SparqPlug : Generating Linked Data from Legacy HTML, SPARQL, and the DOM

Peter Coetzee
Imperial College London

Tom Heath
Talis Information Ltd.

Enrico Motta
Knowledge Media Institute

sparqplug

SparqPlug : Generating Linked Data from Legacy HTML, SPARQL, and the DOM

- The Problem
- Current Approaches
- SparqPlug's Background
- SparqPlug's Approach
- Linked Data
- Anatomy of a Job
- Maintenance
- Wrap-Up

The Problem

- Bootstrapping the Web of Data
- Inertia for webmasters to convert
- Risks of doing so blindly – Good Linked Data!
- Difficult and time-consuming
  - Triplify
  - SquirrelRDF
  - etc

Current Approaches

- Piggy Bank & Thresher: Easy to use screen scrapers → RDF Silo
- Sponger & Triplr: Requires a marked up source
- SPAT: Great approach, implementations??
- XSLT with XQuery: *Another* language to learn, could be more expressive and flexible

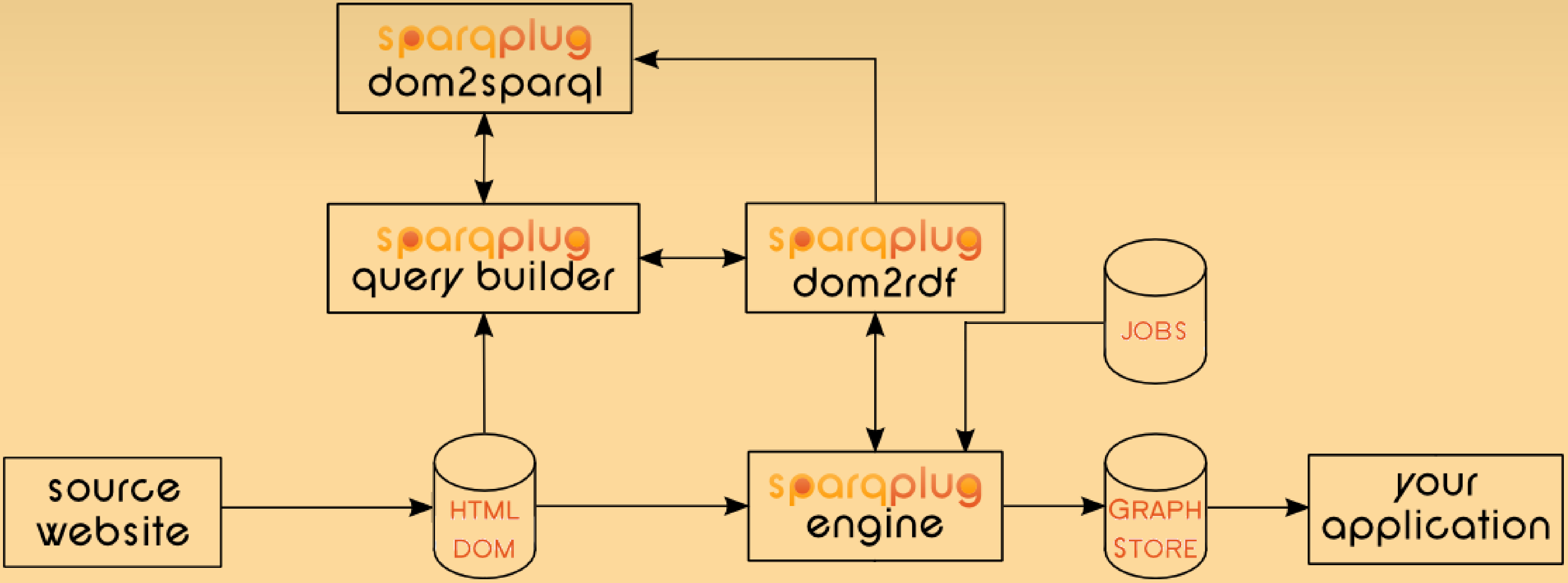
SparqPlug's Background

- Developed in Summer of 2007
- Funded by OpenKnowledge Project, development took place at KMi
- Currently hosted at KMi
- Built on Java, Jena, Tomcat, MySQL, NG4J
- **http://sparqplug.rdfize.com/**

SparqPlug's Approach

- Tidy and DOM2RDF
- Query the DOM *directly* with SPARQL
  - All the expressivity of a declarative query language
  - Proprietary extensions – e.g. Property Functions
- DOM2SPARQL
- Let SparqPlug manage the entire process, from extraction to de-referencing

SparqPlug's Approach



Linked Data

- Content Negotiation handled automatically
- URIs generated in a separate namespace and forwarded through Tomcat to the SparqPlug application
- Property Functions to help process data
- SPARQL endpoint automatically created for each data set

Anatomy of a Job

- You give:
  - Prototypical Query (SPARQL)
  - Link Query (SPARQL)
  - Graph Name Generator (RegExp)
- We create:
  - Maintenance data
  - Linked Data constructs
  - RDF!

Maintenance

- Source graph hashed at SPARQL CONSTRUCT time
- Hash then checked periodically for updated data
- Graph regenerated and UNION'd with existing RDF in each named graph

Wrap-Up

- SparqPlug offers a *simple*, partially *automated* and *scalable* solution to the problem of creation and maintenance of RDF data from an arbitrary HTML data source
- **http://sparqplug.rdfize.com/**
- Questions?
- peter @ coetzee . org