Publishing Relational Databases as Linked Data

Oktie Hassanzadeh
University of Toronto

March 2011

CS 443: Database Management Systems - Winter 2011

Outline

- Part 1: How to Publish Linked Data on the Web
  - 6 Steps in Publishing Linked Data
- Part 2: How to Publish Relational Databases as Linked Data
  - Mapping Relational Databases to RDF
  - Following Linked Data Principles & Guidelines
  - Mapping Tools & D2R Server

Part 1:
How to Publish Linked Data on the Web

Slides by:
Dr. Tom Heath
Platform Division
Talis Information Ltd

tom.heath@talis.com
http://tomheath.com/id/me

Presentation at SSSW2009, Cercedilla, Spain
http://tomheath.com/slides/2009-07-cercedilla-how-to-publish-linked-data.pdf

UofT: DB Group

Scenario

- Online whisky shop: Wiskii.com
- New business venture, founded by Jeff
- For the whisky connoisseur
- Detailed background information from experts
- Contributions from customers
- Custom web app, relational backend
- Simultaneous publication in HTML and RDF

6 Steps to Publishing Linked Data

1. Understand the Principles
2. Understand your Data
3. Choose URIs for Things in your Data
4. Setup Your Infrastructure
5. Link to other Data Sets
6. Describe and Publicise your Data

1. Understand the Principles

Linked Data Principles: Redux

- Use URIs as names for things
  - anything, not just documents
  - you are not your homepage
  - information resources and non-information resources
- Use HTTP URIs
  - globally unique names, distributed ownership
  - allows people to look up those names
- Provide useful information in RDF
  - when someone looks up a URI
- Include RDF links to other URIs
  - to enable discovery of related information

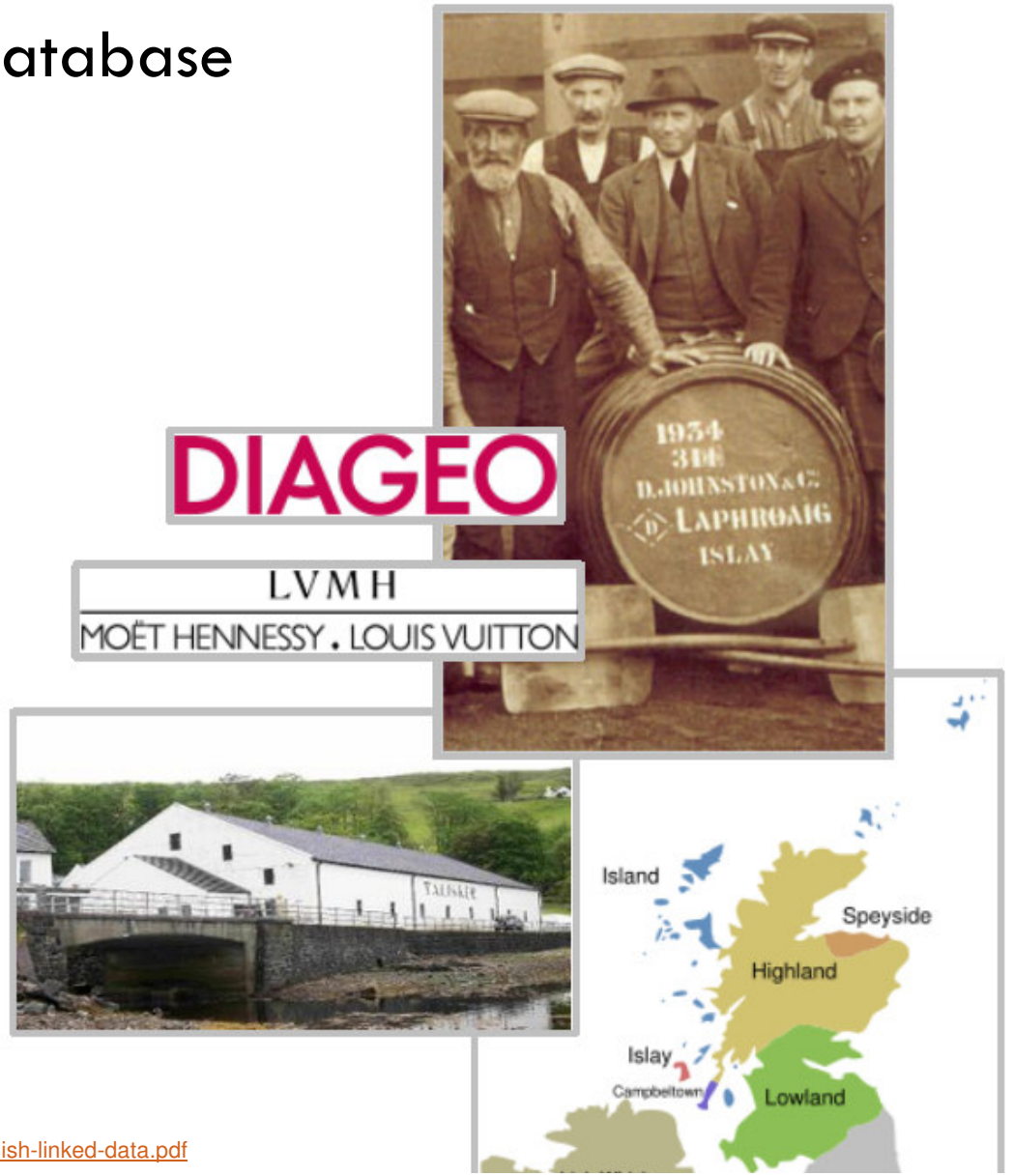
2. Understand your Data

2. Understand Your Data

- What are the key things present in your data?
- People?
  - Places?
  - Books?
  - Films?
  - Musicians?
  - Concepts?
  - Photos?
  - Comments?
  - Reviews?
  - ...

2. Understand Your Data

- Things in the *Wiskii.com* database
  - Distilleries
  - Regions and Locations
  - Founders
  - Owners
  - Brands
  - Products
  - Photos
  - Reviews
  - Comments
  - Prices/Offers



2. Understand Your Data

- What vocabularies can be used to describe these?
  - Principles
    - Reuse, don't reinvent
    - Mix liberally
  - Potential Ontologies/Vocabularies
    - Geo
    - GoodRelations
    - FOAF
    - Review
    - SIOC
    - Whisky

3. Choose URIs for Things in Your Data

3. Choosing URIs: Principles

- Use HTTP URIs
- Keep out of other peoples' namespaces
  - http://www.imdb.com/title/tt0441773/
  - http://www.imdb.com/title/tt0441773/thing
  - http://myfilms.com/tt0441773
  - http://myfilms.com/tt0441773/html
- Abstract away from implementation details
  - http://dbpedia.org/resource/Berlin
  - http://www4.wiwiss.fu-berlin.de:2020/demos/dbpedia/cgi-bin/resources.php?id=Berlin
- Hash or Slash
  - http://mydomain.com/foaf.rdf#me
  - http://mydomain.com/id/me

3. Choosing URIs: Common Patterns

- http://dbpedia.org/resource/New_York_City ← Thing
- http://dbpedia.org/data/New_York_City ← RDF data
- http://dbpedia.org/page/New_York_City ← HTML page

- http://revyu.com/people/tom ← Thing
- http://revyu.com/people/tom/about/rdf ← RDF data
- http://revyu.com/people/tom/about/html ← HTML page

- http://kmi.open.ac.uk/people/tom/ ← Thing
- http://kmi.open.ac.uk/people/tom/rdf ← RDF data
- http://kmi.open.ac.uk/people/tom/html ← HTML page

- http://mydomain.com/thing ← Thing
- http://mydomain.com/thing.rdf ← RDF data
- http://mydomain.com/thing.html ← HTML page

3. Choosing URIs: Wiskii.com

- http://wiskii.com/regions/speyside

- http://wiskii.com/distilleries/talisker
- http://wiskii.com/brands/talisker

- http://wiskii.com/products/talisker-10-yo
- http://wiskii.com/products/glenmorangie-lasanta

- http://wiskii.com/people/william-matheson

- http://wiskii.com/photos/58

- http://wiskii.com/reviews/271

3. Choosing URIs: Wiskii.com

- http://wiskii.com/distilleries/talisker
- http://wiskii.com/distilleries/talisker/rdf
- http://wiskii.com/distilleries/talisker/html

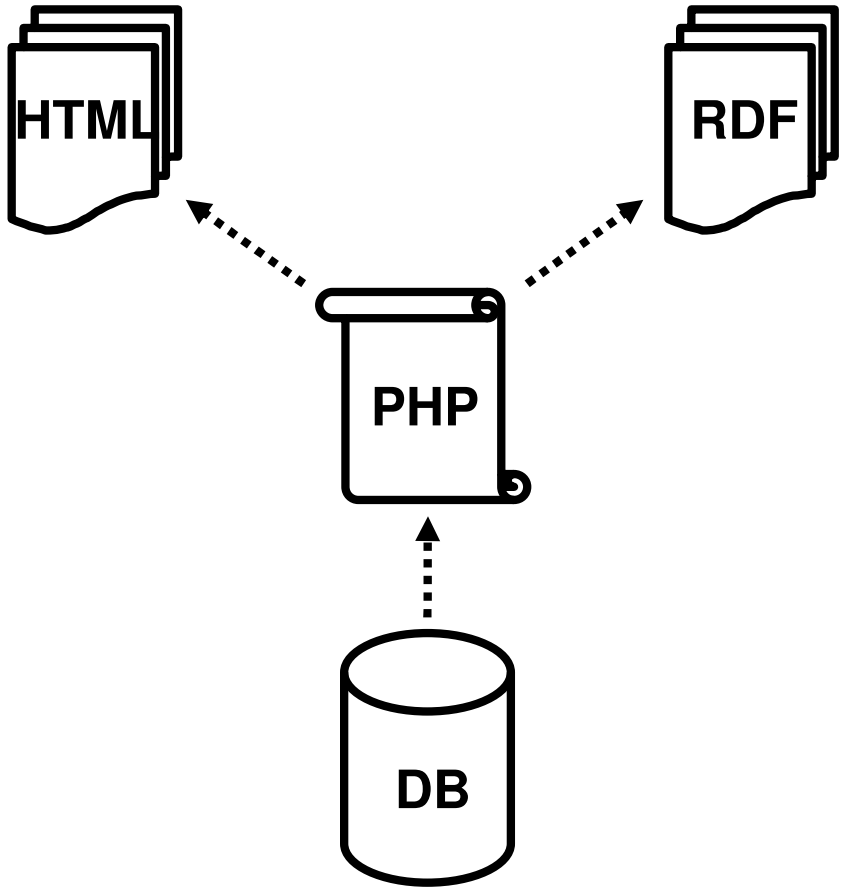
- http://wiskii.com/brands/talisker
- http://wiskii.com/brands/talisker/rdf
- http://wiskii.com/brands/talisker/html

- http://wiskii.com/people/william-matheson
- http://wiskii.com/people/william-matheson/rdf
- http://wiskii.com/people/william-matheson/html

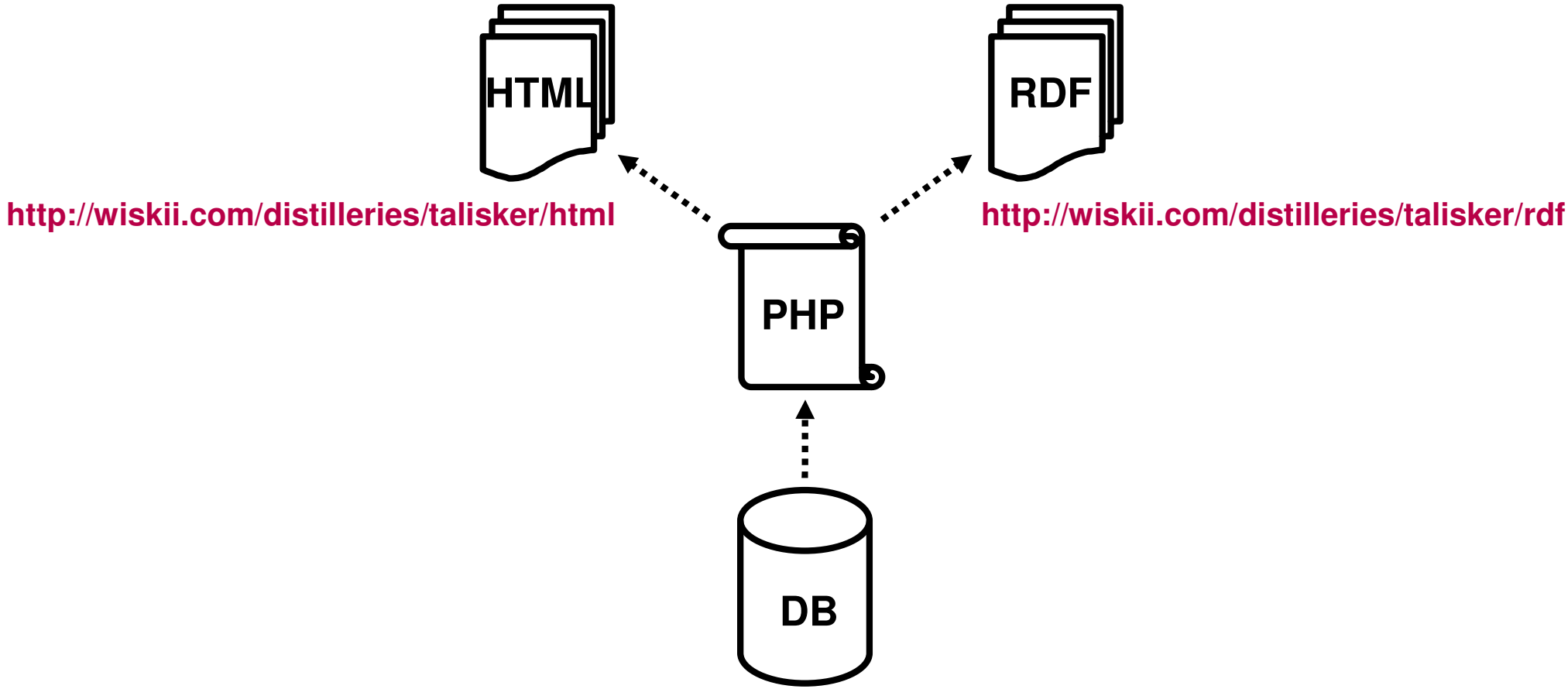
- http://wiskii.com/photos/58

4. Setup Your Infrastructure

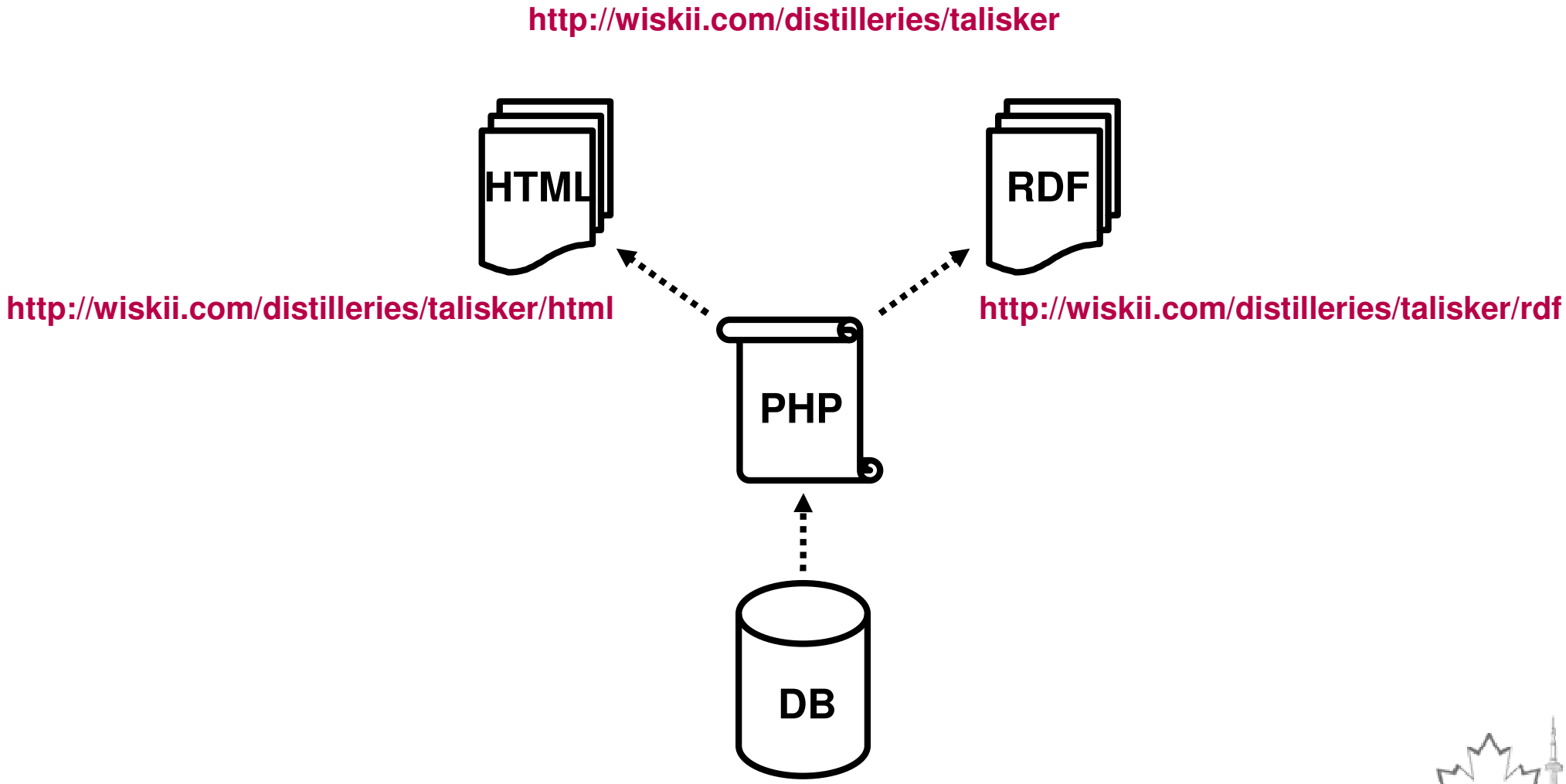
4. Setup Your Infrastructure



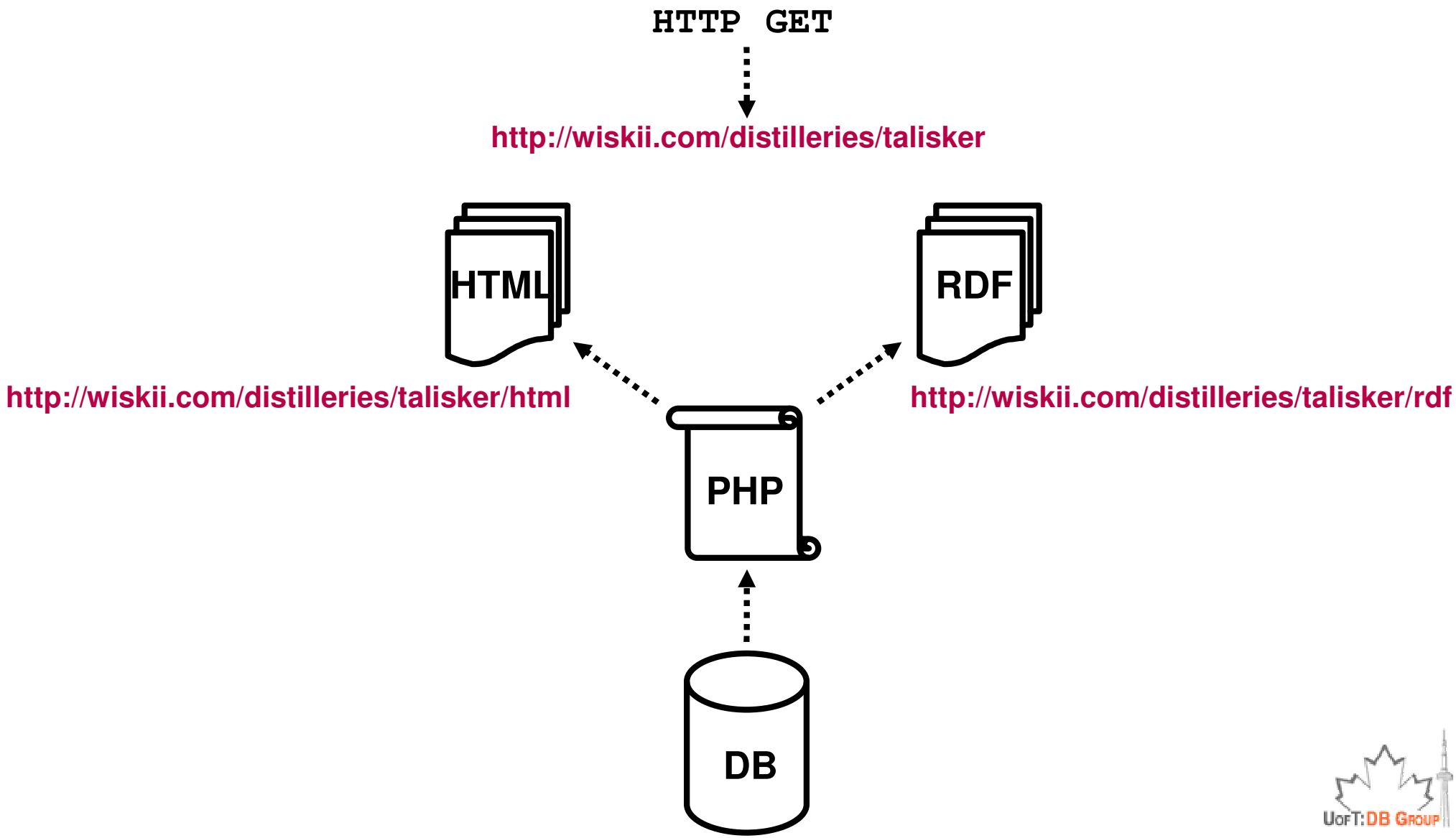
4. Setup Your Infrastructure



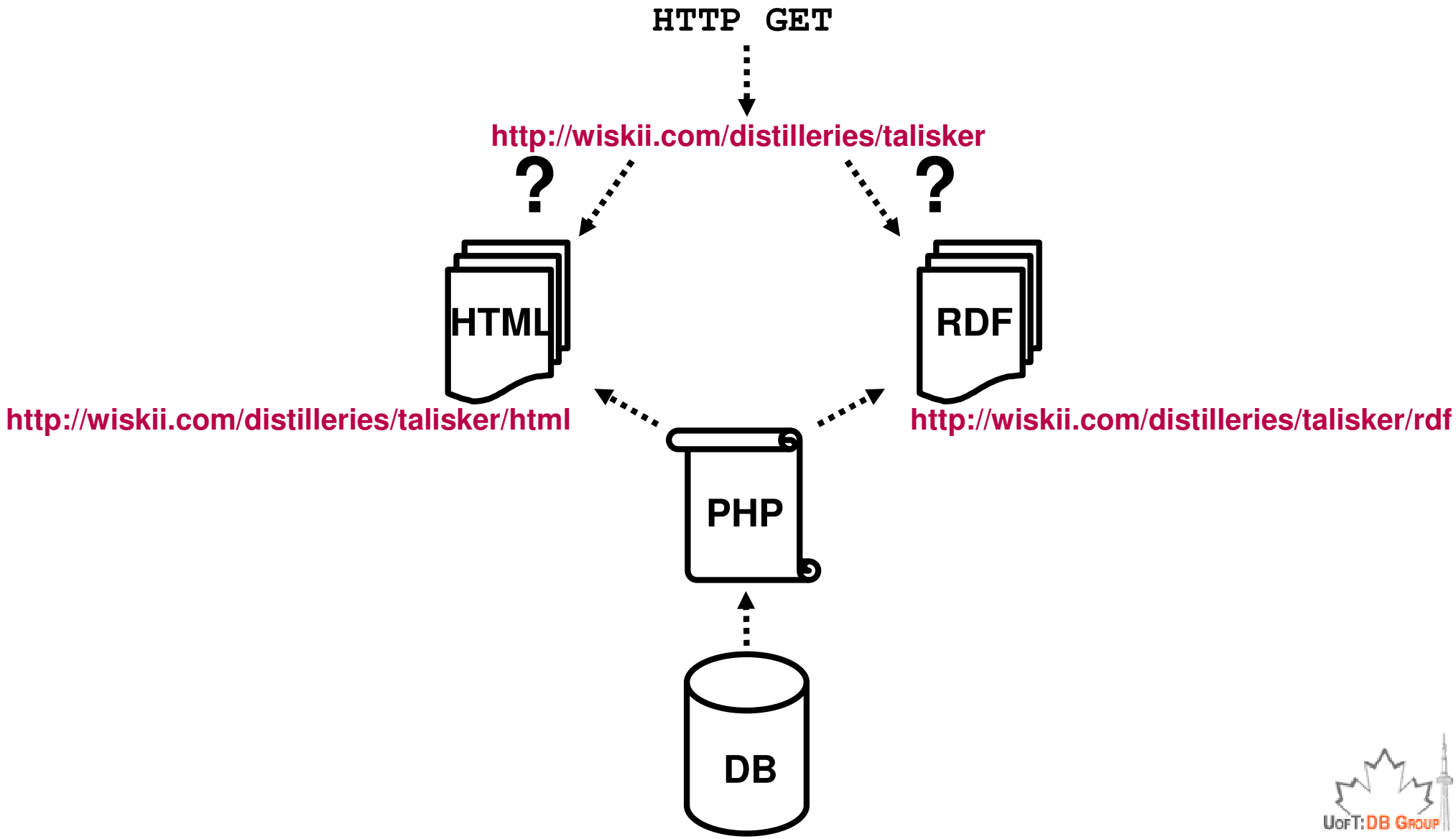
4. Setup Your Infrastructure



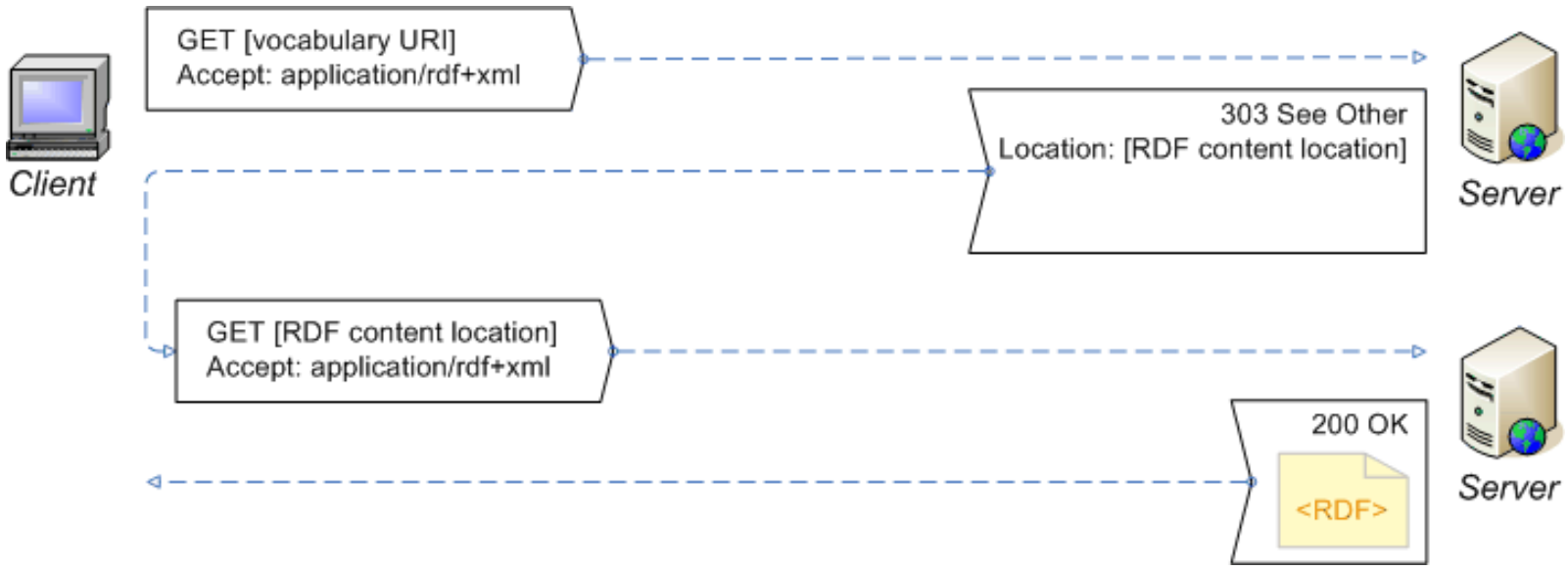
4. Setup Your Infrastructure



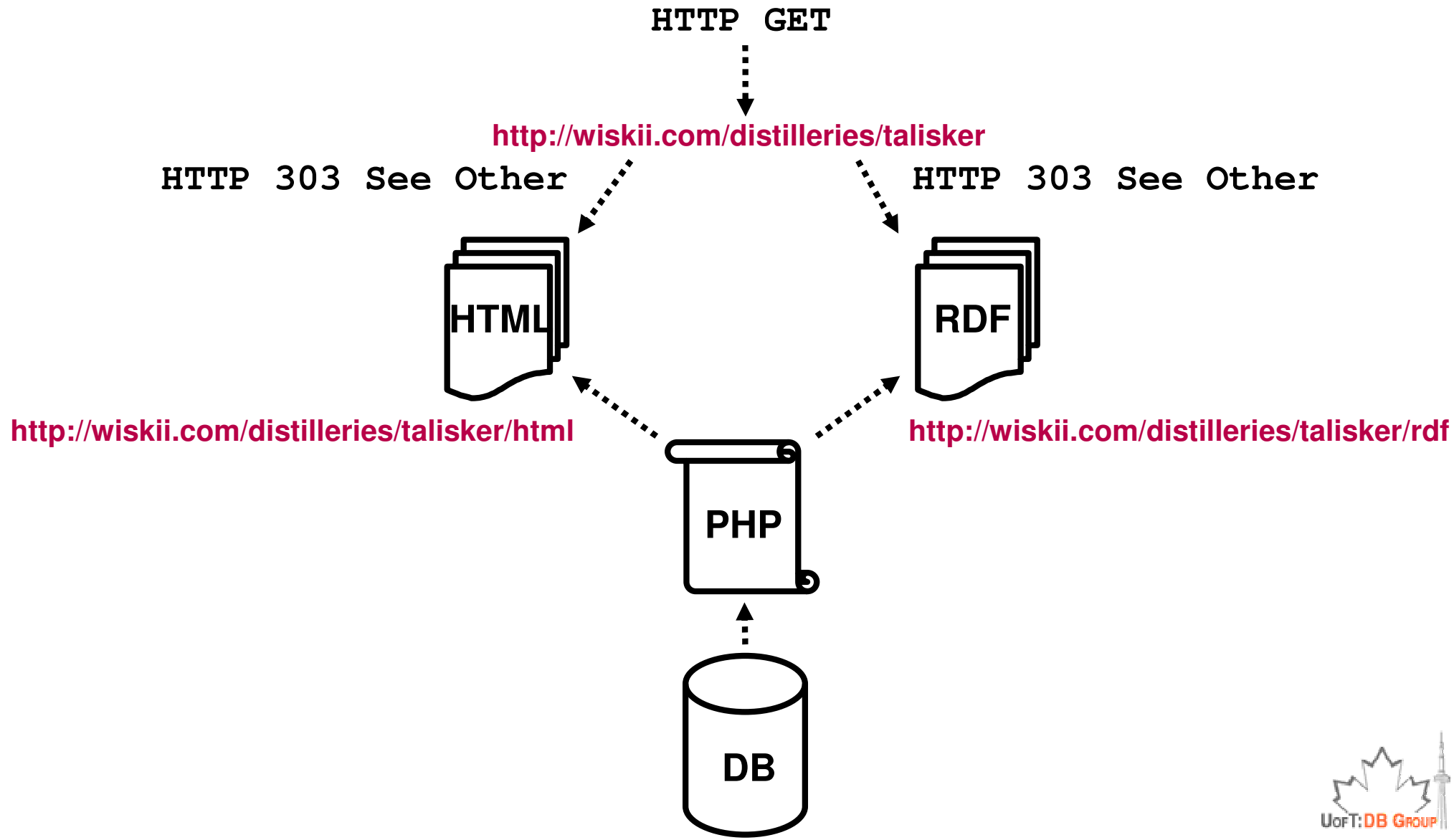
4. Setup Your Infrastructure



Content Negotiation



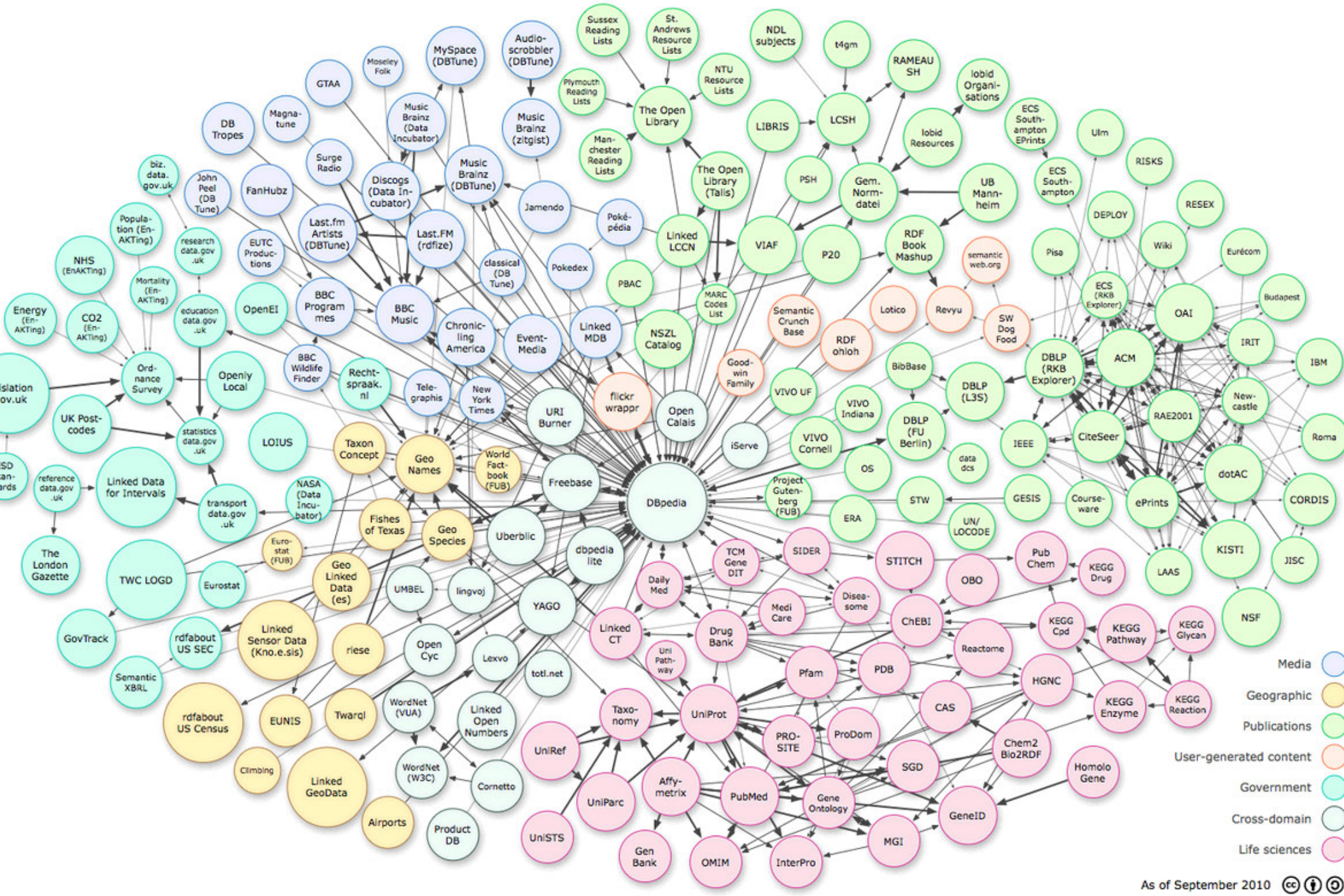
4. Setup Your Infrastructure



4. Setup Your Infrastructure

- Code samples for ConNeg and 303 Redirects
  - http://linkeddata.org/tools
- Useful tools for debugging
  - Firefox Extensions
    - Modify Headers, LiveHTTPHeaders
  - cURL
    - http://dowhatimean.net/2007/02/debugging-semantic-web-sites-with-curl
- You don't have to roll your own!
  - See Toolbox section below and http://linkeddata.org/tools

5. Link to Other Data Sets



Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences

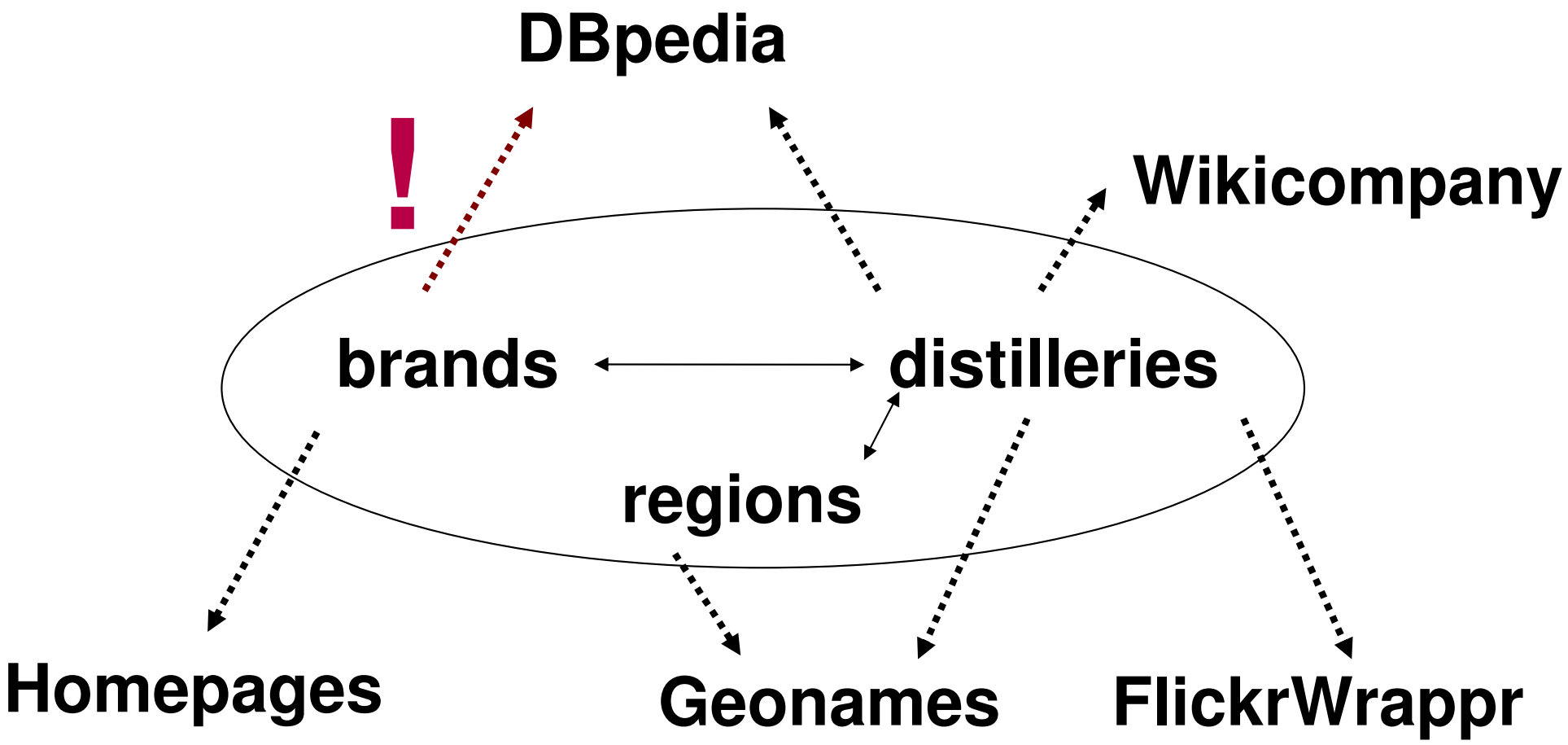
As of September 2010

Slides by Tom Heath, available at http://tomheath.com/slides/2009-07-cercedilla-how-to-publish-linked-data.pdf

5. Link to other Data Sets

- Popular Generic Predicates for Linking
  - owl:sameAs
  - foaf:homepage
  - foaf:topic
  - foaf:based_near
  - foaf:maker/foaf:made
  - foaf:depiction
  - foaf:page
  - foaf:primaryTopic
  - rdfs:seeAlso

5. Link to other Data Sets



Slides by Tom Heath, available at http://tomheath.com/slides/2009-07-cercedilla-how-to-publish-linked-data.pdf

5. Link to other Data Sets

- Basic Linking Approaches
  - Common Key Matching
    - e.g. ISBN, Wikipedia Article ID, Musicbrainz IDs
  - String Matching
    - e.g. comparing labels using string similarity measures
  - Graph Matching
    - Do these two things have the same label, type and coordinates
- Linking Frameworks
  - Silk: Volz et al., LDOW2009
  - LinQuer: Hassanzadeh et al., PVLDB2009
- Aim for reciprocal links

Based on slides by Tom Heath, available at http://tomheath.com/slides/2009-07-cercedilla-how-to-publish-linked-data.pdf

6. Describe and Publicise your Data

6. Describe and Publicise your Data

- Help others discover and index your data
  - Send pings to Sindice and pingthesemanticweb.com
  - Provide a Semantic Sitemap for your Data Set
  - Provide a voiD description of your Data Set

- Apply a license or waiver to your data set
  - Protects consumers of your data => encourages reuse
  - Using Open Database License (ODbL) or releasing into the public domain by applying PDDL or CC0 waivers is encouraged
    - http://opendatacommons.org/
  - Creative Commons licences are also applicable
    - More focus recently on licensing data/databases http://wiki.creativecommons.org/Data

Based on slides by Tom Heath, available at http://tomheath.com/slides/2009-07-cercedilla-how-to-publish-linked-data.pdf

Summary

1. Understand the Principles
2. Understand your Data
3. Choose URIs for Things in your Data
4. Setup Your Infrastructure
5. Link to other Data Sets
6. Describe and Publicise your Data

Part 2: How to Publish Relational Databases as Linked Data

- Mapping Relational Databases to RDF
- Following Linked Data Principles & Guidelines
- Mapping Tools & D2R Server

Mapping Relational Databases to RDF

Building RDF Graphs out of Relational Data

Book

ID	Author	Title	Publisher	Year
ISBN 0-00-6511409-X	id_xyz	The Glass Palace	id_qpr	2000

Author

ID	Name	Homepage
id_xyz	Ghosh, Amitav	http://www.amitavghosh.com

Publisher

ID	Publisher's name	City
id_qpr	Harper Collins	London

Example from slides by Ivan Herman, available at http://www.w3.org/2010/Talks/0622-SemTech-IH/

Building RDF Graphs out of Relational Data

Book

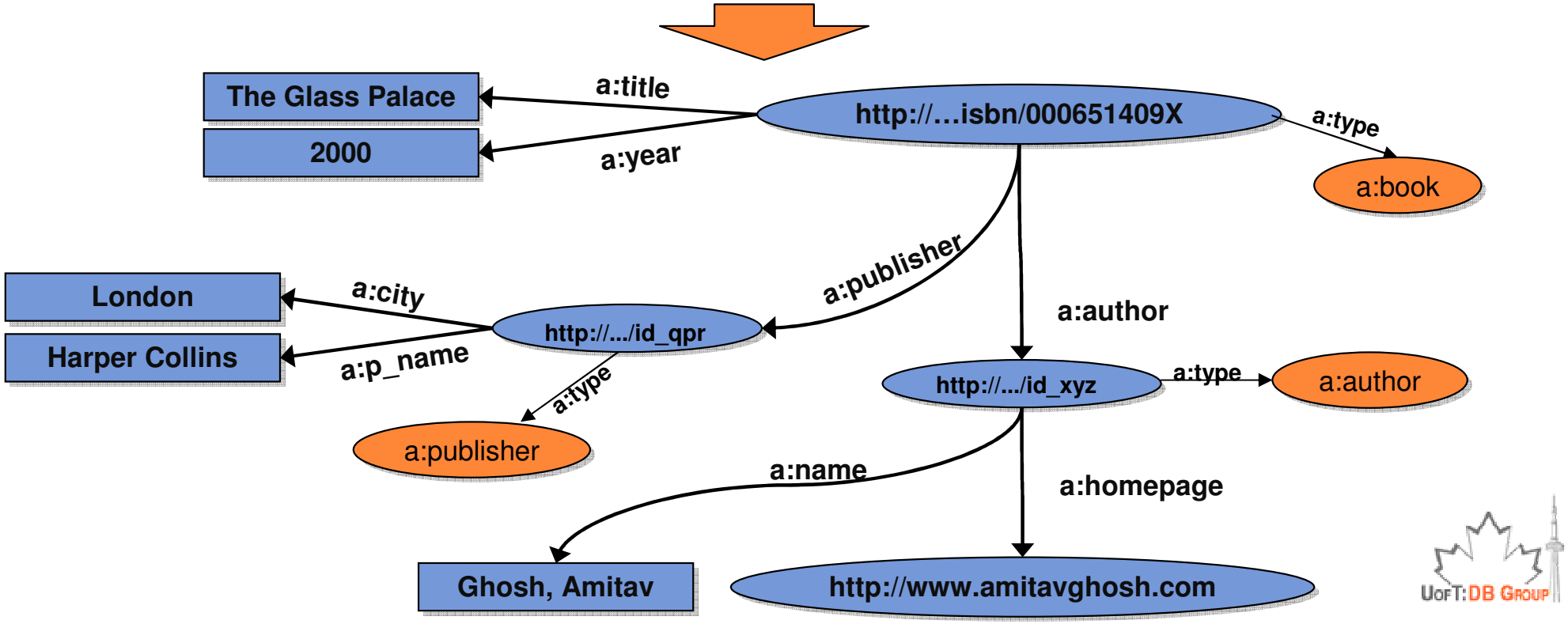
ID	Author	Title	Publisher	Year
ISBN 0-00-6511409-X	id_xyz	The Glass Palace	id_qpr	2000

Author

ID	Name	Homepage
id_xyz	Ghosh, Amitav	http://www.amitavghosh.com

Publisher

ID	Publisher's name	City
id_qpr	Harper Collins	London



Example from slides by Ivan Herman, available at http://www.w3.org/2010/Talks/0622-SemTech-IH/

RDB2RDF Process

- Mapping the relational schema to a custom/existing ontology/vocabulary
  - Identifying "things" (or *object types*)
    - E.g., "book", "author", "publisher"
      - Here, each relational table is mapped to an object type
  - Identifying predicates
    - E.g., "a:author", "a:name", "a:homepage"
- Creating RDF using the mapping
  - Creating instances (or objects), assigning unique IDs (or URIs)
    - E.g., each record in table "Book" is mapped to an object of type "a:book", assigned with a custom URI ending with the ISBN of the book (primary key of the table)
  - Can be performed once in an offline process, or on-the-fly in an online fashion
- Managing the output RDF data
  - Providing efficient translation process & SPARQL query processing capability

Comparison of RDB2RDF Approaches

- Different mapping approaches can be used
  - Mappings Creation
    - Automatic (table-to-class) or Manual/Semi-automatic (domain semantics-driven)
  - Mapping Representation & Accessibility
    - Representation language & access mechanism
  - Mapping Implementation
    - Static or Dynamic
  - Query Implementation
    - SPARQL => RDF or SPARQL => SQL => RDB
  - Application Domain
    - Generic or domain-specific
  - Data Integration
    - The ability to integrate data from multiple sources

The above aspects are components of the comparison framework provided by W3C's RDB2RDF Incubator Group's survey, January 2009

Following Linked Data Principles & Guidelines

Following Linked Data Guidlines

- Remember the guidelines
  - Choose “cool” HTTP URIs
  - Reuse, don't reinvent; Mix liberally
  - Link to other data sets

Following Linked Data Guidlines

- Choose "cool" HTTP URIs
  - http://dbpedia.org/resource/Berlin vs. http://www4.wiwiss.fu-berlin.de:2020/demos/dbpedia/cgi-bin/resources.php?id=Berlin
- Reuse, don't reinvent; Mix liberally
  - foaf:name vs. a:p_name, foaf:homepage vs. a:homepage, rdf:type vs. a:type
  - How to find existing vocabulary terms?
    - Look at similar data sets
      - Search sindice.com
    - Use UMBEL Subject Finder
- Link objects (instances) to other data sets
  - Use owl:sameAs & rdfs:seeAlso predicates to link to other linked data sources with "the same" or "related" objects; Use foaf:page to link to other HTML pages about the object
  - Challenge: How to find "the same" or "related" instances on the (Linked Data) Web?

Discovering Links to Existing Sources

- Linking Approaches
  - Common Key Matching
    - Matching based on common keys
      - E.g. matching ISBN numbers of the books, or Wikipedia Article IDs
    - Matching locations based on geographic coordinates
  - Label Matching
    - Comparing labels using string similarity measures
      - E.g., object/page with title/label "The Shining (film)" on DBpedia/Wikipedia is the same as movie object with title "The Shining" on LinkedMDB
    - Comparing labels using semantic similarity measures
      - E.g., "UofT" is the same "University of Toronto", or a drug named "Tylenol" is the same another drug "Acetaminophen" (scientific name of brand name Tylenol)
  - Graph/Ontology Matching
    - Compare labels, schema elements (e.g., types), and related objects (e.g., matching papers if they have the same set of authors)

Link Discovery over Relational Data

Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org

Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

Patient Visits (PV)

Visit	Diagnosis	Therapy	Location
VID777	Thalassaemia	Prescription: Hydroxyurea	Westchester Medical Center

PubMed (PM)

ID	Title
14988152	Complications of beta-thalassemia major in North America

Wikipedia/DBpedia Articles (DP)

URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

Link Discovery over Relational Data

Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org

Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

sameAs

Patient Visits (PV)

Visit	Diagnosis	Therapy	Location
VID777	Thalassaemia	Prescription: Hydroxyurea	Westchester Medical Center

sameAs

PubMed (PM)

ID	Title
14988152	Complications of beta-thalassemia major in North America

sameAs

Wikipedia/DBpedia Articles (DP)

URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

Link Discovery over Relational Data

Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org

Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

isATypeOf

sameAs

sameAs

Patient Visits (PV)

Visit	Diagnosis	Therapy	ocation
VID777	Thalassaemia	Prescription: Hydroxyurea	Westchester Medical Center

PubMed (PM)

ID	Title
14988152	Complications of beta-thalassemia major in North America

sameAs

Wikipedia/DBpedia Articles (DP)

URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

UofT:DB Group

Link Discovery over Relational Data

Clinical Trials (CT) from ClinicalTrials.gov/LinkedCT.org

Trial	Condition	Intervention	Location	Reference
NCT00336362	Beta-Thalassemia	Drug: Hydroxyurea	Columbia University	PubMed ID: 14988152
NCT00579111	Hematologic Diseases	Drug: Campath	Texas Children's Hospital	PubMed ID: 3058228

isATypeOf

isCloseTo

Patient Visits (PV)

Visit	Diagnosis	Therapy	Location
VID777	Thalassaemia	Prescription: Hydroxyurea	Westchester Medical Center

PubMed (PM)

ID	Title
14988152	Complications of beta-thalassemia major in North America

sameAs

Wikipedia/DBpedia Articles (DP)

URI	Title	Category
http://en.wikipedia.org/wiki/Thalassemia	Thalassemia	Blood_disorders
http://en.wikipedia.org/wiki/Hydroxyurea	Hydroxyurea	Chemotherapeutic_agents
http://en.wikipedia.org/wiki/Alemtuzumab	Alemtuzumab	Cancer_treatments

UofT: DB Group

Link Discovery over Relational Data

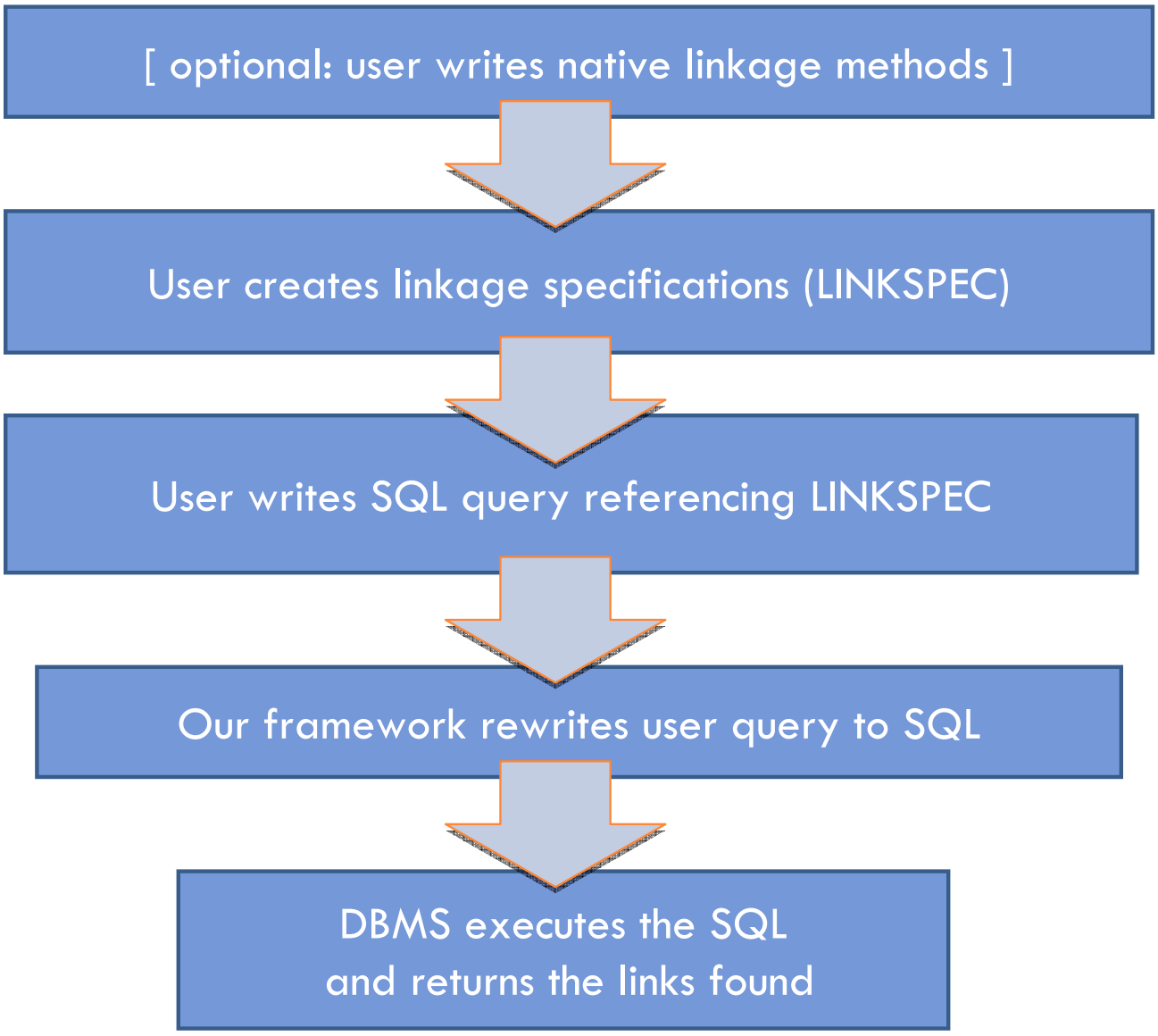
- Major Challenges
  - What types of links can be found?
    - Based on:
      - String errors or differences
      - Semantic relationship or equivalence
      - Both string errors and semantic equivalence
  - How to specify the linkage requirements
    - Easy to use and generic, applicable to various domains
  - How to find the links with the specified requirements
    - Implementation algorithms
      - Easy to adopt in existing data sources
    - Efficiency
      - How to compute string/similarity scores between all source and target records

Our Solution: LinQuer

- Generic, extensible and easy-to-use toolkit for linkage
  - Linkage Specification Language
    - LinQL: an SQL-like language for specification of requirements
      - Simple, easy to use, and extensible
  - SQL Implementation
    - LinQL is translated into standard SQL queries
      - Ease of use and applicability to existing relational data sources

http://dblab.cs.toronto.edu/project/linquer/

LinQuer Framework Overview



LinQuer Framework Overview

[optional: user writes native linkage methods]

User creates linkage specifications (LINKSPEC)

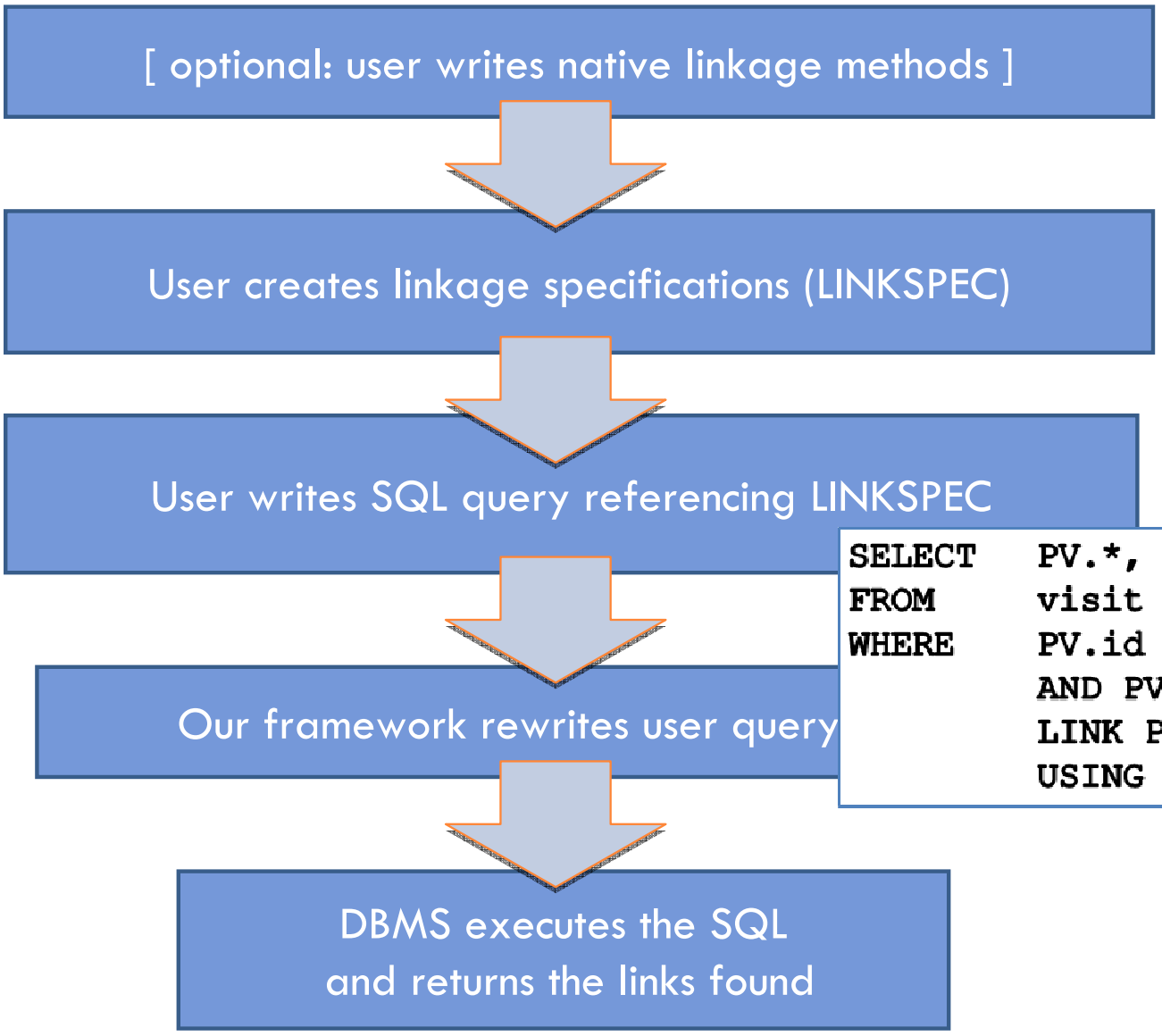
```
CREATE LINKSPEC myJaccard1
AS weightedJaccard (0.5, 2, 50)
```

User writes SQL query referencing LINKSPEC

Our framework rewrites user query to SQL

DBMS executes the SQL
and returns the links found

LinQuer Framework Overview




```
SELECT   PV.*, CT.*
FROM     visit PV, trial CT
WHERE    PV.id = 1234 AND CT.city="New York"
         AND PV.diagnosis = CT.condition
         LINK PV.diagnosis WITH CT.condition
         USING myJaccard1 LINKLIMIT 10
```

LinQuer Framework Overview

[optional: user writes native linkage methods]

User creates linkage specifications (LINKSPEC)

User writes SQL query referencing

Our framework rewrites user que

DBMS executes the SQL
and returns the links found

```
WITH    scores(tid1, tid2, score) AS (
        SELECT tid1, tid2,
               (SI.sinter/(BSUMW.sumw+QSUMW.sumw-SI.sinter))
               AS score
        FROM   (SELECT  BTW.tid AS tid1,QT.tid AS tid2,
                        SUM(BTW.weight) AS sinter
                FROM    (SELECT * FROM visit_diagnosis_weights
                        WHERE id = 1234) AS BTW,
                        trials_condition_tokens AS QT
                WHERE   BTW.token = QT.token
                GROUP BY BTW.tid, QT.tid) AS SI,
                (SELECT  *
                FROM    visit_diagnosis_sumweights
                WHERE   id = 1234 ) AS BSUMW,
                (SELECT  Q.tid, SUM(BTW.weight) AS sumw
                FROM    trials_condition_tokens Q,
                        visit_diagnosis_tokenweights AS BTW
                WHERE   Q.token = BTW.token
                GROUP BY Q.tid ) AS QSUMW
        WHERE  BSUMW.tid=SI.tid1 and SI.tid2 = QSUMW.tid )
SELECT PV.*, CT.*, s.score
FROM   visit AS PV, trials AS CT, scores AS S
WHERE  PV.id = 1234 AND CT.city='NEW YORK' AND
       S.tid1=PV.visitid AND S.tid2=CT.trial AND S.score>0.5
```

Tools for RDB2RDF Mapping and Linked Data Publication

UofT: DB Group

RDB2RDF Tools

- Several tools and frameworks exist, with different characteristics
  - W3C's RDB2RDF Incubator Group's survey contains a complete list of existing systems
- Some popular tools include
  - D2RQ and D2R Server
  - OpenLink Virtuoso's RDF Views
  - Triplify
- Some directly follow Linked Data principles
  - For those that only generate RDF, there are tools that can create Linked Data interfaces for SPARQL endpoints
    - E.g., Pubby http://www4.wiwiss.fu-berlin.de/pubby/

D2R Server

- Based on D2RQ
  - A declarative language to describe mappings between relational database schema and RDF-S/OWL ontologies
  - Providing RDF view over relational data
    - In: any (JDBC) database, Out: RDF (Jena API, SPARQL endpoint)
- Provides Linked Data view over relational sources
  - Following Linked Data principles
    - http://data.linkedmdb.org/resource/film/2014 redirects to:
      - http://data.linkedmdb.org/page/film/2014 in HTML browsers
      - http://data.linkedmdb.org/data/film/2014 in RDF browsers
    - RDF description contains all the predicates that have the URI as object or subject along with any metadata
    - HTML view shows a user-friendly view of the predicates
- All of these are done on-the-fly
  - Based on the D2RQ mapping specification file
- Semi-automatic mapping creation

Virtuoso RDF View

- RDB data represented as virtual RDF graphs without physical creation of RDF datasets
- RDF views are composed of *quad map patterns*
  - Define the mapping from a set of RDB columns to triples
  - Represented in the Virtuoso Meta-Schema Language (MSL), which also supports SPARQL-style notations
  - Manual creation + an additional tool for automatic Linked Data Generation & Deployment
- More powerful toolkit
  - But this means more training is required to be able to understand and use all the features of the system

Triplify

- A quick and easy way to produce and publish linked data
- Very lightweight
  - less than 500 lines of code, currently in PHP
- Based on a configuration file
  - More complex, containing SQL queries
  - Manual creation: the user needs to write the mapping from scratch
- Not very scalable
  - Currently aimed at small to medium web applications

Example Mapping Using D2R Server

D2R Server tutorial available at:
http://sw.cs.technion.ac.il/d2rq/tutorial

References

http://linkeddata.org

- Linked Data: Evolving the Web into a Global Data Space.
  By Tom Heath & Christian Bizer
  Available online at http://linkeddatabook.com/editions/1.0/
- LinkedData.org - http://linkeddata.org/
- Linking Open Data Project Wiki
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
- W3C's RDB2RDF Incubator Group
  http://www.w3.org/2005/Incubator/rdb2rdf/
- D2R Server & D2RQ's Documentation
  http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
  http://www4.wiwiss.fu-berlin.de/bizer/d2rq/spec/