From Bibliography to Network Analysis, Data Analytics, Image recognition and Machine Learning: The Journey of the Iberian Books Project

Dr John B Howard, Prof. Sandy Wilkinson
University College Dublin

Coalition for Networked Information
11 December 2018
Iberian Books: 2006-2018

• Bibliographic survey of books printed in the Spanish and Portuguese sphere of influence from the Age of Discovery

• Undertaken in two phases (1472-1600; 1601-1700)

• Funded by the Andrew W. Mellon Foundation

Principal objectives:

• Finding: to allow researchers to find works linked to authors or printers/publishers, or on specific themes, and to locate surviving copies

• Understanding: to move forward our understanding of the contours, character and rhythms of the book industry

• Conservation: To offer strategic input into the conservation of the printed heritage

• Sustainability: To produce a sustainable and accessible reference resource through interoperability and enabling secondary use of data
A partner project of the **USTC** (St Andrews)
Methodology & Data Sources

• Online and printed library catalogues
• Other major bibliographies (e.g. of dramatists, or of places of printing)
• Auction catalogues
• Visits to significant collections without published catalogue
• Other resources: indices of forbidden books, inquisition records, booksellers’ catalogues, & post-mortem inventories
• Aggregation, emendation, and harmonisation of bibliographic data
Information keyed from printed sources

Library Catalogue Files (MARC)

Web Data Extraction

Data processed and then compared to records held

Core relational database
Library curation workflow ...

• Data conversion (focus on interoperability)
  – Analysis, metadata mapping (MODS, MADS), quality assurance processes
  – Verification of linked data references

• Image processing:
  – Thumbnails
  – Phase 2: include pages with graphical elements & associated metadata

• Geocoding, geospatial linked data

• Interlinked systems & services:
  – repository, index
  – PostgreSQL/postGIS, IIIF, image matching software
Overview of the repertory

1472-1700 TOTALS

- Ca. 131,000 items
- 651,000 copies
- 43,200 links to digital copies
- Ca. 2,500 libraries
Welcome to Iberian Books

The objective of Iberian Books is to produce a foundational listing of all books published in Spain, Portugal and the New World or printed elsewhere in Spanish or Portuguese during the Golden Age (1472-1700).

The Iberian Books Project

Iberian Books is a research project based at the School of History, University College Dublin. From 2010 to 2018, it was funded through two generous grants awarded by the Andrew W. Mellon Foundation under their Scholarly Communications and Information Technology Scheme.

The objective of Iberian Books is to produce a foundational list of all

Summer 2018 Updates

Iberian Books repository updated

After more than a decade of work by the Iberian Books team, we are delighted to announce that the database has now reached its terminal date of 1700, and that the data will be published as part of the UCD Digital Library by the end of the Summer. This more than doubles the previous

https://iberian.ucd.ie/
Analytics
Oración panegírica, que en la festividad del jueves santo, se celebro en la iglesia parroquial de santo Thome de los cavalleros, este año de mil seisientos y noventa y siete

Navarro, Manuel [Author]
Esteves, María, viuda de Lucas Pérez [Printer]
How Publishing in Spain, Portugal and the New World Evolved before 1701

This viz allows you to better understand how the publishing industry evolved in the Iberian Peninsula and in Latin America before 1701. You can choose to look at the number of items printed in different cities, or at the number of printed sheets used to produce these items. You will notice just how dispersed the industry was in Spain, with many centres of printing, and just how concentrated it was in Portugal, Mexico and Peru. For Spain, in the age of the Incarnation, Salamanca with its prestigious university produced more items than any other city. In the first half of the sixteenth century, Seville became the leading centre of print - a mantle handed back again to Salamanca until the 1580s.

From 1590, Madrid became the most significant centre of print, a position it would continue to hold throughout the seventeenth century. By 1680-1700, Madrid accounted for 20% of all books published in Spain. A quarter of the items printed in the country (where a place of publication is known), with Barcelona producing a quarter. In the seventeenth century, printing was known to have occurred in around 130 locations in Spain, though only around 26 of these locations are known to have printed more than 100 items.

The bottom section of the viz allows you to create line graphs (based on the number of items printed or on the sheets used to produce each item) for every known publishing location.

Two points of caution. For publishing centres like Antwerp and Lyon – outside of the Iberian Peninsula or its associated overseas territories – the data is limited to books in Spanish or Portuguese only. Secondly, we have kept only those items (those without any known location) and these appear on the map at the centre of each country.
Unanswered questions

• The printer for half of all items recorded in *Iberian Books* is unknown

• The place of printing for a third of items in *Iberian Books* is unknown

• Uncertainties continue to exist with regard to undated imprints

• Beyond the common, declarative aspects of traditional bibliography, what additional evidence can be brought to bear on these questions?
  – Recent research on type and typography
graphical elements
Rationale for Exploring Graphical Elements

• Largely unexploited as evidence for understanding the repertory of early printed books

• Allow exploration of long-term patterns of sharing, selling, and inheriting of ornaments and illustrations

• To facilitate research into the visual culture of the early-modern Iberian world. Woodcut and engraved illustrations are a largely neglected area of historical and art historical research; other areas will benefit too

• To promote understanding of the use of visual culture in the colonial acculturation of indigenous or non-literate peoples
Tratado de las drogas y medicinas de las Indias Orientales con sus plantas debuxadas al biuo

Acosta, Cristóbal [Author]
Victoria, Martín de [Printer]

IB Identifier 51036
In collection ➔ Iberian Books
† thumbnail image used with permission

Bibliographic record

An exemplar has been referenced in a major analytical bibliography, or inspected book in hand, by a member of the Iberian Books project team or a collaborating institution.

Responsibility Burgos, por Martín de Victoria, 1578.
Citation/reference IB: 41
Languages Spanish
Genre Book
Subject Medical
Place of origin Burgos
Location http://n2t.net/ark:/87925/drs1.iberian.2862
Methodology

• Establish list of digitized materials; download digital works
• Each book split into constituent pages
• Pages without any ornamentation or illustration discarded (automated, with manual QA)
• Classification of ornaments (manual)
• Indexing of page images
• Initial dataset: ca. 230,000 pages with graphical components
Four approaches to resource discovery

1. *Iberian Books*: Items that have been digitized and contents analysed for graphical components
   - Direct links to image matching application
   - If IIIF APIs supported, links to Mirador

   - Metadata search, with links to Iberian Books and to Ornamento


4. *Mirador ImageMatch*
   - Use Ornamento plugin to search graphical elements from any early printed book accessible via IIIF
1. Iberian Books → Ornamento ImageMatch

Indexed graphical elements in this document

The following graphical elements have been identified in this document and are indexed in Ornamento

- Graphical element: Illustrations — IconClass: 49M41
  Find similar graphical content in other imprints using Ornamento
  View document in the Mirador viewer: https://lif.archiveland.org/lif/bub_gb_e81ZvAh3qCG
  © Google Books. Reproduced here on a non-commercial basis.

- Graphical element: Illustrations — IconClass: 49M41
  Find similar graphical content in other imprints using Ornamento
  View document in the Mirador viewer: https://lif.archiveland.org/lif/bub_gb_e81ZvAh3qCG
  © Google Books. Reproduced here on a non-commercial basis.

- Graphical element: Illustrations — IconClass: 49M41
  Find similar graphical content in other imprints using Ornamento
  View document in the Mirador viewer: https://lif.archiveland.org/lif/bub_gb_e81ZvAh3qCG
  © Google Books. Reproduced here on a non-commercial basis.

- Graphical element: Illustrations — IconClass: 49M41
  Find similar graphical content in other imprints using Ornamento
  View document in the Mirador viewer: https://lif.archiveland.org/lif/bub_gb_e81ZvAh3qCG
  © Google Books. Reproduced here on a non-commercial basis.

- Graphical element: Illustrations — IconClass: 49M41
  Find similar graphical content in other imprints using Ornamento
  View document in the Mirador viewer: https://lif.archiveland.org/lif/bub_gb_e81ZvAh3qCG
  © Google Books. Reproduced here on a non-commercial basis.

- Graphical element: Illustrations — IconClass: 49M41
  Find similar graphical content in other imprints using Ornamento
  View document in the Mirador viewer: https://lif.archiveland.org/lif/bub_gb_e81ZvAh3qCG
  © Google Books. Reproduced here on a non-commercial basis.

File: Choose file: No file chosen
or URL: 

Return to Ornamento

$1036

Full size image
2. Ornamento Search → Ornamento ImageMatch

Welcome to Ornamento

Ornamento is a pilot project, exploring the value of creating a European repository of ornamentation and illustration contained in early-printed books. Currently, it contains close to a quarter of a million ornate letters, ornaments, borders, musical notation, diagrams, and illustrations drawn from Iberian print before 1701. You can explore these by searching metadata or via image-matching technology.

Using Ornamento

There are four ways to use Ornamento to find related images.

Search Ornamento

Click on Search Ornamento to open the Advanced Search box. Here you can search by image classification (ornaments, ornate letters, etc.), IconClass category (4SA122: armorial bearings, etc.), ornamental letter (A, B, etc.), Country, or by Title/keyword, IB edition number, place of printing, and date. Please note that not all images in Ornamento have been categorised.

Ornamento ImageMatch

Part of the Ornamento pilot project is the implementation of the ImageMatch technology, developed by Professor Andrew Zisserman and Relja Arandjelovic of the University of Oxford Department of Engineering Science’s Visual Geometry Group, in collaboration with Giles Bergel (Oxford Faculty of English), Alexandra Franklin and Richard Ovenden (Bodleian Library).

Development and maintenance of VGG Image Search Engine (VSE) is...
3. Ornameto ImageMatch direct image search

Return to Ornameto

List of images in the database

Images 1 to 20 of 229509. Next
Liber de contemptu mundi, sive de miseria humanae conditionis

Innocentius III [Author]
Posa, Pere [Printer]

An exemplar has been referenced in a major analytical bibliography, or inspected book in hand, by a member of the IberoBooks project team or a collaborating institution.
4b. Mirador ImageMatch → Annotation
Observations from pilot application

- Ornate letters – almost now all identified (150,000 A’s B’s C’s &c)
- 1,200 printers’ devices identified (some provisionally)
- 1,350 portraits of individuals now identified
- At least 1st level Iconclass designations, but how much work should be done on illustrations?
Known Issues and Next Steps (1)

- Ca. 1.8M known images of early printed books from major repositories
  - Only 20% available via IIIF
  - Some IP concerns
- Size of current sample provide proof of concept, but not the level of critical mass required
- Technical:
  - VGG Image Search Engine (VISE)
    - Data management issues
    - No inherent metadata management utility
    - Questions regarding scalability and performance
  - Image storage and access
Known Issues and Next Steps (2)

• Methodology refinement
  – Identification of pertinent pages/graphics
  – Image description and classification
    – Vocabularies (imageClass, Getty, etc.)
    – Machine learning approaches
    – Object recognition challenges
  – Quality assurance methodologies
  – End-user interactivity

• Metadata representation
  – Current approach: postgresQL database
  – Descriptive records as annotations: indexed in local Anótála Annotation Store
  – Challenges in notifications of third-party annotations: e.g., LDN

• Additional use cases:
  – Numismatics, bindings, watermarks, seals, bookstamps, object recognition, etc.