The British National Bibliography
Who Uses Our Linked Data?

Corine Deliot, Neil Wilson
The British Library

Luca Costabello, Pierre-Yves Vandenbussche
Fujitsu Research, Ireland

SWIB-2016, Bonn - 30 November 2016
Overview

• **Context:**
  - The British Library Metadata Services
  - The British National Bibliography (BNB)
  - The Linked Open BNB
  - Linked Open Data: Some Challenges

• **The project:**
  - The British Library/Fujitsu collaboration
  - The RDF analytics platform

• **What we learnt about usage of the BNB**

• **The value of RDF Analytics**
The British Library Act records our role as “national centre for… bibliographical & other information services”

British Library Metadata Services

- Originally offered priced services & evolved through many technologies
- Began to offer open data in 2010 & Linked Open Data in 2011
- Collection Metadata Strategy published in 2015
The British National Bibliography

3.7m entries for UK and Republic of Ireland publications on all subjects in all languages, 1950-to-date

- Reusable publication dataset - *not a unique institutional catalogue*
- Permissive Licence – *CC0*
- Includes: *People, Places, Dates, Subjects*
- Consistent - *over 60 years*
The Linked Open BNB

- Datasets – Books & Serials & VoID descriptions accessible at:
  - BNB Linked data platform: [http://bnb.data.bl.uk](http://bnb.data.bl.uk)
  - SPARQL endpoint: [http://bnb.data.bl.uk/sparql](http://bnb.data.bl.uk/sparql)
  - SPARQL editor: [http://bnb.data.bl.uk/flint-sparql](http://bnb.data.bl.uk/flint-sparql)
  - Bulk downloads [http://www.bl.uk/bibliographic/download.html](http://www.bl.uk/bibliographic/download.html)
    - Serializations available: RDF/XML, N-Triples
  - Updated monthly

Usage terms: [http://creativecommons.org/licenses/by-sa/3.0/](http://creativecommons.org/licenses/by-sa/3.0/)"
Linked Open Data: Some Challenges

• Scarce resources (human & financial)
  ➢ *is it worth continuing to provide the service?*
  ➢ *where best to focus our efforts?*

• Limited user feedback
  ➢ *Who uses our data and what for?*
  ➢ *How can we best support those users?*

• Lack of linked data-specific analytics tools
Current Monitoring of BNB Data Uses

Statistics:

- e.g. Number of hits on the SPARQL endpoint
- e.g. Number of downloads on the British Library webpage
- e.g. Basic web logs analysis reports
Current Monitoring of BNB Data Uses

BNB data used in pilot projects
- *e.g.* Linked Open BNB data used as test data for a semantic search demonstrator.
- *e.g.* data provided to Microsoft to assist in their research into linking structured data

BNB data used in tutorials
British Library - Fujitsu collaboration

Metadata Publication as Linked Data

• Who is using our data?
• Which data?
• How to optimise our publication?

Linked Data Analytics Solution

• 10+ years experience in Linked Data
• Scalable & innovative analytics
Distinctive Features

• SPARQL-specific metrics

• Fine-grained analytics for each category of RDF resource (instances, classes, properties & graphs)

• Native support for RDF dereferencing (303 pattern)

• Visitor session detection

• SPARQL queries complexity classification (light/heavy)

• Human vs Machine classification
System Overview

Data Publisher Logs → ETL (Log Ingestion, Filters) → Data Warehouse → MOLAP Unit → Web UI

Metrics Extraction
Analytics for Linked Data Publishers

Overview

Content
- Requests Count
- Response Code

Audience
- Location
- User Agent
- Visitors
- Sessions

Behaviour
- Data Access

Request Count (All Time)

252.8K
Requests

Request Count (Selected Time Frame)

8.0K
Requests

Number Of Requests Between 2015-04-23 And 2015-04-30

Number of requests

Map of continents

www.bl.uk
What Did We Learn?

252 K out of 44 M requests (13 months) were kept.

- Overall request flow is stable

April 2014 (first complete month)
  - ~18k Requests (HTTP + SPARQL)

April 2016
  - ~24k Requests (HTTP + SPARQL)

- New users keep coming in

- Bounce rate: 48%
Instances, Classes & Properties

In the top 5 instances:

- The hobbit / J. R. R. Tolkien
  - 6,092 requests
  - 2,115

- Lewis, C. S. (Clive Staples), 1898-1963
  - 1,485 requests
  - 1,429

In the top 5 classes:

- http://purl.org/dc/terms/BibliographicResource
  - 2,115

- http://purl.org/ontology/bibo/Author
  - 1,429

- http://purl.org/ontology/bibo/Book
  - 1,307

- http://purl.org/vocab/bio/0.1/birth
  - 591

- http://bnb.data.bl.uk/resource/Author
  - 486

In the top 5 properties:

- http://purl.org/ontology/bibo/isbn10
  - 27,781

- http://purl.org/dc/terms/title
  - 15,646

- http://www.w3.org/1999/02/22-rdf-syntax-ns#type
  - 15,268

- http://www.w3.org/2000/01/rdf-schema#label
  - 10,179

- http://purl.org/dc/terms/creator
  - 7,590
Locations

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>🇺🇸</td>
<td>84,278</td>
</tr>
<tr>
<td>2</td>
<td>🇬🇧</td>
<td>55,032</td>
</tr>
<tr>
<td>3</td>
<td>🇩🇪</td>
<td>25,045</td>
</tr>
</tbody>
</table>
## User Categories

### Visitors - Academia

<table>
<thead>
<tr>
<th>Rank</th>
<th>Institution Name</th>
<th>Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Karlsruhe Institute of Technology</td>
<td>408</td>
</tr>
<tr>
<td>2</td>
<td>University of Leeds</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>Imperial College</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>University of Wisconsin</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>University of Liverpool</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>Cardiff University</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>Vienna University of Economics and Business</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>University of Manchester</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>University of the Arts London</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>University of Sheffield</td>
<td>22</td>
</tr>
<tr>
<td>11</td>
<td>University of Glasgow</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>University of Oxford</td>
<td>21</td>
</tr>
<tr>
<td>13</td>
<td>The Open University</td>
<td>20</td>
</tr>
<tr>
<td>14</td>
<td>University of St Andrews</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>University of Birmingham</td>
<td>19</td>
</tr>
<tr>
<td>16</td>
<td>University of Southampton</td>
<td>19</td>
</tr>
<tr>
<td>17</td>
<td>University of Reading</td>
<td>19</td>
</tr>
<tr>
<td>18</td>
<td>Newcastle University</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>University of Strathclyde</td>
<td>18</td>
</tr>
<tr>
<td>20</td>
<td>University of Bristol</td>
<td>18</td>
</tr>
</tbody>
</table>

### Visitors - Government

<table>
<thead>
<tr>
<th>Rank</th>
<th>Institution Name</th>
<th>Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Department of Public Expenditure &amp; Reform</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>Library of Congress</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>Met Office</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>Department of Defence</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>U.S. National Library of Medicine</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>National Library of Australia</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>UK Cabinet Office</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>Natural Resources Wales</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>U.S. Department of State</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>Dorset Council</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>Indian Railways</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>East Dunbartonshire Council</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>Dunedin City Council</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>Isle of Anglesey County Council</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>State Government of Victoria</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>Walsall Council</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>Forestry Commission</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>North Tyneside Council</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>Leeds City Council</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>Devon County Council</td>
<td>2</td>
</tr>
</tbody>
</table>
Access

■ SPARQL accounts for **29%** of total requests*.

■ Direct human access accounts for **62%** of total requests*.

■ Sharp increase in requests from Software Libraries (95x)

*Percentages and counts are approximate and subject to change.
User Agent & Sessions

- Software Libraries have bigger, deeper, and longer sessions.

**Software Libs**

**Sessions Count**
- 3,760 Sessions

**Average Daily Depth**
- 11.05 Unique Resources

**Average Daily Size**
- 24 Resources

**Average Daily Duration**
- 01h:03m:04s

**Sessions Count**
- 59,074 Sessions

**Average Daily Depth**
- 1.98 Unique Resources

**Average Daily Size**
- 2 Resources

**Average Daily Duration**
- 00h:26m:53s

**Browsers**
Value of RDF Analytics
For The British Library

- **Offers better understanding of Linked Open BNB usage**
  - At greater levels of granularity than previously possible
  - Via more user friendly visualisations

- **Supports business case for service continuity**

- **Assists resource balancing for user support activities**

- **Informs dialogue with existing platform provider**

- **Informed tender specification**
Links

Demo site: http://52.49.205.156/analytics

Contacts:

British Library: metadata@bl.uk

Fujitsu Ireland: luca.costabello@ie.fujitsu.com

Free data services: http://www.bl.uk/bibliographic/datafree.html

Downloads: Linked data; Open data; Researcher format (.CSV) http://www.bl.uk/bibliographic/download.html

Thank you
## Metrics

### Content Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requests count</td>
<td>Includes global count &amp; break-downs (i.e.: graphs, classes, instances, predicates)</td>
</tr>
</tbody>
</table>

### Protocol Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Access Protocol</strong></td>
<td>The separate counts of HTTP lookups &amp; SPARQL queries. This information is helpful to determine whether visitors prefer HTTP lookups or if they rather execute SPARQL queries (over a given time frame).</td>
</tr>
<tr>
<td><strong>SPARQL Query type</strong></td>
<td>The counts of SPARQL verbs. It includes for example the count of SELECT, ASK, DESCRIBE, &amp; CONSTRUCT queries.</td>
</tr>
<tr>
<td><strong>SPARQL Query Complexity</strong></td>
<td>Indicates the number of “light” &amp; “heavy” SPARQL queries sent to the triplestore.</td>
</tr>
<tr>
<td><strong>HTTP Methods Count</strong></td>
<td>The count of how many requests have been issued for the most popular HTTP verbs (GET, POST, HEAD).</td>
</tr>
<tr>
<td><strong>Request Errors Count</strong></td>
<td>The count of HTTP &amp; SPARQL response codes occurred in a time frame. We distinguish between:</td>
</tr>
<tr>
<td></td>
<td>- Misses: HTTP 404 Not Found errors. This measure is useful to understand whether visitors are looking for resources which are not currently included in the dataset.</td>
</tr>
<tr>
<td></td>
<td>- Other client-side errors: other HTTP 4xx errors. This is important for a series of reasons, e.g. measuring how many malformed SPARQL queries have been issued (HTTP 400), or to detect whether visitors attempt to access forbidden RDF resources (HTTP 403).</td>
</tr>
<tr>
<td></td>
<td>- Server-side errors: the count of HTTP 5xx error codes. Important to identify server-side misconfiguration, or estimate whether repeated SPARQL queries trigger errors in the underlying triplestore.</td>
</tr>
</tbody>
</table>
### Audience Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td><em>Country &amp; city of origin of a visitor.</em></td>
</tr>
<tr>
<td>Network provider</td>
<td>The visitor host network.</td>
</tr>
<tr>
<td>Language</td>
<td>The preferred language requested by a visitor. Such information is extracted from the <code>Accept-Language</code> HTTP header (for HTTP lookups) &amp; by extracting xsd language-tagged string literals in SPARQL queries.</td>
</tr>
<tr>
<td>User Agent type</td>
<td>The visitor user agent type. It can belong to the following categories: - <strong>Software Library</strong> (e.g. Jena, Python sparql-client, etc.) - <strong>Browser &amp; Mobile Browser</strong> (Chrome, Safari, etc.) - <strong>Other</strong> (e.g. email clients)</td>
</tr>
<tr>
<td>Visitor Type</td>
<td>The nature of the visitor, that can be either: - human (e.g. manually-written SPARQL queries, one-time HTTP lookups) - machine (bot, crawlers, semantic web services, etc.)</td>
</tr>
<tr>
<td>New vs Returning visitors</td>
<td>New visitors vs visitors that have performed at least one visit before.</td>
</tr>
<tr>
<td>External Referrer</td>
<td>When dereferencing an RDF resource, the HTTP request might contain a third-party URI that identifies the resource “linking” to the data store.</td>
</tr>
<tr>
<td>Sessions count</td>
<td>The global count of all sessions for all visitors.</td>
</tr>
<tr>
<td>Session size</td>
<td>The number of requests sent by a visitor during a session (requests might be a mix of HTTP lookups &amp; SPARQL queries).</td>
</tr>
<tr>
<td>Session depth</td>
<td>The number of distinct RDF resources (graphs, classes, properties, instances) requested by a visitor during a session.</td>
</tr>
<tr>
<td>Session duration</td>
<td>The duration of a session.</td>
</tr>
<tr>
<td>Average session size</td>
<td>The average size of the sessions detected over a given time frame</td>
</tr>
<tr>
<td>Average Session depth</td>
<td>The average depth of the sessions detected over a given time frame.</td>
</tr>
<tr>
<td>Average session duration</td>
<td>The average duration of the sessions detected over a given time frame.</td>
</tr>
<tr>
<td>Bounce Rate</td>
<td>Indicates the percentage of sessions that contain only one resource request (whether this is an HTTP lookup or a SPARQL query).</td>
</tr>
</tbody>
</table>