job_title	position_type	department_name	department_ID	start_date
2	2254	Assistant Professor I	Faculty	Energy Sciences Department	101028	2003
3	2217	Assistant Professor II	Faculty	Energy Sciences Department	101028	2008
4	1968	Assistant Professor III	Faculty	Energy Sciences Department	101028	2002
5	2303	Assistant Professor IV	Faculty	Energy Sciences Department	101028	2003
6	3903	Assistant Professor V	Faculty	Energy Sciences Department	101028	2005
7	2146	Assistant Professor VI	Faculty	Energy Sciences Department	101028	1999
8	2895	Associate Curator of South Library	Faculty	Energy Sciences Department	101028	2005
9	2682	Associate Curator of Central Library	Faculty	Energy Sciences Department	101028	2010
10	3112	Professor of Kinetic Energy	Faculty	Energy Sciences Department	101028	2008
11	3095	Professor of Elastic Energy	Faculty	Energy Sciences Department	101028	2008
12	2389	Curator of South Library	Faculty	Energy Sciences Department	101028	2008

VIVO Data Ingest Guide

http://www.vivoweb.org/data-ingest-guide

Step #1: Create a Local Ontology

Data Ingest Menu

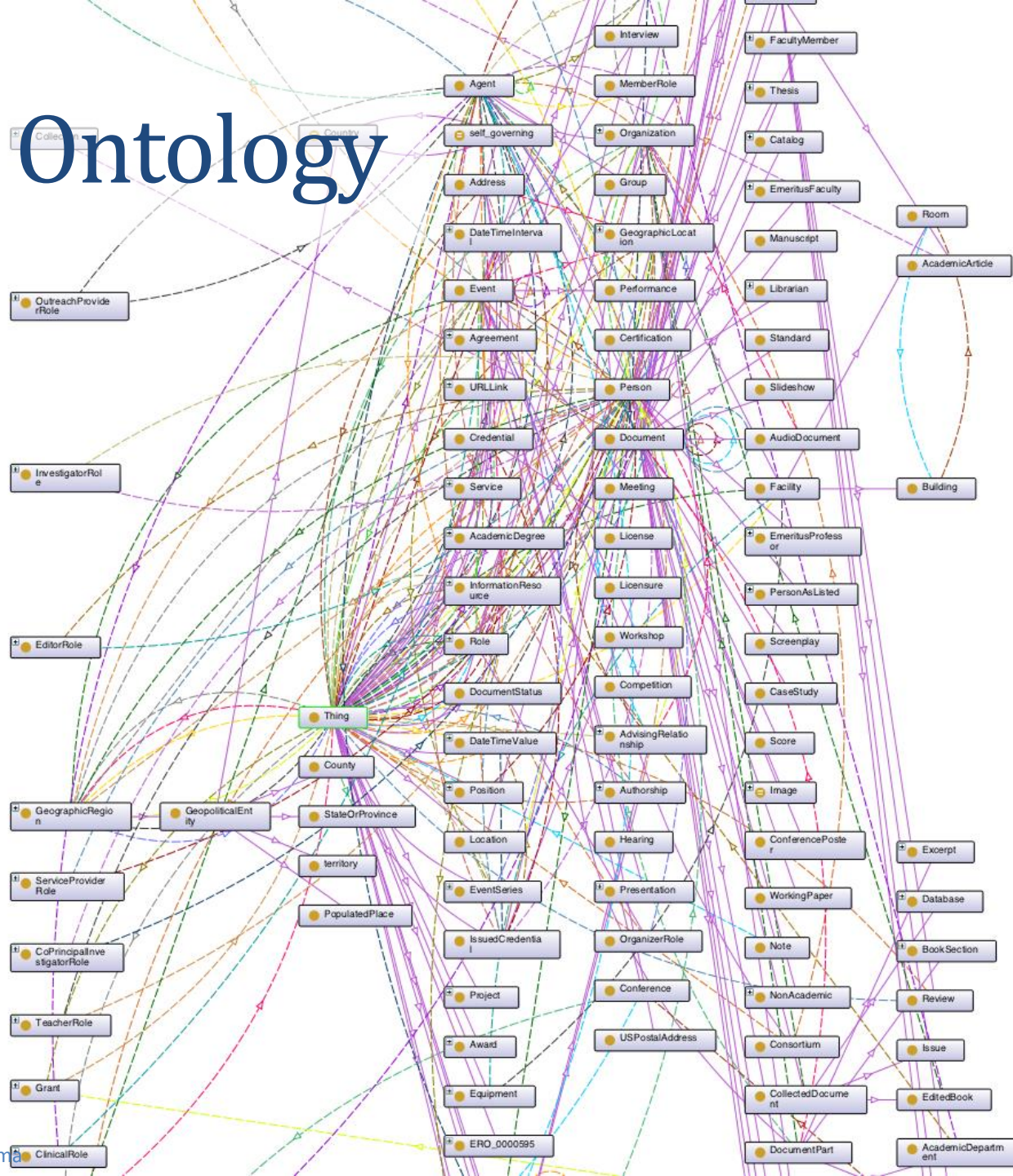Step#2: Create Workspace Models

Step#3: Pull External Data File into RDF

Step# 4: Map Tabular Data onto Ontology

Step#5: Construct the Ingested Entities

Step#6: Load to Webapp

VIVO Data Ingest Guide

http://www.vivoweb.org/data-ingest-guide

Step #1: Create a Local Ontology

Data Ingest Menu

Step#2: Create Workspace Models

Step#3: Pull External Data File into RDF

Step# 4: Map Tabular Data onto Ontology

Step#5: Construct the Ingested Entities

Step#6: Load to Webapp

VIVO Ontology

VIVO Data Ingest Guide

http://www.vivoweb.org/data-ingest-guide

Step #1: Create a Local Ontology

Data Ingest Menu

Step#2: Create Workspace Models

Step#3: Pull External Data File into RDF

Step# 4: Map Tabular Data onto Ontology

Step#5: Construct the Ingested Entities

Step#6: Load to Webapp

Step#5: Construct the Ingested Entities

Write the following SPARQL query

```
Construct {
    ?person <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
            <http://vivoweb.org/ontology/core#FacultyMember> .
    ?person <http://www.w3.org/2000/01/rdf-schema#label> ?fullname .
    ?person <http://xmlns.com/foaf/0.1/firstName> ?first .
    ?person <http://vivoweb.org/ontology/core#middleName> ?middle .
    ?person <http://xmlns.com/foaf/0.1/lastName> ?last .
    ?person <http://vitro.mannlib.cornell.edu/ns/vitro/0.7#moniker> ?title .
    ?person <http://vivoweb.org/ontology/core#workPhone> ?phone .
    ?person <http://vivoweb.org/ontology/core#workFax> ?fax .
    ?person <http://vivoweb.org/ontology/core#workEmail> ?email .
    ?person <http://localhost/vivo/ontology/vivo-local#peopleID> ?hrid .
}
Where {
    ?person <http://localhost/vivo/ws_ppl_name> ?fullname .
    ?person <http://localhost/vivo/ws_ppl_first> ?first .
    optional { ?person <http://localhost/vivo/ws_ppl_middle> ?middle . }
    ?person <http://localhost/vivo/ws_ppl_last> ?last .
    ?person <http://localhost/vivo/ws_ppl_title> ?title .
    ?person <http://localhost/vivo/ws_ppl_phone> ?phone .
    ?person <http://localhost/vivo/ws_ppl_fax> ?fax .
    ?person <http://localhost/vivo/ws_ppl_email> ?email .
    ?person <http://localhost/vivo/ws_ppl_person_ID> ?hrid .
}
```

Constructs the people entities

SPARQL Ingest Is Difficult

```
Construct {
    ?person <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://vivoweb.org/ontology/core#FacultyMember> .
    ?person <http://www.w3.org/2000/01/rdf-schema#label> ?fullname .
    ?person <http://xmlns.com/foaf/0.1/firstName> ?first .
    ?person <http://vivoweb.org/ontology/core#middleName> ?middle .
    ?person <http://xmlns.com/foaf/0.1/lastName> ?last .
    ?person <http://vitro.mannlib.cornell.edu/ns/vitro/0.7#moniker> ?title
        .
    ?person <http://vivoweb.org/ontology/core#workPhone> ?phone .
    ?person <http://vivoweb.org/ontology/core#workFax> ?fax .
    ?person <http://vivoweb.org/ontology/core#workEmail> ?email .
    ?person <http://localhost/vivo/ontology/vivo-local#peopleID> ?hrid .
}
Where {
    ?person <http://localhost/vivo/ws_ppl_name> ?fullname .
    ?person <http://localhost/vivo/ws_ppl_first> ?first .
    optional { ?person <http://localhost/vivo/ws_ppl_middle> ?middle . }
    ?person <http://localhost/vivo/ws_ppl_last> ?last .
    ?person <http://localhost/vivo/ws_ppl_title> ?title .
    ?person <http://localhost/vivo/ws_ppl_phone> ?phone .
    ?person <http://localhost/vivo/ws_ppl_fax> ?fax .
    ?person <http://localhost/vivo/ws_ppl_email> ?email .
    ?person <http://localhost/vivo/ws_ppl_person_ID> ?hrid .
}
```

```
Construct {
    ?org <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://xmlns.com/foaf/0.1/Organization> .
    ?org <http://localhost/vivo/ontology/vivo-local#orgID> ?deptID .
    ?org <http://www.w3.org/2000/01/rdf-schema#label> ?name .
}
Where
{
    ?org <http://localhost/vivo/ws_org_org_ID> ?deptID .
    ?org <http://localhost/vivo/ws_org_org_name> ?name .
}
```

```
Construct {
    ?position <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://vivoweb.org/ontology/core#FacultyPosition> .
    ?position <http://vivoweb.org/ontology/core#startYear> ?year .
    ?position <http://www.w3.org/2000/01/rdf-schema#label> ?title .
    ?position <http://vivoweb.org/ontology/core#titleOrRole> ?title .
    ?position <http://vivoweb.org/ontology/core#positionForPerson> ?person .
    ?person <http://vivoweb.org/ontology/core#personInPosition> ?position .
}
Where {
    ?position <http://localhost/vivo/ws_post_department_ID> ?orgID .
    ?position <http://localhost/vivo/ws_post_start_date> ?year .
    ?position <http://localhost/vivo/ws_post_job_title> ?title .
    ?position <http://localhost/vivo/ws_post_person_ID> ?posthrid .
    ?person <http://localhost/vivo/ws_ppl_person_ID> ?perhrid .
    FILTER((?posthrid)=(?perhrid))
}
```
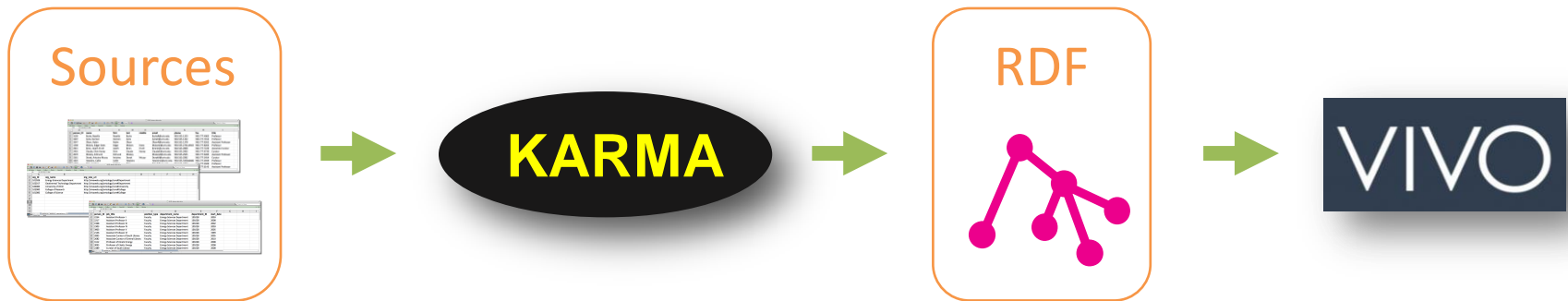
```
Construct {
    ?position <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
        <http://vivoweb.org/ontology/core#FacultyPosition> .
    ?position <http://vivoweb.org/ontology/core#startYear> ?year .
    ?position <http://www.w3.org/2000/01/rdf-schema#label> ?title .
    ?position <http://vivoweb.org/ontology/core#titleOrRole> ?title .
    ?org <http://vivoweb.org/ontology/core#organizationForPosition> ?position .
    ?position <http://vivoweb.org/ontology/core#positionInOrganization> ?org .
}
Where {
    ?position <http://localhost/vivo/ws_post_start_date> ?year .
    ?position <http://localhost/vivo/ws_post_job_title> ?title .
    ?position <http://localhost/vivo/ws_post_department_ID> ?postOrgID .
    ?org <http://localhost/vivo/ws_org_org_ID> ?orgID .
    FILTER((?postOrgID)=(?orgID))
}
```

Harvester Data Ingest

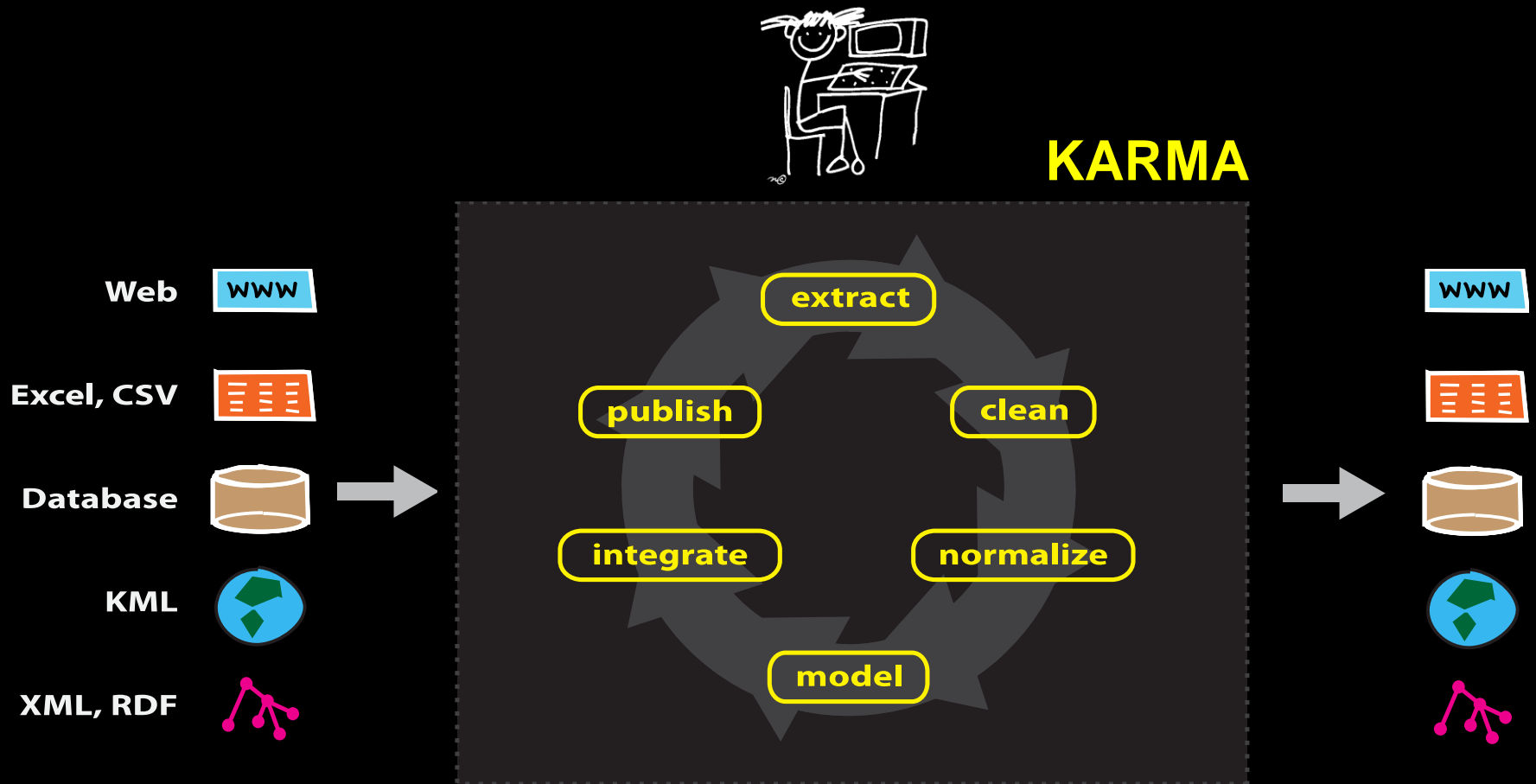## Program in XSLT

```
<core:positionInOrganization>
    <rdf:Description rdf:about="{$baseURI}org/org{$orgID}">
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Organization"/>
    <xsl:if test="not( $this/db-CSV:DEPARTMENTID = '' or $this/db-CSV:DEPARTMENTID = 'null' )">
        <score:orgID><xsl:value-of select="$orgID"/></score:orgID>
    </xsl:if>
    <xsl:if test="not( $this/db-CSV:DEPARTMENTNAME = ''
                  or $this/db-CSV:DEPARTMENTNAME = 'null' )">
        <rdfs:label><xsl:value-of select="$this/db-CSV:DEPARTMENTNAME"/></rdfs:label>
    </xsl:if>
    <core:organizationForPosition rdf:resource=
         "{$baseURI}position/positionFor{$personid}from{$this/db-CSV:STARTDATE}"/>
    </rdf:Description>
</core:positionInOrganization>
```

Karma Approach

Sources → KARMA → RDF → VIVO

Overall Karma Effort

Using Karma to Ingest Data into VIVO

KARMA

Web

Excel, CSV

Database

KML

XML, RDF

extract

publish

clean

integrate

normalize

model
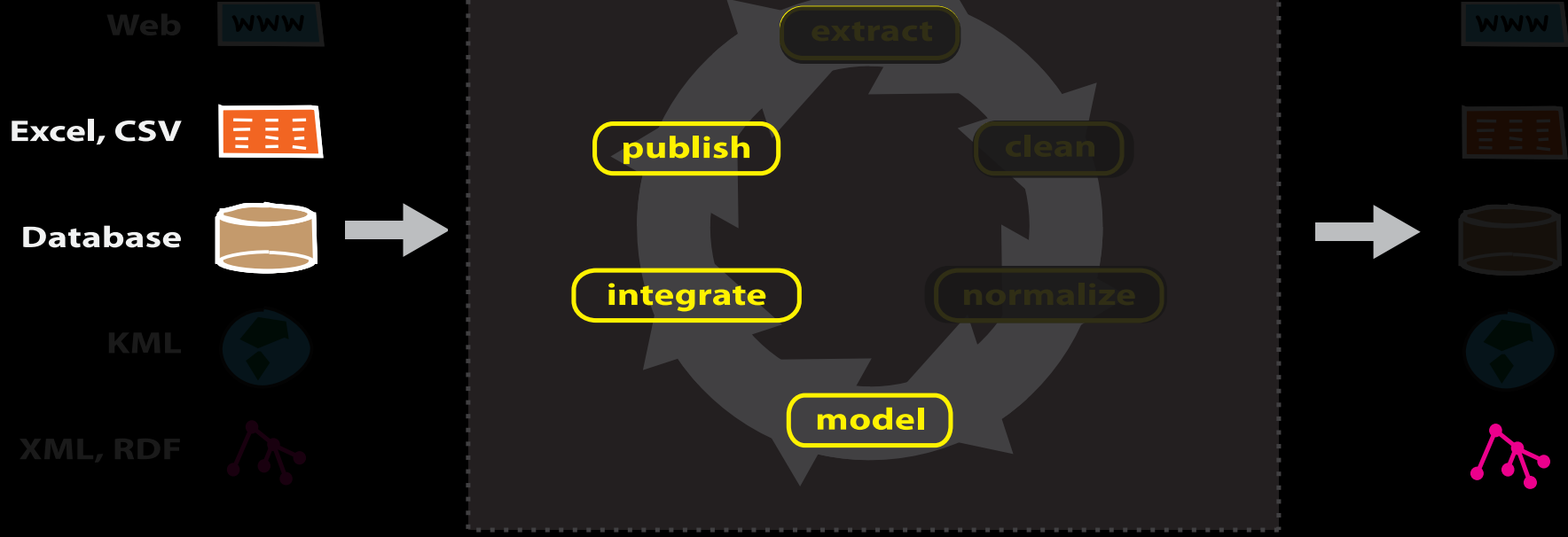
Karma Benefits

~~Programming~~

Interactive

Easy

Fast

Karma Workspace

Karma Models: Semantic Types

Semantic Types

Capture semantics of the values in each column
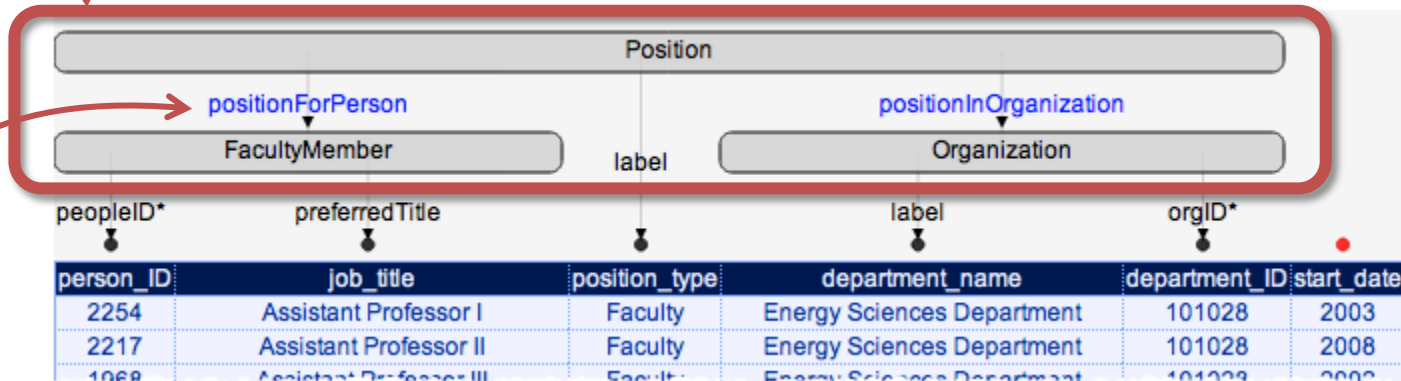in terms of classes and properties in the ontology
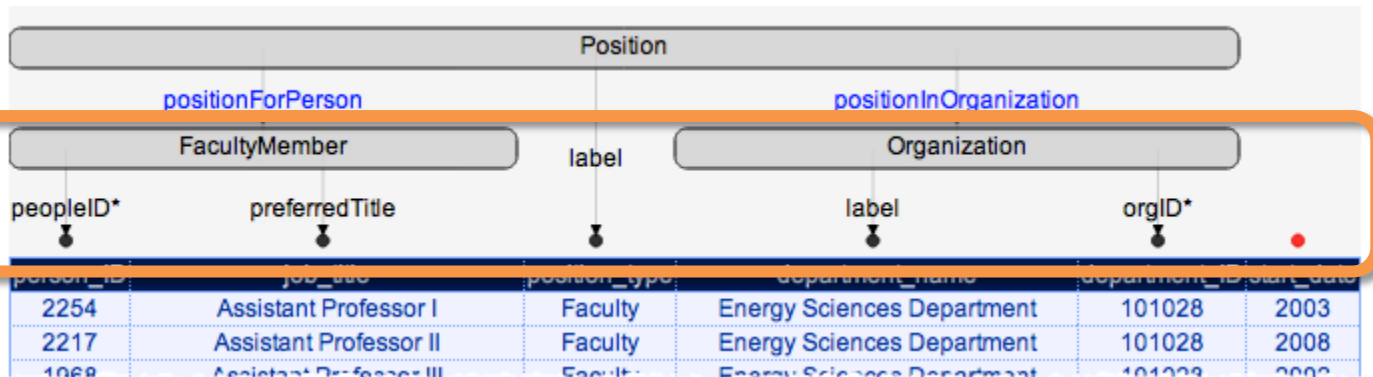
the peopleID of a FacultyMember

the label of an Organization

**Karma learns to recognize semantic types
each time the user assigns one manually**

Karma Models: Relationships

Relationships

Capture the relationships among columns
in terms of classes and properties in the ontology

the relationship between Position and FacultyMember is positionForPerson

**Karma automatically computes relationships
based on the object properties defined in the ontology**

Using Karma to ingest data samples from the "Data Ingest Guide"

Karma Demo

Conclusions

Conclusions

- Generic data-to-ontology-to-RDF mapping tool

- Easy to use: interactive, no programming

- Used Karma to populate USC VIVO instance

- Open source: you can use it too

From Simon Gaeremynck, Sakai Foundation

More Information

- http://youtu.be/EQcMc4TrfuE
  - Using Karma to ingest VIVO data

- http://isi.edu/integration/karma
  - Publications and videos
  - Software download (open source)

- Contacts:
  - pszekely@isi.edu
  - knoblock@isi.edu