

The British National Bibliography

Who Uses Our Linked Data?

Corine Deliot, Neil Wilson

The British Library

Luca Costabello, Pierre-Yves Vandenbussche

Fujitsu Research, Ireland

FUJITSU

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**

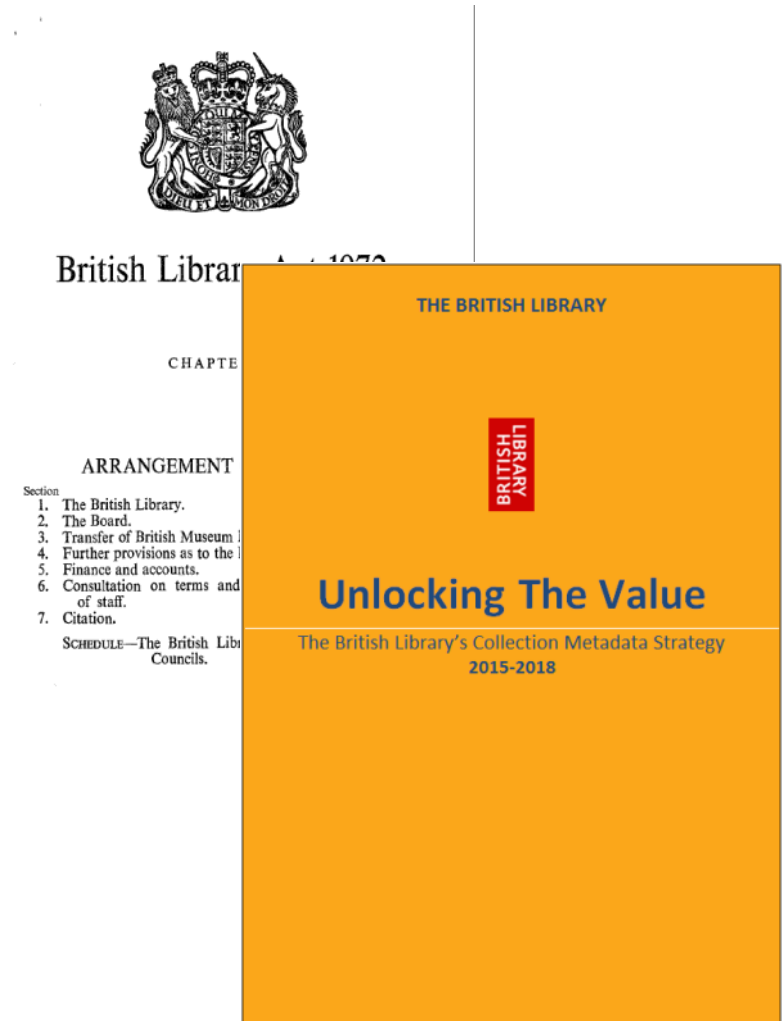
- **The value of RDF Analytics**

British Library Metadata Services

**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*

LIBRARY
HSIIRB

COLLECTION METADATA

The British National Bibliography

Home > Collection Metadata > The British National Bibliography

The British National Bibliography

The national bibliography records the publishing activity of the United Kingdom and the Republic of Ireland and as such is a measure of their intellectual output. This has traditionally included printed publications and more recently has been extended to electronic publications following the extension of legal deposit to this class of material in 2003.

New books and serials have been recorded in the British National Bibliography (BNB) since 1950. The BNB is the single most comprehensive listing of UK titles. UK and Irish publishers are obliged by law to send a copy of all new publications, including serial titles, to the **Legal Deposit Office** of the British Library. This material is catalogued by experienced staff in accordance with international **standards** for resource description and access. This work is done in partnership with the five other British and Irish libraries allowed by law the privilege of legal deposit, under the **Legal Deposit Libraries Shared Cataloguing Programme** (LDLSCP).

The BNB also contains details of forthcoming books. Under the **Cataloguing-in-Publication Programme** (CIP) information on new titles appears up to 16 weeks ahead of the announced publication date. Advance information on well over 50,000 titles each year is provided in this way.

The coverage of the BNB has always been selective (see **exclusions policy**) with the emphasis being on mainstream monographs available through normal book buying channels.

The availability of BNB records was traditionally shown by the BNB/MARC hit-rate derived from the currency survey carried out by the UK Office for Library and Information Networking (UKOLN), but this was discontinued from 1st April 2005 (further information is available by following the link above).

All of BNB is available for searching **here**. Our **Z39.50 service** allows downloading records in MARC21 for free.

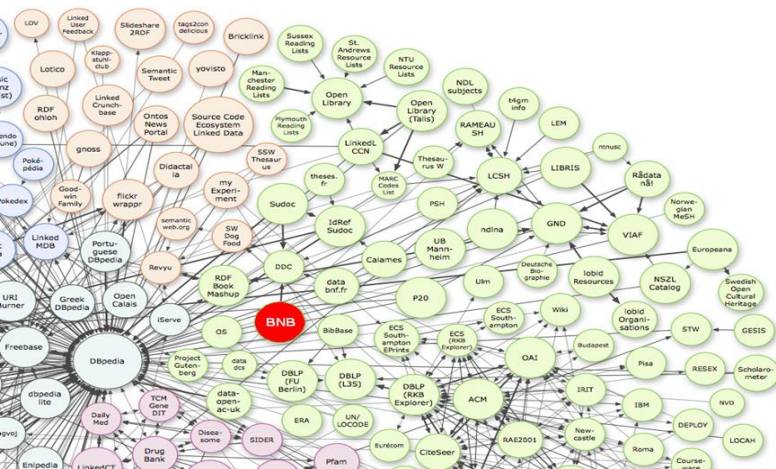
Sections:

- British National Bibliography
 - Search the BNB
 - This week's new BNB records
 - The Legal Deposit Libraries Shared Cataloguing Programme
 - The Cataloguing-in-Publication Programme
 - The British National Bibliography Exclusions
 - Structure of the BNB number
 - Downloading MARC 21 records
- Data Services
- Standards
- News
- Contact us

This page contains links to Adobe PDF files. Accessibility solutions and free 'Reader' software are available from Adobe.

The Linked Open BNB

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



“Linking Open Data cloud diagram, by Richard Cyganiak & Anja Jentzsch. <http://lod-cloud.net/>”

Usage terms: <http://creativecommons.org/licenses/by-sa/3.0/>

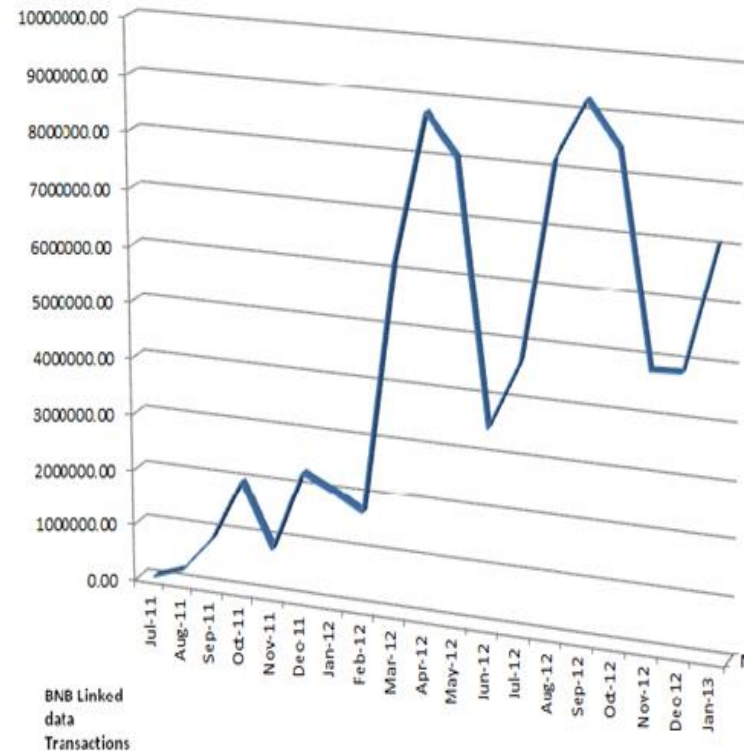
Linked Open Data: Some Challenges

- **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

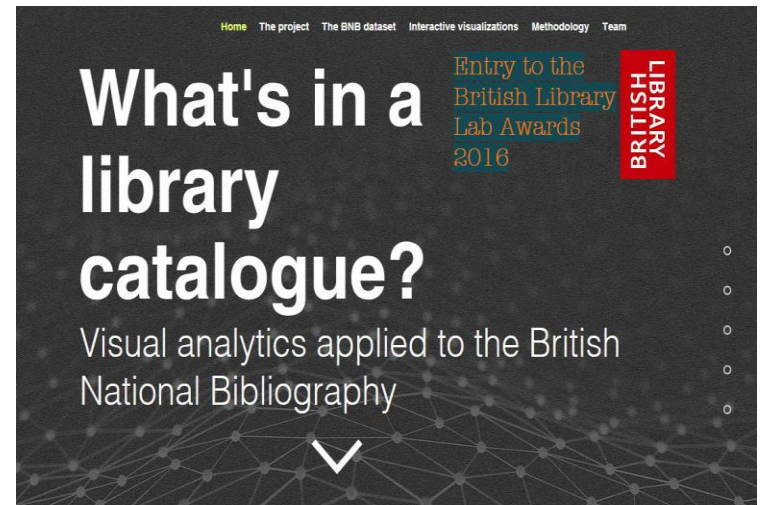
- e.g. Linked Open BNB data used as test data for a semantic search demonstrator.
- e.g. data provided to Microsoft to assist in their research into linking structured data

BNB data used in tutorials

- e.g.
http://www.meanboyfriend.com/overdue_ideas/2014/10/using-an-api-hands-on-exercise/ - Owen Stephens

 **libraries hacked** @librarieshacked · Oct 28
In fact that's books published in Bath, about Bath is this one. double good. data.bathhacked.org/Heritage/BNB-B...

 **libraries hacked** @librarieshacked · Oct 28
books about Bath automatically being extracted into city data portal from british national bibliography. impressive data.bathhacked.org/Heritage/BNB-B...

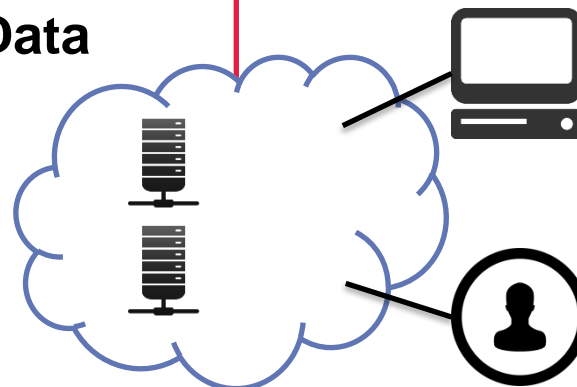


This project presents a research method and a tool, called the Network Coincidence Analysis (NCA) framework, and applies it to the BNB dataset providing a method to delve into the data's inherent relationships, discover associations and make comparisons. NCA is a visual analytics framework that

British Library - Fujitsu collaboration

Metadata Publication as Linked Data

- **Who** is using our data?
- **Which** data?
- **How** to optimise our publication?



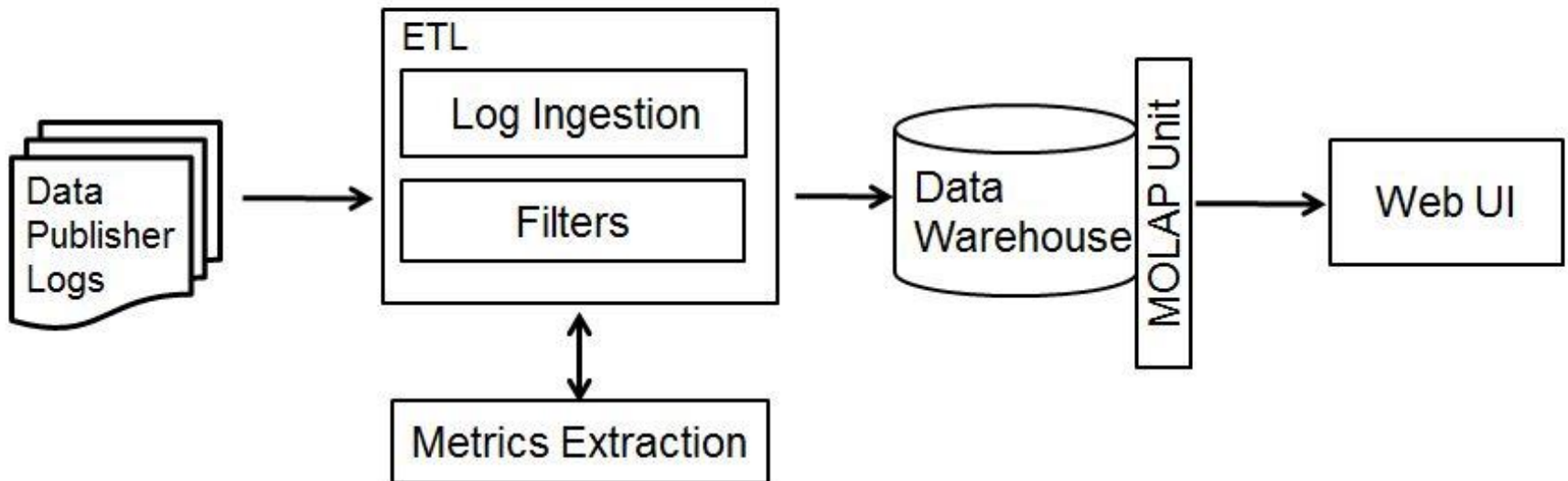
Linked Data Analytics Solution

- 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



Analytics for Linked Data Publishers



Analytics for Linked Data

British Library ▾



Overview

Content

Requests Count

Response Code

Audience

Location

User Agent

Visitors

Sessions

Behaviour

Data Access

Overview

Request Count (All Time) ⓘ

252.8K

Requests

Request Count ⓘ
(Selected Time Frame)

8.0K

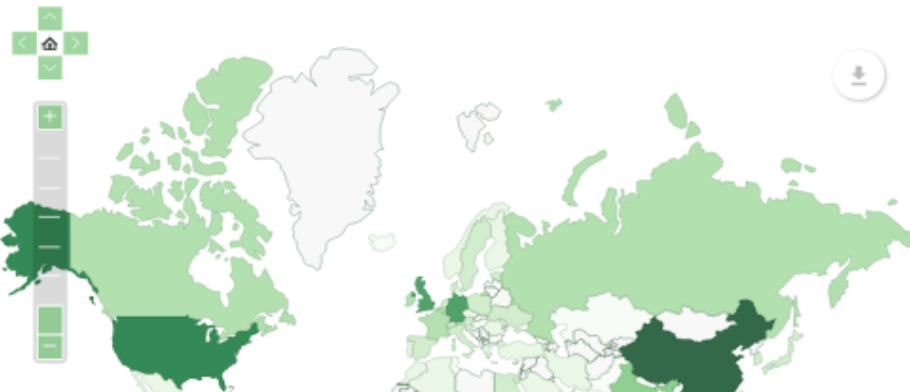
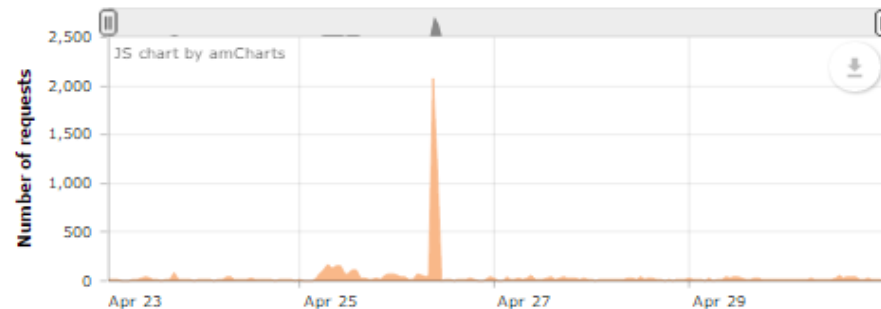
Requests

Protocols (2) ▾

User Agents (3) ▾

2015-04-23 - 2015-04-30

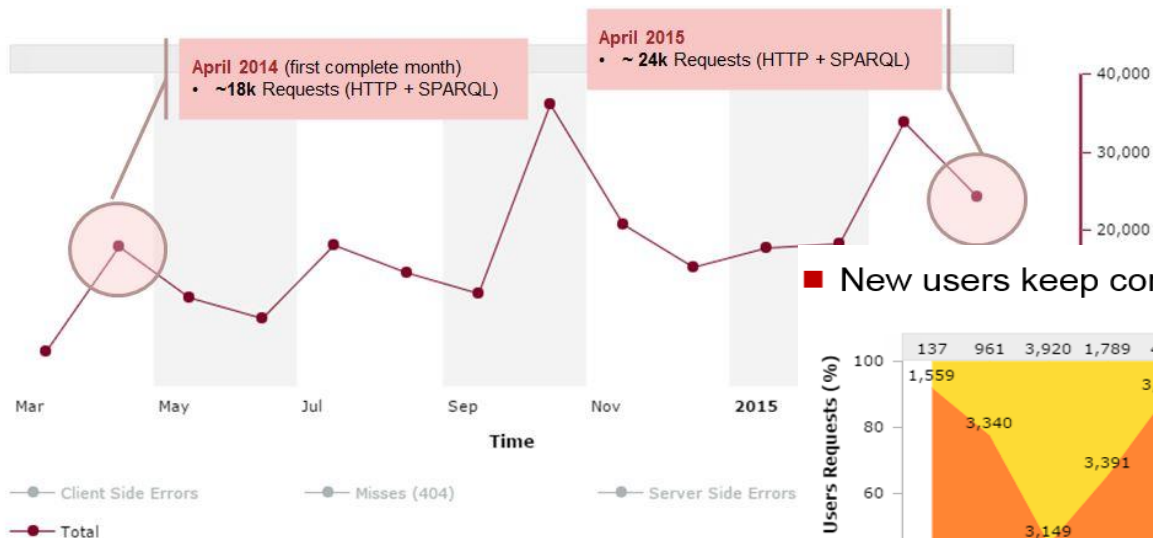
Number Of Requests Between 2015-04-23 And 2015-04-30 ⓘ



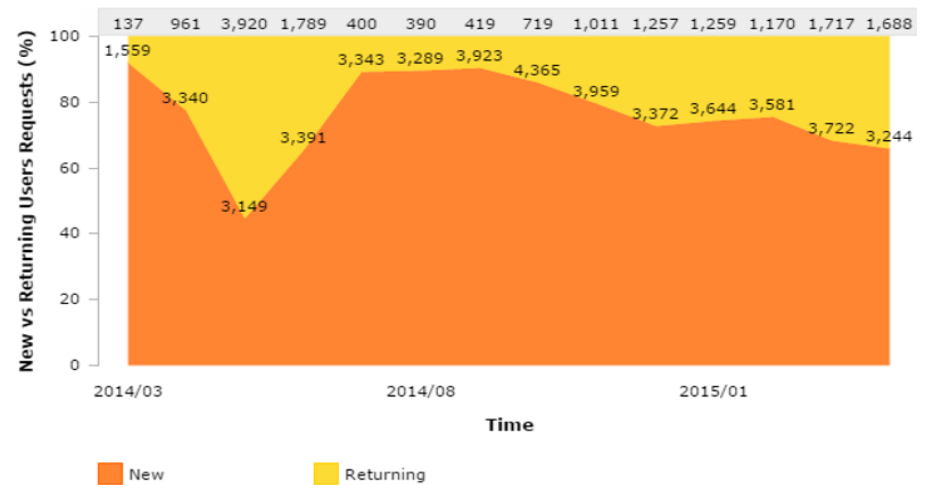
What Did We Learn?

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

■ Overall request flow is stable



■ New users keep coming in



■ Bounce rate: **48%**

Instances, Classes & Properties

In the top 5 instances

The hobbit / J. R. R. Tolkien

<http://bnb.data.bl.uk/id/resource/009910399>

6,092 requests

Lewis, C. S. (Clive Staples), 1898-1963

[http://bnb.data.bl.uk/id/person/LewisC3%28Clive Staples%291898-1963](http://bnb.data.bl.uk/id/person/LewisC3%28Clive%20Staples%291898-1963)

1,485 requests

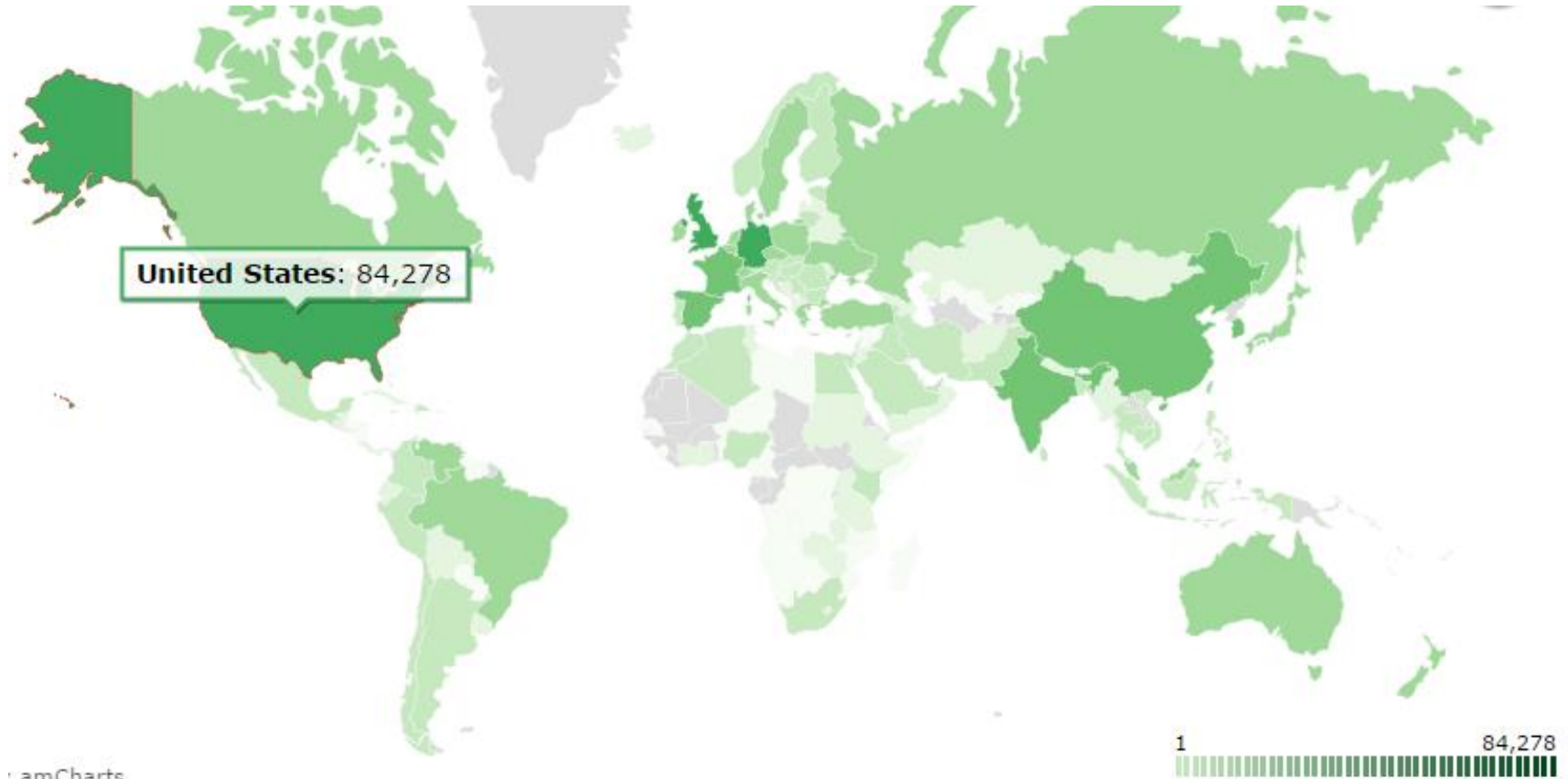
In the top 5 classes

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	476




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























amCharts

1		84,278	33.3%
2		55,032	21.7%
3		25,045	9.9%

User Categories

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

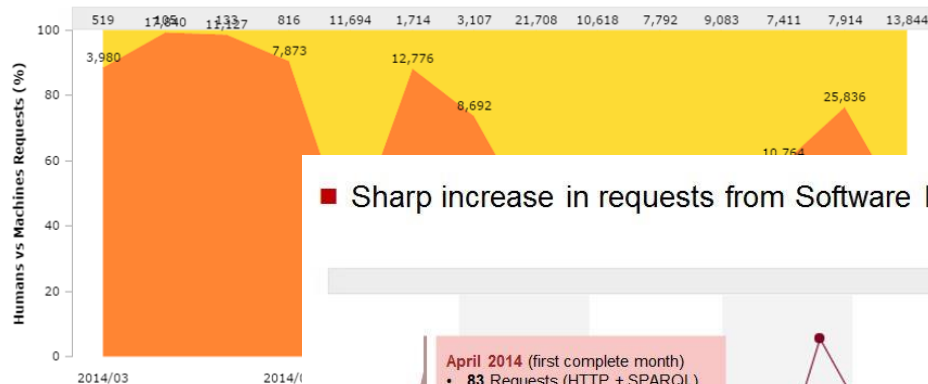
	Visitors - Government		Sessions
1	Department of Public Expenditure & Reform		29
2	Library of Congress		12
3	Met Office		9
4	Department of Defence		7
5	U.S. National Library of Medicine		6
6	National Library of Australia		6
7	UK Cabinet Office		6
8	Natural Resources Wales		5
9	U.S. Department of State		4
10	Dorset Council		4
11	Indian Railways		4
12	East Dunbartonshire Council		4
13	Dunedin City Council		4
14	Isle of Anglesey County Council		3
15	State Government of Victoria		3
16	Walsall Council		3
17	Forestry Commission		3
18	North Tyneside Council		3
19	Leeds City Council		3
20	Devon County Council		2

Access

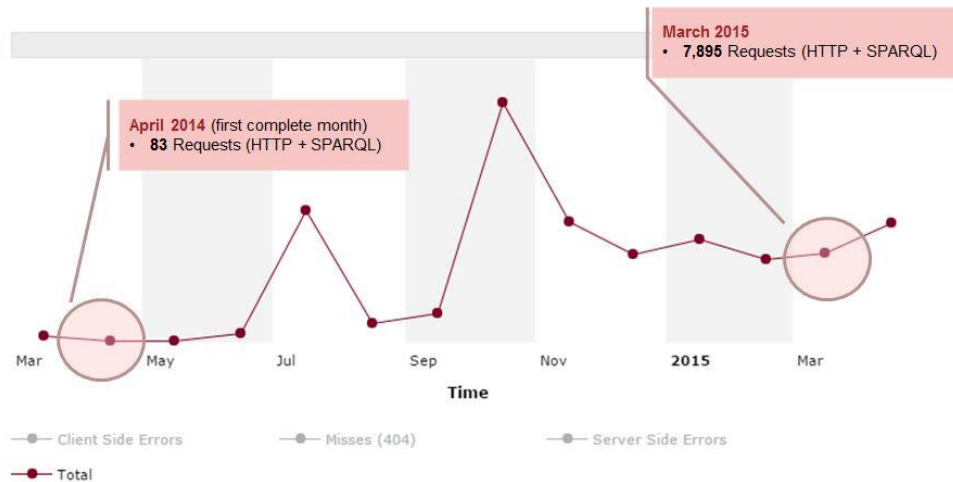
■ SPARQL accounts for **29%** of total requests*.



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



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



User Agent & Sessions

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



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>

Collection Metadata Strategy <http://www.bl.uk/bibliographic/pdfs/british-library-collection-metadata-strategy-2015-2018.pdf>

Thank you

Content Metrics

Metric	Description
Requests count	Includes global count & break-downs (i.e.: graphs, classes, instances, predicates)

Protocol Metrics

Metric	Description
Data Access Protocol	<i>The separate counts of HTTP lookups & 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).</i>
SPARQL Query type	The counts of SPARQL verbs. It includes for example the count of SELECT, ASK, DESCRIBE, & CONSTRUCT queries.
SPARQL Query Complexity	Indicates the number of “light” & “heavy” SPARQL queries sent to the triplestore.
HTTP Methods Count	The count of how many requests have been issued for the most popular HTTP verbs (GET, POST, HEAD).
Request Errors Count	<p>The count of HTTP & SPARQL response codes occurred in a time frame. We distinguish between:</p> <ul style="list-style-type: none">- 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.- 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).- 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.

Audience Metrics

Metric	Description
Location	<i>Country & city of origin of a visitor.</i>
Network provider	The visitor host network.
Language	The preferred language requested by a visitor. Such information is extracted from the <code>Accept-Language</code> HTTP header (for HTTP lookups) & by extracting xsd language-tagged string literals in SPARQL queries.
User Agent type	The visitor user agent type. It can belong to the following categories: <ul style="list-style-type: none"> - Software Library (e.g. Jena, Python sparql-client, etc.) - Browser & Mobile Browser (Chrome, Safari, etc.) - Other (e.g. email clients)
Visitor Type	The nature of the visitor, that can be either: <ul style="list-style-type: none"> - human (e.g. manually-written SPARQL queries, one-time HTTP lookups) - machine (bot, crawlers, semantic web services, etc.)
New vs Returning visitors	New visitors vs visitors that have performed at least one visit before.
External Referrer	When dereferencing an RDF resource, the HTTP request might contain a third-party URI that identifies the resource “linking” to the data store.
Sessions count	The global count of all sessions for all visitors.
Session size	The number of requests sent by a visitor during a session (requests might be a mix of HTTP lookups & SPARQL queries).
Session depth	The number of distinct RDF resources (graphs, classes, properties, instances) requested by a visitor during a session.
Session duration	The duration of a session.
Average session size	The average size of the sessions detected over a given time frame
Average Session depth	The average depth of the sessions detected over a given time frame.
Average session duration	The average duration of the sessions detected over a given time frame.
Bounce Rate	Indicates the percentage of sessions that contain only one resource request (whether this is an HTTP lookup or a SPARQL query).