

A Content Development Plan for Datasets at the British

Library

Jez Cope, Data Services Lead

Overview

- 1. Introduction
- 2. Content Development at the British Library
- 3. A CDP for Datasets

Introduction

What this talk **is not**

- A description of the British Library's strategy for datasets
- The British Library's official position

What this talk **is**

- Observations that shape my current thinking
- Heavily influenced by ideas from many others
- A provocation and conversation starter

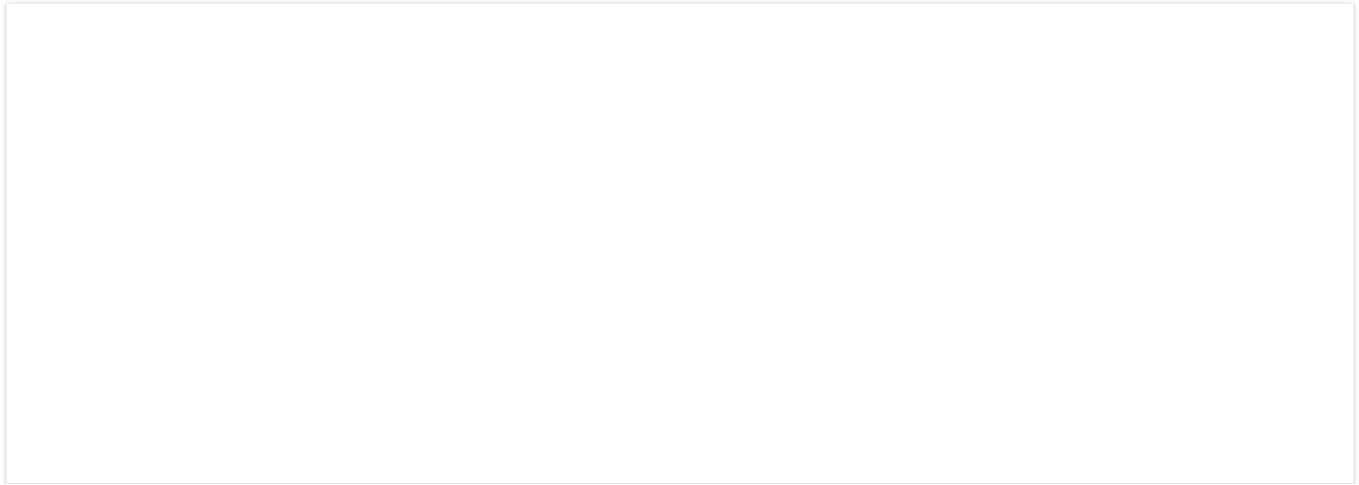
The British Library

- National Library of the United Kingdom
- One of 6 Legal Deposit libraries
 - Includes non-print since 2013

“Our Mission: We make our intellectual heritage accessible to everyone, for research, inspiration

Datasets at the BL

- Not a traditional focus of the British Library!
- Over a decade of work in various areas
- Data, Data Science, ML/AI increasingly



“We will build on this world-leading research, working in partnership to improve access to our data and new

DataCite

- British Library a founding partner in DataCite in 2009/10
- Consistent growth over 10 years in the UK

Strengths

- Lots of existing datasets available
 - Collection metadata (BNB)
 - Digitised books, manuscripts, etc.

Strengths

- Wealth of expertise and enthusiasm to draw on
- Recent research e.g. into Emerging Formats
- Broad national & international network

Challenges

- Plenty of good practice, but scattered
- Primary focus on print (and print-like) content
- Priority is maintenance of the existing collection

Content Development at the British Library

Key principles

- There is **one** British Library collection
- **Legal Deposit** is the foundation of our collection building

Overarching Content Strategy

- Long timescale
- Cross-cutting principles
- Governance and decision-making structure

Content Development Plans (CDPs)

- Shorter timescale
- Driven by evidence of user need
- Two types

A CDP for Datasets

- Bring datasets into Content Strategy framework
- Buy-in from senior stakeholders
- Embed into our core collecting practice

Observations

Without data, we are not seeing the whole picture

- Part of our **cultural heritage** is data
- Part of the **scholarly record** is data

**Innovations in research methodology have
outstripped our capacity to provide data to those
using them**

Data derived from the collection is

1. Valuable
2. Expensive to (re)create, and
3. Worth preserving

The rights environment is complex

- Datasets not covered by Non-Print Legal Deposit
- Changing this needs primary legislation
- Legal Deposit complicates even open-licensed

Datasets need specialised discovery mechanisms

Priorities

1. Data already in the collection
2. Data derived from existing collection items
3. *Data not currently collected*

(Potential) actions

Audit and integrate existing datasets more closely into the national collection

- Establish documentation using datasheets for datasets[†] as standard practice
- Connect to known external sources of data

Further develop training & guidance for curators

- Affordances of cultural heritage data & how researchers use it
- Subject-specific selection criteria & priorities

Clarify rights and licensing options

- “As open as possible; as closed as necessary”
- Allow open licenses to override Legal Deposit restrictions

Collaborate across the sector

- Don't reinvent the wheel
- Don't hoard expertise
- Involvement in RDA, AI4LAM, LIBER Data

Research & development

- What infrastructure is needed to collect datasets at a national scale?
- How can we facilitate computational access

What do you think?

- What is the role of a National Library?
- What do you & your users need?

Thanks!

- Rachael Kotarski; Torsten Reimer
- Blanka Matkovic; Hannah Liebeschuetz;
Research Infrastructure Services

Any questions?