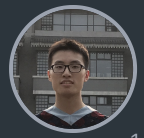
OldVisOnline

Curating a Dataset of Historical Visualizations


Yu Zhang[1]

Ruike Jiang[2]

Liwenhan Xie[3]

Yuheng Zhao[4]

Can Liu[2]

Tianhong Ding[5]

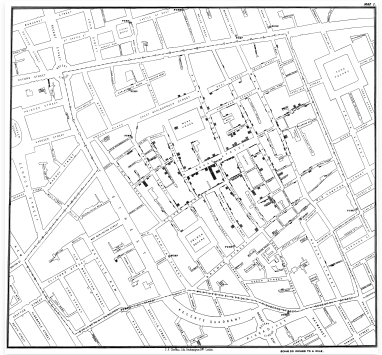
Siming Chen[4]

Xiaoru Yuan[2]

[1]University of Oxford [2]Peking University [3]Hong Kong University of Science and Technology [4]Fudan University [5]Huawei

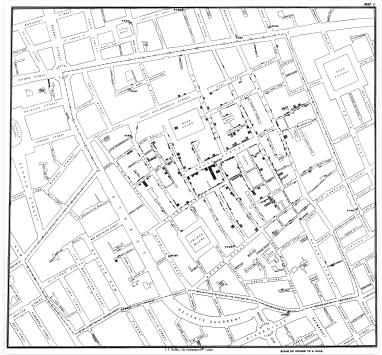
Historical Visualizations (or Old Visualizations)

Historical Visualizations (or Old Visualizations)

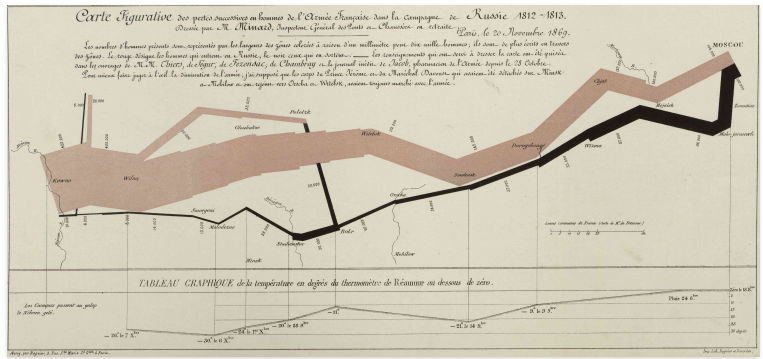


John Snow's cholera map, 1855
(image in public domain)

Historical Visualizations (or Old Visualizations)

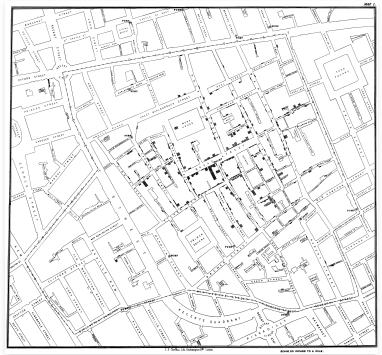


John Snow's cholera map, 1855
(image in public domain ⓒ)

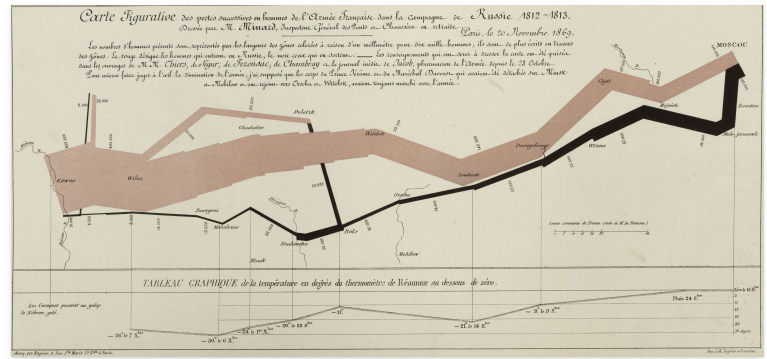


Charles Minard's map of Napoleon's Russian campaign, 1869
(image in public domain ⓒ)

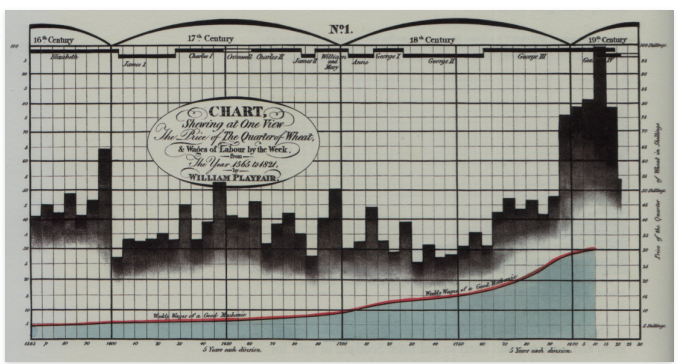
Historical Visualizations (or Old Visualizations)



John Snow's cholera map, 1855
(image in public domain)

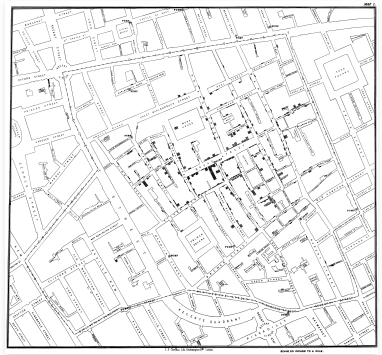


Charles Minard's map of Napoleon's Russian campaign, 1869
(image in public domain)

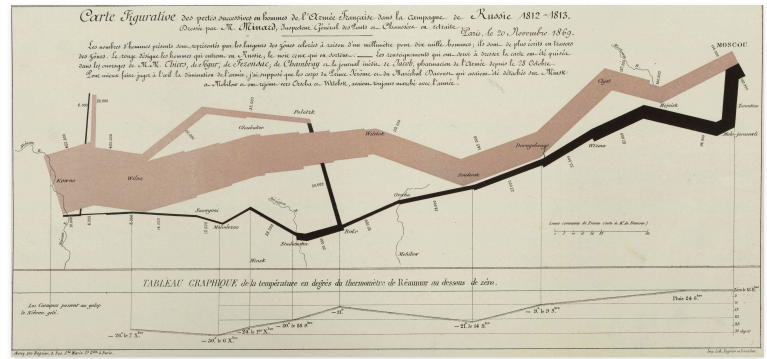


William Playfair's chart of wheat price and wage, 1784
(image in public domain)

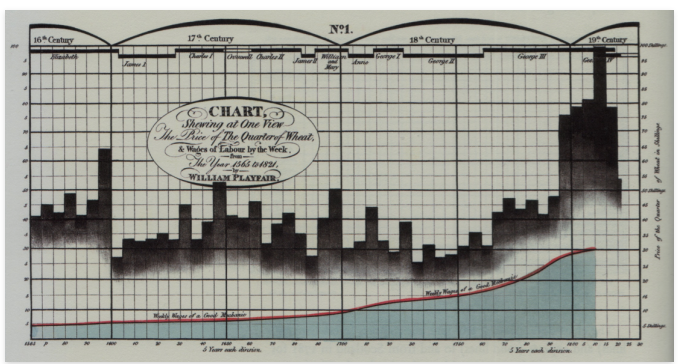
Historical Visualizations (or Old Visualizations)



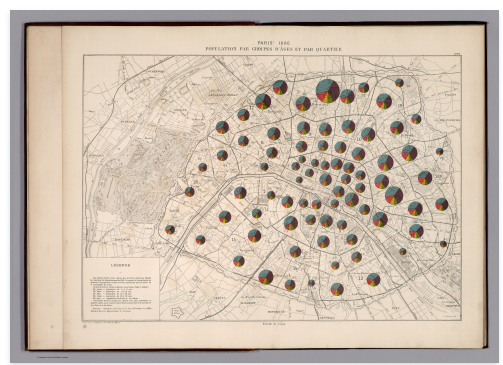
John Snow's cholera map, 1855
(image in public domain)



Charles Minard's map of Napoleon's Russian campaign, 1869
(image in public domain)



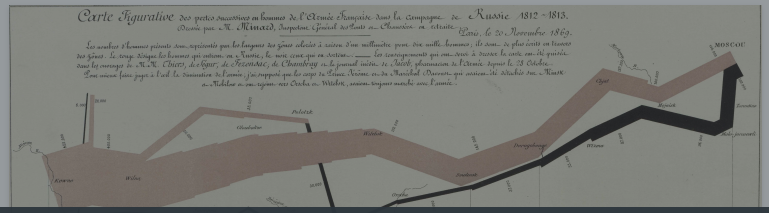
William Playfair's chart of wheat price and wage, 1784
(image in public domain)



Jacques Bertillon's thematic map with pie-chart glyphs, 1886
(image in public domain)

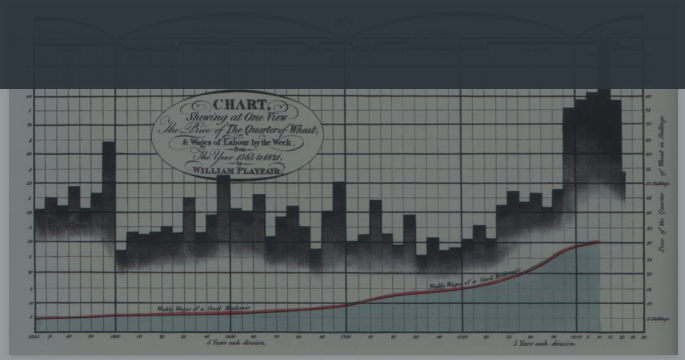
Historical Visualizations (or Old Visualizations)

John Snow's cholera map, 1855

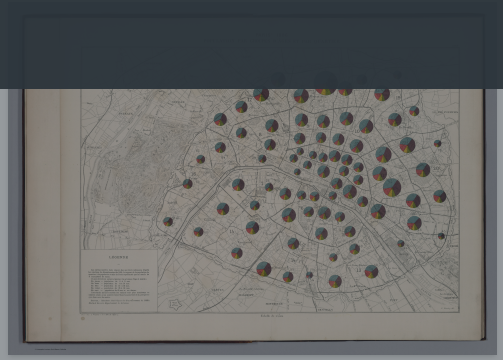


Charles Minard's map of Napoleon's Russian campaign, 1869
(image in public domain)

- How many historical visualizations have survived to date?
- What can we use historical visualizations for?



William Playfair's chart of wheat price and wage, 1784
(image in public domain)



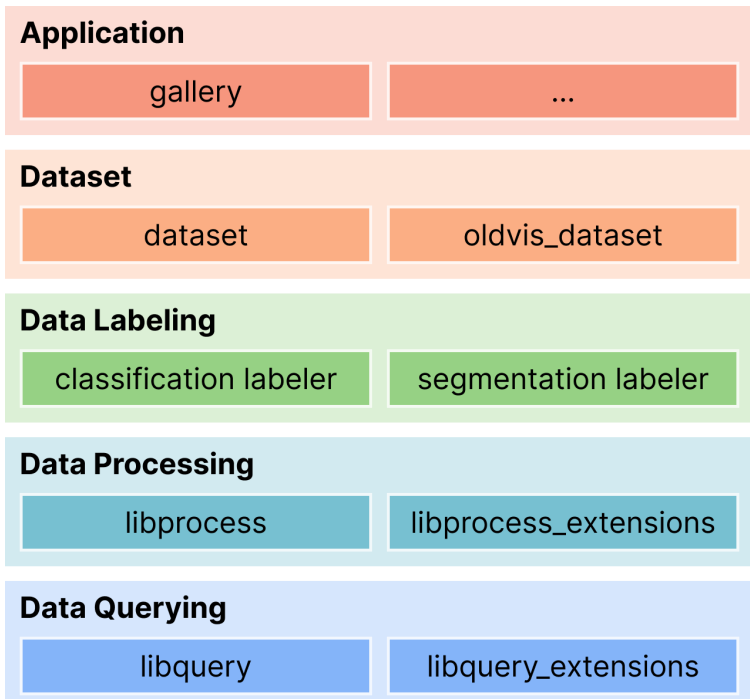
Jacques Bertillon's thematic map with pie-chart glyphs, 1886
(image in public domain)

How many historical visualizations have survived to date?

Finding historical visualizations (images + metadata)

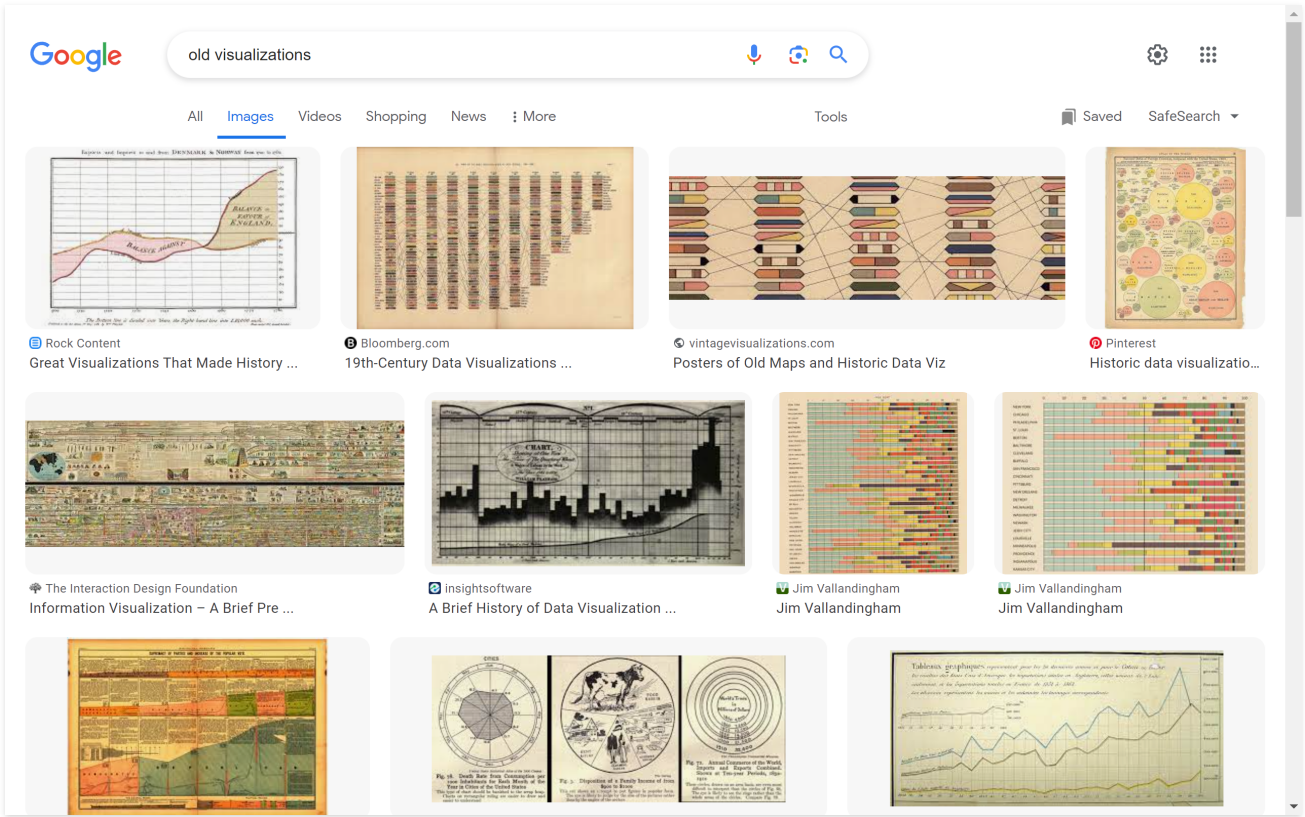
Overview

The OldVis project stack



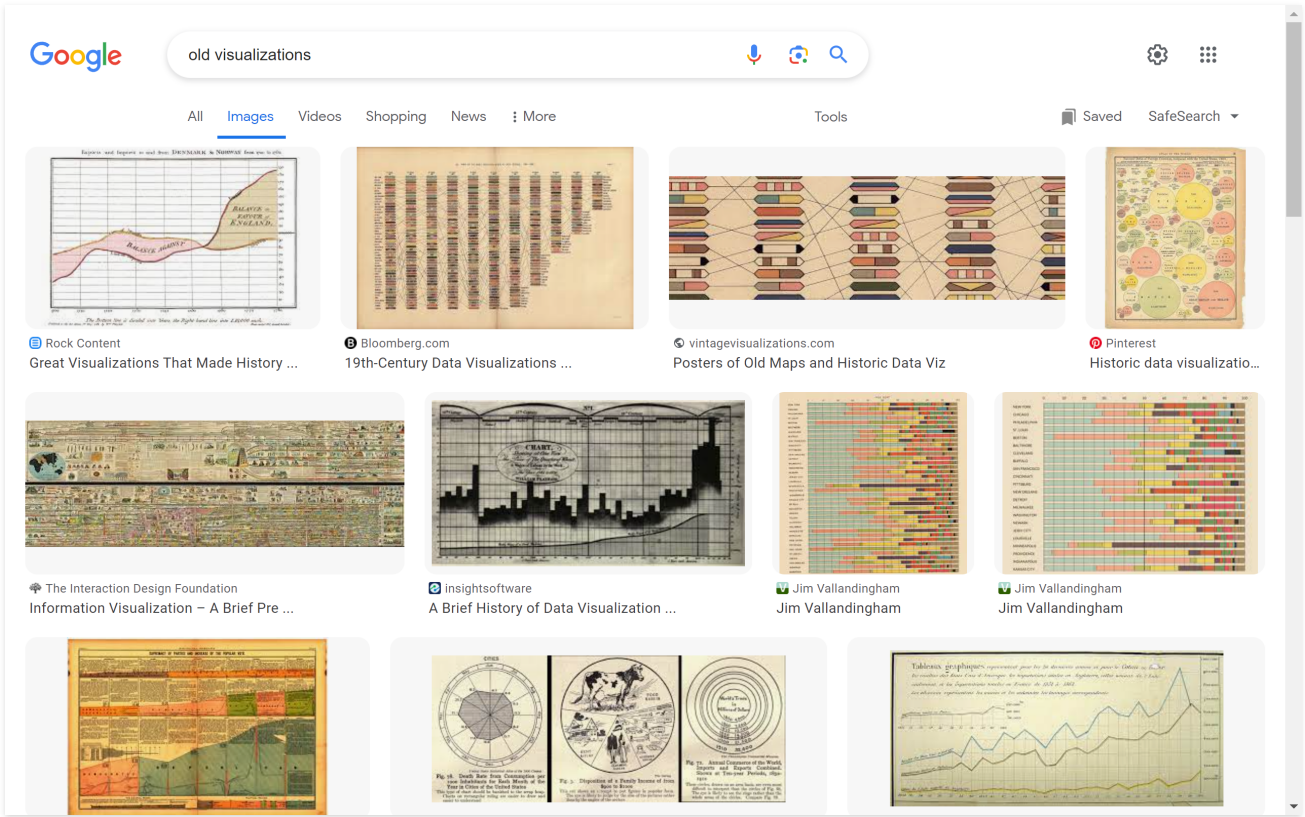
Data Source

General-purpose image search engines?



Data Source

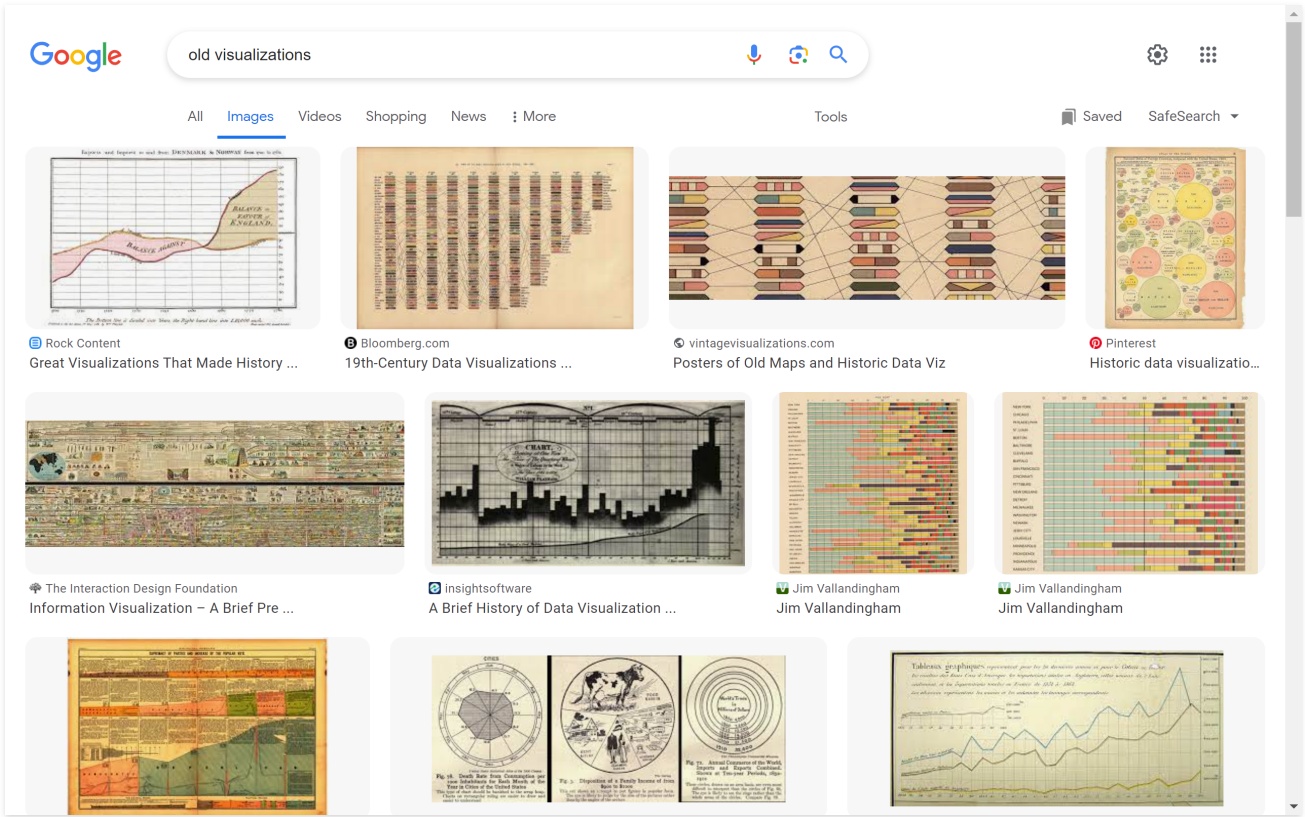
General-purpose image search engines?



- Lack metadata

Data Source

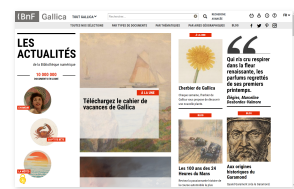
General-purpose image search engines?

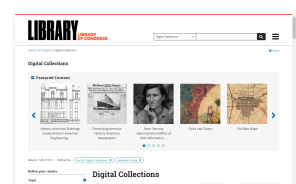


- Lack metadata
- Duplication

Digital Libraries

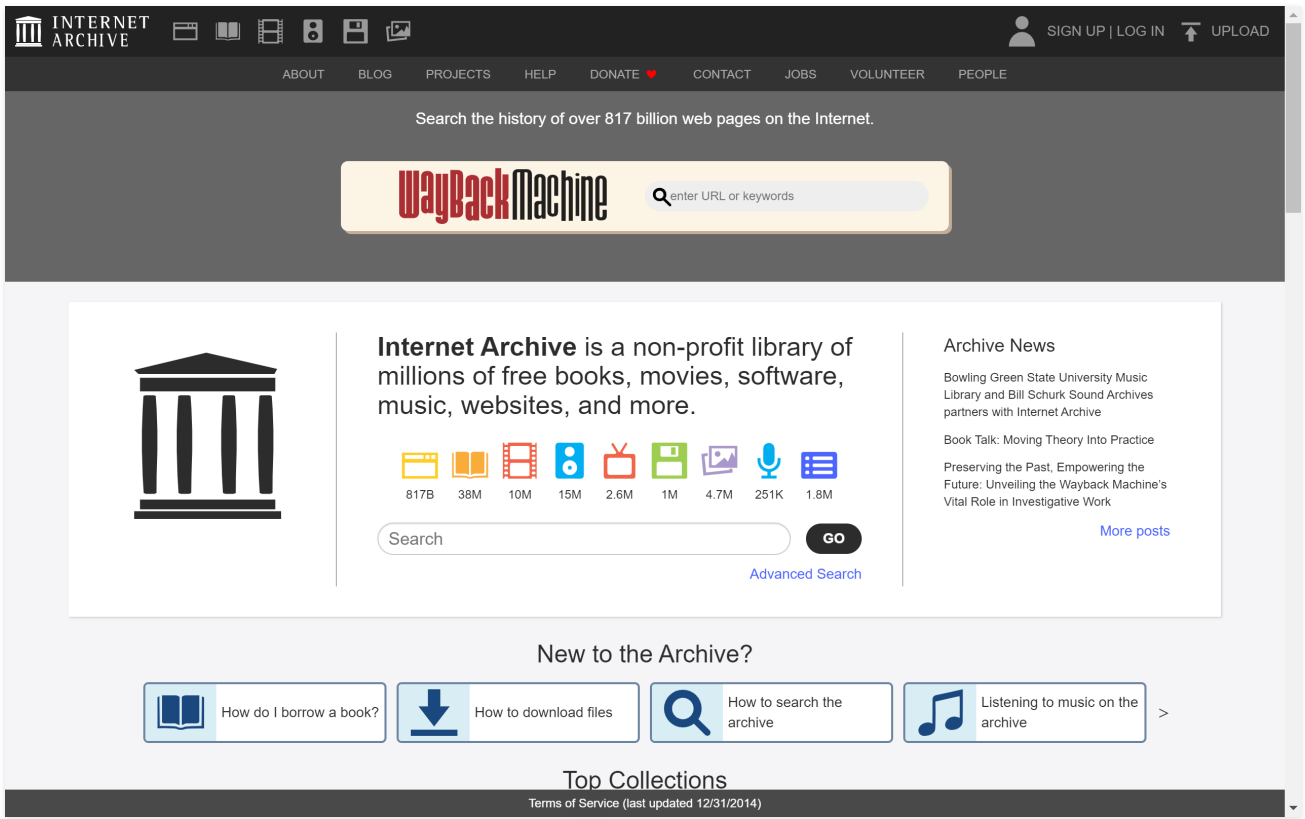
David Rumsey Map Collection (https://www.davidrumsey.com/)

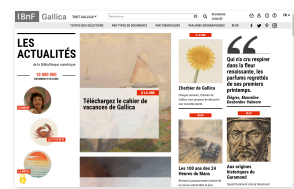


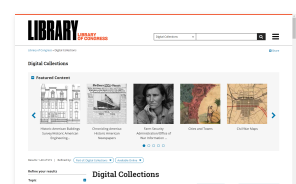


Digital Libraries

Internet Archive (https://archive.org/)

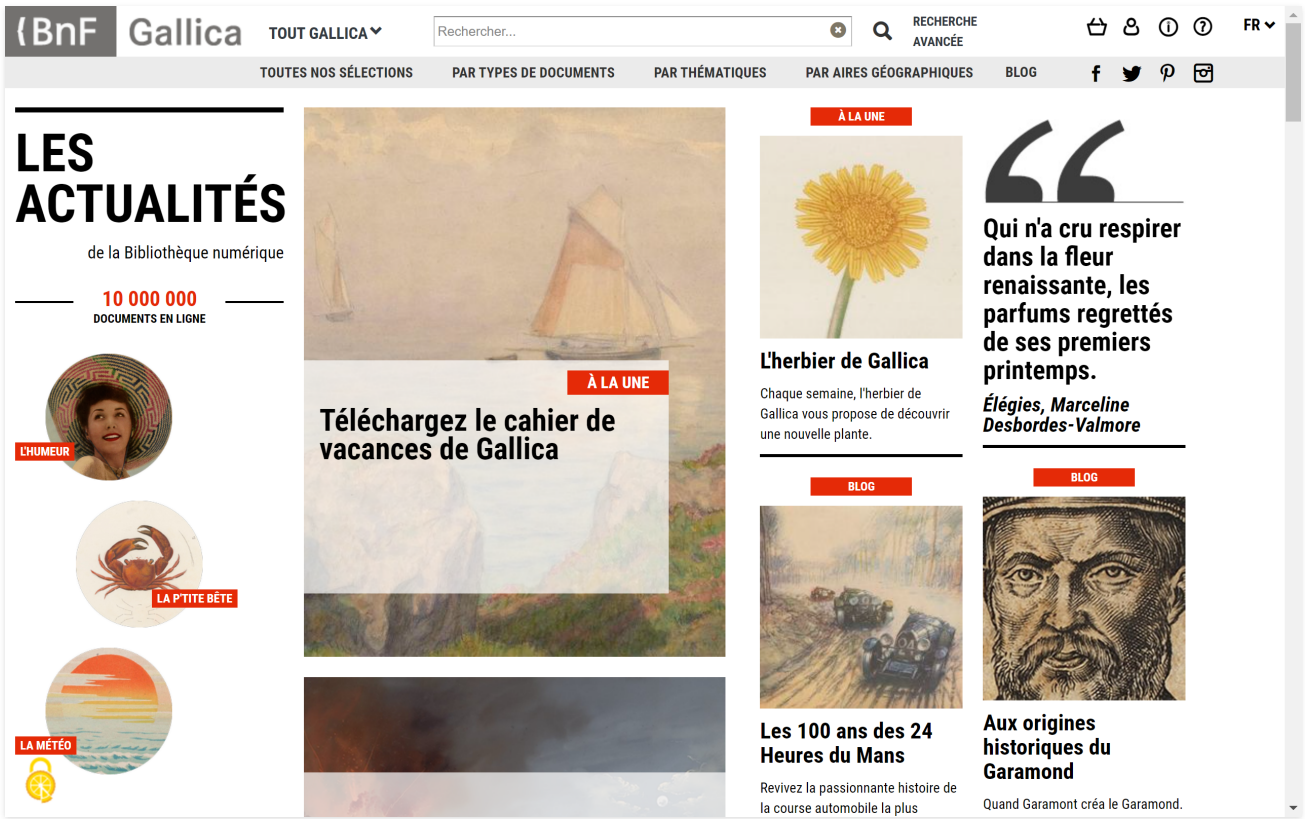


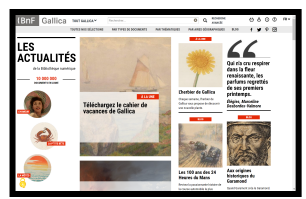


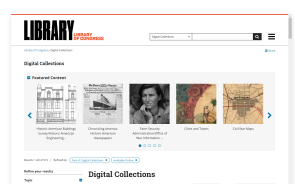


Digital Libraries

Gallica (https://gallica.bnf.fr/)

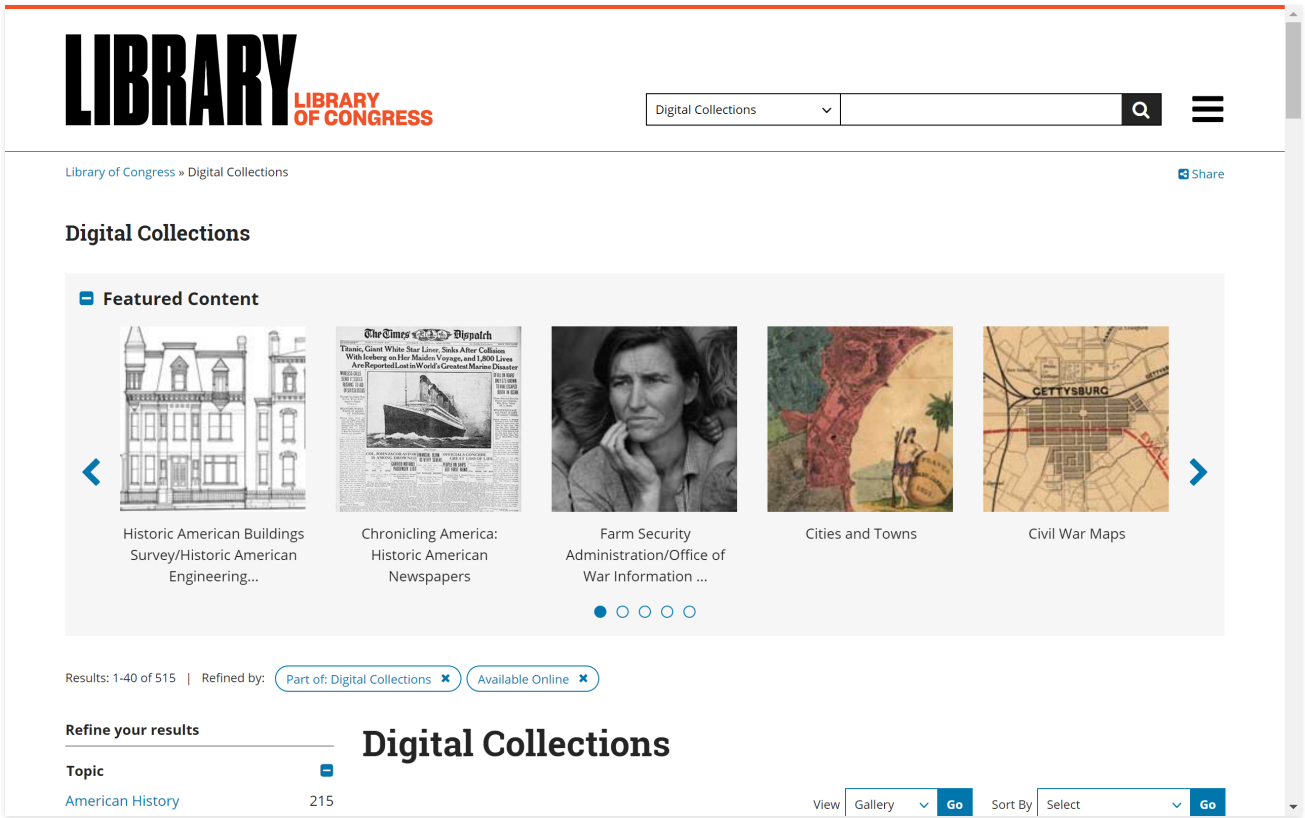


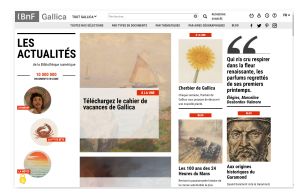


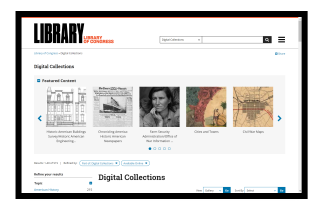


Digital Libraries

Library of Congress (https://www.loc.gov/collections/)







Data Obtaining

Querying, parsing, and cleaning

https://github.com/oldvis/libprocess

pypi v0.1.4 | code style black | commitizen friendly

Installation

```
pip install libprocess
```

Usage Example

```
from libprocess import DavidRumseyMapCollection

base_url = "https://www.davidrumsey.com"
service = "/luna/servlet/as/search?"
query = "q=type=chart"
querier = DavidRumseyMapCollection("./metadata/", "./imgs/")
querier.fetch_metadata([f"{base_url}{service}{query}"])
querier.process_metadata(save_path="processedMetadata.json")
```

Data Obtaining

Querying, parsing, and cleaning

https://github.com/oldvis/libprocess

pypi v0.1.4 | code style black | commitizen friendly

Installation

```
pip install libprocess
```

Usage Example

```
from libprocess import DavidRumseyMapCollection

base_url = "https://www.davidrumsey.com"
service = "/luna/servlet/as/search?"
query = "q=type=chart"
querier = DavidRumseyMapCollection("./metadata/", "./imgs/")
querier.fetch_metadata([f"{base_url}{service}{query}"])
querier.process_metadata(save_path="processedMetadata.json")
```

Data Obtaining

Querying, parsing, and cleaning

https://github.com/oldvis/libprocess

pypi v0.1.4 | code style black | commitizen friendly

Installation

```
pip install libprocess
```

Usage Example

```
from libprocess import DavidRumseyMapCollection

base_url = "https://www.davidrumsey.com"
service = "/luna/servlet/as/search?"
query = "q=type=chart"
querier = DavidRumseyMapCollection("./metadata/", "./imgs/")
querier.fetch_metadata([f"{base_url}{service}{query}"])
querier.process_metadata(save_path="processedMetadata.json")
```

Data Obtained

Number of images with corresponding metadata

Digital Collection	#Images
Gallica	195585
Internet Archive	82225
British Library	39312
Library of Congress	27742
David Rumsey Map Collection	25563
Telefact	264
Alabama Maps	142
Total	**370883**

Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

- Bibliographic information

Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

- Bibliographic information
- Image file properties

Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

- Bibliographic information
- Image file properties
- Copyright information

Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

- Bibliographic information
- Image file properties
- Copyright information

Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

- Bibliographic information
- Image file properties
- Copyright information

Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

- Bibliographic information
- Image file properties
- Copyright information

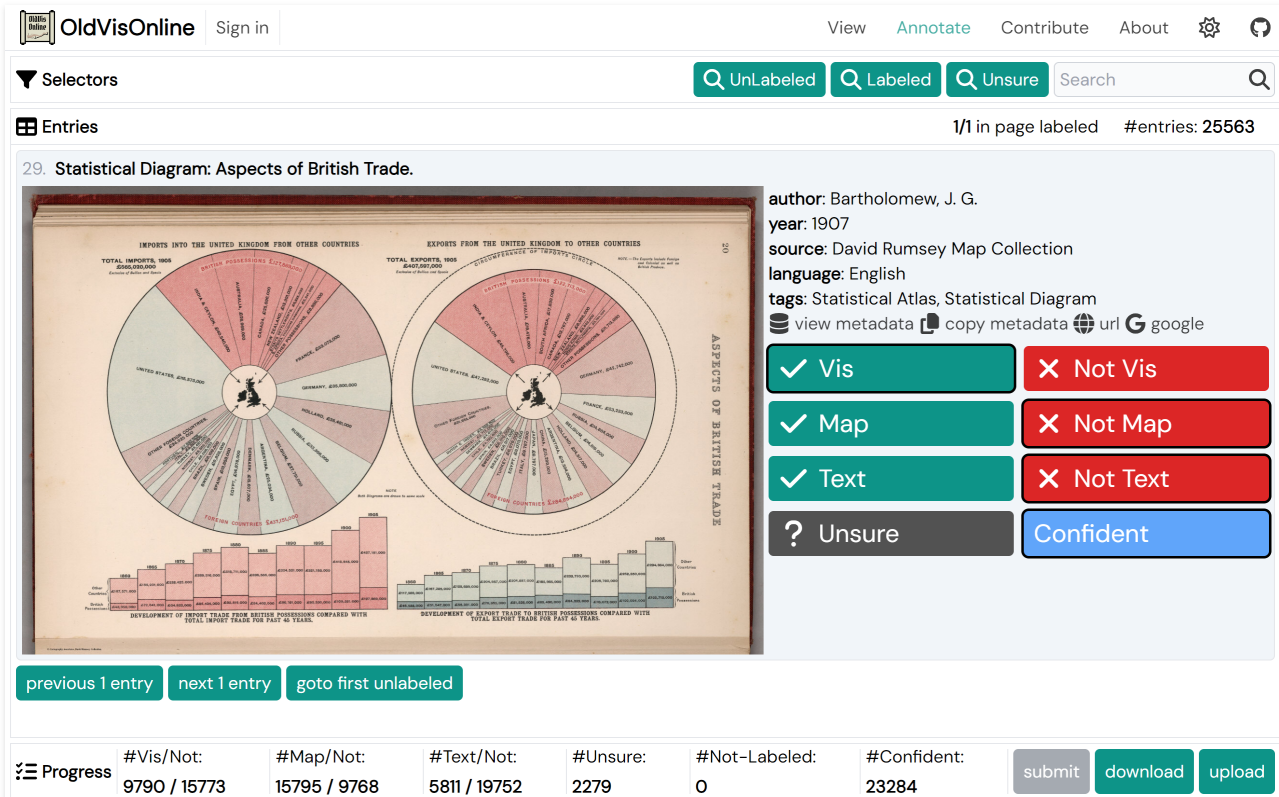
Metadata Schema

```
interface Entry {
  uuid: string
  authors: string[]
  displayName: string
  publishDate: Date | [Date, Date],
  downloadUrl: string
  viewUrl: string
  languages: string[]
  tags: string[]
  abstract: string
  md5: string
  phash: string
  resolution: [number, number]
  fileSize: number
  rights: string
  source: {
    name: string
    url: string
    accessDate: Date
  }
}
```

- Bibliographic information
- Image file properties
- Copyright information

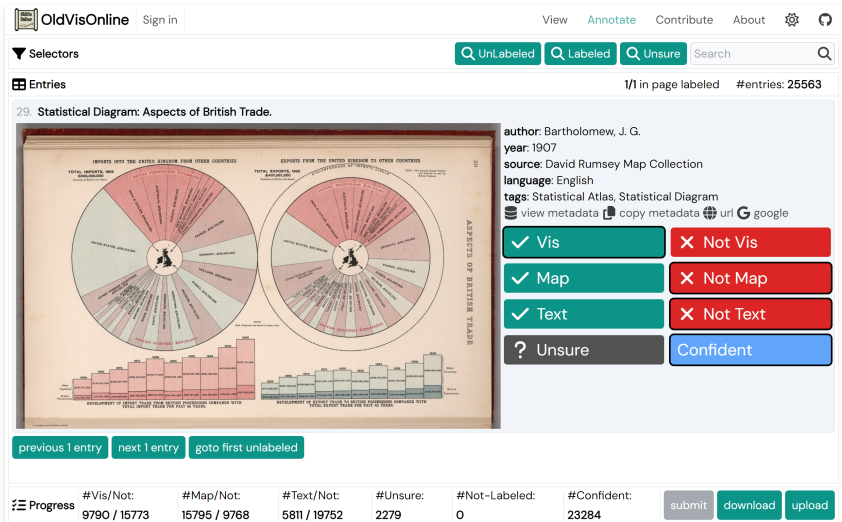
Labeling System

Distinguish visualization images from other images



Labeling System

Distinguish visualization images

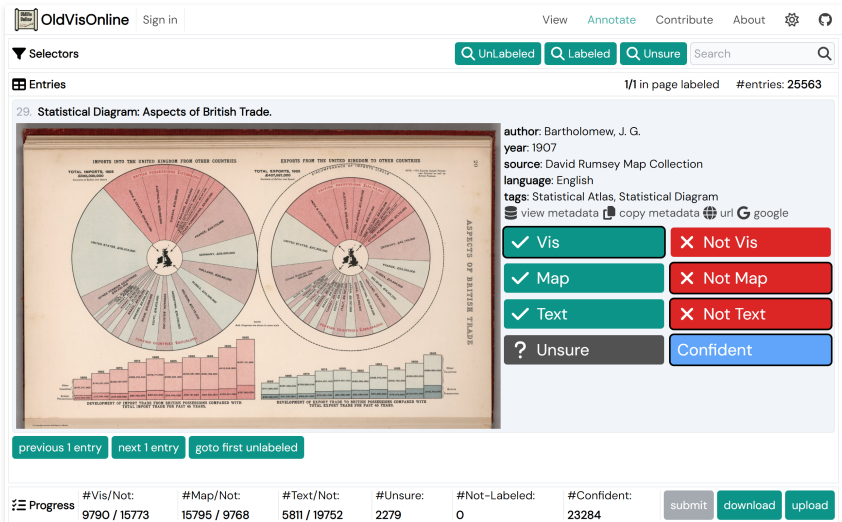


Semi-Automatic Labeling

The workflow

Labeling System

Distinguish visualization images



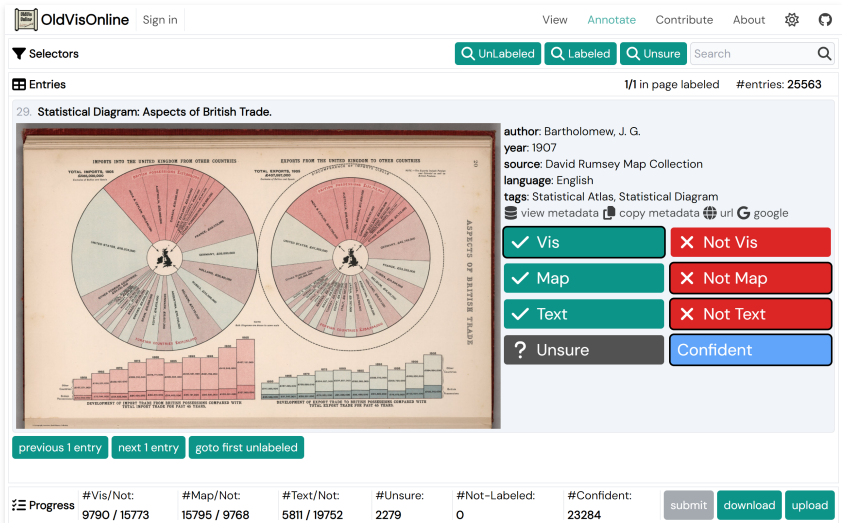
Semi-Automatic Labeling

The workflow

1. Manually annotate 25K images

Labeling System

Distinguish visualization images



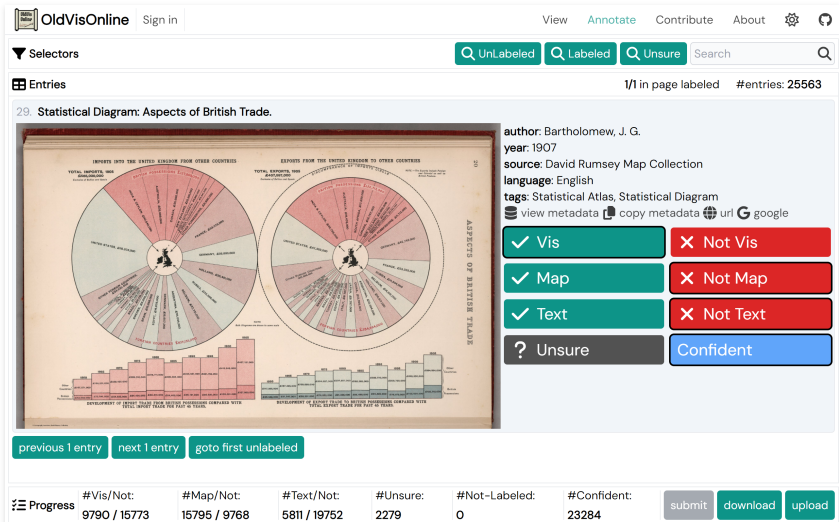
Semi-Automatic Labeling

The workflow

1. Manually annotate 25K images
2. Finetune VGG-16 for text/map/vis detection

Labeling System

Distinguish visualization images



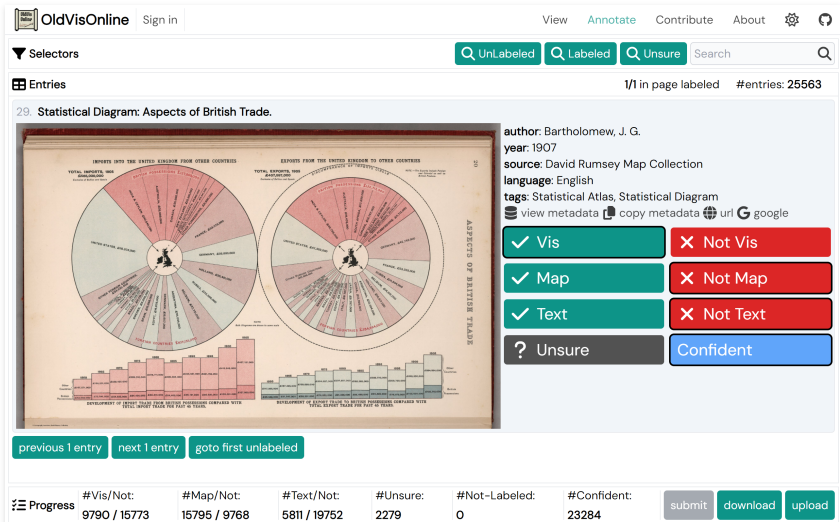
Semi-Automatic Labeling

The workflow

1. Manually annotate 25K images
2. Finetune VGG-16 for text/map/vis detection
3. Predict on the remaining 345K images

Labeling System

Distinguish visualization images



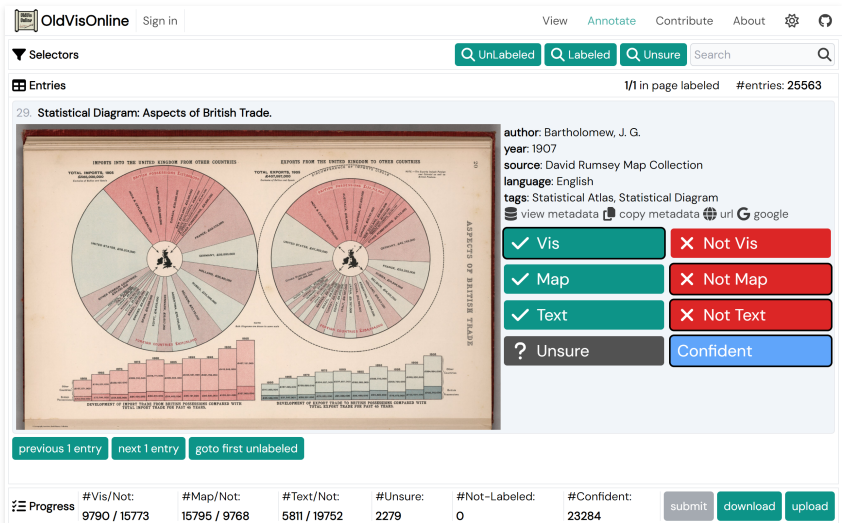
Semi-Automatic Labeling

The workflow

1. Manually annotate 25K images
2. Finetune VGG-16 for text/map/vis detection
3. Predict on the remaining 345K images
4. Manually screen and correct the predictions

Labeling System

Distinguish visualization images



Semi-Automatic Labeling

The workflow

1. Manually annotate 25K images
2. Finetune VGG-16 for text/map/vis detection
3. Predict on the remaining 345K images
4. Manually screen and correct the predictions
5. Filter visualizations by publish date

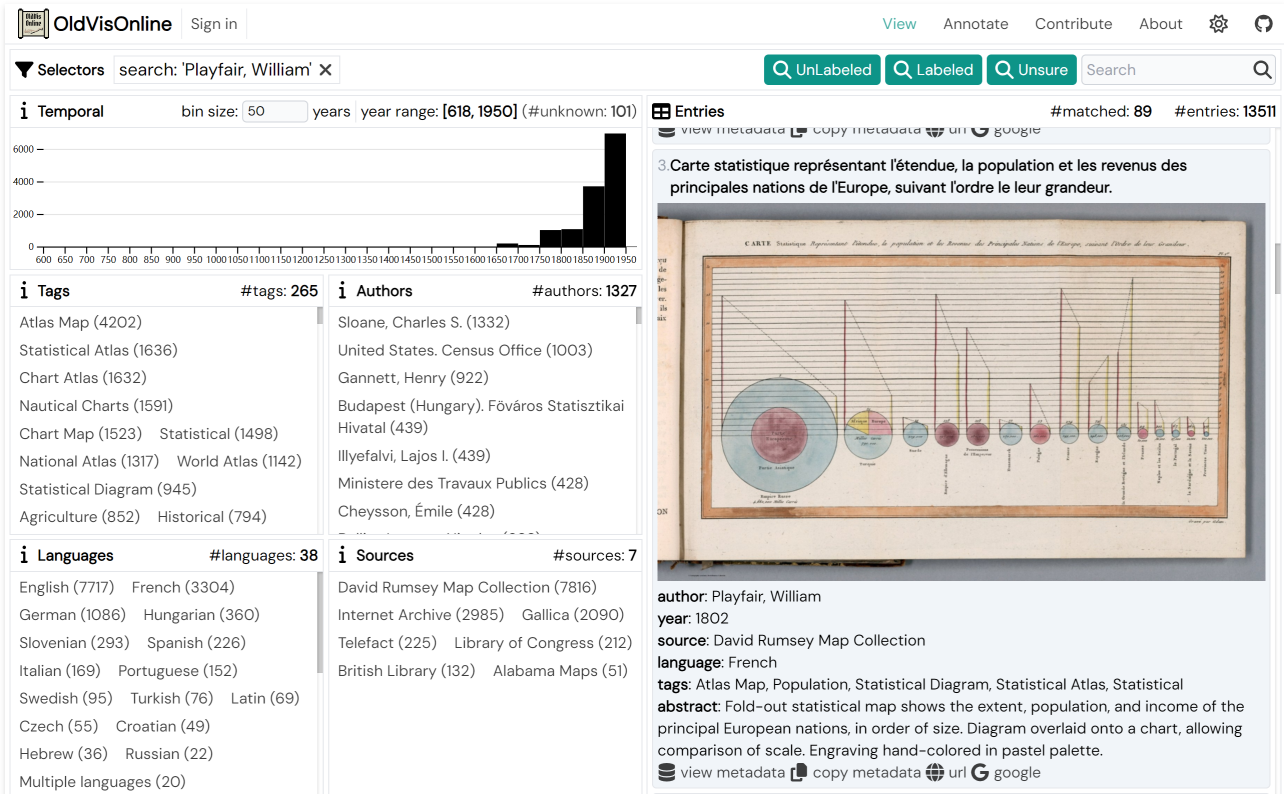
Outcome

13K images with metadata

Digital Library	#Images	#Vis	Before 1950
David Rumsey Map Collection	25563	9790	7816
Internet Archive	82225	6885	2985
Gallica	195585	2183	2090
Library of Congress	27742	299	212
Telefact	264	225	225
British Library	39312	137	132
Alabama Maps	142	52	51
Total	370883	19571	**13511**

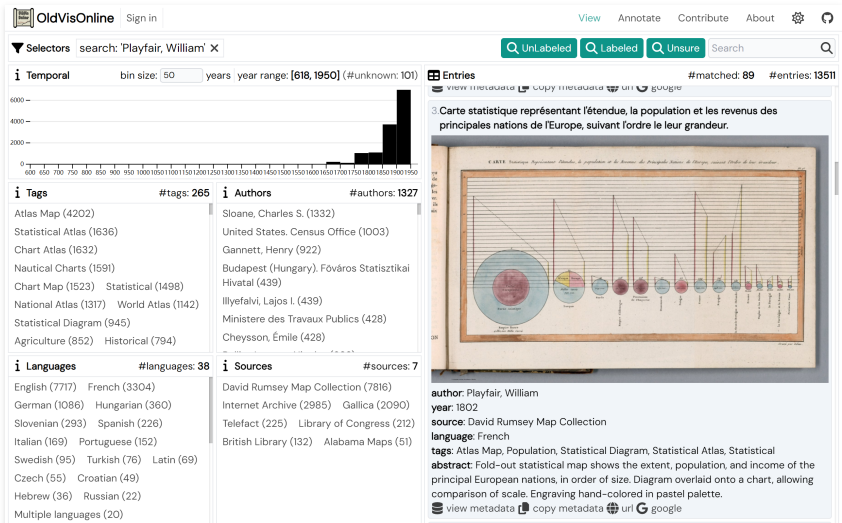
Online Gallery

https://github.com/oldvis/gallery



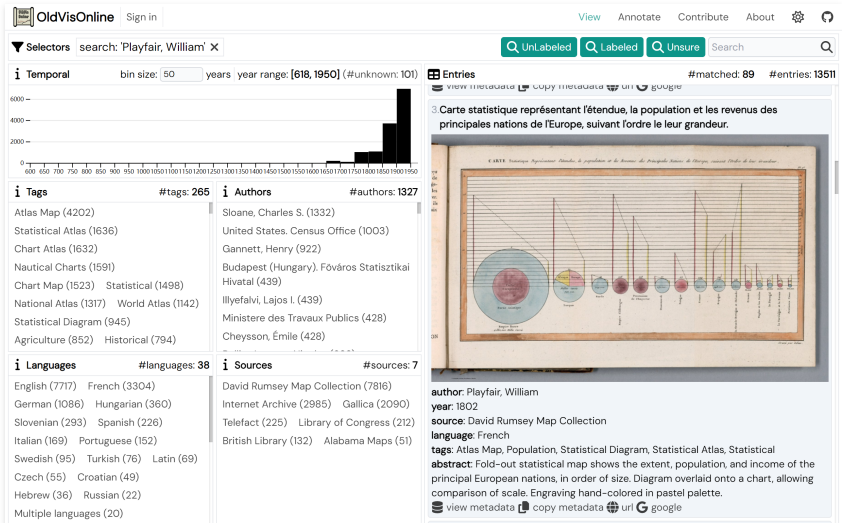
Online Gallery

https://github.com/oldvis/gallery



Online Gallery

https://github.com/oldvis/gallery



Dataset Downloading

https://github.com/oldvis/oldvis_dataset

pypi v0.1.1 | code style black | commitizen friendly

Installation

```
pip install oldvis_dataset
```

Usage Example

```
from oldvis_dataset import visualizations, fetch_images

visualizations.download("./visualizations.json")
fetch_images("./visualizations.json", "./images/")
```

What can we use historical visualizations for?

Usage scenarios of historical visualizations

Inspiration for Contemporary Design

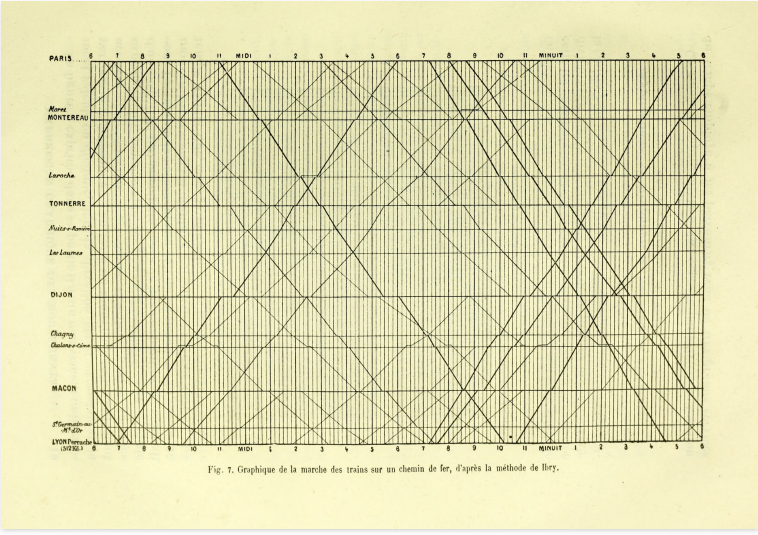


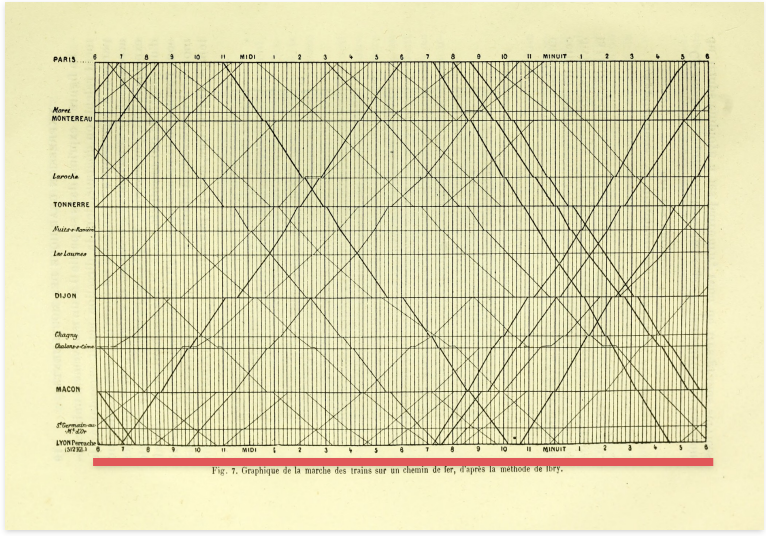
Fig. 7. Graphique de la marche des trains sur un chemin de fer, d'après la méthode de Ibry.

Étienne-Jules Marey's schedule chart[1]
(image in public domain ⊘)

[1] **La Méthode Graphique dans les Sciences Expérimentales et Principalement en Physiologie et en Médecine**, Étienne-Jules Marey, 1878

[2] **ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories**, Panpan Xu et al., IEEE Transactions on Visualization and Computer Graphics, 2017

Inspiration for Contemporary Design



Étienne-Jules Marey's schedule chart[1]
(image in public domain ⊘)

[1] **La Méthode Graphique dans les Sciences Expérimentales et Principalement en Physiologie et en Médecine**, Étienne-Jules Marey, 1878

[2] **ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories**, Panpan Xu et al., IEEE Transactions on Visualization and Computer Graphics, 2017

Inspiration for Contemporary Design

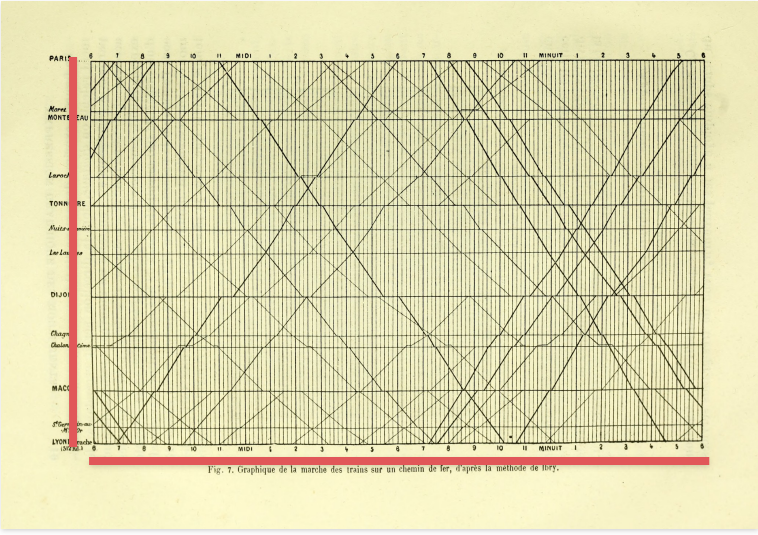


Fig. 7. Graphique de la marche des trains sur un chemin de fer, d'après la méthode de Iory.

Étienne-Jules Marey's schedule chart[1]
(image in public domain ⊘)

[1] **La Méthode Graphique dans les Sciences Expérimentales et Principalement en Physiologie et en Médecine**, Étienne-Jules Marey, 1878

[2] **ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories**, Panpan Xu et al., IEEE Transactions on Visualization and Computer Graphics, 2017

Inspiration for Contemporary Design

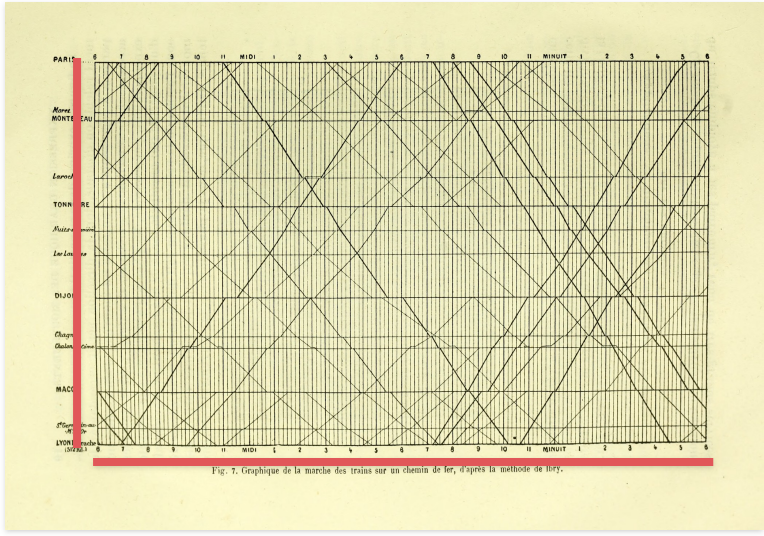
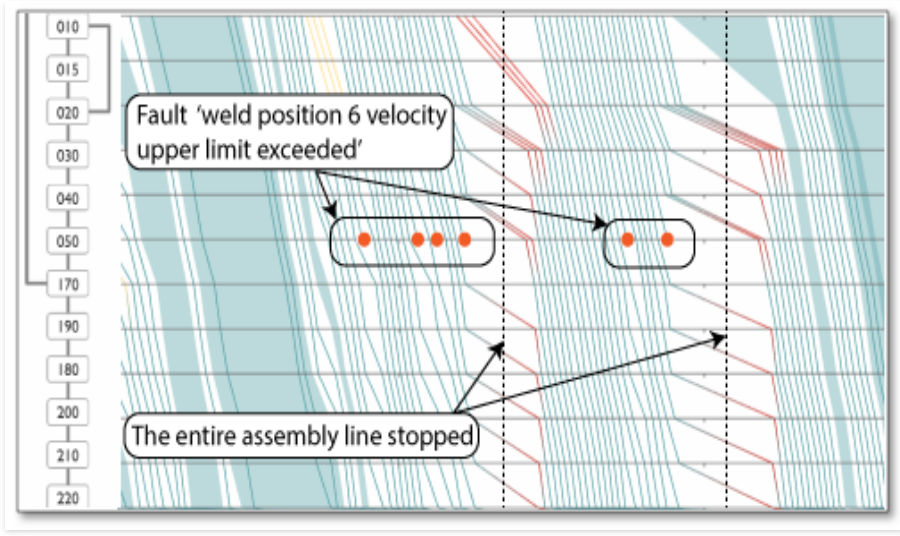


Fig. 7. Graphique de la marche des trains sur un chemin de fer, d'après la méthode de Mary.

Étienne-Jules Marey's schedule chart[1]
(image in public domain ⊜)

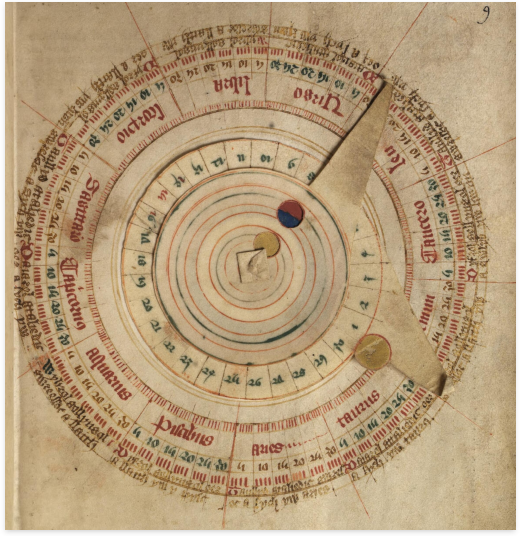


Xu et al.'s redesign[2]
(image © 2017 IEEE)

[1] **La Méthode Graphique dans les Sciences Expérimentales et Principalement en Physiologie et en Médecine**, Étienne-Jules Marey, 1878

[2] **ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories**, Panpan Xu et al., IEEE Transactions on Visualization and Computer Graphics, 2017

Inspiration for Contemporary Design

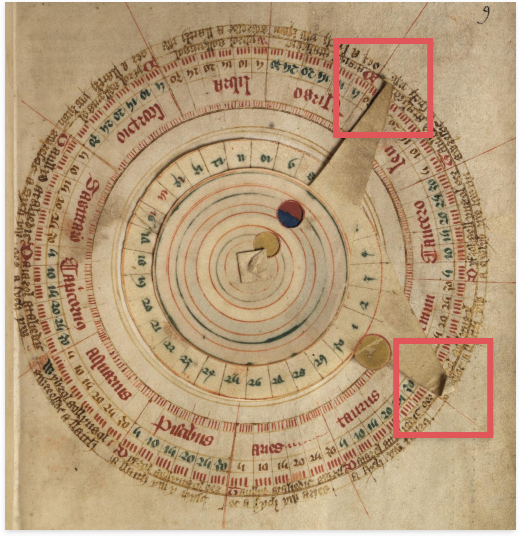


Gutun Owain's volvelle[1]
(image in public domain ⓒ)

[1] **Hen Almanac Cymreig & C.**, Gutun Owain, 1489

[2] **Vol²velle: Printable Interactive Volume Visualization**, Sergej Stoppel and Stefan Bruckner, IEEE Transactions on Visualization and Computer Graphics, 2017

Inspiration for Contemporary Design

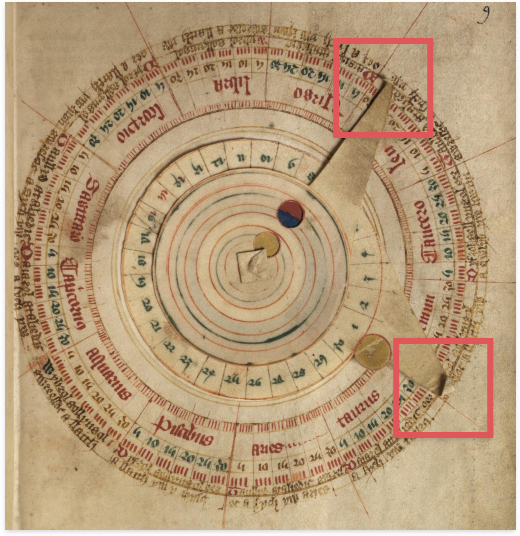


Gutun Owain's volvelle[1]
(image in public domain ⊘)

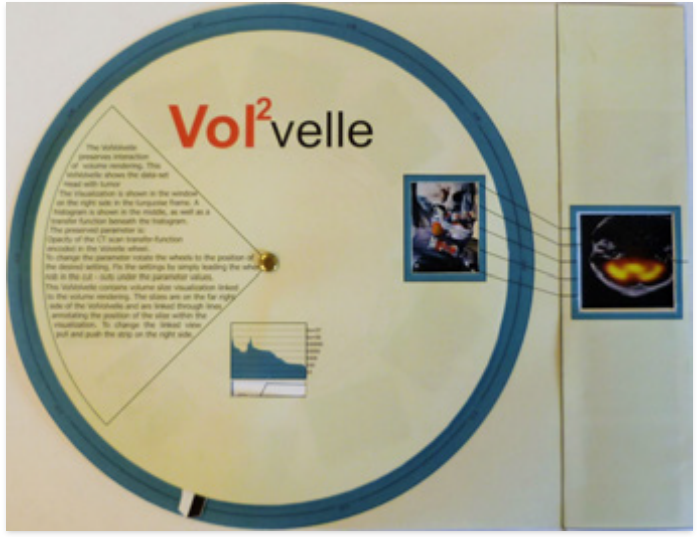
[1] **Hen Almanac Cymreig & C.**, Gutun Owain, 1489

[2] **Vol²velle: Printable Interactive Volume Visualization**, Sergej Stoppel and Stefan Bruckner, IEEE Transactions on Visualization and Computer Graphics, 2017

Inspiration for Contemporary Design



Gutun Owain's volvelle[1]
(image in public domain ⓪)

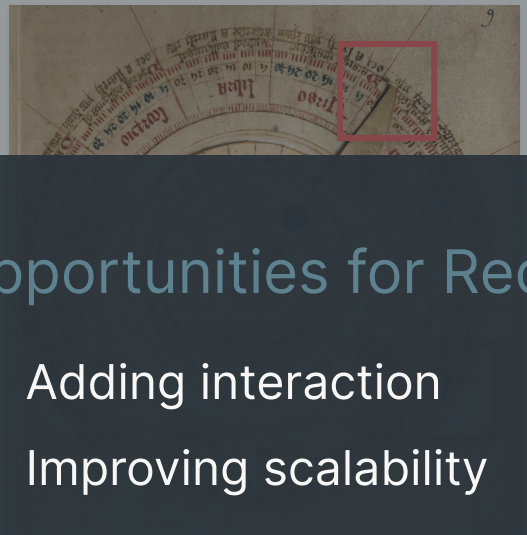


Stoppel and Bruckner's redesign[2]
(image © 2017 IEEE)

[1] **Hen Almanac Cymreig & C.**, Gutun Owain, 1489

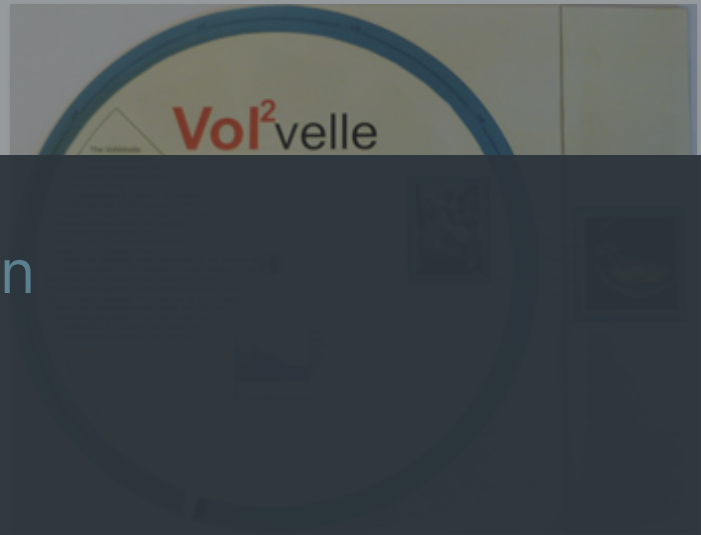
[2] **Vol²velle: Printable Interactive Volume Visualization**, Sergej Stoppel and Stefan Bruckner, IEEE Transactions on Visualization and Computer Graphics, 2017

Inspiration for Contemporary Design



Gutun Owain's volvelle[1]

(image in public domain)



Stoppel and Bruckner's redesign[2]

(image © 2017 IEEE)

Opportunities for Redesign

- Adding interaction
- Improving scalability

[1] **Hen Almanac Cymreig & C.**, Gutun Owain, 1489

[2] **Vol²velle: Printable Interactive Volume Visualization**, Sergej Stoppel and Stefan Bruckner, IEEE Transactions on Visualization and Computer Graphics, 2017

Textual Criticism

How many cholera cases are plotted on John Snow's Map[1]?

Textual Criticism

How many cholera cases are plotted on John Snow's Map[1]?

- 579[2]
- 578[3, 4]

[1] **On the Mode of Communication of Cholera**, John Snow, 1855

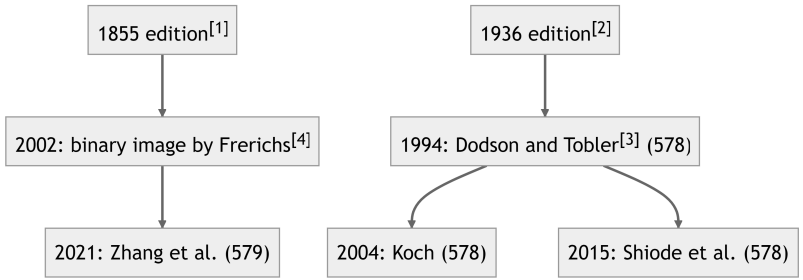
[2] **MI3: Machine-Initiated Intelligent Interaction for Interactive Classification and Data Reconstruction**, Yu Zhang et al., ACM Transactions on Interactive Intelligent Systems, 2021

[3] **The Map as Intent: Variations on the Theme of John Snow**, Tom Koch, Cartographica, 2004

[4] **The mortality rates and the space-time patterns of John Snow's cholera epidemic map**, Narushige Shiode et al., International Journal of Health Geographics, 2015

Textual Criticism

The lineage of the data sources



[1] **On the Mode of Communication of Cholera**, John Snow, page 44-45, 1855

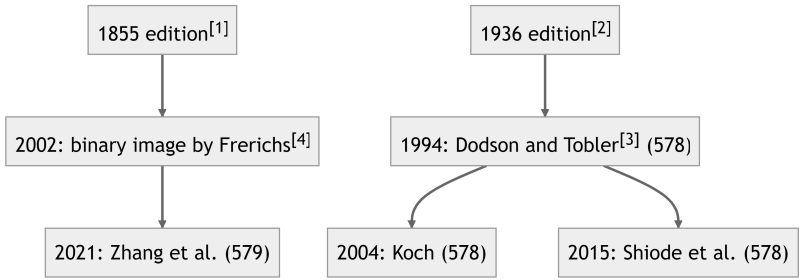
[2] **Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D.**, John Snow, page 44-45, 1936

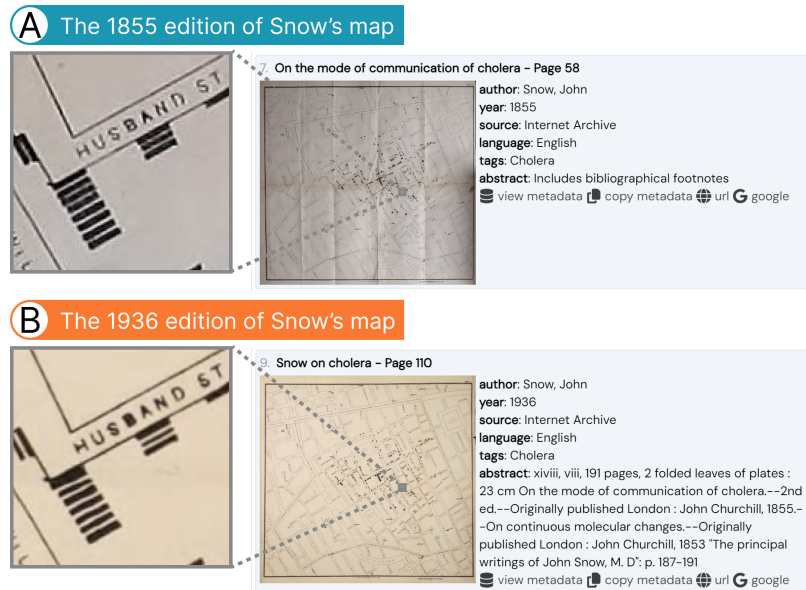
[3] **Snow's Cholera Map**, Rusty Dodson and Waldo Tobler, 1994

[4] **John Snow - a historical giant in epidemiology**, Ralph R. Frerichs, 2002

Textual Criticism

The lineage of the data sources





[1] **On the Mode of Communication of Cholera**, John Snow, page 44-45, 1855

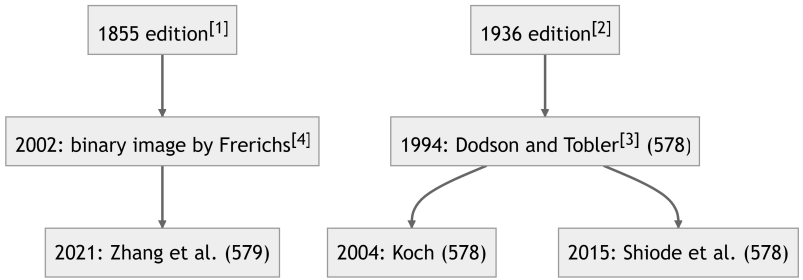
[2] **Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D.**, John Snow, page 44-45, 1936

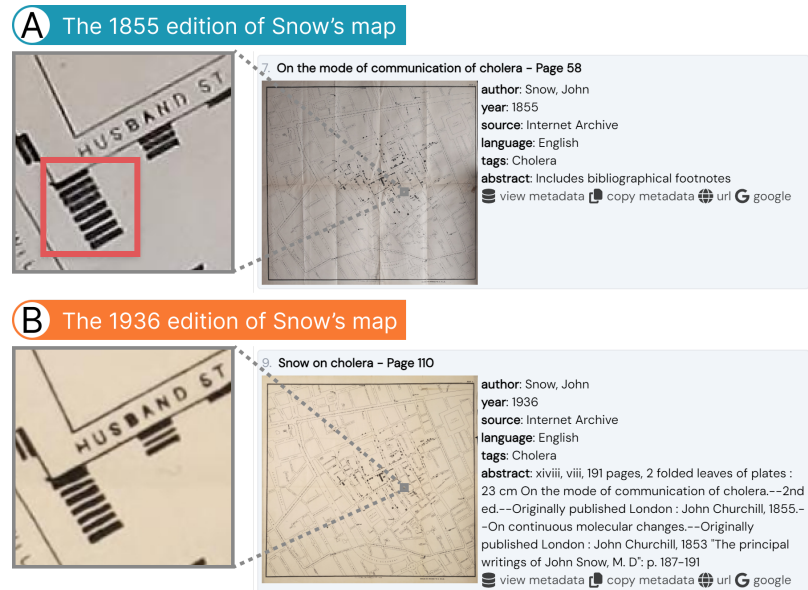
[3] **Snow's Cholera Map**, Rusty Dodson and Waldo Tobler, 1994

[4] **John Snow - a historical giant in epidemiology**, Ralph R. Frerichs, 2002

Textual Criticism

The lineage of the data sources





- 1855 edition[1]: **8** in the highlighted area (**579** in total)

[1] **On the Mode of Communication of Cholera**, John Snow, page 44-45, 1855

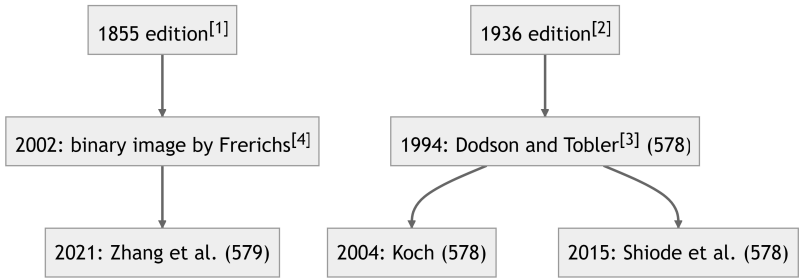
[2] **Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D.**, John Snow, page 44-45, 1936

[3] **Snow's Cholera Map**, Rusty Dodson and Waldo Tobler, 1994

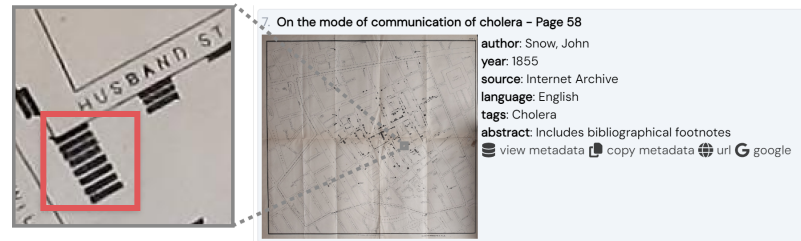
[4] **John Snow - a historical giant in epidemiology**, Ralph R. Frerichs, 2002

Textual Criticism

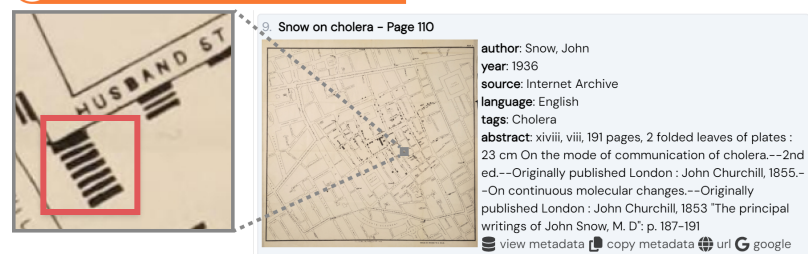
The lineage of the data sources



A The 1855 edition of Snow's map



B The 1936 edition of Snow's map



- 1855 edition[1]: **8** in the highlighted area (**579** in total)
- 1936 edition[2]: **7** in the highlighted area (**578** in total)

[1] **On the Mode of Communication of Cholera**, John Snow, page 44-45, 1855

[2] **Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D.**, John Snow, page 44-45, 1936

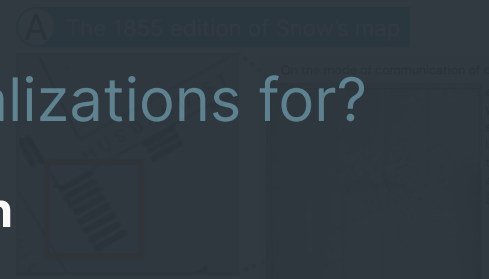
[3] **Snow's Cholera Map**, Rusty Dodson and Waldo Tobler, 1994

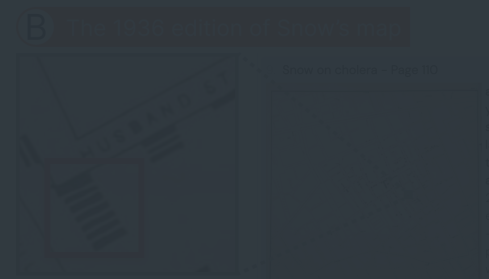
[4] **John Snow - a historical giant in epidemiology**, Ralph R. Frerichs, 2002

Textual Criticism

What can we use historical visualizations for?

- **Inspiration for contemporary design**
- **Textual criticism**
- Historical data extraction
- Design space analysis
- Metadata analysis
- Metadata research





- 1855 edition[1]: **8** in the highlighted area (**579** in total)
- 1936 edition[2]: **7** in the highlighted area (**578** in total)

[1] **On the Mode of Communication of Cholera**, John Snow, page 44-45, 1855

[2] **Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D.**, John Snow, page 44-45, 1936

[3] **Snow's Cholera Map**, Rusty Dodson and Waldo Tobler, 1994

[4] **John Snow - a historical giant in epidemiology**, Ralph R. Frerichs, 2002

Textual Criticism

What can we use historical visualizations for?

- **Inspiration for contemporary design**
- **Textual criticism**
- Historical data extraction
- Design space analysis
- Metadata analysis
- Metadata research

[1] **On the Mode of Communication of Cholera**, John Snow, page 44-45, 1855

[2] **Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D.**, John Snow, page 44-45, 1936

[3] **Snow's Cholera Map**, Rusty Dodson and Waldo Tobler, 1994

[4] **John Snow - a historical giant in epidemiology**, Ralph R. Frerichs, 2002

Textual Criticism

What can we use historical visualizations for?

- **Inspiration for contemporary design**
- **Textual criticism**
- Historical data extraction
- Design space analysis
- Metadata analysis
- Metadata research

[1] **On the Mode of Communication of Cholera**, John Snow, page 44-45, 1855

[2] **Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D.**, John Snow, page 44-45, 1936

[3] **Snow's Cholera Map**, Rusty Dodson and Waldo Tobler, 1994

[4] **John Snow - a historical giant in epidemiology**, Ralph R. Frerichs, 2002

Summary

- A dataset of 13K historical visualizations
  - Seven digital collections as data sources
  - Packages for data collection and processing
  - Semi-automatic labeling
- Usage scenarios
  - Inspiration for contemporary design
  - Textual criticism
  - And more in our paper

Summary

- A dataset of 13K historical visualizations
  - Seven digital collections as data sources
  - Packages for data collection and processing
  - Semi-automatic labeling
- Usage scenarios
  - Inspiration for contemporary design
  - Textual criticism
  - And more in our paper

Dataset & Code: github.com/oldvis

Paper: arxiv.org/abs/2308.16053