



**HAL**  
open science

# DATA MINING HISTORICAL NEWSPAPERS METADATA

Jean-Philippe Moreux

► **To cite this version:**

Jean-Philippe Moreux. DATA MINING HISTORICAL NEWSPAPERS METADATA. Document Analysis Systems 2016, Apr 2016, Santorin, Greece. , DAS 2016 Proceedings. hal-01389462

**HAL Id: hal-01389462**

**<https://bnf.hal.science/hal-01389462>**

Submitted on 2 Dec 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

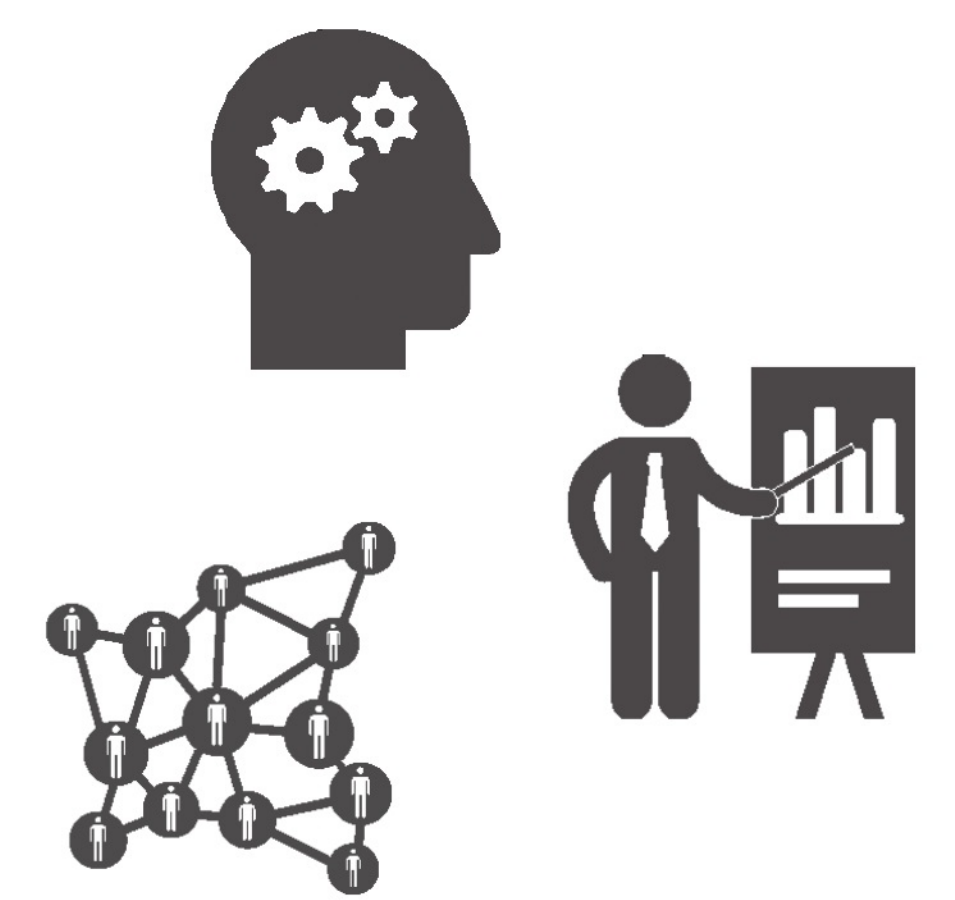


# DATA MINING HISTORICAL NEWSPAPERS METADATA

## Old News Teaches History

Would you like to provide:

- \* **Researchers** (DH, History of Press, Information Science) with quantitative metadata on press layout and content?
- \* **Digital Curators** with insights on their collections?
- \* **Digitization Program Managers** with statistics?



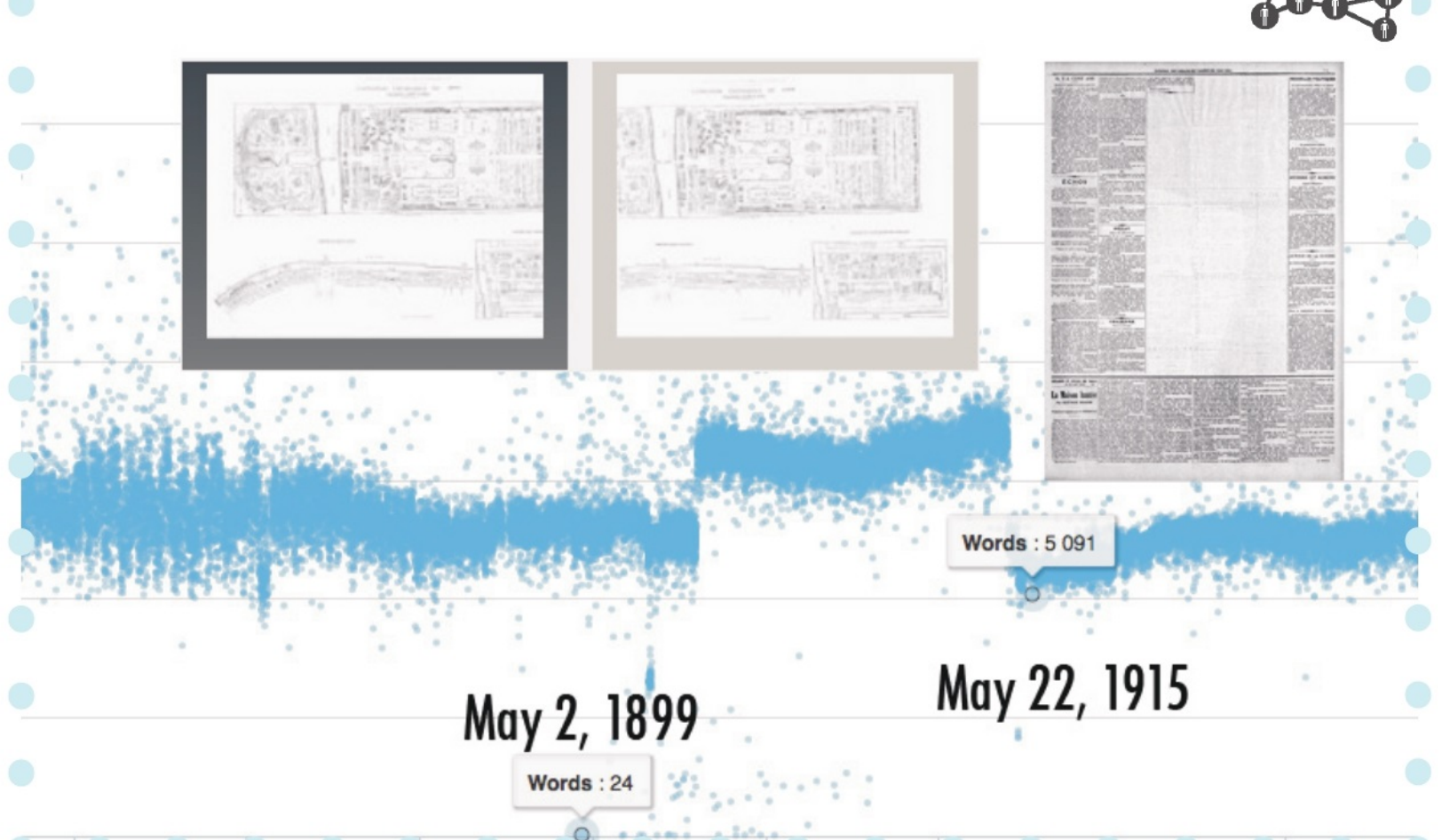
Just **datamine** and **dataviz** your METS/ALTO digitized press collection!

### THE TEST BED

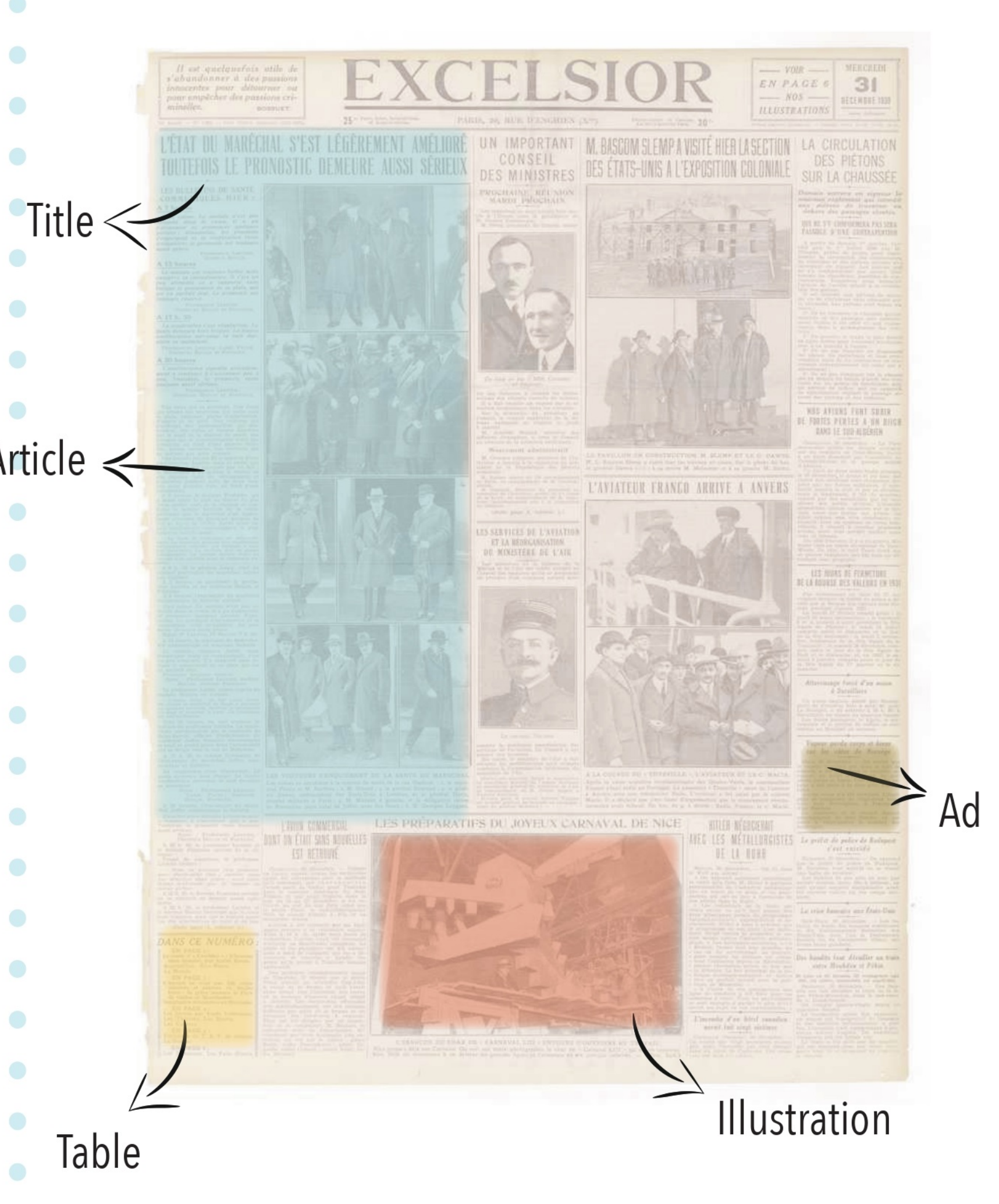


**Europeana Newspapers project (2012-2015):** 880K OLR'ed pages from BnF collection (1814-1944)

### UNCOMMON ISSUES



### USE OF ILLUSTRATION

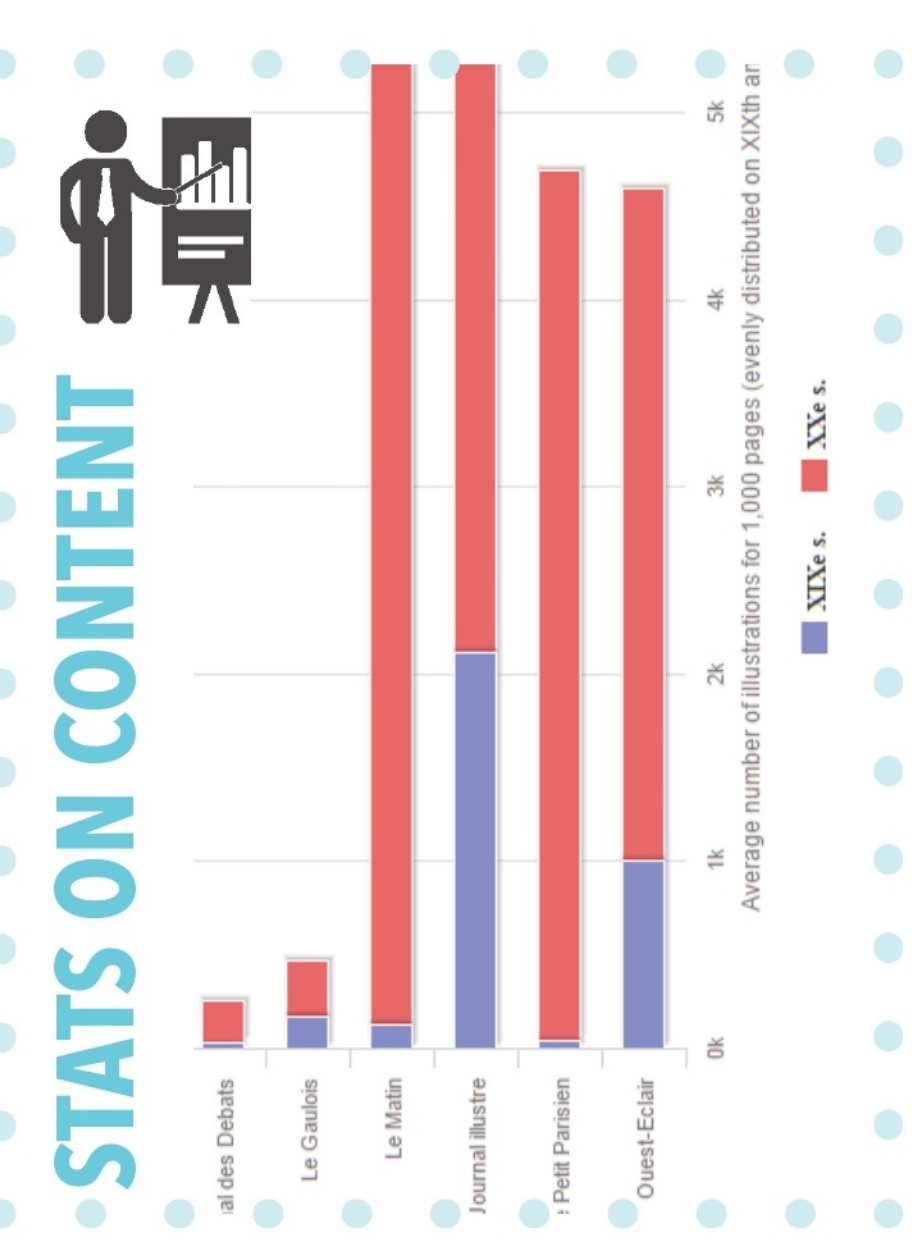


corpus  
1 Tb

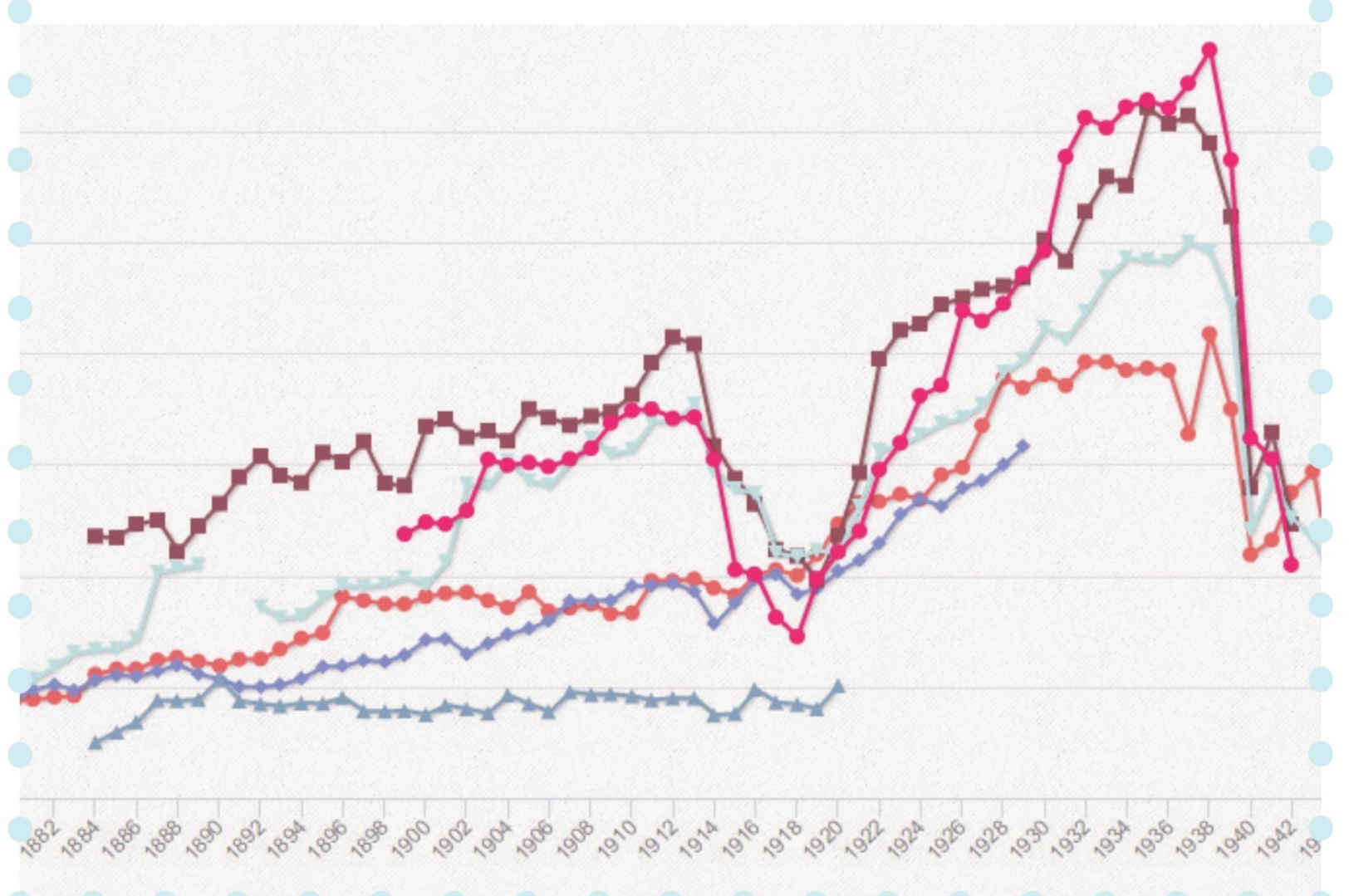
XSLT



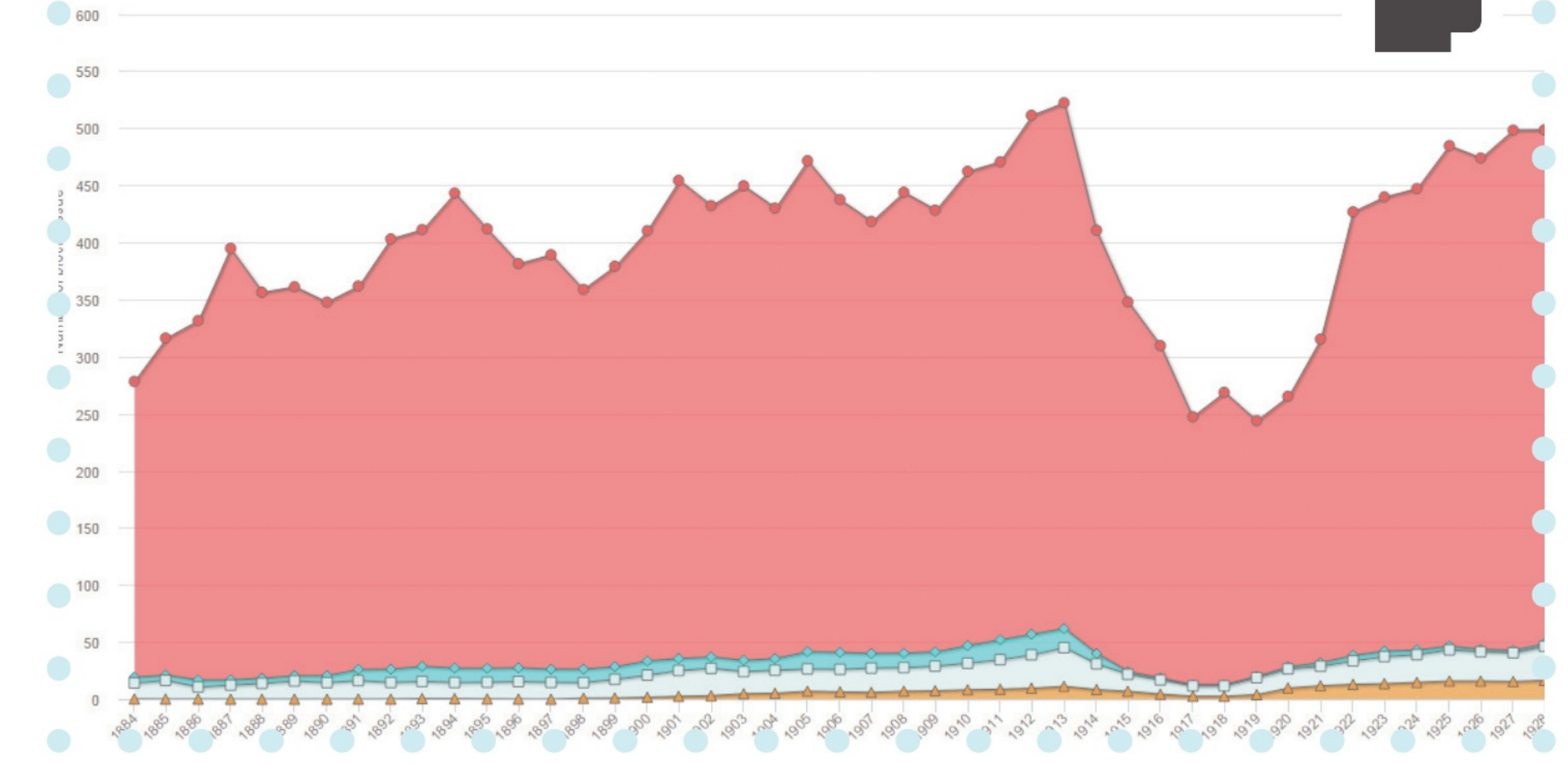
derived metadata  
80 Mb  
5.5 M of metadata values



### FROM GAZETTE TO MODERN DAILY



### CONTENT TYPES



Europeana newspapers

Dive into the details of Europe's historic newspapers.

Jean-Philippe Moreux, BnF

[www.europeana.eu](http://www.europeana.eu) & [www.theeuropeanlibrary.org/tel4/newspapers](http://www.theeuropeanlibrary.org/tel4/newspapers)



Staatsbibliothek zu Berlin  
Preußischer Kulturbesitz



This project runs from February 2012 to February 2015. It is led by the Staatsbibliothek zu Berlin and co-funded by the European Commission under the Competitiveness and Innovation Framework Programme. <http://ec.europa.eu/ict-pip>