

DH DAY // 27.01.2026



TRANSCRIPTION TECHNIQUES

Approaches, tools,
and challenges

Rafael Uriarte
Alessandra Failla

Source materials and challenges

Handwritten Documents

Inconsistent handwriting styles, archaic scripts, and ink bleeding through paper.

Printed Documents

Faded or blurry ink, non-standard historical fonts, complex layout, and physical damage to pages.

Annotated Material

Distinguishing between original printed text and handwritten notes that may overlap or run into the margins.

PISA.

Camposanto.

28149	Andrea da Firenze (Bonaiuti). La conversione di S. Ranieri.	N
28150	— La partenza di S. Ranieri per la Palestina.	N
28151	— Miracoli e tentazioni di S. Ranieri.	N
28152	Antonio Veneziano. Ritorno a Pisa e miracoli di S. Ranieri . . .	N
28153	— — Particolare: S. Ranieri smaschera l'oste	N
28154	— — — S. Ranieri invitato a desinare dai canonici del Duomo.	N
* 28155	— — — Tre teste di personaggi	N
* 28156	— — — Un Santo nel fregio dell'affresco.	N
28157	— La morte di S. Ranieri (frammento)	N
* 28158	— — Particolare: Testa di donna	N
28159	Bonamico Buffalmacco (ora attribuito a Francesco da Volterra e aiuti). L'Ascensione (parte superiore)	N
28160	— La Crocifissione.	N
28161	— La Resurrezione	N
28162	— L'incredulità di S. Tommaso	N
28163	Francesco da Volterra (ora attribuito a Giovanni di Taddeo Gaddi). Le sventure di Giobbe.	N
28164	— — Particolare: Parte inferiore sinistra	N
28165	— — — Parte inferiore destra.	N
28166	— — — Due teste	N
28167	— — — Due te	N
28168	— — — Tre te	N
28169	— Convito di Giobbe.	N
28170	— Dio permette a Luc	N
28171	— — Particolare: I	N
28172	— — — I Sabe	N
28173	Giotto (Scuola). Framment	N
28174	— — con due teste	N
28175	— — con mezza fig	N
28176	Gozzoli Benozzo. La vend	N
28177	— — Particolare:	N
28178	— — — Le nu	N

Datum	Ino. N.	Künstler	Gegenstand	Ort	Signatur
1000		Perugino	Handl. zw. Hebe u. Kriemhild	Paris, Louvre	Ma XV
1001		Loggion, Costa	Mythol. Szene	?	?
1002		Manegna	Parnass	?	?
1003		Giov. Bellin	Mda m. 2 Hlg.	Petersburg Eremit.	?
1004			Adam	Rom, Gall. Borghese	
1005		?	Eva	?	?
1006		Giorgione	Die Feldmesser	Wien, Belvedere	Ma XVI
1007		?	Judith	Petersburg Eremit.	?
1008		?	Mda m. 2 Hlg.	Madrid, Prado	?
1009		Rondinelli	Madonna	Rom, Gall. Doria	?
1010		?	Mda m. Joh.	?	?
1011		Palma Vecchio	Mda, 2 Hlg., Stifter	Chromkoll.	?
1012		?	Mda m. Hlgem	Petersburg Eremit.	?
1013		?	Ank. d. d. Hlgem	?	?
1014		Isauro Bassano	Handl. d. d. Hlgem	Madrid, Prado	?
1015		Francesco di St. Croce	Heilich. vor Christus	Petersburg Eremit.	?
1016		Filippo Mazzola	Porträt e. Maynes	Rom, Gall. Doria	?
1017			Porträt e. Hannes	?	?
1018		Andrea del Sarto	Disputation	Florenz, Pitti	?
1019		Botticelli	Trauerndes Mädch.	Rom	?
1020		Bernardino Licinio	Porträt e. Fran	Dresden, Gall.	Ma XVII

Sources: Anderson,
Domenico: "Fotografie
di D. Anderson.
Catalogo X. Pisa -
Lucca - Arezzo. E
Dintorni: S. Sepolcro -
Città di Castello, ecc.",
1928, p. 5 (above);
Photothek of the KHI,
Inventarbuch I (right).

Fundamentals: Layout analysis and Segmentation

Layout Analysis (Region Detection)

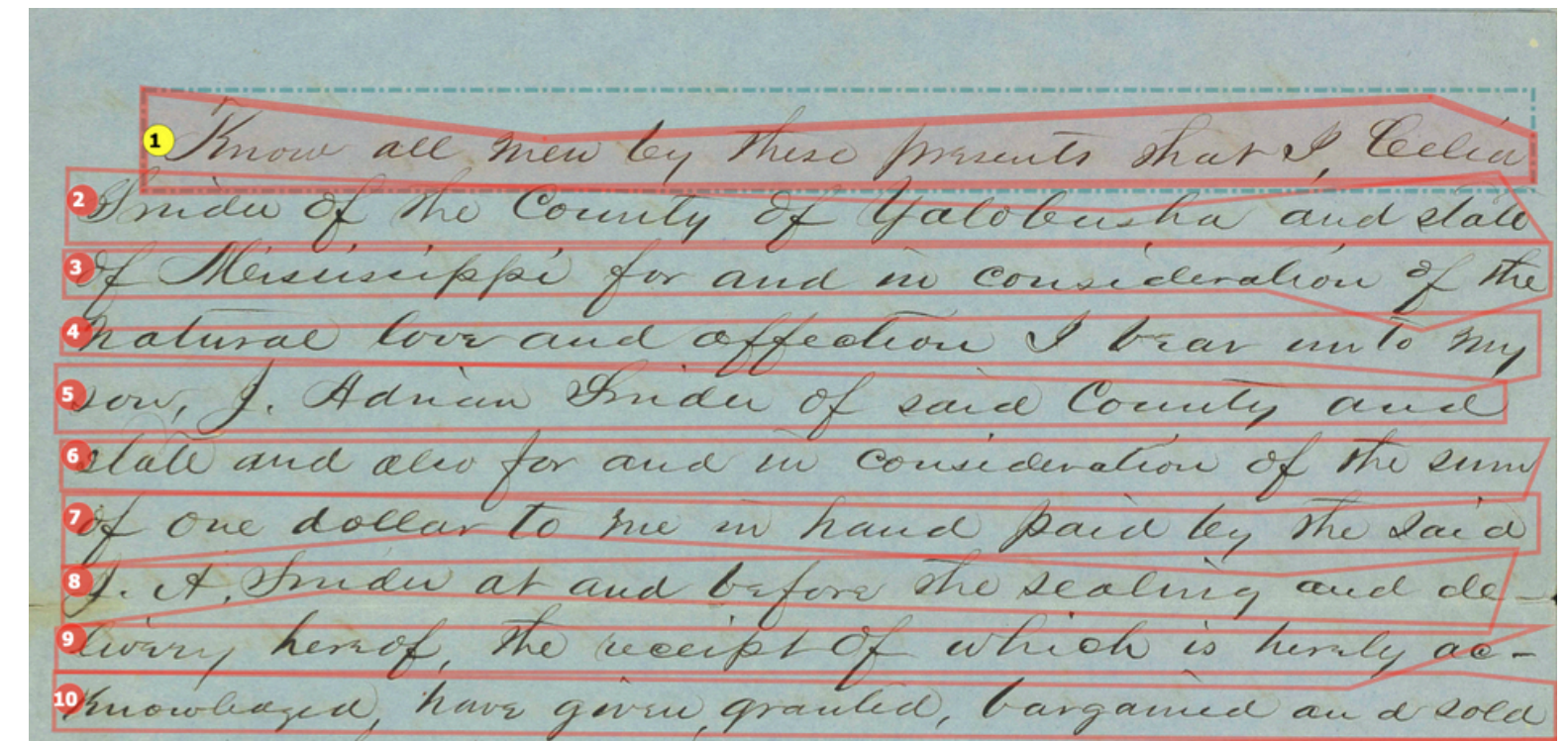
The AI identifies different "zones" on a page, such as text blocks, images, headings, and margins.

Segmentation

The process of cutting the text blocks into individual lines and words so the computer can read them one by one.

Challenges

Complex layouts, like multi-column pages or scattered notes in a diary, often confuse the AI, leading to text being read in the wrong order.



Source: <https://scholarblogs.emory.edu/ecds/transcribe-with-transkribus/>

Traditional Recognition Frameworks

Optical Character Recognition (OCR) for printed text

How it works

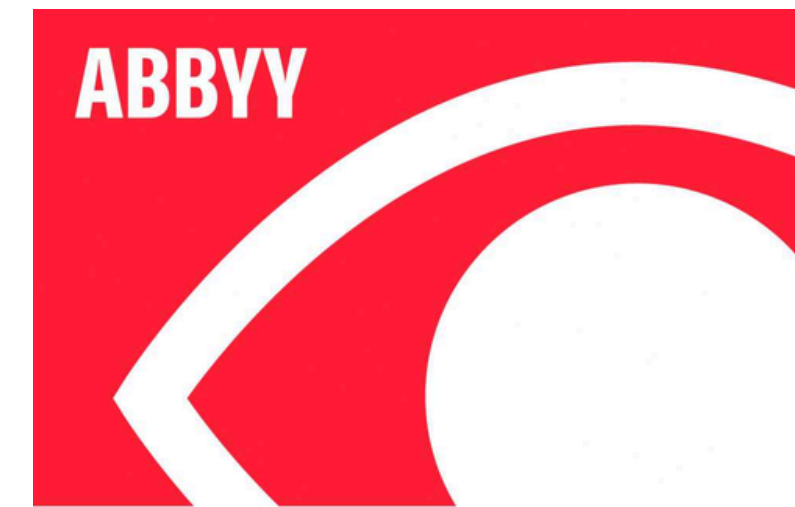
The software looks for patterns that match standard, modern fonts.

Standard Tools

- **ABBYY FineReader** A commercial tool used to keep the original page look and create searchable PDFs.
- **Tesseract** A free, open-source tool often used for large-scale library projects.

Limitations

These tools often struggle with historical decorative fonts, complex layout, or paper that is stained or wrinkled.



FineReader PDF



Tesseract OCR

Handwritten Text Recognition (HTR) and Transkribus

Beyond Pattern Matching

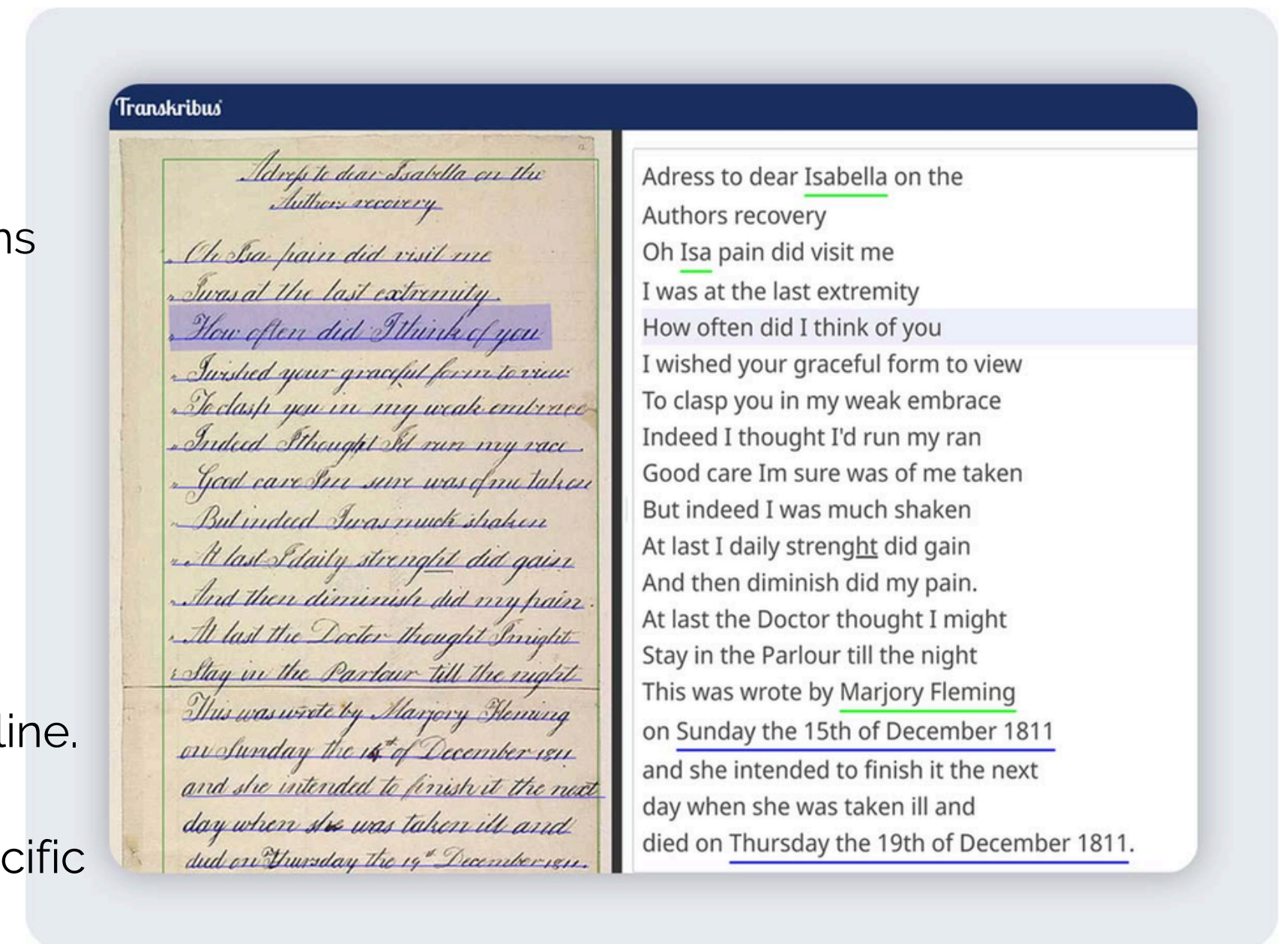
Modern HTR uses "context" to guess words, much like how humans read a sentence to understand a messy word.

Transkribus

A specialized platform for historical documents. It allows users to manage archives and train custom AI models.

Collaborative Transcription

This allows a team of researchers to work on the same project online. One person can transcribe, another can check the work, and everyone's corrections help "teach" the AI to get better at that specific handwriting.



Source: <https://www.transkribus.org/>

Ground Truth and Model Performance

GROUND TRUTH

Defining Ground Truth

A small set of pages (50–100) transcribed perfectly by a human to show the AI what the "correct" result looks like.

TRAINING

The Training Process

The AI compares its guesses to the Ground Truth and learns from its mistakes.

CONTEXT

The Context Challenge

AI models are specialized rather than universal; a tool trained on one specific type of document may not work well when the language, time period, or writing style changes.

Generative AI and Large Language Models (LLMs)

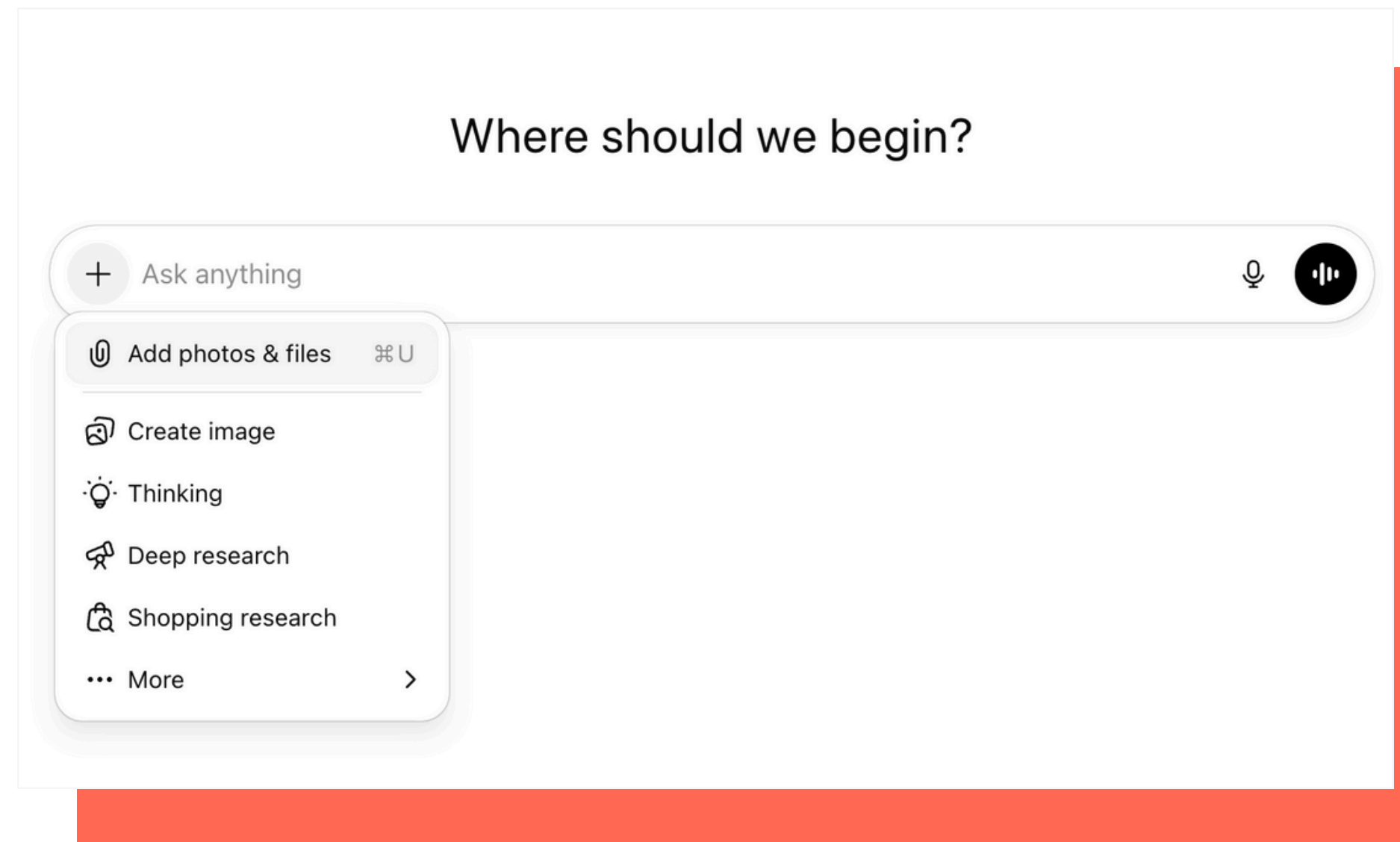
Large Language Models (LLMs) as Transcription Tools

Current Models

ChatGPT, Gemini, Claude

The "Zero-Shot" Approach

You can get a transcription immediately by simply uploading an image to the chat interface and asking the AI to read it. You do not need to "train" the model yourself.



Instructional Design (Prompt Engineering)

Context

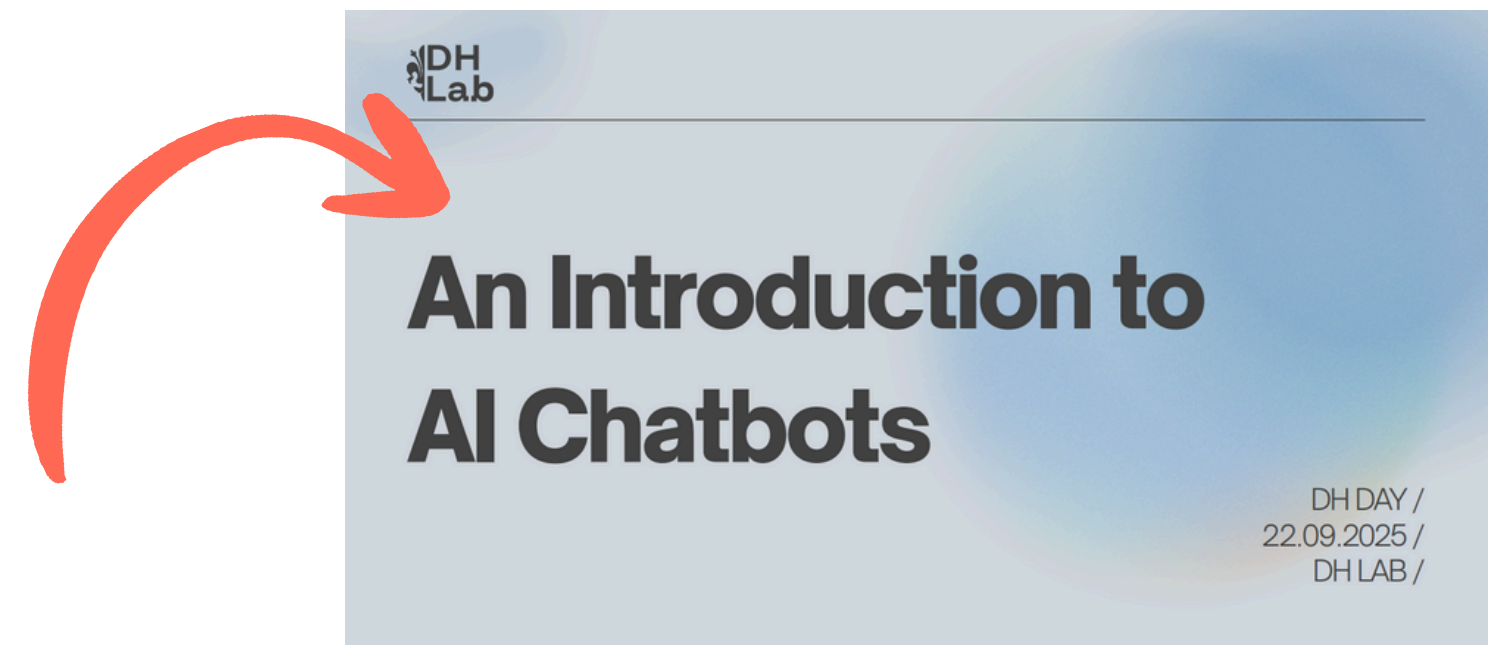
Provide a context description of the document
(e.g. 19th century inventory book in Italian)

Task Specificity

Give clear rules for tricky elements, such as how to handle abbreviations, symbols, or letters that look similar.

Output Structuring

Ask the AI to provide the text in a specific way, such as a simple list, a table, a structured forma (e.g. JSON), or coded formats like TEI or XML.



Source: An Introduction to AI Chatbots (dhlabs.khi.fi.it > Resources > Download Section)

```
{  
  "page": 7,  
  "negative_number": "21901",  
  "artist": "Albani Francesco",  
  "raw_transcription": "21901 Albani Francesco - La Galatea. P.",  
  "catalogue_title": "Dresda. Riproduzioni"  
}
```


Improving Results: Learning by Example

The Value of Examples

Providing the AI with 2–3 lines of text you have already verified acts as a "cheat sheet." This helps e.g. the AI understand the specific handwriting style and your preferred formatting.

Keep it Simple

Clear, direct instructions are usually better than very long, complicated descriptions. The AI performs best when it has a focused task and a few good examples to follow.

Ino.-N.	Künstler.	Gegenstand.	Ort.
7360	Fra Filippo Lippi - Schule	Madonna u Kind	Rom Galleria Colonna
7361	Filippino Lippi	4 Heilige	Lucca S. Michele



"7360| Fra Filippo Lippi - Schule |
Madonna u Kind | Rom Galleria Colonna"

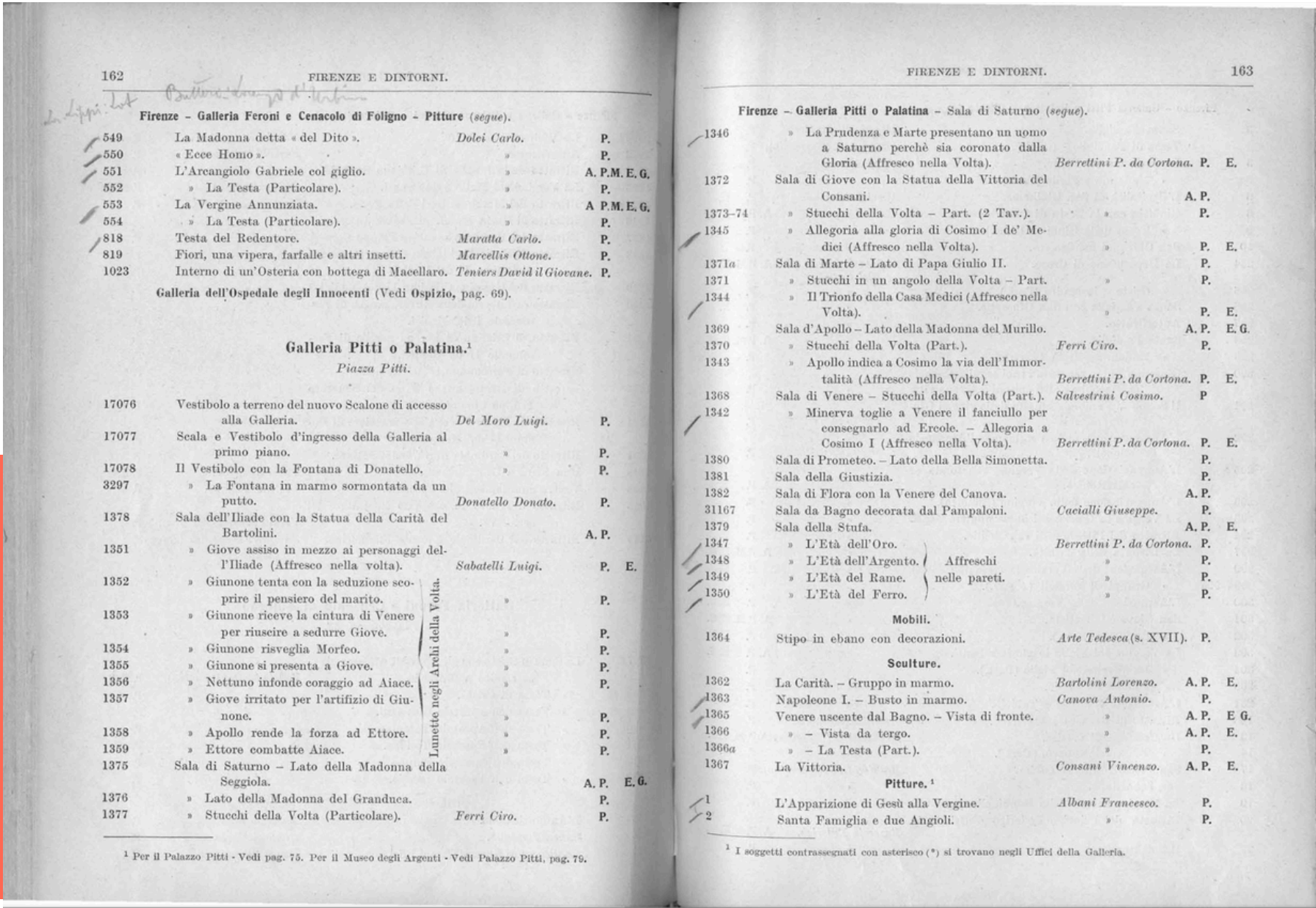
Case Studies and Evaluation

Case Study: Photographer Sales Catalogues



The Material

Photothek's collection, each contains a list of photographs (generally negative number + short image content description)



Source: Alinari: "Firenze e dintorni. Riproduzioni Fotografiche", 1916, p. 2, 94.

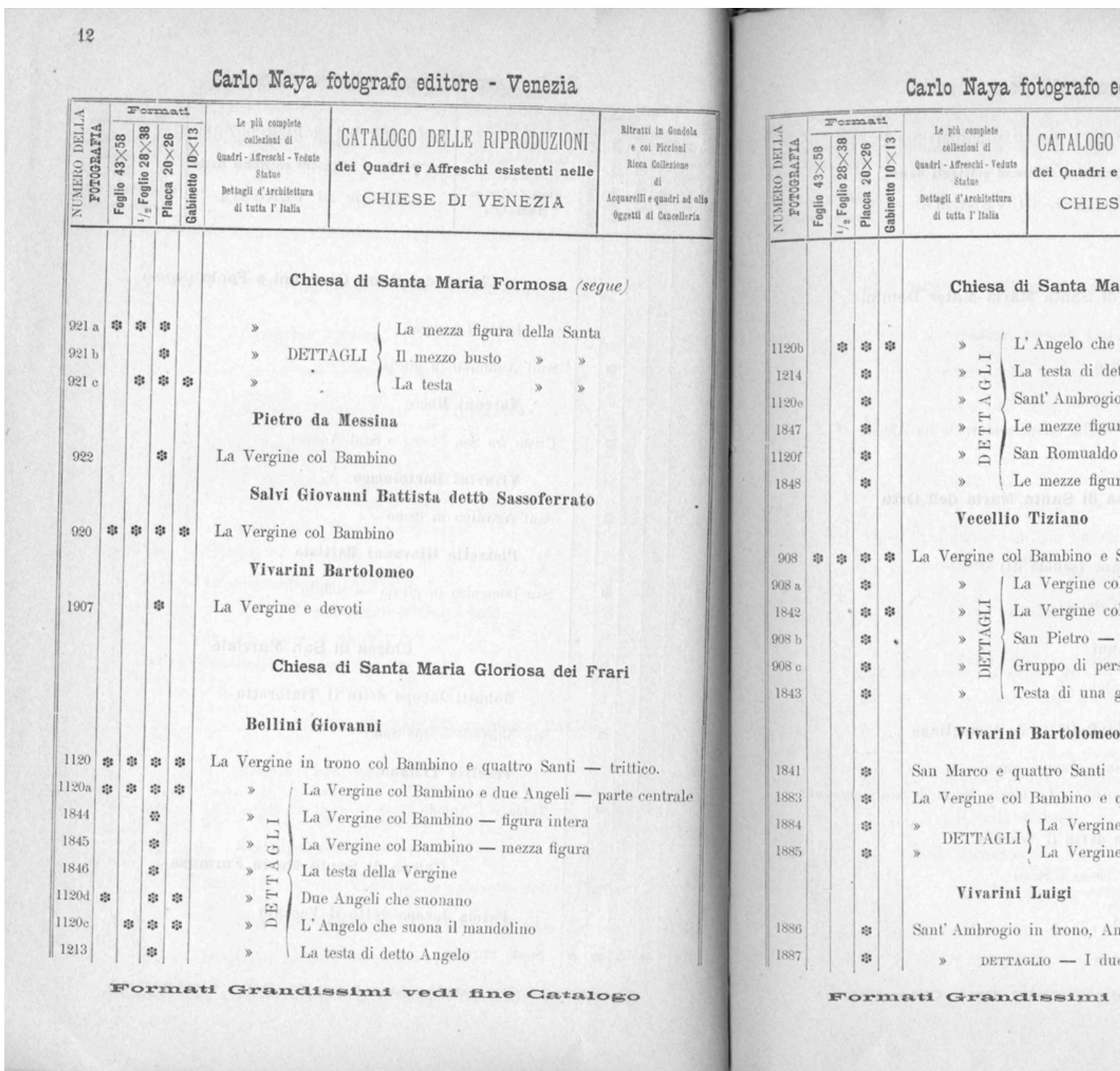
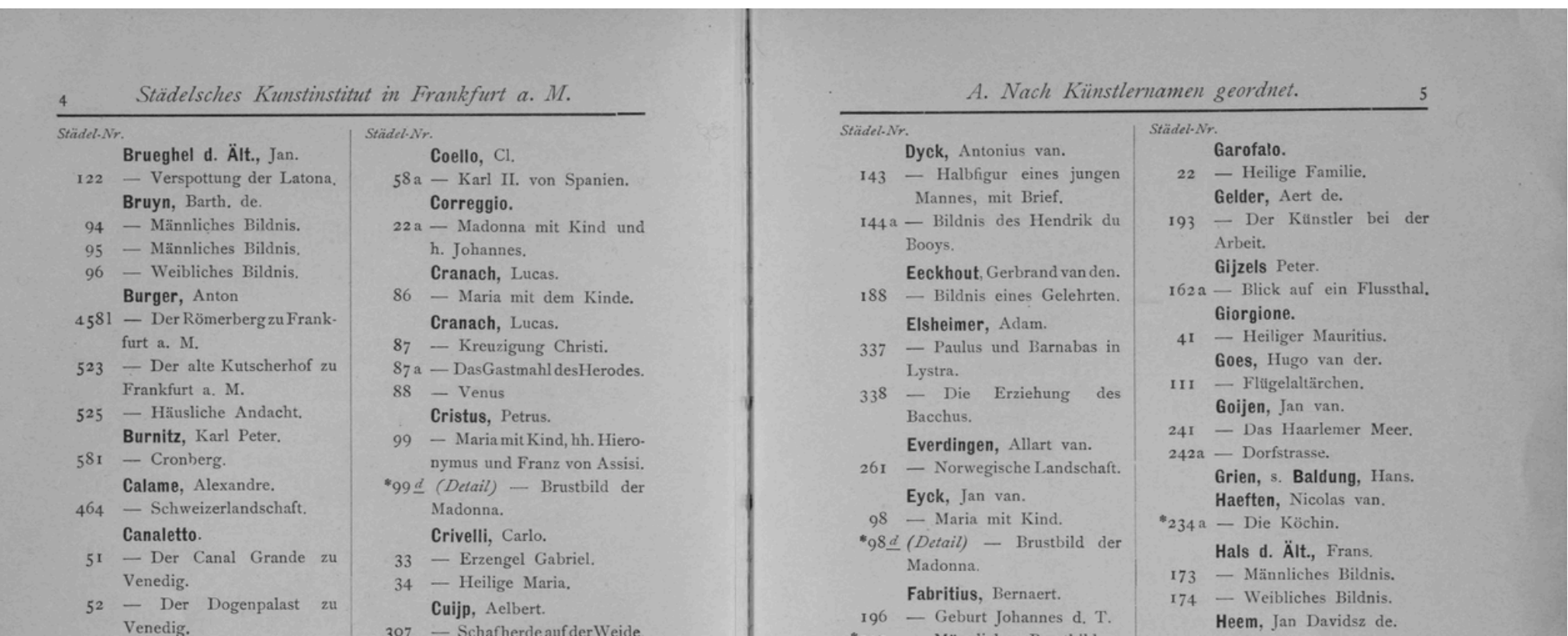
Case Study: Photographer Sales Catalogues



Challenges

- Printed lists with variable structure and layout (columns, tables, linear)
- Variable length (Between 10 and 200 pages per catalogue)
- Distinction between pages with and without photograph entries
- Inconsistencies, abbreviations, ditto marks, etc.

Source: Bruckmann, Friedrich: "Bruckmanns Pigmentdrucke der Gemälde-Galerie des Städelschen Kunstinstitutes in Frankfurt a. M.", 1899, p. 4.



Source: Naya, Carlo: "Catalogo Generale die Quadri e Affreschi Esistenti nelle Chiese di Venezia riprodotti col Sistema Isocromatico", 1900, p. 9.

Case Study: Photographer Sales Catalogues

Workflow (Gemini API)

Instead of typing each page into a chat, we use an "API" (a direct connection) to send hundreds of images to the AI at once. The AI reads them and automatically organizes the information into a digital file.

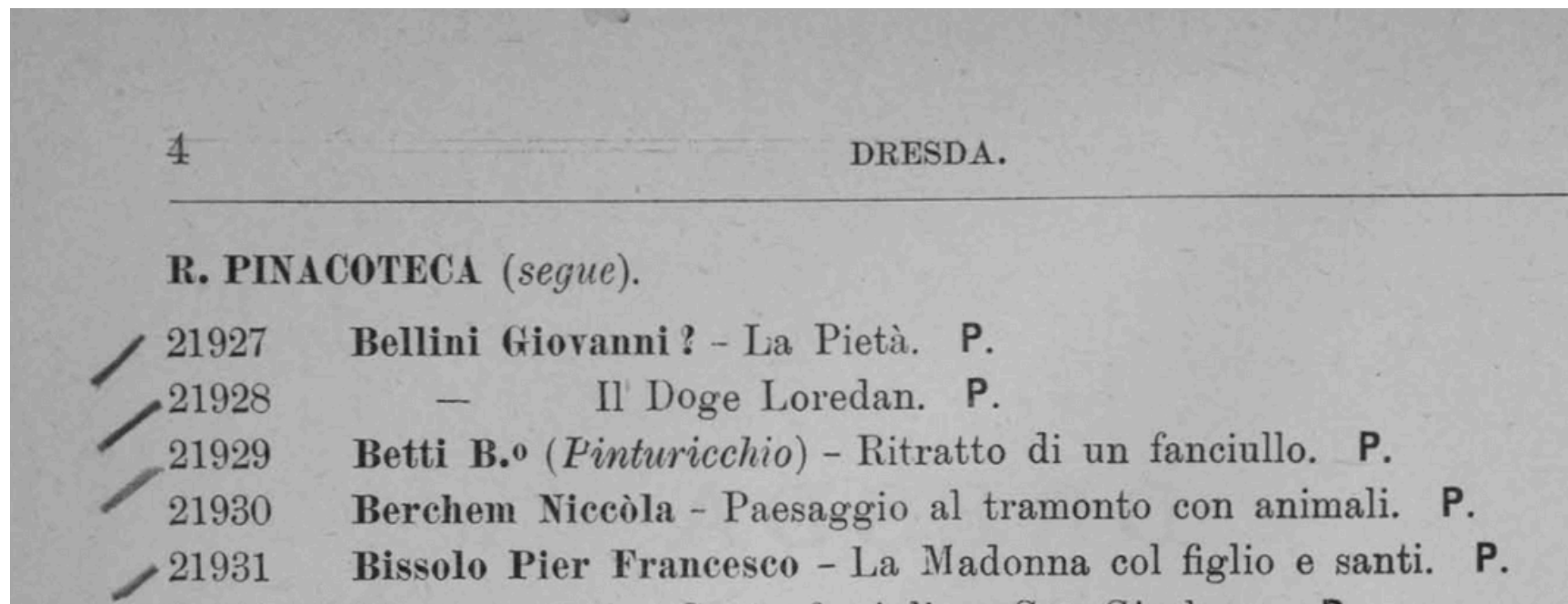
Goal

To create a searchable database where we can instantly sort entries by artist, date, or category.



```
{
  "page": 4,
  "negative_number": "21927",
  "location": [
    "DRESDA.",
    "R. PINACOTECA (segue).",
  ],
  "artist": "Bellini Giovanni?",
  "subject": [
    "Bellini Giovanni?",
    "La Pietà"
  ],
  "raw_transcription": "21927 Bellini Giovanni? - La Pietà. P.",
  "resolved_transcription": "21927 Bellini Giovanni?, La Pietà. DRESDA. R. PINACOTECA (segue).",
  "formats": "P.",
  "notes": "",
  "catalogue_title": "Dresda. Riproduzioni",
  "filename": "Alinari_1906_2",
  "publisher": "ALINARI",
  "year": "1906",
  "places": [
    {
      "id": "cebe9b5c-1d41-4f08-8239-e9cee52b8e4a",
      "geoNamesId": 3176959,
      "itName": "Firenze",
      "enName": "Florence",
      "deName": "Florenz",
      "createdAt": "2023-11-29T23:45:48.219Z",
      "updatedAt": "2023-11-29T23:45:48.219Z"
    }
  ],
  "prompt_version": "extraction_prompt_v1",
  "prompt_author": "author1",
},
```


Case Study: Photographer Sales Catalogues



Source: Alinari: "Dresda, Dintorni", 1906, p. 5.

```
{
  "page": 4,
  "negative_number": "21927",
  "location": [
    "DRESDA.",
    "R. PINACOTECA (segue).",
  ],
  "artist": "Bellini Giovanni?",
  "subject": [
    "Bellini Giovanni?",
    "La Pietà"
  ],
  "raw_transcription": "21927 Bellini Giovanni? - La Pietà. P.",
  "resolved_transcription": "21927 Bellini Giovanni?, La Pietà. DRESDA. R. PINACOTECA (segue).",
  "formats": "P.",
  "notes": "",
  "catalogue_title": "Dresda. Riproduzioni",
  "filename": "Alinari_1906_2",
  "publisher": "ALINARI",
  "year": "1906",
  "places": [
    {
      "id": "cebe9b5c-1d41-4f08-8239-e9cee52b8e4a",
      "geoNamesId": 3176959,
      "itName": "Firenze",
      "enName": "Florence",
      "deName": "Florenz",
      "createdAt": "2023-11-29T23:45:48.219Z",
      "updatedAt": "2023-11-29T23:45:48.219Z"
    }
  ],
  "prompt_version": "extraction_prompt_v1",
  "prompt_author": "author1",
}
```

Choosing the Right Tool

Effort vs. Results

Is it worth spending weeks training a custom model in Transkribus, or is a quick AI chat good enough for a few pages?

The Need for Human Oversight

All AI can "hallucinate" (make up words that aren't there). For academic work, a human must always check the final transcript for accuracy.

Size of the Collection:

Small collections: Use chat interfaces (ChatGPT/Gemini).
Large archives: Use automated API processing or Transkribus models.

Main Takeaway

Match your tool to your project.
Use OCR for modern print,
Transkribus for large sets of the same handwriting, and LLMs for quick results or complex data.

THANK YOU!

<https://www.transkribus.org/>

<https://chatgpt.com/>

<https://gemini.google.com/app>

<https://claude.ai/>