MANCHESTER 1824
The University of Manchester

COMP38120
Workshop 10

Data on the Web

Riza Batista-Navarro
Sandra Sampaio
Goran Nenadic

MANCHESTER 1824
The University of Manchester

Next five workshops

- Linked Data principles and RDF
- Linked data design and engineering
- Publishing linked data
- Consuming and aggregating linked data
- Guest lecture(s)

The Web Today: a Web of Documents

- a global file system full of documents
- limitations
  - data is weakly structured
  - data not fully connected (data islands)

MANCHESTER 1824
The University of Manchester

The Web Today: Low degree of structure

- documents are written in **HTML**
  - meant to structure documents NOT data
  - data is intermingled with surrounding text
  - designed for humans
  - BUT hard for machines to understand the data

[source http://www.thefarside.com/]

[source http://www.thefarside.com/]

- What we say to Web agents

" Visit *** the syllabus *** *for further information on COMP38120. "*

- What they "hear"

" blah *** blah blah blah *** *blah blah blah blah blah blah "*

MANCHESTER 1824
The University of Manchester

The Web Today: Low degree of structure

- **microformats** (http://microformats.org)
  - HTML for publishing structured data on specific types of entities, e.g., people, events

```
<article class="h-recipe">
  <h1 class="p-name">Bagels</h1>

  <ul>
    <li class="p-ingredient">Flour</li>
    <li class="p-ingredient">Sugar</li>
    <li class="p-ingredient">Yeast</li>
  </ul>

  <p>Takes <time class="dt-duration" datetime="1H">1 hour</time>,
     serves <data class="p-yield" value="4">four people</data>.</p>

  <div class="e-instructions">
    <ol>
      <li>Start by mixing all the ingredients together.</li>
    </ol>
  </div>
</article>
```

  - embedded in webpages

MANCHESTER 1824
The University of Manchester

The Web Today: Low degree of structure

- **microformats**
  - only a limited number of entities can be represented

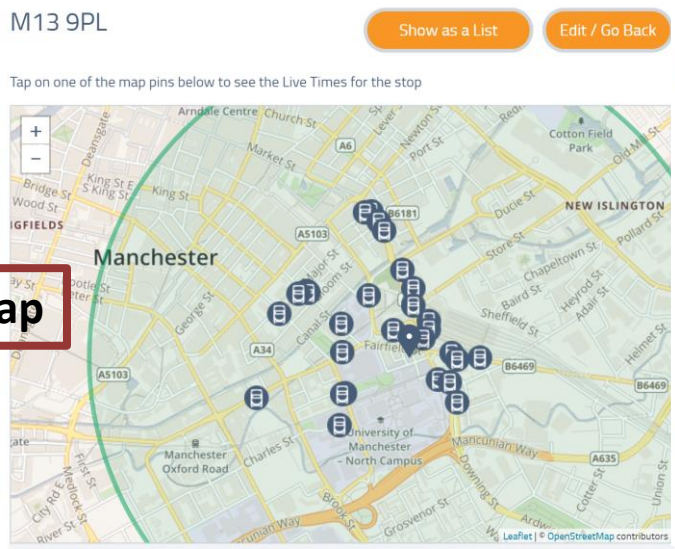
h-adr
h-card
h-entry
h-event
h-feed
h-geo
h-item
h-listing
h-product
h-recipe
h-resume
h-review
h-review-aggregate

  - limited support for linking between entities

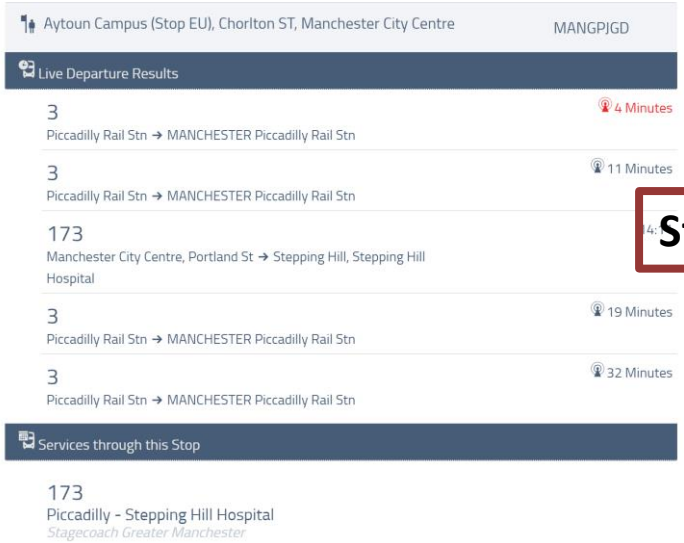
The Web Today: Low degree of structure

- **Web APIs**
  - query-based programmatic access to structured data over HTTP
  - enables creation of **mashups:** applications that combine data from multiple sources

traveline



OpenStreetMap



First Bus

Stagecoach

MANCHESTER 1824
The University of Manchester

The Web Today: Low degree of structure

- **Web APIs**
  - return highly structured data
  - each one is specialised
  - integration means having to conform with each API's specifications, writing custom code

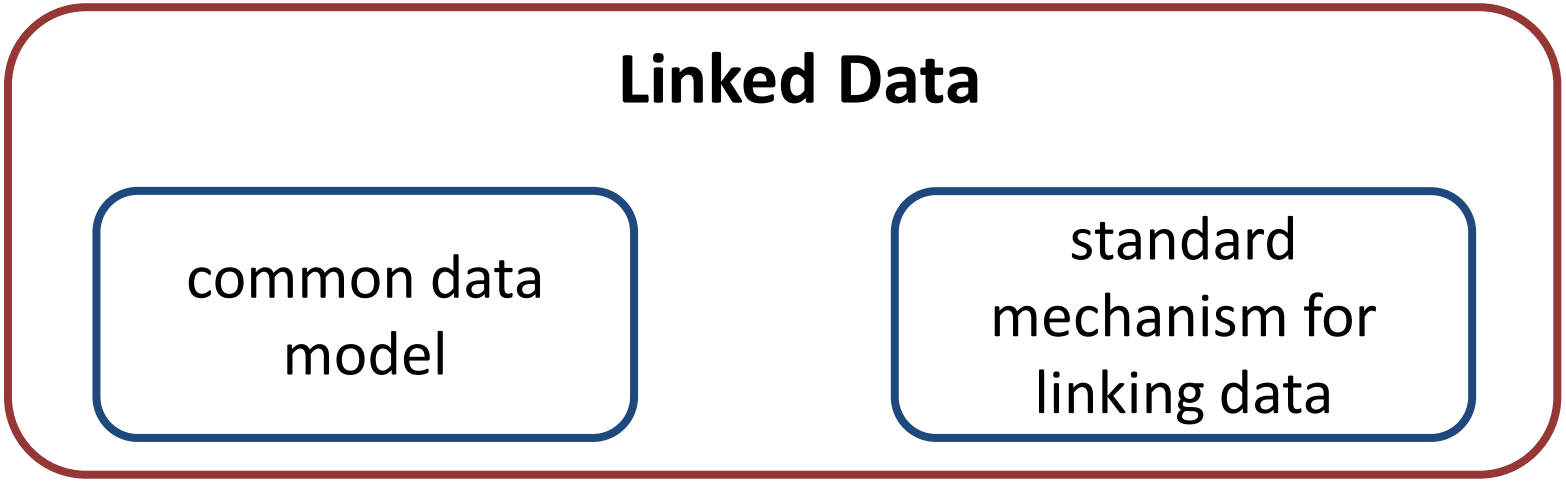
MANCHESTER 1824
The University of Manchester

The Web Today: Data Islands

- **Web APIs**
  - return highly structured data

    BUT not always with links for finding related data
  - use identifiers valid within a local context only
  - no standard mechanism to refer to entities across different data sets
    - leads to isolated fragments
    - unlike in HTML documents: anchor tag and href attribute can be followed by browsers and crawlers

What was needed

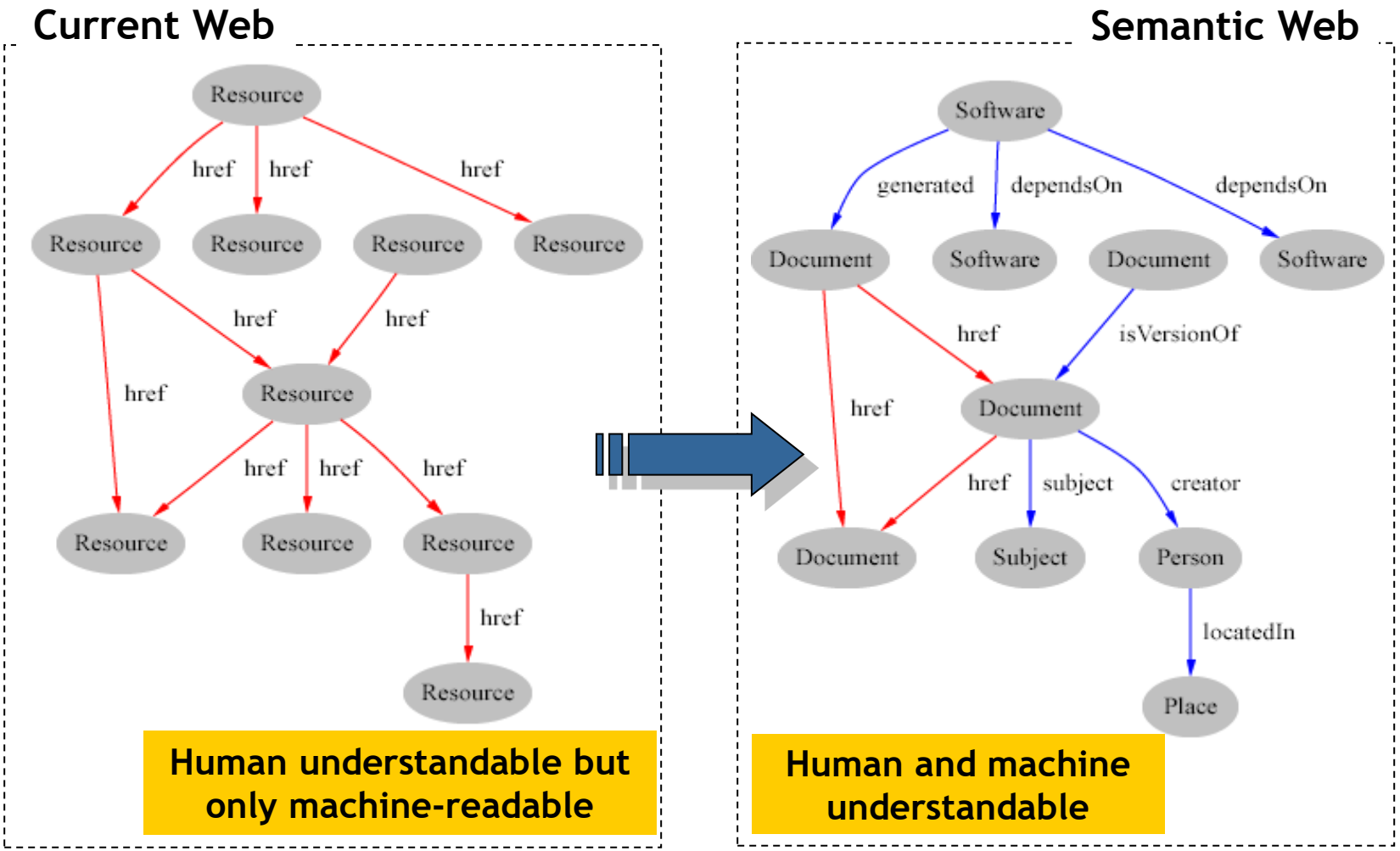


led to the creation of

Web of Data
(aka Semantic Web)

MANCHESTER 1824
The University of Manchester

The Semantic Web



MANCHESTER 1824
The University of Manchester

The Semantic Web

- What do we gain?
  - enable machines to understand data
  - maximised interconnections between data
  - reusability, reduced redundancy

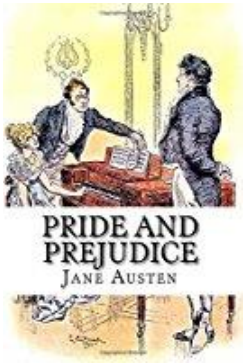
MANCHESTER 1824
The University of Manchester

How do we get there? Linked Data Principles

- based on the successful infrastructure behind the Web of Documents
  - globally unique addressing system: URIs
  - universal access mechanism: HTTP
  - common format for documents: HTML
  - hyperlinks between documents

Linked Data Principles

1. Use **Uniform Resource Identifiers (URIs)** as names for things (i.e., **resources**)
   - documents, real-world entities, abstract concepts
2. Use **HTTP URIs** so that they can be looked up (**dereferenced**) using a universal access mechanism



Jane Austen, 1813

https://isbndb.com/book/1547254742

Linked Data Principles

3. Use a common format, i.e., a **data model** – the **Resource Description Framework (RDF)**
4. Use **typed hyperlinks** to other URIs to connect things

URIs as names for (almost) every resource

- Individuals

  http://dbpedia.org/resource/Barack_Obama
- Types

  http://schema.org/Person
- Properties

  http://dbpedia.org/ontology/almaMater
- Values of properties

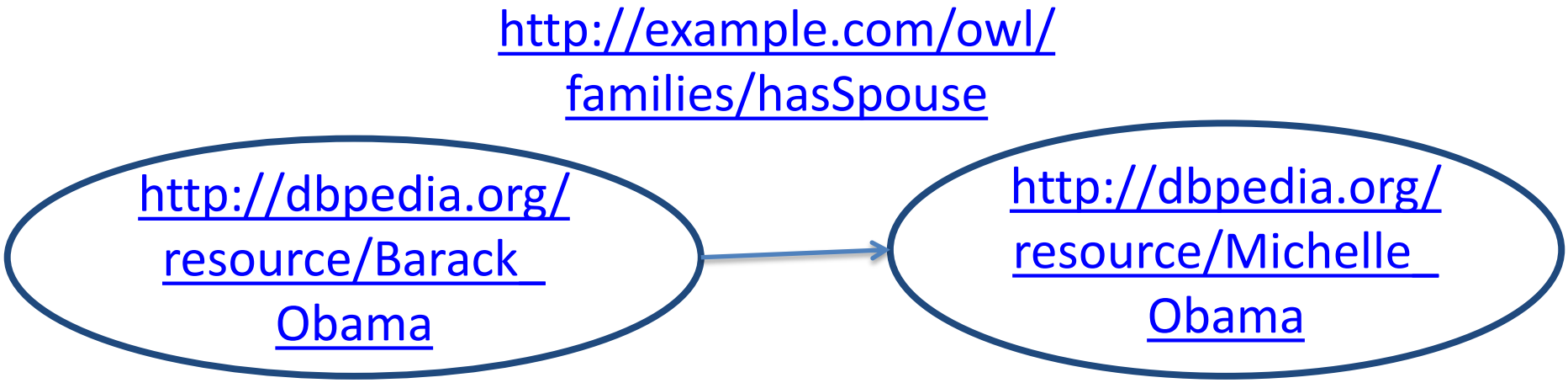
  http://dbpedia.org/resource/Harvard_Law_School
- Relation types

  http://example.com/owl/families/hasSpouse

URIs as names for (almost) every resource

- a single URI may define many different resources
- to identify a single **fragment**, we use the **hash** notation
  - e.g., http://example.org/index#person

URIs as names for (almost) every resource



Using HTTP URIs

- advantages
  - decentralised creation of globally unique names
  - not just a name but also allows access to information describing the entity (as long as **dereferenceable**)

Using HTTP URIs

- **dereferenceable** HTTP URIs
  - web clients can look up the URI using HTTP and retrieve a description of the resource identified by the URI
  - descriptions: embodied in the form of Web documents
    - for humans: HTML
    - for machines: RDF
  - Note:
    - the resource description is not the same as the resource itself
    - the URI of the Web document containing the description is different from the URI of the resource

The RDF Data Model

- First, recall relational database tables

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

The RDF Data Model

- rows represent things

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

The RDF Data Model

- columns represent properties

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

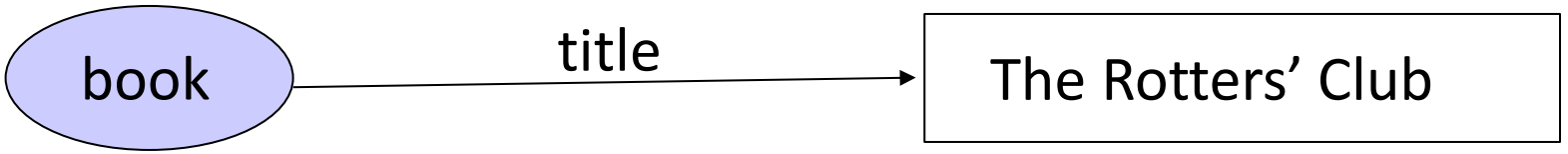
The RDF Data Model

- intersections represent properties of things

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

The RDF Data Model

- graph-based representation



more generally:

triple = (subject, ~~property~~ predicate, ~~value~~ object)
where object can be either a literal value OR another thing (URI)

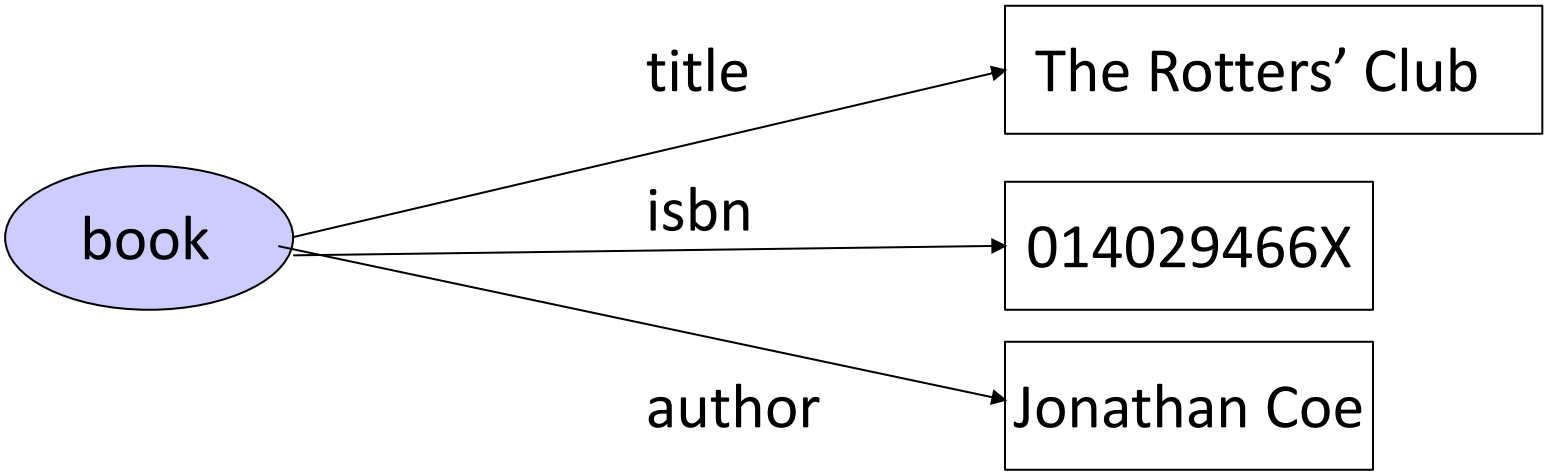
The RDF Data Model

- selecting multiple properties

isbn	title	author	publisherId	pages
0743267478	Q&A	Vikas Swarup	1435	336
014029466X	The Rotters' Club	Jonathan Coe	1546	416
…	…	…	…	…
..	…	…	…	…

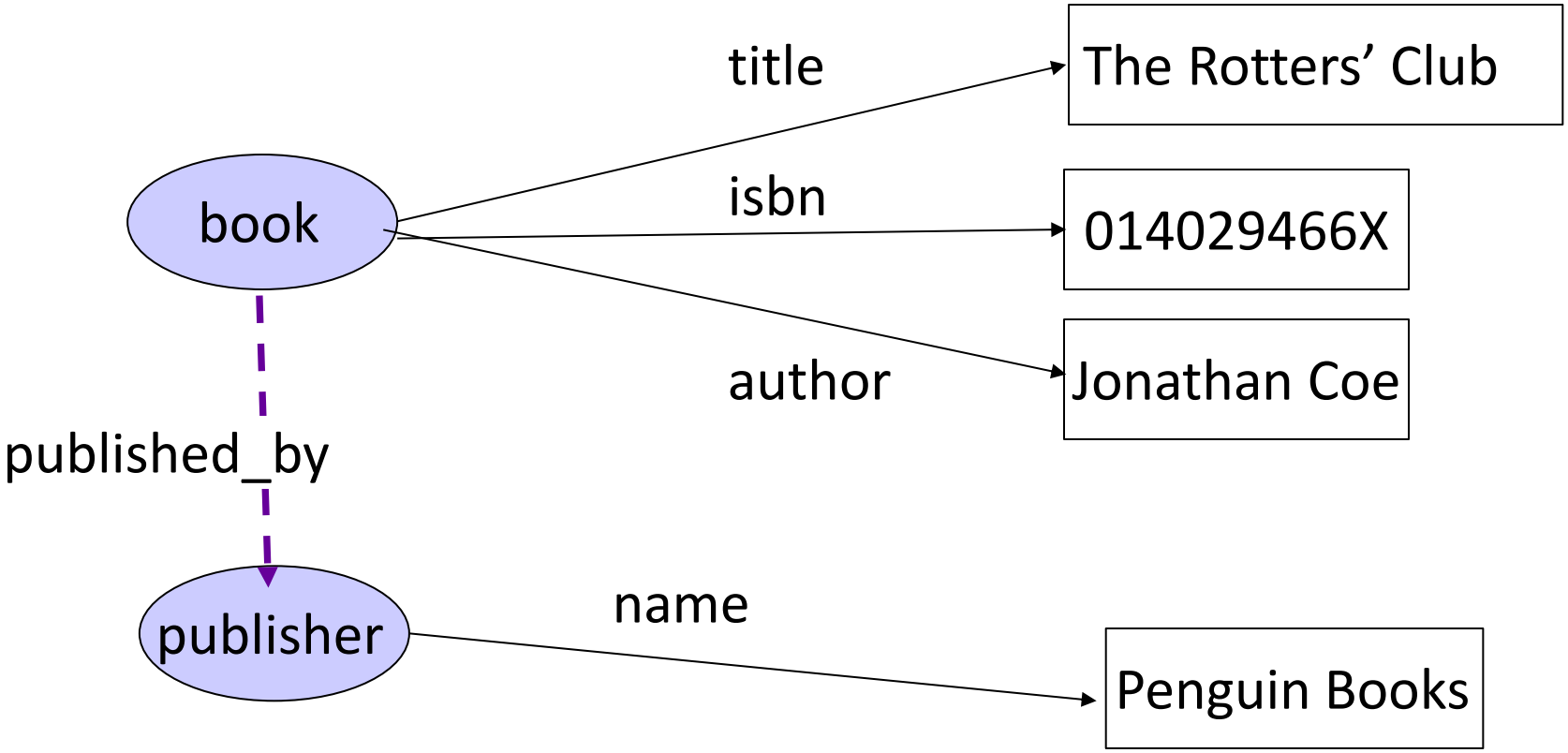
The RDF Data Model

- multiple properties in a graph



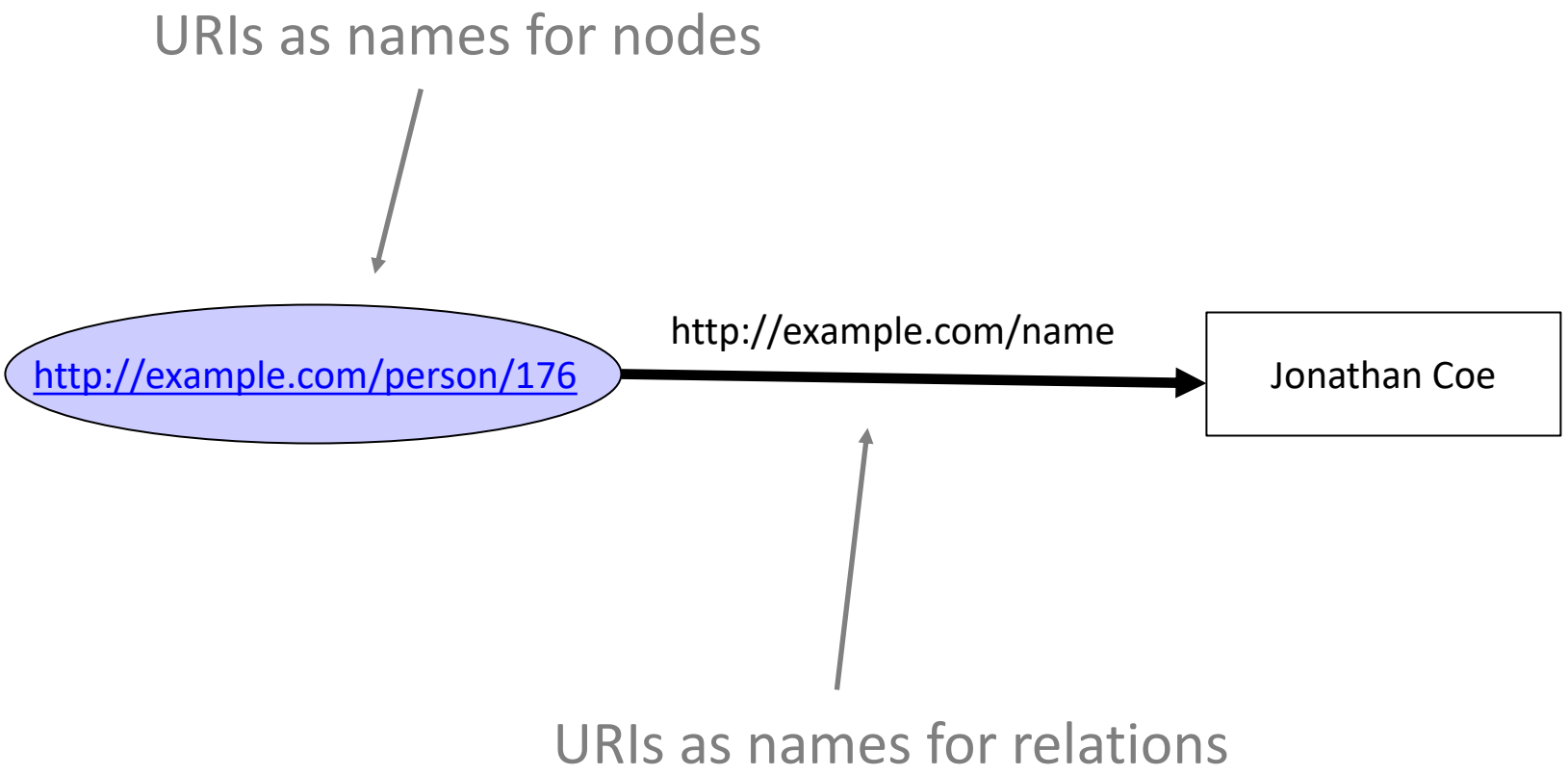
The RDF Data Model

- relations between things



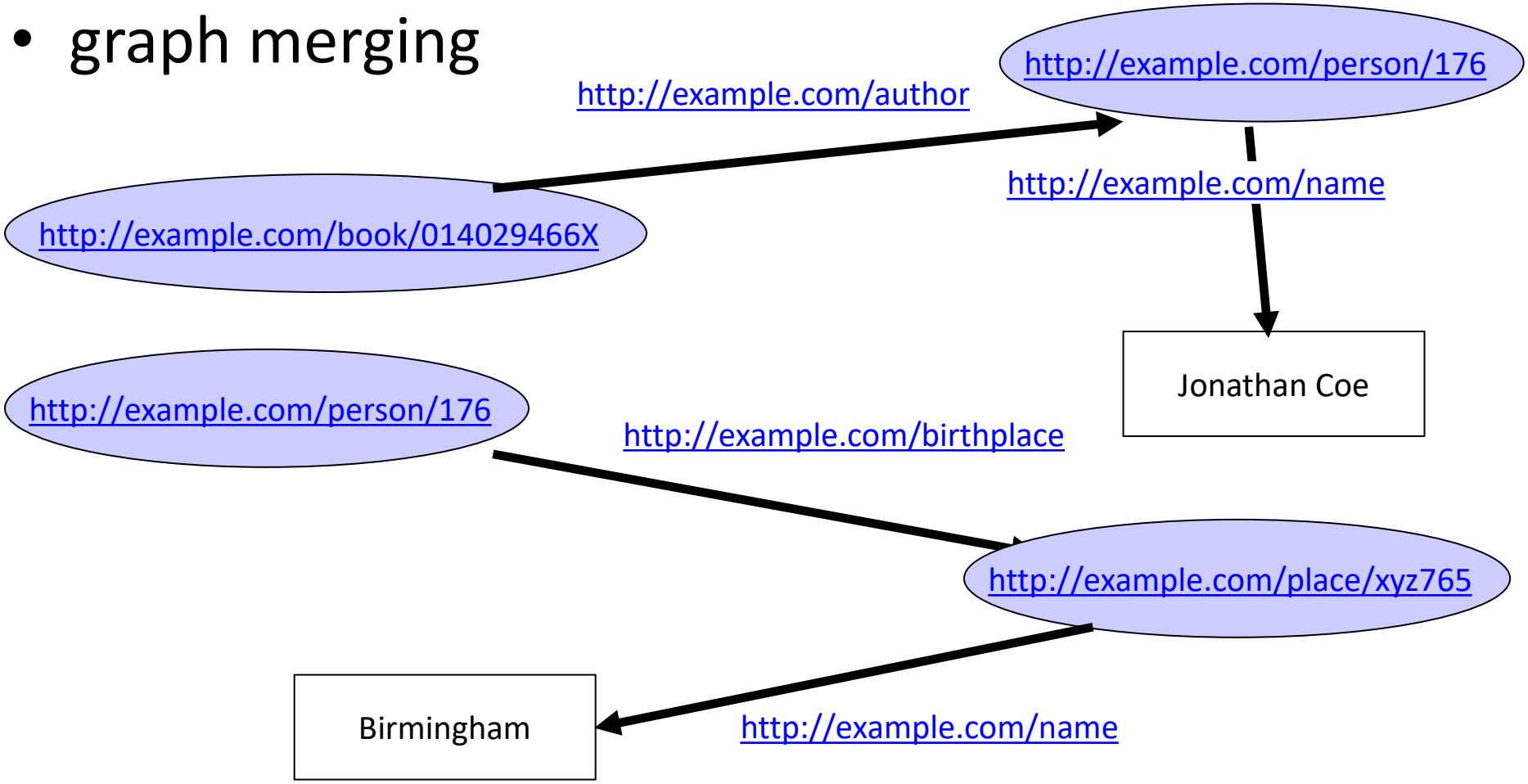
The RDF Data Model

- using URIs



The RDF Data Model

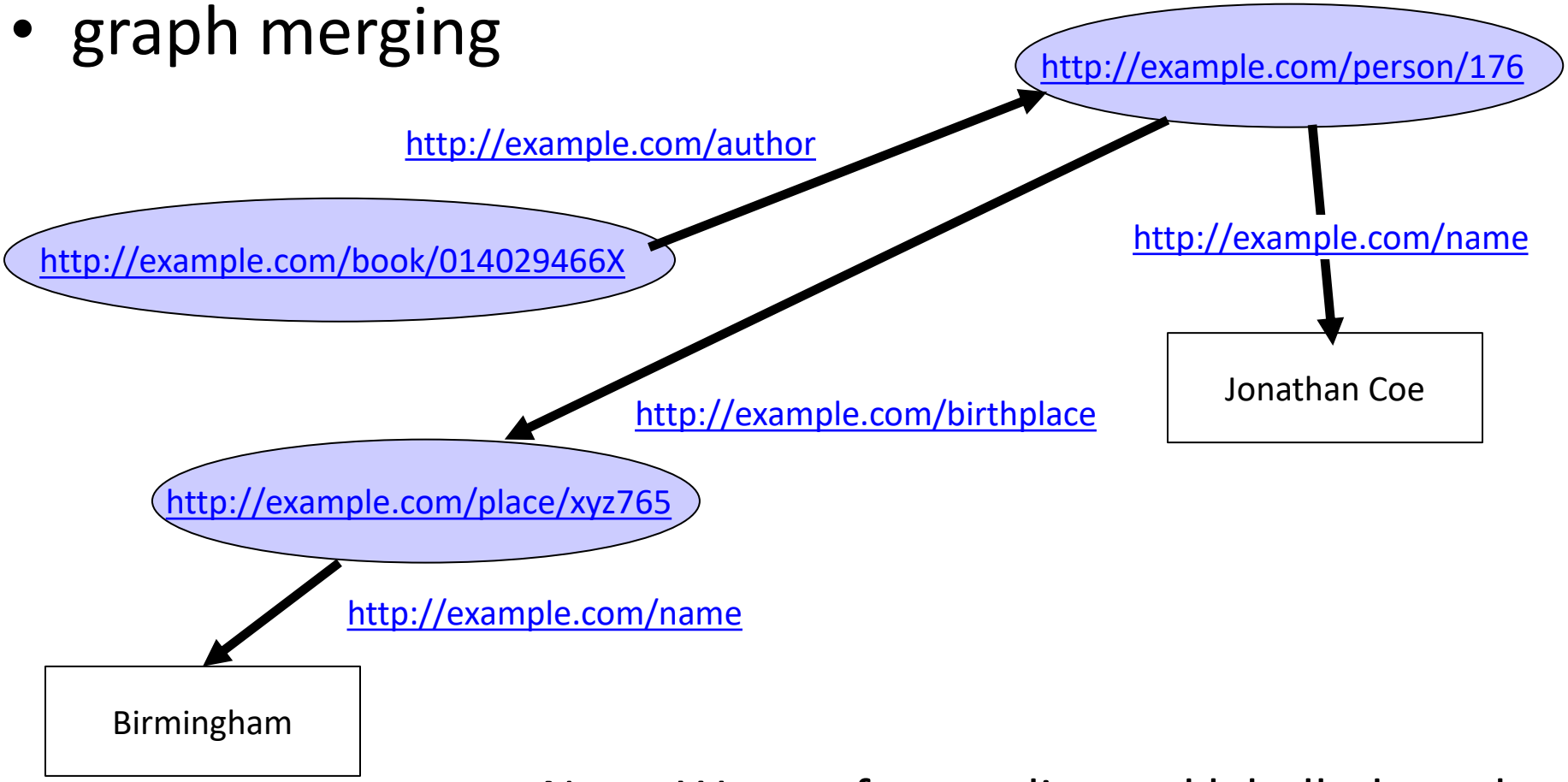
- graph merging



Note: the subject of one statement can be the object of another.

The RDF Data Model

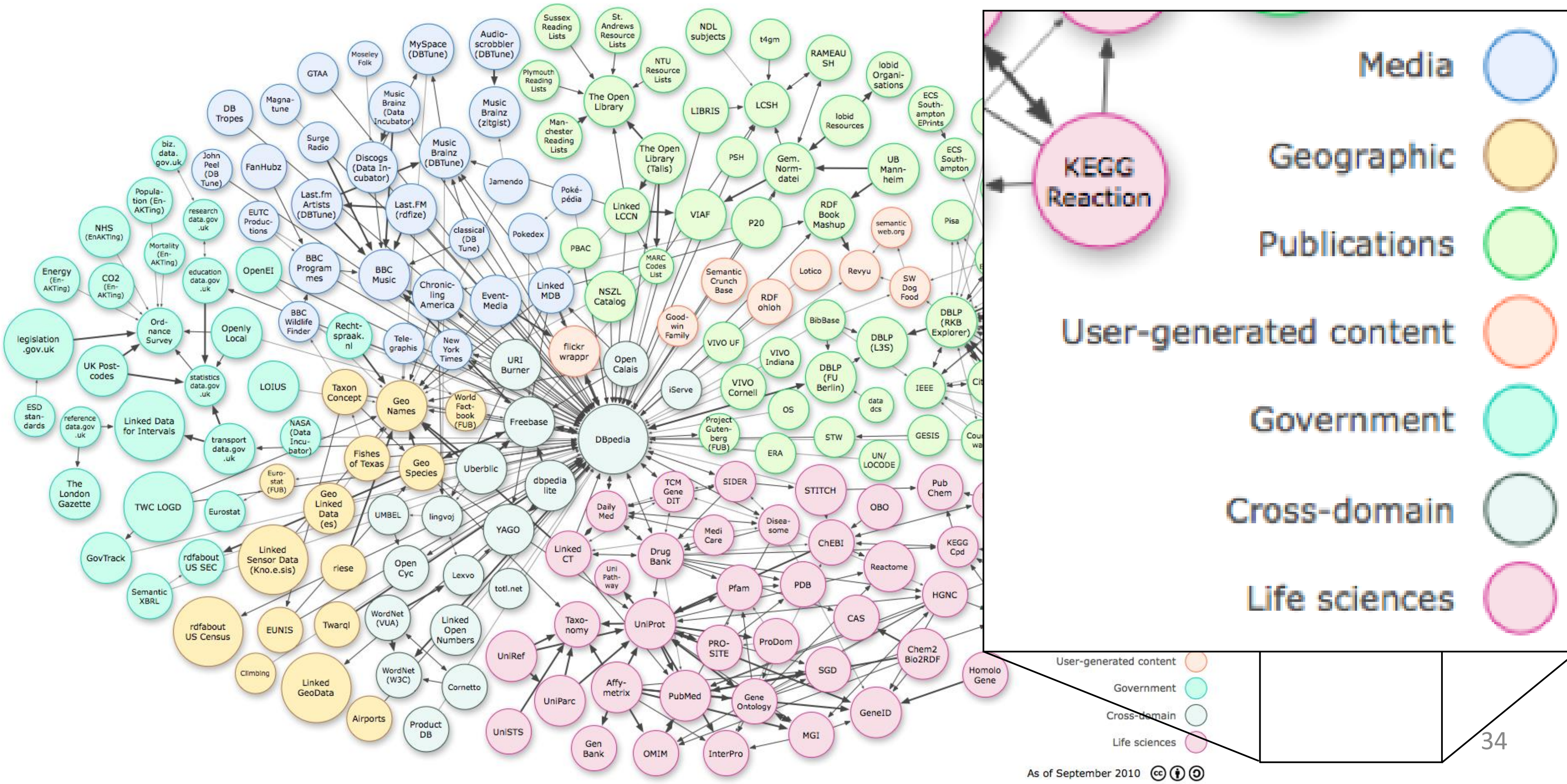
- graph merging



Note: We can form a directed labelled graph.

The University of Manchester

Are we there yet?
How big is the Semantic Web right now?



MANCHESTER 1824
The University of Manchester

Are we there yet?
How big is the Semantic Web right now?

- Task 1: Exploring the coverage and content of the Linked Data space (5 minutes)

RDF Data Model

- a W3C recommendation (https://www.w3.org/RDF)
- a formal specification of concepts, terms, and relationships

The RDF Data Model: Types of triples

- Literal Triples
  - object is a string, number or a date
  - describe properties of resources, e.g., birthdate
  - types
    - **plain**: a string that sometimes comes with a language tag

      `"any text", "How are you?"@en-GB`
    - **typed**: a string combined with a datatype URI which identifies the datatype of the literal (according to the XML Schema)

      `"2001-10-26+02:00"^^xsd:date,`
      `"hello"^^xsd:string,"1"^^xsd:integer`

MANCHESTER 1824
The University of Manchester

The RDF Data Model: Types of triples

- Literal Triples
  - can only be the object of a triple

```
⟨<http://example.org/#john>,
      <http://example.org/#hasName>,
            "John Smith"^^xsd:string⟩
```

MANCHESTER 1824
The University of Manchester

The RDF Data Model: Types of triples

- RDF Links
  - describe relationship between two resources
  - consist of 3 URIs, one for each of subject, predicate and object

```
⟨<http://example.com/book/014029466X>,
      <http://example.com/hasAuthor>,
            <http://example.com/person/176>⟩
```

MANCHESTER 1824
The University of Manchester

The RDF Data Model: Types of triples

- RDF Links can be either…
  - internal
    - connect resources within the same Linked Data source
    - subject and object URIs are in the same namespace
  - external
    - connect resources served by different Linked Data sources
    - subject and object URIs are in different namespaces
    - crucial for the Semantic Web; the glue that connects data islands

Benefits of the RDF Data Model

1. Global; enables anybody to refer to anything thanks to HTTP URIs.
2. Additional information on a resource can be retrieved by means of dereferencing.
3. Different data sources can be connected with RDF Links.
4. Information from disparate sources can be combined by merging triples into one graph.
5. Information expressed using different schemata can be represented in one graph.

MANCHESTER 1824
The University of Manchester

Connecting disparate data sources using RDF Links

- Relationship links
- Identity links
- Vocabulary links

MANCHESTER 1824
The University of Manchester

Connecting data sources using RDF Links

- Relationship links: point at related things in other data sources

```
⟨<http://dbpedia.org/resource/Gemma_Atkinson>,
      <http://xmlns.com/foaf/0.1/based_near>,
            <http://sws.geonames.org/2643123>⟩
```

Connecting data sources using RDF Links

- Identity links
  - point at **URI aliases** used by other data sources to identify the same resource
  - enables retrieving more descriptions
  - uses the **`owl:sameAs`** predicate from the Web Ontology Language (OWL)

```
⟨<http://dbpedia.org/resource/Manchester>,
    <http://www.w3.org/2002/07/owl#sameAs>,
        <http://sws.geonames.org/2643123>⟩
```

More on URI aliases

- multiple URIs identifying the same entity
- Why is this desirable?
  - variety of opinions: to allow different views on a resource to be expressed
  - traceability: to allow users of Linked Data to find a particular publisher's view on a resource
  - decentralisation: to eliminate the need for one authority to assign URIs; no single point of failure

Connecting data sources using RDF Links

- Vocabulary links
  - serve as a bridge between the schemata used by different data sources
  - predicates: `owl:equivalentClass`, `owl:equivalentProperty`, `rdfs:subClassOf`, `rdfs:subPropertyOf`, `skos:broadMatch`, `skos:narrowMatch`

⟨**<http://example.british.namespace/Lecturer>**,
<http://www.w3.org/2002/07/owl#equivalentClass>,
<http://example.american.namespace/AsstProf>⟩

RDF Serialisation

- How do we write RDF/publish RDF graphs?
- using serialisation formats
  - RDF/XML
  - N-Triples
  - Turtle

RDF/XML

- `Description` element describes a resource
- `about` attribute names the resource
- properties (RDF predicates) are represented as nested elements inside a `Description`
  - names of nested elements are property URIs
- object can be either a literal or a URI specified using the `resource` attribute

RDF/XML example


```
<Description about="some.uri/person/sean_bechhofer">
   <hasName>Sean K. Bechhofer</hasName>
   <hasColleague
         resource="some.uri/person/uli_sattler"/>
</Description>

<Description about="some.uri/person/uli_sattler">
   <o:hasHomePage>http://www.cs.mam.ac.uk/~sattler
   </o:hasHomePage>
</Description>

<Description about="some.uri/person/carole_goble">
  <o:hasColleague
         resource="some.uri/person/uli_sattler"/>
</Description>
```

MANCHESTER 1824
The University of Manchester

RDF/XML

- Task 2: XML Basics (5 minutes)
- Task 3: Understanding RDF statements (5 minutes)

N-Triples

- simple line-based, plain-text serialisation
- full URIs are enclosed in angle brackets (<>) and full stop at the end of the line signals end of the triple

```
<http://example.org/bob#me>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
          <http://xmlns.com/foaf/0.1/Person> .
<http://www.wikidata.org/entity/Q12418>
                <http://purl.org/dc/terms/title>
                                    "Mona Lisa" .
```

- often used for exchanging large amounts of RDF data that do not fit into memory (as they can be parsed one line at a time)

Turtle

- Terse RDF Triple Language
- an extension of N-Triples
- supports namespace prefixes
- often used for writing RDF triples by hand

Turtle example


```
BASE <http://example.org/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX schema: <http://schema.org/>
PREFIX wd: http://www.wikidata.org/entity/

<bob#me>
    a foaf:Person ;
    foaf:knows <alice#me> ;
    schema:birthDate "1990-07-04"^^xsd:date ;
    foaf:topic_interest   wd:Q12418 .
```

MANCHESTER 1824
The University of Manchester

Turtle example


```
BASE <http://example.org/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX schema: <http://schema.org/>
PREFIX wd: http://www.wikidata.org/entity/

<bob#me>
    a foaf:Person ;
    foaf:knows <alice#me> ;
    schema:birthDate "1990-07-04"^^xsd:date ;
    foaf:topic_interest wd:Q12418 .
```

<http://example.org/bob#me>

rdf:type

<http://schema.org/birthDate>

MANCHESTER 1824
The University of Manchester

To summarise…

- Linked Data
  - a set of best practices for publishing and interlinking structured data on the Web
  - applies the same infrastructure used by the Web of Documents
  - enables Web of Data (aka Semantic Web) which is understandable by both humans and machines
  - underpinned by a graph-based model: RDF

To summarise…

- RDF
  - uses triples: <subject, predicate, object>
    - subject: resource
    - predicate: property
    - object: resource or literal
  - uses HTTP URIs
  - allows for merging of triples into one graph

MANCHESTER 1824
The University of Manchester

Your own summary

Three of the most important things you learnt during today's workshop:

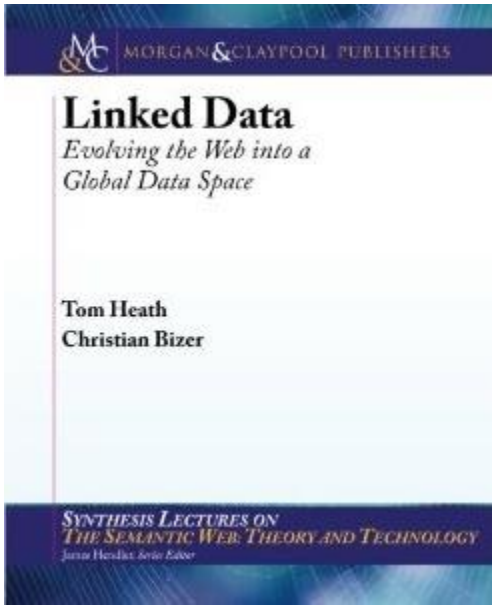
1.

2.

3.

Further reading (and listening)

- Tom Heath and Christian Bizer (2011) **Linked Data: Evolving the Web into a Global Data Space** (1st edition). Morgan & Claypool.
  http://linkeddatabook.com/editions/1.0/



- Tim Berners-Lee's 2009 TED Talk: **Introduction to Linked Open Data**

http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

Creating RDF data

- Task 4: Building some linked data (5 minutes)
- Friend of a Friend (FOAF)
  - "machine-readable vocabulary (ontology) describing people, their activities and relations to other people and objects"
  - uses RDF and OWL
  - http://www.foaf-project.org/docs

MANCHESTER 1824
The University of Manchester

Acknowledgments

These slides are partially based on:

- **Linked Data and RDF (COMP60421 slides)** by Sean Bechhofer

  http://studentnet.cs.manchester.ac.uk/pgt/2013/COMP60421/

- **An Introduction to the Semantic Web for GIS Practitioners** by Emanuele Della Valle

  http://applied-semantic-web.org/slides/2011/05/SemanticWeb4GIS.ppt

- **An Introduction to Linked Data**, Tom Heath

  http://tomheath.com/slides/2009-02-austin-linkeddata-tutorial.pdf