arXiv:2604.03553v3 [cs.AI] 18 Aug 2026

Chronos: The AI Co-Historian

Lorenz Hufe[*,1,2] **Niclas Griesshaber**[*,2] **Gavin Greif** [*,2] **Sebastian Oliver Eck**[2]

Pieter Francois[2,5] **Wojciech Samek**[1,3,4] **Christian Schroeder de Witt**[2] **Philip Torr**[2]

[1]Fraunhofer HHI [2]University of Oxford [3] Technische Universität Berlin
[4] BIFOLD – Berlin Institute for the Foundations of Learning and Data
[5] The Alan Turing Institute
[*]Equal contribution
Correspondence:
lorenz.hufe@hhi.fraunhofer.de

Abstract

AI is increasingly supporting, accelerating, and automating scientific discovery across subjects. Yet, the adoption of AI in historical research remains limited due to the lack of specialised solutions for historians. To change this, we introduce Chronos, an AI Co-Historian designed to support historians. It allows researchers to create and customize research workflows through natural-language interaction and share these as Chronos-Extensions with others. Chronos specifically addresses the need of historians for a tool that is specialised, non-technical, highly customizable, and facilitates extensive task evaluation. As a first extension, we introduce Chronos-Extract, which enables researchers to automate the targeted extraction of information from image scans of historical sources. We benchmark Chronos-Extract on three historical source corpora and find that it achieves high task-accuracy across primary sources spanning three centuries and diverse languages, layouts, and typefaces. Chronos is openly available and ready for historians to use on their own primary and secondary sources.

1 Introduction

In many areas of AI for Science, progress can be measured with quantitative objectives, such as benchmark performance in machine learning and algorithm discovery, target properties of generated materials, or reaction yield in chemistry [Novikov et al., 2025, Zeni et al., 2025, Boiko et al., 2023]. This makes autonomous scientific discovery attractive, since systems can optimize toward measurable improvements. Historical research is different: its most important contributions can rarely be captured by a single quantitative metric, but must be judged qualitatively by expert historians. We therefore argue that AI for History should prioritise human-AI collaboration over fully autonomous historical research.

Historians visit archives, search for, transcribe, contextualise, and interpret primary sources to build their arguments. These primary sources are the core of historical research. Extracting knowledge from them is not only a technical step but an interpretative process that requires domain-specific expertise. Current AI systems do not yet fully support historians in this process and as a result receive little adoption in their research workflows.

To increase AI adoption among historians, we introduce `Chronos`: the first AI Co-Historian. In particular, we make the following three contributions in support of our vision for advancing AI for History:

Preprint.

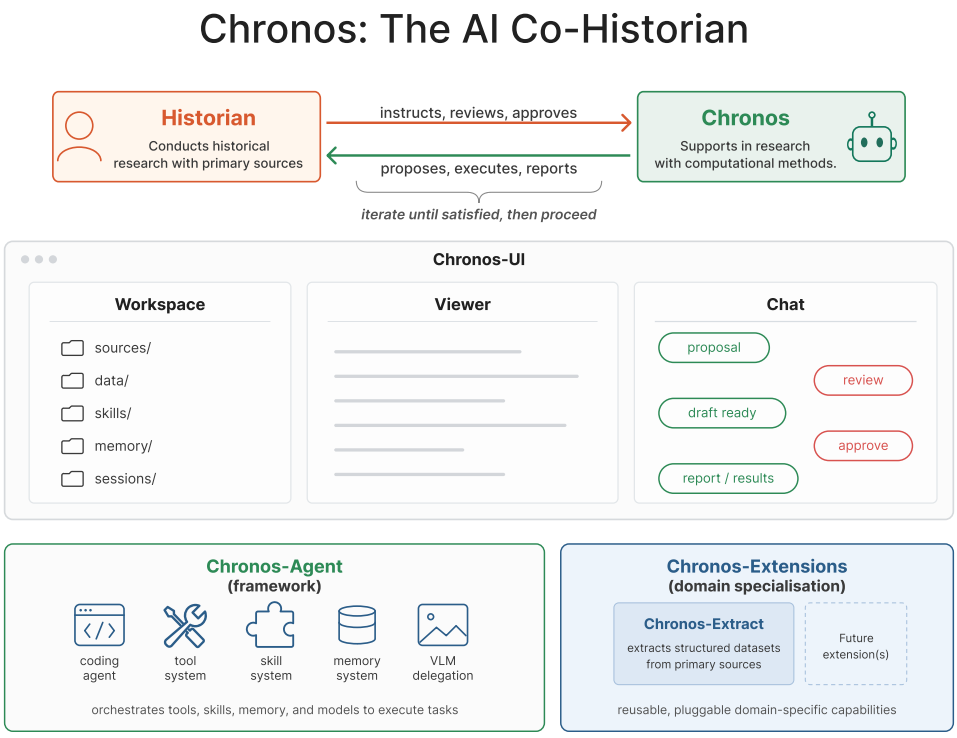


Figure 1: `Chronos`: an AI Co-Historian that connects historians, digitized primary sources, and reusable computational workflows. Historians interact with the `Chronos-Agent` through the `Chronos-UI`, using a dedicated source viewer to inspect evidence, refine instructions, validate outputs, and approve workflows. The agent combines tools, skills, memory, a coding agent, and a VLM subagent to propose, execute, and report computational steps over primary sources. Successful workflows can be packaged as reusable `Chronos-Extensions`, allowing methods to be shared and adapted across projects. `Chronos-Extract` is the first such extension and turns scanned primary sources into structured datasets for downstream historical analysis.

1. We introduce `Chronos`, the first AI Co-Historian that provides a flexible framework for designing, adapting, executing, and sharing AI-assisted research workflows through natural language interaction. `Chronos` enables historians to partially automate tasks across their research process while remaining human-in-the-loop.
2. We present `Chronos-Extract`, a modular `Chronos-Extension` for constructing structured datasets from digitised primary sources. By shifting from manual pipeline implementation to natural language workflow interaction, `Chronos-Extract` enables the rapid creation and adaptation of VLM-based extraction pipelines tailored to heterogeneous historical sources.
3. We evaluate `Chronos-Extract` on three widely used historical corpora in economic history spanning different languages, layouts, and typefaces, demonstrating strong performance. Our approach achieves $F_1 > 0.99$ on two of the three datasets, matching or exceeding results previously attainable only with corpus-specific pipelines.

The code can be accessed at: https://github.com/ai-historian/chronos

2 Related Work

Extracting information from images is a challenge faced across multiple research domains. In machine learning, a large body of literature addresses the technical challenges of extracting structured data from documents, including optical character recognition (OCR) [Fu et al., 2024, Liu et al., 2024], document understanding and information extraction [Sui et al., 2024, Ke et al., 2025, Li et al.,

2025a,b, Ouyang et al., 2025, Yang et al., 2025, Zhang et al., 2025b], and reasoning over document content [Huang et al., 2025, Yu et al., 2025].

Building on these advances, researchers in the digital humanities have applied such techniques to historical corpora, creating the data for computational analyses and the exploration of new research questions [Michel et al., 2010, Tyrkkö and Mäkinen, 2022]. Likewise, the social sciences have begun to use these methods to analyse visual data and construct new large-scale economic history datasets [Dell et al., 2023, Carlson et al., 2024, Silcock et al., 2024, Dell, 2025].

The arrival of multimodal large language models (MLLMs), and in particular vision-language models (VLMs) [Liu et al., 2023], has further expanded the scope of document extraction. Unlike traditional OCR pipelines, which primarily convert visual text into machine-readable strings, VLMs can jointly perform transcription and semantic parsing, mapping visual document content directly into structured outputs. This enables more flexible and general-purpose extraction workflows, reducing the need for task-specific post-processing pipelines such as named entity recognition or rule-based field extraction.

Many studies using VLMs, however, focus on a single source type or a narrowly defined extraction task. For example, economists have used LLMs to digitise historical tables [Bäcker-Peral et al., 2025], economic historians have constructed datasets from patent registers and firm reports [Griesshaber and Streb, 2025, Jayes, 2025, Xie et al., 2025], while musicologists have demonstrated that VLMs can extract structured data from structurally complex and diverse sources such as digitised concert programmes [Eck and Page, 2026]. Recent work has begun to apply (M)LLMs to extract information from diverse museum and archival collections [Schimmenti et al., 2024, Reusens et al., 2025, Toth et al., 2025, Vafaie et al., 2025], and institutions such as the Philadelphia Fed have begun to deploy them on their archival sources [Moulton and Severen, 2025].

While these efforts demonstrate the growing applicability of VLMs across diverse historical data sources, prior work typically contributes domain-specific pipelines designed around the layout, notation, and structure of individual corpora. Although effective, such pipelines require substantial manual effort to design and adapt to new sources.

We take a complementary approach: rather than introducing a corpus-specific pipeline, we make extraction pipelines easier for historians to create, customize, and reuse. `Chronos` enables historians to specify extraction tasks in natural language, which are then translated into VLM-based workflows by the `Chronos-Agent`. In this sense, `Chronos-Extract` provides a reusable "pipeline cookbook" that enables the rapid construction and adaptation of extraction pipelines across heterogeneous sources.

3 Chronos: The AI Co-Historian

`Chronos` consists of two core components: `Chronos-Agent`, which accelerates and automates historical research tasks, and `Chronos-UI`, a user interface that enables historians to supervise and interact with the agent. `Chronos-Agent` has an extension system that allows users to load, share, and create extensions. This design allows `Chronos-Agent` to solve domain-specific historical tasks. For example, `Chronos-Extract` is an extension that specialises `Chronos-Agent` to support historians in the extraction of structured data from primary sources.

`Chronos-Agent` is built on pi[1], a general-purpose, model-agnostic coding agent framework. The agent has access to the tools listed in Table 1, which fall into three categories: (a) tools for navigating the host machine's operating system, (b) tools for interacting with `Chronos-UI`, and (c) tools for querying a VLM subagent.

`Chronos-UI` builds on the open-source Integrated Development Environment (IDE) Visual Studio Code. Its conceptual design is shown in Figure 1, and a screenshot of the implemented interface is provided in Appendix Section A. It is intended to help historians in organising their research environment. `Chronos-UI` provides functions for (i) initializing the standardised workspace structure (ii) importing and preprocessing primary sources and (iii) starting an agent session.

Skills. Historians perform many tasks that are highly specific to their chosen topic, region, and period. Skills provide a way to formalise these domain-specific research procedures in natural language [Zhang et al., 2025a], similar to how a historian would explain a task to a collaborator. This

[1] https://github.com/badlogic/pi-mono

Table 1: Chronos-Agent tool inventory.

Tool	Category	Purpose
read, write, edit	OS	File I/O
grep, find, ls	OS	File discovery
bash	OS	Shell exec
list_pages	OS	List source pages
show_page	UI	Show page in viewer
show_text	UI	Show text in viewer
change_source	UI	Switch source of viewer
analyse_page	VLM	Query VLM
follow_up_question	VLM	Continue dialogue
analyse_pages_batch	VLM	Batch VLM query

improves the performance of Chronos-Agent on such tasks. Chronos-Agent discovers provided skills automatically and applies them if appropriate. Chronos-Agent can also help construct new skills through conversation: the historian describes a task in natural language, the agent proposes a procedure, and this is iteratively refined through human-AI collaboration. Skills are portable, allowing historians to share or adapt procedures across projects. More information about the skill implementation are given in Appendix Section B.

Viewer. Chronos-UI provides a page viewer alongside the conversational interface as shown in Figure 1. When the Chronos-Agent analyses a page or the historian asks to inspect one, the scanned image appears in the viewer, allowing the historian to verify the agent's work directly and can be accessed by the agent using the tool described in Table 1.

Delegated Visual Inspection. Chronos-Agent can query information about specific primary source pages without loading their images into its own context by delegating to a non-agentic VLM. This design choice was motivated as we observed significant performance degradation when images were loaded natively into the agent's context. By decoupling visual inspection from the reasoning loop, the agent can choose the model best suited for each task.

Chronos-Extensions. A Chronos-Extension is a reusable bundle of skills that extends Chronos for a specific domain or task by packaging the procedures, prompts, and dependencies needed for a specialised workflow.

4 Chronos-Extract: Information Extraction from Historical Sources

Chronos-Extract is a Chronos-Extension, which enables historians to construct custom extraction processes through natural-language interaction with Chronos. Chronos-Extract was developed within Chronos, exclusively through natural language interactions with Chronos-Agent. Below we describe the Chronos-Extract workflow, shown in Figure 2.

Step 1: Locating the relevant pages. Historians rarely need an entire primary source, but only the pages relevant to their research question. The first skill (locate-section) addresses this step by employing a binary search over all page images to find the relevant section or, if available, by first consulting the table of contents. To ensure the page range is identified correctly, the Chronos-Agent automatically opens the proposed start and end pages in the page viewer for the historian to verify.

Step 2: Designing the extraction prompt. Once the relevant pages have been identified, the challenge is to construct an extraction prompt that captures the unique structure and conventions of the primary source. The skill (draft-prompt) searches for structural cues (e.g., abbreviation tables or legends), transcribes them, and incorporates them into a base prompt with source-specific rules and column definitions. Chronos-Agent tests this prompt on a random page using its VLM subagent. Next, the agent presents the extracted entries alongside the original scan to the historian and checks for ambiguities. If issues are identified—for example, that secondary text encodes additional attributes that should be captured in a dedicated column—the agent can revise the prompt until the extraction quality is deemed satisfactory by the historian.

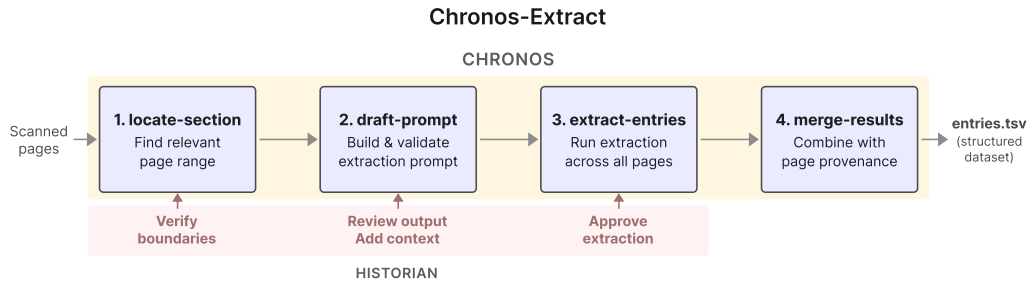


Figure 2: `Chronos-Extract` **workflow.** (1) Locate relevant pages, verified by the historian. (2) Draft and interactively refine an extraction prompt. (3) Execute batch extraction across pages after user approval. (4) Merge results into a structured dataset.

Step 3: Batch extraction. Before starting the batch extraction `Chronos-Agent` summarises the page range, displays the final prompt, justifies the model choice, and estimates the cost. Only after the historian's approval does it proceed with the `ask_pages_batch` tool call. Once completed, `Chronos-Agent` examines the results by checking for missing files, column consistency, and common artefacts such as unintended Markdown formatting in the VLM output.

Step 4: Merging with provenance. For the extracted data to be of value to the historian, the separate files must be concatenated into a dataset suitable for analysis. The final skill (`merge-results`) performs this step by combining the per-page TSV files and prepending a `page_id` column so each row can be traced back to its source page. This step is fully automated and requires no historian input.

5 Experiments

To benchmark `Chronos-Extract`, we evaluate it on three historical source corpora that are both relevant to ongoing historical research and accompanied by gold data ground truth data: German Patent Registers [Griesshaber and Streb, 2025], French Trade Directories [Abadie et al., 2022c], and American Newspapers [Dell et al., 2023]. Together, these corpora span three centuries, several languages and typefaces, and a wide range of document layouts, from structured registers to dense multi-column newspapers. They therefore provide a realistic test bed for evaluating whether `Chronos-Extract` can generalize across heterogeneous primary sources used in historical research.

For each dataset, we define the target record set as $Y = \{r_1, \ldots, r_N\}$ and the prediction set as $\hat{Y} = \{\hat{r}_1, \ldots, \hat{r}_M\}$. Each record is defined as $r_i = (t_i, \mathbf{v}_i)$, where t_i is the full text used for entry-level matching and $\mathbf{v}_i$ denotes structured variables associated with the record.

5.1 Datasets

American Newspapers. Historical newspapers are a valuable source for historians and are a particularly demanding test case for information extraction systems, as a large amount of text is densely packed into a single image scan, typically arranged in multi-column layouts with varying typefaces, print degradation, and article text that spans non-contiguous regions. We evaluate `Chronos-Extract` on ten fully annotated pages that were created by Dell et al. [2023] to benchmark the accuracy of a large-scale US newspaper dataset they constructed. Figure 5 depicts one of the newspaper scans for which annotations exist. Since this dataset does not provide structured variables, we only use the full text entries t_i for record matching and transcription evaluation.

French Trade Directories. Trade directories are a valuable source for studying urban societies, their occupational structures, and social networks. We evaluate `Chronos-Extract` on a corpus of 8,765 manually corrected directory entries drawn from 78 pages across 18 French directories with different layouts published between 1798 and 1861. Figure 6 depicts an image scan of one of the French directory types. [Abadie et al., 2022b,a, Tual et al., 2023, Duménieu et al., 2023]. Each directory record contains a `full_text` field and six structured tags: `Person`, `Activity`, `Location`, `House No.`, `Title`, and `Facility`.

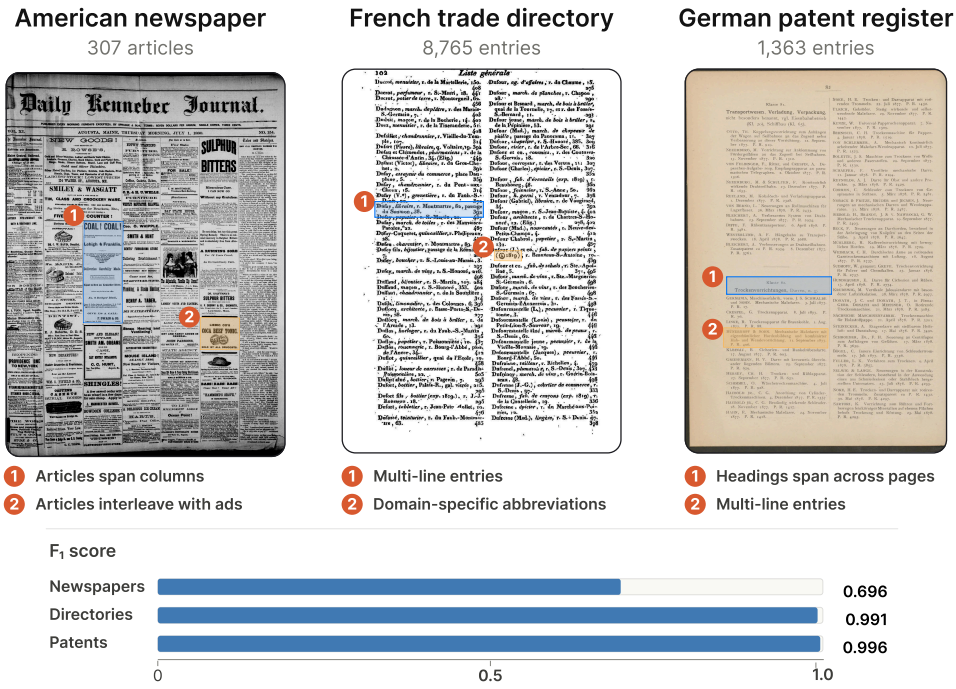


Figure 3: **Heterogeneity of historical sources.** Example documents from newspapers, trade directories, and patent registers illustrate variation in layout (e.g., multi-column text, interleaved content), encoding conventions (e.g., abbreviations), and typography. These differences require domain-specific interpretation and lead to varying extraction performance (bottom). The source images are shown at high resolution and can be zoomed in for detailed inspection.

German Patent Registers. Historians consult patent registers to study innovation and technological change. We evaluate `Chronos-Extract` on a ground truth dataset consisting of 41 image scans published between 1877 and 1918 by Germany's Imperial Patent Office. Patent registers differ from other historical sources in that they often appear in a structured, double-column list format and include many elements on a page that historians may not wish to extract, such as lists of patent numbers or subcaptions of technological classes. These patent registers were partially printed in Fraktur typefaces, which pose a significant challenge for OCR systems [Springmann et al., 2018]. Each patent record contains a `full_entry` field and five structured fields: `patent_id`, `assignee`, `location`, `title`, and `date`.

5.2 Experimental setup

We structure the evaluation of `Chronos-Extract` around the following three questions.

5.2.1 Does Chronos-Extract identify the records present in the source documents?

To evaluate whether `Chronos-Extract` identifies the records present in the source documents, we focus on the full-text fields t_i and $\hat{t}_j$. For any two strings x and $\hat{x}$, we define the normalised Levenshtein similarity as

$$\mathcal{S}(x, \hat{x}) = 1 - \frac{\mathrm{Lev}(x, \hat{x})}{\max(|x|, |\hat{x}|)}.$$

We compute pairwise similarities $\mathcal{S}(t_i, \hat{t}_j)$ for all target and predicted records, and use the Hungarian algorithm to obtain the maximum-similarity matching A. We discard all matched pairs below the threshold $\tau = 0.8$, yielding the set of successful record matches

$$\mathcal{M} = \{(r_i, \hat{r}_j) \mid (i, j) \in A,\ \mathcal{S}(t_i, \hat{t}_j) \geq \tau\}.$$

Given $\mathcal{M}$, we compute precision, recall, and F_1. We additionally report the absolute counts of missed records, i.e. unmatched target records, and extra records, i.e. unmatched predictions.

5.2.2 How accurately does Chronos-Extract transcribe the records?

To evaluate transcription accuracy, we compute character error rate (CER) on concatenated text fields, following OCR evaluation practice [Heidenreich et al., 2026]. We report two variants.

First, $\mathrm{CER}_{\mathcal{M}}$ measures transcription quality on successfully matched records only. Specifically, we concatenate the target texts t_i for all $(i, j) \in \mathcal{M}$, concatenate the corresponding predictions $\hat{t}_j$ in the same order, and then compute CER between the two resulting strings.

Second, $\mathrm{CER}_{\mathcal{C}}$ measures corpus-level transcription quality. Here, we compute CER between the concatenation of all target text fields $t_1, \ldots, t_N$ and the concatenation of all predicted text fields $\hat{t}_1, \ldots, \hat{t}_M$. Unlike $\mathrm{CER}_{\mathcal{M}}$, $\mathrm{CER}_{\mathcal{C}}$ also reflects missed records, hallucinated records, and ordering errors.

5.2.3 How accurately does Chronos-Extract extract structured variables from the records?

For directories and patents, records include structured variables in addition to their full-text field. We evaluate variable extraction on successfully matched records $(i, j) \in \mathcal{M}$. Let v_{ik} denote the value of field k in target record r_i, and let $\hat{v}_{jk}$ denote the corresponding predicted value in $\hat{r}_j$. Before comparison, we apply a normalization function $\nu(\cdot)$ that collapses whitespace and removes trailing punctuation and enclosing quotes.

We report two match rates for each field k. A strict match requires exact equality after normalization,

$$\nu(v_{ik}) = \nu(\hat{v}_{jk}).$$

A relaxed match allows minor transcription differences and requires normalized Levenshtein similarity above the same threshold used for record matching,

$$\mathcal{S}(\nu(v_{ik}), \nu(\hat{v}_{jk})) \geq \tau.$$

We report strict and relaxed match rates separately for each directory tag and patent field.

5.3 Results

We initialize `Chronos` with Opus 4.6 as the orchestrator and Gemini 3.1-Pro-Preview as the VLM for batch extraction. We use Gemini 3.1-Pro-Preview because it is the strongest Gemini model at the time of writing, and Semnani et al. [2025] found Gemini models to perform best on historical font understanding. In collaboration with `Chronos`, we construct one dataset-specific extraction prompt per benchmark on a holdout document to generate our results. The resulting prompts are shown in Appendix E. During the execution of the workflow we auto accepted all tool calls and did not provide additional guidance during the extraction stages.

Retrieval quality. `Chronos-Extract` achieves near-perfect record-level retrieval on the German Patent Registers and French Trade Directories, with $F_1 > 0.99$ on both datasets, but performs substantially worse on American Newspapers, where it reaches $F_1 = 0.696$ (Table 2). This suggests that language and historical typography are not the main bottlenecks in our evaluation: both German and French sources are retrieved reliably, and the challenging Fraktur subset of the patent registers remains close to the Antiqua subset. Instead, retrieval performance appears to be driven primarily by layout complexity. The structured two-column layouts of patents and directories allow `Chronos-Extract` to identify records reliably, whereas the dense multi-column newspaper pages contain interleaved articles, headlines, advertisements, and non-contiguous article continuations. These layout features lead to substantially more missed records and lower recall.

Surprisingly `Chronos-Extract` *finds entries missing from the human annotated ground truth.* Table 2 details the number of false positives and negatives retrievals. `Chronos-Extract` achieves excellent entry-level matching on the German Patent Registers and French City Directories, leaving just 32 of 10,128 target set entries unmatched. Crucially, manual inspection of the 144 unmatched predicted entries $\hat{t}$ reveals they are not hallucinations and predominantly valid entries absent from the ground truth, indicating `Chronos-Extract` identified entries that were missed by human annotators.

Transcription Quality. The transcription quality for successfully matched pairs at $\tau \geq 0.8$ is very strong (Table 2). The matched record character error rate ($\mathrm{CER}_{\mathcal{M}}$) averages 0.02 across all sources. Patents printed in modern Antiqua achieve $\mathrm{CER}_{\mathcal{M}} = 0.007$, while the historical Fraktur

Table 2: Extraction performance by domain and source at matching threshold $\tau = 0.8$. FN (false negative): t_i entries missed by the prediction. FP (false positive): $\hat{t}_i$ prediction entries produced by `Chronos-Extract` absent from ground truth. $\mathrm{CER}_{\mathcal{M}}$: on matched entries only. $\mathrm{CER}_{\mathcal{C}}$: on all of the corpus.

Source	N	F_1	$\mathrm{CER}_{\mathcal{M}}$	$\mathrm{CER}_{\mathcal{C}}$	FN	FP
Patents	1363	0.996	0.020	0.028	9	3
Antiqua	542	0.997	0.007	0.013	3	0
Fraktur	821	0.995	0.029	0.038	6	3
Directories	8765	0.991	0.020	0.044	23	144
Bottin1	1480	0.994	0.014	0.029	1	18
Bottin3	872	0.989	0.011	0.036	6	13
Cambon	665	0.997	0.015	0.034	0	4
Deflandre	738	0.970	0.034	0.082	11	34
Didot	2816	0.994	0.017	0.046	0	35
DidotBottin	776	0.990	0.016	0.027	1	14
Duverneuil	790	0.989	0.048	0.062	4	14
Favre	230	0.991	0.029	0.042	0	4
La Tynna	307	0.989	0.014	0.048	0	7
Notables	91	0.995	0.018	0.028	0	1
Newspapers	506	0.696	0.018	0.531	201	65
Articles	307	0.759	0.018	0.523	102	28
Headlines	199	0.595	0.021	0.757	99	37

Table 3: Variable-level match rates on matched entry pairs. *Strict*: exact string match after normalization. *Relaxed*: normalised Levenshtein similarity ≥ 0.8. N: number of aligned variable pairs evaluated.

	Variable	Strict	Relaxed	N
Patent fields	ID	0.999	0.999	1354
	Person	0.691	0.968	1354
	Location	0.671	0.967	1354
	Description	0.374	0.923	1354
	Date	0.989	0.999	1354
Directory tags	Person	0.837	0.909	8742
	Activity	0.771	0.900	6533
	Location	0.735	0.952	8724
	House No.	0.952	0.954	8466
	Title	0.053	0.063	917
	Facility	0.047	0.049	506

fonts raise this slightly to 0.029 indicating that historical fonts can impact transcription quality. Matched newspaper articles too only show a small error rate of $\mathrm{CER}_{\mathcal{M}} = 0.018$. This indicates the subpar newspaper F_1-score is a layout-driven detection failure, rather than a transcription quality concern. Corpus-level transcription quality is substantially lower for the American Newspapers, where $\mathrm{CER}_{\mathcal{C}} = 0.531$, compared to 0.028 for patents and 0.044 for directories. This large gap is caused by the many record-level misclassifications: missed records, extra predicted records, and ordering errors all contribute directly to $\mathrm{CER}_{\mathcal{C}}$, even though the successfully matched newspaper records are transcribed accurately.

Variable-Level Accuracy. Table 3 reports variable-level accuracy on matched patent and directory entries. For the patent registers, `Chronos-Extract` achieves consistently strong performance. Patent IDs and dates are extracted almost perfectly, and relaxed match rates exceed 0.92 for all fields, including longer description fields where exact string agreement is more sensitive to minor transcription differences. The gap between strict and relaxed scores for person, location, and description fields therefore primarily reflects small textual deviations rather than failures to recover the relevant information.

For the French Trade Directories, performance is high on the main semantic fields: person names, activities, locations, and house numbers all reach relaxed match rates of at least 0.90. The main weaknesses are `Title` and `Facility`. These fields are substantially harder because they are often

encoded by compact, source-specific symbols or abbreviations rather than ordinary alphabetic text, as illustrated in Figure 3 leading to degraded performance. In the ground truth, 80.5% of `Title` spans contain at least one non-ASCII Unicode character, and 44.2% of `Facility` spans do so.

6 Discussion and Limitations

`Chronos` suggests that natural-language workflow construction can make computational history more accessible while keeping historians directly involved in task design and validation. Beyond the extraction results reported here, this points toward a broader form of reusable research infrastructure: historical workflows that can be shared, adapted, and improved across projects.

Our evaluation deliberately focused on `Chronos-Extract` rather than `Chronos` as a general AI Co-Historian because historical research involves a wide range of interpretative tasks that are difficult to benchmark in its entirety. We therefore evaluated `Chronos-Extract` as a proxy: structured dataset construction is an important component of computational historical research, but it captures only one part of what historians do.

Our results show that `Chronos-Extract` performs well on structured registers and directories, but struggles with complex newspaper layouts. Historical sources remain highly heterogeneous, spanning low-resource or extinct languages, handwriting, damaged scans, marginalia, and domain-specific conventions. Expanding evaluation is difficult because high-quality ground truth requires costly manual annotation, often by multiple annotators [Dell et al., 2023]. Moreover, VLMs can hallucinate records, especially in complex layouts, which can corrupt downstream historical datasets. Benchmarking and human validation are therefore indispensable, particularly for low-resource languages where VLM performance remains weaker [Liu et al., 2026].

Finally, the deployment of AI systems for historical research raises questions of cost, sovereignty, and control. The most capable current VLMs are proprietary and typically require sending digitised primary sources to commercial data centers, which can be problematic for sensitive, restricted, or institutionally governed sources. Their use also creates financial concerns, since higher performance often requires higher spending. However, open-source models such as introduced Semnani et al. [2025] offer a promising path toward reducing these barriers. If sufficiently capable, they could lower costs, improve transparency, and allow archives, universities, and public institutions to run extraction systems under their own governance.

As historians increasingly outsource dataset construction and other historical tasks to AI systems, they entrust them with shaping the basis on which historical inquiry rests. This could introduce errors, biases, and failure modes methodologically important: AI systems do not merely accelerate historical work, but can influence which records are extracted, omitted, or transformed. For this reason, `Chronos` is designed around human-in-the-loop workflow construction, where historians specify goals, inspect outputs, and approve workflows. Nevertheless, future AI co-historians will require careful evaluation, transparent validation procedures, and continued reflection on how AI-mediated datasets affect historical knowledge production.

7 Conclusion

In this paper, we introduced `Chronos`, an AI Co-Historian that enables historians to create, customize, and share AI-assisted research workflows through natural-language interaction. As a first extension, `Chronos-Extract` addresses one of the central bottlenecks in computational history: extracting structured datasets from digitized primary sources. Our experiments show that `Chronos-Extract` can achieve strong performance across heterogeneous documents spanning three centuries, while remaining flexible enough to be adapted to different source types and research questions. More broadly, `Chronos` points toward a reusable infrastructure for historical research, where workflows, extensions, and datasets can be shared and improved by the community. By reducing the technical and practical barriers to large-scale source processing, `Chronos` can expand what historians are able to study, while keeping domain expertise central to the formulation of questions, validation of outputs, and interpretation of evidence.

References

Nathalie Abadie, Stéphane Baciocchi, Edwin Carlinet, Joseph Chazalon, Pascal Cristofoli, Bertrand Duménieu, and Julien Perret. A benchmark of named entity recognition approaches in historical documents: Application to 19th century french directories. In *Document Analysis Systems*, volume 13237 of *Lecture Notes in Computer Science*, pages 445–460. Springer, 2022a.

Nathalie Abadie, Stéphane Baciocchi, Edwin Carlinet, Joseph Chazalon, Pascal Cristofoli, Bertrand Duménieu, and Julien Perret. A dataset of french trade directories from the 19th century (FTD), 2022b.

Nathalie Abadie, Edwin Carlinet, Joseph Chazalon, and Bertrand Duménieu. A Benchmark of Named Entity Recognition Approaches in Historical Documents — Application to 19th Century French Directories. Springer, May 2022c.

Verónica Bäcker-Peral, Vitaly Meursault, and Christopher Severen. Can llms credibly transform the creation of panel data from diverse historical tables? *arXiv preprint arXiv:2505.11599*, 2025.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Jacob Carlson, Tom Bryan, and Melissa Dell. Efficient ocr for building a diverse digital history. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8105–8115, 2024.

Melissa Dell. Deep learning for economists. *Journal of Economic Literature*, 63(1):5–58, 2025.

Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. American stories: A large-scale structured text dataset of historical us newspapers. *Advances in Neural Information Processing Systems*, 36:80744–80772, 2023.

Bertrand Duménieu, Edwin Carlinet, Nathalie Abadie, and Joseph Chazalon. Entry separation using a mixed visual and textual language model: Application to 19th century french trade directories. *arXiv preprint arXiv:2302.08948*, 2023.

Sebastian Oliver Eck and Kevin Page. Multimodal large language model-assisted metadata extraction from historical concert programmes (1872–1928). In *Proceedings of the 13th International Conference on Digital Libraries for Musicology (DLfM 2026)*, New York, NY, USA, 2026. Association for Computing Machinery. Accepted; proceedings not yet published.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.

Niclas Griesshaber and Jochen Streb. Multimodal LLMs for historical dataset construction from archival image scans: German patents (1877–1918). *arXiv preprint arXiv:2512.19675*, 2025.

Hunter Heidenreich, Ben Elliott, Olivia Dinica, and Yosheb Getachew. Gutenocr: A grounded vision-language front-end for documents, 2026. URL https://arxiv.org/abs/2601.14490.

Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning. *arXiv preprint arXiv:2505.17163*, 2025.

Jonathan Jayes. Like moths to a flame: an individual level approach to technological change in 20th century sweden. 2025.

Wenjun Ke, Yifan Zheng, Yining Li, Hengyuan Xu, Dong Nie, Peng Wang, and Yao He. Large language models in document intelligence: A comprehensive survey, recent advances, challenges, and future trends. *ACM Trans. Inf. Syst.*, 44(1), November 2025. ISSN 1046-8188. doi: 10.1145/3768156.

Renyu Li, Antonio Jimeno Yepes, Yao You, Kamil Pluciński, Maximilian Operlejn, and Crag Wolfe. Score: A semantic evaluation framework for generative document parsing. *arXiv preprint arXiv:2509.19345*, 2025a.

Zichao Li, Aizier Abulaiti, Yaojie Lu, Xuanang Chen, Jia Zheng, Hongyu Lin, Xianpei Han, Shanshan Jiang, Bin Dong, and Le Sun. READoc: A unified benchmark for realistic document structured extraction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21889–21905, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1128.

Bonan Liu, Zeyu Zhang, Bingbing Meng, Han Wang, Hanshuo Zhang, Chengping Wang, Daji Ergu, and Ying Cai. Omniocr: Generalist ocr for ethnic minority languages. *arXiv preprint arXiv:2602.21042*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2010. doi: 10.1126/science.1199644.

Daniel Moulton and Christopher Severen. Harvesting historical data with llms. *Economic Insights*, 10(4):1–6, 2025.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. pages 24838–24848, 2025.

Manon Reusens, Amy Adams, and Bart Baesens. Large language models to make museum archive collections more accessible. *AI & SOCIETY*, 2025. ISSN 1435-5655. doi: 10.1007/s00146-025-02227-8.

Andrea Schimmenti, Valentina Pasqual, Francesca Tomasi, Fabio Vitali, and Marieke van Erp. Structuring authenticity assessments on historical documents using llms, 2024.

Sina J. Semnani, Han Zhang, Xinyan He, Merve Tekg"urler, and Monica S. Lam. CHURRO: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 2025.

Emily Silcock, Abhishek Arora, Luca D'Amico-Wong, and Melissa Dell. Newswire: A large-scale structured database of a century of historical news. *Advances in Neural Information Processing Systems*, 37:49768–49779, 2024.

Uwe Springmann, Christian Reul, Stefanie Dipper, and Johannes Baiter. Ground truth for training OCR engines on historical documents in German fraktur and early modern latin. *Journal of Language Technology and Computational Linguistics*, 33(1):97–114, 2018.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635752.

Gabor Mihaly Toth, Richard Albrecht, and Cedric Pruski. Explainable ai, llm, and digitized archival cultural heritage: a case study of the grand ducal archive of the medici. *AI & SOCIETY*, 2025. ISSN 1435-5655. doi: 10.1007/s00146-025-02238-5.

Solenn Tual, Nathalie Abadie, Joseph Chazalon, Bertrand Duménieu, and Edwin Carlinet. A benchmark of nested named entity recognition approaches in historical structured documents. In *International Conference on Document Analysis and Recognition*, pages 313–328. Springer, 2023.

Jukka Tyrkkö and Ilkka Mäkinen. Culturomic explorations of literary prominence using google books: A pilot study. *Knygotyra*, 78:111–139, 2022. ISSN 0204-2061. doi: 10.15388/Knygotyra.2022.78.108.

Mahsa Vafaie, Sven Hertling, Inger Banse-Strobel, Kevin Dubout, and Harald Sack. End-to-end information extraction from archival records with multimodal large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6075–6083, 2025.

Yunting Xie, Matti La Mela, and Fredrik Tell. Multimodal llm-assisted information extraction from historical documents: The case of swedish patent cards (1945-1975) and chatgpt. In *The 9th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2025), March 5–7, 2025, Tartu, Estonia*, pages 1–15. University of Oslo Library, 2025.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, et al. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21744–21754, 2025.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 837–847, October 2025.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Barry Zhang, Keith Lazuka, and Mahesh Murag. Equipping agents for the real world with Agent Skills. Anthropic Engineering Blog, October 2025a. URL https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-skills. Accessed: 2026-04-16.

Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*, 2025b.

A Chronos-UI Screenshot

Figure 4 shows the `Chronos-UI` during an agent session. The interface combines a standard workspace structure with a conversational agent panel and a source viewer. This allows historians to inspect primary-source scans, review intermediate outputs, and provide feedback without leaving the research environment. The viewer is central to the human-in-the-loop design of `Chronos`: whenever the agent analyses a source page or proposes a decision, the corresponding scan can be displayed for direct verification by the historian.

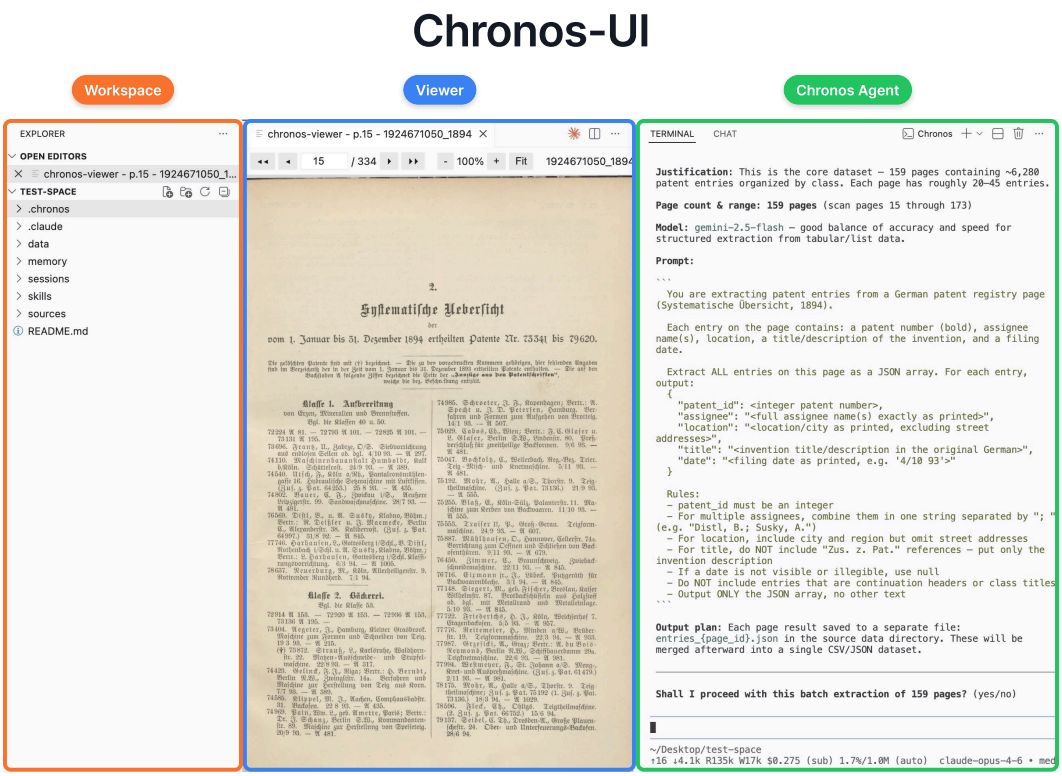


Figure 4: `Chronos-UI`. Screenshot of the `Chronos` interface, showing the workspace, source viewer, and conversational agent panel. The interface allows historians to inspect primary sources, interact with the agent, and validate outputs within a single research environment.

B Skill Prerequisite Mechanism

Listing 1: Front matter of a skill that finds the table of contents of a primary source.

```
---
name: toc-finder
description: Finds the table of contents
requires: none
---
You are an expert analyst of historical
documents (18th-20th century).

Step 1 -- Read the per-source memory file...
Step 2 -- Orient: call list_pages...
Step 3 -- Find the Table of Contents...
Step 3.1 -- ...
...
```

Each `SKILL.md` begins with a YAML front matter block containing three fields: `name` (a human-readable label), `description` (a one-line summary surfaced in the interface), and `requires` (a comma-separated list of artefact files that must exist before the skill will run). Listing 1 shows a minimal example.

The `requires` field is the mechanism that turns independent skills into ordered pipelines. When the agent is asked to execute a skill, it checks for each listed file before proceeding and halts with a clear message if any is missing. For instance, the `batch-extract` skill in the pipeline shown in Figure 2 declares `requires: batch_prompt.md`, ensuring that the `prompt-construction` skill has completed successfully and produced its artefact before batch extraction begins. This lightweight dependency mechanism enables multi-step workflows without requiring a separate orchestration layer: the historian or agent simply invokes skills in sequence, and the prerequisite check enforces the correct ordering.

C Dataset Links

Table 4 provides links to the source collections and annotated datasets used in our evaluation.

Table 4: **Links to benchmark datasets.** The table lists the public source collections and ground truth datasets used to evaluate `Chronos-Extract`.

Dataset	Link type	Link
German Patent Registers	Digitized source collection	Mannheim University Library
French Trade Directories	Ground truth dataset	Zenodo record
American Newspapers	Large-scale source dataset	American Stories
American Newspapers	Fully annotated evaluation pages	American Stories Training

D Example Skill: `locate-section`

Listing D shows an adapted version of the `locate-section` skill for a German historical source, used as Step 1 of the extraction pipeline in Figure 2. The skill is a single Markdown file that the agent reads and executes step by step; there is no accompanying code. It illustrates the level of detail at which domain procedures are captured: operational constraints (one page at a time, binary search discipline), source-specific knowledge (Berichtigungen and Änderungen as named section types, the printed-vs-tool page ID distinction), an embedded prompt for the vision subagent, and a human-verification loop that gates the final save. Other skills in the pipeline follow the same structural template.

Listing 2: Complete text of the `locate-section` skill.

```
---
name: locate-section
description: Locate the page range of the alphabetical name list in a historical German city directory,
      plus any Berichtigungen or Aenderungen sections.
---
You are an expert analyst of historical German documents (18th-20th century).

## Setup
The source directory contains PNG scans accessed via tools. To preserve your context, you do not view
     pages yourself -- delegate to ‘analyze_page‘, a specialist vision agent that returns summaries or
      answers questions about a single page.
The only valid end state is a verified result file written to the source’s ‘data/‘ directory. If
     uncertain about any finding, resolve it with additional tool calls before writing.

## Rules
**One page at a time.** Never issue concurrent ‘analyze_page‘ calls. Fire one call, wait, reason, then
      pick the next page.
**Inspect before reporting.** Every page number you report must be one you personally viewed via ‘
     analyze_page‘. Never trust a table of contents without verification.
**Page ID != printed page number.** Tool page IDs are sequential file indices. Printed page numbers
     inside the scan are different (e.g. printed "p. 347" may be tool ID 412). Always use tool IDs in
     tool calls and in the output.
**Gaps are normal.** Scans may have missing pages or non-consecutive IDs. ID N+1 is not necessarily
     physically adjacent to ID N. Widen your interval if a boundary search returns unexpected content.
```

```
**Verify both sides of every boundary.**
- Start: candidate page contains name-list entries; the page before does not.
- End: candidate page is still name-list; the page after is not.
**Strict binary search.** Given `[lo, hi]` where `lo` is outside and `hi` is inside the section,
     compute `mid = (lo + hi) // 2`, view it, and narrow. Repeat until `hi - lo == 1`. Do not guess
     multiple adjacent pages in parallel.
**Memory discipline.** Read `DOCUMENT_MEMORY.md` at the start; it may contain findings from prior runs.
      Update it every ~10 steps so partial results survive interruption. Prefer `edit` over `write`.
     Keep it concise.

## Task
Find the page range of the main alphabetical name list (residents, citizens, or subscribers) and
     identify any Berichtigungen (corrections) or Aenderungen (amendments).
**First occurrence only.** Some directories repeat the name list in multiple orderings (e.g. by
     surname, then by street). Report only the first. Use the alphabetical ordering itself to detect
     when a second pass begins.
**Berichtigungen & Aenderungen** may precede, follow, or be embedded within the main list. Search
     actively -- null is only acceptable after a targeted search, not by assumption.

## Procedure
**Step 1 -- Read memory.** Read `DOCUMENT_MEMORY.md` for prior findings.
**Step 2 -- Orient.** Call `list_pages`. Note first and last page IDs.
**Step 3 -- Table of contents and offset.** Scan the first ~15 pages with `analyze_page` to locate the
      Inhaltsverzeichnis. If found, read the printed page number for the start of the name list, then
     find the corresponding tool page ID. Compute `offset = tool_page_id - printed_page_number` and
     use it to estimate the end boundary from the TOC.
Use this prompt for every `analyze_page` call during search:
> Classify and summarise this page for the purpose of finding the name list in a historical German city
      directory.
> Step 1 -- Classify: assign one category:
>   A) Listing of names, addresses, or personal information
>   B) Table of contents
>   C) Other (advertisements, narrative text, notices, preface, etc.)
> Step 2 -- Report:
>   - A: state in the first sentence that the page contains a name/address listing. Specify the
     ordering. Give the first and last entry verbatim. List fields.
>   - B: state in the first sentence that the page is a table of contents. Report it completely,
     preserving structure and printed page numbers.
>   - C: state in the first sentence that the page does NOT contain a name listing. Describe the
     content.
**Step 4 -- Coarse scan (fallback).** If the TOC is missing or unusable, call `analyze_page` every ~50
      pages to bracket the section manually.
**Step 5 -- Binary search.** For each boundary, bracket `[lo, hi]` and perform strict binary search at
      the midpoint. No parallel guesses. Stop when one side is confirmed outside and the other inside.
**Step 6 -- Confirm.** Re-run `analyze_page` on both pages of each boundary pair. If anything looks
     off, return to Step 5.
**Step 7 -- Berichtigungen / Aenderungen.** Targeted search: 3 pages immediately before the start
     boundary, 3 pages after the end boundary, the first 5 pages, and the last 5 pages.
**Step 8 -- Update memory.** Write a concise summary to `DOCUMENT_MEMORY.md` (page ranges, document
     structure, anything useful for future runs). Preserve existing content.
**Step 9 -- Save.** Write `name_ranges.json` to the source's `data/` directory using the schema below,
      with `human_verified_start` and `human_verified_end` set to `false`. Proceed to Step 10.
**Step 10 -- Human verification.** For each boundary in turn:
  1. Call `show_page` with the boundary page ID.
  2. Ask the user: "This is the [first/last] page of the name list (page X). Does this look correct?"
  3. Wait for the user's reply -- do not proceed without it.
  4. If confirmed: mark the corresponding `human_verified_*` field `true`. If rejected: ask follow-up
     questions, adjust the boundary, re-save, and repeat until confirmed.
  5. After both boundaries are confirmed, save once more with the verified flags. This final save is
     the only acceptable end state.

## Output schema
Write `name_ranges.json` to the source's `data/` directory:
{
  "source": "<source folder name>",
  "name_list": {
    "start_page":  <int>,
    "end_page":    <int>,
    "description": <string>,
    "human_verified_start": <bool>,
    "human_verified_end":   <bool>
  },
  "amendments": {
    "berichtigungen": {
      "start_page":  <int | null>,
      "end_page":    <int | null>,
      "description": <string | null>
    },
    "aenderungen": {
```

```
      "start_page":  <int | null>,
      "end_page":    <int | null>,
      "description": <string | null>
    }
  },
  "notes": <string>
}
All page numbers are tool page IDs. Use ‘null‘ for sections actively searched and not found.
```

E Prompts

This section provides prompts and skills used by `Chronos-Extract`. They illustrate how extraction workflows can be specified in natural language and adapted to different historical source types. Each prompt encodes domain-specific rules, output formats, and constraints that guide the VLM during extraction.

Listing 3: Extraction prompt (`benchmarking/prompts/american-newspapers-v4.md`) for historical newspapers (Figure 5).

````
# American Newspapers - Extraction Prompt

Extract every **headline** and **article body** from this historical newspaper page.

## What to extract

- **Headlines**: The title text above an article. Multi-line headlines are common. Do not include
      bylines or author names.
- **Articles**: The body text of news reports, editorials, correspondence columns, poems, fiction,
      obituaries, and feature pieces. Start with the first sentence of the body - do not include the
      headline or byline.

Use model `gemini-3.1-pro-preview` with batch processing.

## What to exclude

Advertisements, classifieds, legal notices, mastheads, banner titles, subscription terms, section
      headers (e.g., "LOCAL NEWS"), page numbers, and decorative elements.

## Formatting rules

1. **Preserve column line breaks** exactly as printed. Do not reflow text or join hyphenated words
      across lines.
2. **Preserve original spelling and punctuation.** Do not modernize or correct.
3. **Continuation markers** (e.g., `[Continued on fourth page.`) should be copied as printed.

## Grouping rules

- **Correspondence columns** (a single correspondent reporting multiple local items in one column
      segment): bundle as **one article**.
- **"Locals" / "personal mentions" pages** where each item has a typographic separator: each item is
      its **own article**.
- **When in doubt, prefer bundling** over splitting.

## Output format

Return a JSON object with an `entries` array. Each entry has a `type` ("headline" or "article") and `
      text`.

```json
{
  "entries": [
    {"type": "headline", "text": "Hotel Life in New York."},
    {"type": "article", "text": "Mr. Fitzwalter was the pastor of an\nimpoverished but genteel colored
       con-\ngregation in Port Royal."},
    {"type": "headline", "text": "The Arrest of Parnell-Details of the\nAffair."},
    {"type": "article", "text": "Charles Stewart Parnell was arrested\nat his hotel in Morrison's..."}
  ]
}
```

## Final output

Consolidate results into a single `.xlsx` file with columns: `page`, `type`, `text`. One row per entry.
````

Listing 4: Extraction prompt (`benchmarking/prompts/french-directories.md`) for French trade directories (Figure 6).

````
# French Trade Directories - Extraction Prompt

First, review the full set of directory page images to familiarize yourself with the layout, typography
      , and entry structure before extracting.

Use model ‘gemini-3.1-pro-preview‘ with batch processing.

Extract every individual business or person entry from this directory page.

Each entry typically contains some combination of: a person or business name, their profession or
      activity, a street address, and a house number. Some entries also include honorary titles or
      facility type information.

## Output Format

Return a JSON object with an ‘entries‘ array. For each entry, provide:

- ‘full_text‘: The complete entry as continuous plain text. Preserve original spelling, abbreviations,
       and punctuation. Do not preserve column line breaks - join lines into flowing text.
- ‘PER‘: Array of person or business name strings found in the entry.
- ‘ACT‘: Array of profession or activity strings (e.g., "pharmaciens", "bottier", "essayeur du
      commerce").
- ‘LOC‘: Array of location or street name strings (e.g., "r. de la Chaussée-d’Antin", "place Dauphine")
      .
- ‘CARDINAL‘: Array of house/street number strings (e.g., "34", "12").
- ‘TITRE‘: Array of title or honor strings (e.g., "(Elig.)", "(Jur.)").
- ‘FT‘: Array of facility type strings (e.g., "fab. à Caen", "dépôt", "boutique").

If an entity type is not present in an entry, use an empty array ‘[]‘.

## Guidelines

- Extract every entry on the page, even short or partially legible ones.
- Preserve original 19th-century French spelling and abbreviations exactly as written (e.g., "r." for
      "rue", "fab." for "fabrique", "pl." for "place", "V.e" for "Veuve").
- Entries may span multiple lines - combine them into one entry.
- Do not modernize spelling, expand abbreviations, or correct apparent OCR-like artifacts in the
      original print.
- Some entries may contain numbers at the very end that are reference numbers, not street numbers -
      include them in ‘full_text‘ but only tag actual street/house numbers as ‘CARDINAL‘.

## Example Output

‘‘‘json
{
  "entries": [
    {
      "full_text": "Dufan et Clémendot, pharmaciens, r. de la Chaussée-d’Antin, 34. (Elig.) 449",
      "PER": ["Dufan et Clémendot"],
      "ACT": ["pharmaciens"],
      "LOC": ["r. de la Chaussée-d’Antin"],
      "CARDINAL": ["34"],
      "TITRE": ["(Elig.)"],
      "FT": []
    }
  ]
}
‘‘‘

## Final Output

After extracting from all pages, consolidate the results into a single ‘.xlsx‘ file for the entire
      source.
````

Listing 5: Extraction prompt (`benchmarking/prompts/german-patents.md`) for German patent registers (Figure 7)

````
# German Patents - Extraction Prompt

First, review the full set of patent page images to familiarize yourself with the layout, typography,
     and entry structure before extracting.

Use model ‘gemini-3.1-pro-preview‘ with batch processing.

Extract every patent entry from this page of the German Patent Office records (Kaiserliches Patentamt /
      Reichspatentamt).

Each entry typically contains: a patent identification number, the name of the inventor or assignee,
     optionally a location (city or address), a description of the invention, and a date.

## Output Format

Return a JSON object with a ‘patents‘ array. For each entry, provide:

- ‘full_entry‘: The complete entry as continuous plain text. Preserve original spelling and
     punctuation. Do not preserve column line breaks - join lines into flowing text.
- ‘patent_id‘: The patent number as a string (e.g., "919", "1312").
- ‘assignee‘: The inventor or applicant name exactly as written (e.g., "OTTO, TH.", "PHILIPPE, E., und
      J. M. GENET").
- ‘location‘: The city or address if present, otherwise ‘null‘. (e.g., "Berlin S.W., Kochstraße 10", "
     Paris", or ‘null‘).
- ‘title‘: The patent title or description of the invention.
- ‘date‘: The date as printed (e.g., "12. September 1877").

## Example Output

‘‘‘json
{
  "patents": [
    {
      "full_entry": "OTTO, TH. Kuppelungsvorrichtung zum Anhängen der Wagen auf Seilbahnen an das
     Zugseil nebst Verbesserung an dieser Vorrichtung. 12. September 1877. P. R. 919.",
      "patent_id": "919",
      "assignee": "OTTO, TH.",
      "location": null,
      "title": "Kuppelungsvorrichtung zum Anhängen der Wagen auf Seilbahnen an das Zugseil nebst
     Verbesserung an dieser Vorrichtung",
      "date": "12. September 1877"
    }
  ]
}
‘‘‘

## Guidelines

- Only extract patent entries that are **complete** on the page. Skip entries that are truncated at
     the top-left or bottom-right of the image (partial entries that continue from a previous or onto
     a next page).
- Preserve German characters exactly as written (ä, ö, ü, ß, s).
- Preserve original spelling and technical terminology - do not translate or modernize.
- Some entries have multiple inventors - include all names in the ‘assignee‘ field as a single string.
- The patent ID may appear in different positions within the entry (beginning, end, or prefixed with "
     P. R.", "D. R. P.", "Nr.", etc.) - extract just the number.

## Final Output

After extracting from all pages, consolidate the results into a single ‘.xlsx‘ file for the entire
     source.
````

F Primary Source Examples

This appendix provides visual examples of the diverse primary sources used in our benchmarking evaluation. Because layout, typography, and image degradation strongly influence extraction performance, these examples illustrate the complexity and heterogeneity of the historical datasets.



Daily Kennebec Journal.

PUBLISHED EVERY MORNING (SUNDAYS EXCEPTED), BY SPRAGUE & SON. TERMS: SEVEN DOLLARS PER ANNUM. SINGLE COPIES, THREE CENTS.

VOL. XI. AUGUSTA, MAINE, THURSDAY MORNING, JULY 1, 1880. NO. 154.

The Kennebec Journal.

Railroads and Steamboats.

Maine Central Railroad.

Change of Time.

Gardiner Accommodation Train,

EASTERN STEAMBOAT CO.

BATH & BOOTHBAY LINE.

FREIGHT TRAINS.

For Boothbay, Ocean Point, Squirrel, Capitol and Mouse Islands.

FOR BOSTON!

STAR of the EAST.

Maine Steamship Co.,

REOPENING

PROVIDENCE LINE TO NEW YORK

STEAMER MASSACHUSETTS,

STEAMER RHODE ISLAND,

FIRE WORKS

NEW GOODS!

—JUST RECEIVED AT—

E. ROWSE'S,

—CONSISTING OF—

Gold and Silver Watches, Ladies' and Gents' Gold Chains, Lockets, Crosses, Charms, Neck Chains, Gold and Plated Sets, Rings, Studs, Sleeve Buttons, Shawl, Lace and Cuff Pins,

Gold and Silver Thimbles,

Silver Plated Tea Sets, Ice Pitchers, Goblets, Cake Baskets, Fruit Stands, Cups, Vases, and Spectacles of all Styles.

LARGEST STOCK! LOWEST PRICES

124 WATER ST., AUGUSTA, ME.

SMILEY & WASGATT

HAVE ON HAND A FINE STOCK OF

TIN, CLASS AND CROCKERY WARE.

Also other Goods in their line. Any one needing

Jig Saws and Designs for Brackets, Etc.,

Will find them here. They are now running a

FIVE CENT COUNTER!

Business Cards.

REMOVAL!

DR. C. W. DAVIS, Dentist,

Wm. McDAVID, Jr., SURGEON AND Janitorial Tatist.

DENTIST.

E. J. Roberts, D. D. S.

Office: First Door to Post Office.

HUNT'S BLOCK, Water St., Augusta.

NEW DEPARTURE!

A FULL LINE OF

Choice Groceries & Provisions,

Low as the Lowest

WM. G. FIFIELD & CO.,

(Successors to DOW & STORER.)

COFFINS CASKETS BURIAL ROBES

Ware Rooms on Bridge St.

Reduced Prices. Very Low.

Ware Rooms on Bridge St., near R. R. Crossing

C. E. TOWN.

COAL! COAL!

Lehigh & Franklin.

Deliveries Carefully Made.

Equal to Best on Kennebec River.

No. 9 Bridges' Block,

GIVE US A CALL.

F. SWAIN & CO.

NEW AND ELEGANT STYLES

SMITH AM. ORGANS

SOLD FOR CASH, OR ON EASY MONTHLY INSTALMENTS.

C. C. HUNT,

Opp. Post Office, Augusta, Me.

SHINGLES!

A few sawed cedar will be at wharf in a few days for sale cheap. Apply to

J. MANCHESTER HAYNES.

BOWDOIN COLLEGE

Good Fitting, Easy and Effective TRUSSES!

SOLD AT

PARTRIDGES' DRUG STORE.

Furniture.

NOTWITHSTANDING The Fire in

FURNITURE.

CHAMBER SETS, PARLOR SUITS AND TABLES,

Upholstery Work a Specialty, Repairing at short Notice.

Coffins, Caskets, Robes,

C. R. & H. U. WELLS.

FURNITURE

WHOLESALE AND RETAIL AT

WHIPPLE'S MANUFACTORY, HALLOWELL.

CURTAINS AND FIXTURES, COVERINGS AND UPHOLSTERERS' SUPPLIES

Ceo. O. WHIPPLE.

SMITH & HUNT

Have taken the store recently occupied by Wm. B. HUNT, opposite the post office, and have opened a first class

GENTS' FURNISHING STORE

—AND—

Tailoring Establishment!

They are prepared to cut and fit to suit in the latest and boldest styles at reasonable prices.

NEW STORE! NEW GOODS!

Come and See.

COAT and PANT MAKER WANTED.

Plumbing & Steam Heating

—BY—

HENRY A. TABER,

163 WATER STREET,

(Opposite American Express Office.)

Steam Heating and Ventilating!

SATISFACTION GUARANTEED!

MOUSE ISLAND!

SEASON 1880.

Now Ready for Visitors.

L. L. Morrison, Purchasing Agent.

General Manager of Hotel.

Col. Chas. A. Nielson, Jr., late of Astor House, New York.

Clerk and Cashier, E. F. DANFORTH.

Cooks and Table First Class.

Steam and Fast free of charge to all regular points.

Hot, Fair Water Baths, everyone should try! Helps Rheumatism, Neuralgia, Headache, Nervously, &c. Open to all, July 15, at Mouse Island, under decision of Bath Boat Club.

ROWING REGATTA.

A DELICIOUS ICE CREAM

CAN BE PROCURED AT

HART'S,

Opposite Post Office.

MISS A. C. MORGAN'S English, French, and German BOARDING SCHOOL for YOUNG LADIES, Portsmouth, N. H.

Good Fitting, Easy and Effective TRUSSES!

SOLD AT

PARTRIDGES' DRUG STORE.

For Sale or to Let.

For Sale

RARE CHANCE to purchase a fine residence in this city. Property offered for sale consists of two-story house, containing 12 rooms, with gas, hot and cold water, and in good condition. Lot 100x126, fruit trees, bushes, &c. Location one of the best in the city. For further particulars call on or address E. W. WHITEHOUSE, A.T.

Augusta, Me., May 4, 1880. may5tf

For Sale or Rent!

HOUSE No. 11, State St. One of the most desirable residences in Augusta. Has a large garden back of the house. Also for sale, a three seated Carriage, nearly new. For particulars inquire of SAMUEL TITCOMB.

FOR SALE!

SEVERAL Fine Residences in good locations in this city, varying in prices from $2500 to $4500; also desirable family cottages in desirable locations, at prices ranging from $450 to $2500, all in the city; also some desirable houses to let, with rent either at will or for a term of years; also several small places in the suburbs of the city, at cash prices.

E. W. WHITEHOUSE, Freemans Bank Building.

apr21tf

Miscellaneous.

NOTICE.

ALL persons who are indebted to the late firm of C. H. & C. A. WADSWORTH, are called upon to call immediately with E. B. THOMPSON, Allen's Block, Augusta.

BAY STATE HOUSE,

382 Hanover Street, Boston, Mass.

Entirely refitted and furnished. Most convenient location for business men in the city. Horse cars to Scollay Square, which have one passing the house every few minutes to all parts of the city. Chain bath, cooking, table and sanitary arrangements perfect.

W. A. WATES, Proprietor. A. J. HEMENWAY, Manager.

sept21eod3m

NOTICE.

I HEREBY notify the public that I have been appointed Superintendent of Burials, for the City of Augusta, and in pursuance of Forest Lawn Cemetery, I am prepared to attend to all matters appertaining to the duties of that office. All orders may be left at my residence, No. 23 Oak Street.

CHARLES A. WADSWORTH

apr28tjs

I have won this Cross . . . I have worn the Finest in the Land . . . I have worn it Night and Day for Long Years . . . I say here is what

The Admirable, Flexible Hip Corset!

Contains 112 Bones. It is the Easiest and most Perfect Fitting corset made. Ladies, you will wear no other after once giving them a trial. Sizes from 18 to 30, in white and colored. A full assortment just received. Also a full line of

LADIES' SUITS AND ULSTERS,

In Linen, Grass Cloth, and Gray Poplins for Summer wear. Suits from $2.00 and upwards. Ulsters from $1.25. All of these goods are New Styles and direct from New York Manufacturers. A lot of new styles and pretty patterns in

Ladies' Print Wrappers,

Shetland Shawls, Circulars, Skirts, Parasols, Handkerchiefs in Silk, Linen and Cotton, and a full line of

DRY AND FANCY GOODS IN GENERAL.

JOHN PARSONS,

152 WATER ST., AUGUSTA.

may17tf&w

STATE AGENCY

For the New England Copy and Law Reporting Co. 603 Crane Lanes, Steward and Groundlers, however the Law Reporting Co. 603 Crane Lanes, Steward and Groundlers, however New Crops greatly improved.

N. B. Waltz Lanes handsomely cleaned at lowest prices.

STEAM FEATHER RENOVATING: Feather Beds, Pillows and curled hair cleansed by steam. The only way that insures a warm, sweet, and thoroughly renovated bed. Also Feathers dressed by steam at the

Steam Dye House & Feather Drying Establishment.

Water St., Augusta, Me., Established 1867.

EMILE BARBIER & CO.

Awarded First Premium at Maine State Fair 1879 for Any kinds of Dress Goods, in the piece, or made into garments, Ribbons, Fringes, Socks, Velvets, Shawls, Kids, Feathers, etc., dyed or cleaned in the best manner and on short notice. Also, Gents garments dyed, cleaned, repaired and pressed ready to wear. Carpets and Lace Curtains cleaned. Velvet trimming dyed without ripping. Goods received and returned promptly by express. Send for circular price list. Prices reduced. Orders left at Dye House, Water St., Augusta, Me., will be called and delivered if desired.

L. M. GROSVENOR, Hillsor, Agent, Hallowell. C. C. CHANDLER, Dry Goods, Agent, Winthrop

may20eod3m&w6

TO LET!

MILLNOKETT LODGE

AT

Ocean Point!

This sea side House is charmingly located upon the south shore, commanding an extensive view. Sea Bathing, Boating, Fishing and Driving unsurpassed by any Sea Shore Resort. It contains sixteen rooms and is supplied with water and other conveniences which render it a desirable home for a family or a few private families. The house is furnished. For further particulars enquire of

STARBIRD'S PHOTOGRAPH ROOMS

125 Water St., Augusta,

may7eodtf

Fine Cigars, Tobacco & Snuffs, at PARTRIDGES' DRUG STORE

SULPHUR BITTERS

Are just what they claim to be—the GREAT BLOOD PURIFIER, invigorating and strengthening the system, driving out humors from the blood, and restoring the patient to perfect health. For all forms of insanity and weakness, filling the veins with pure rich, life-giving fluid, making, indeed, a new man of the invalid or sufferer with not the blessings of life.

Miraculous Cure. I CAN NOW GO

Without my Crutches.

NORTH BRIDGEWATER, Mass., May 28, 1880.

Gentlemen:—I was taken sick with Rheumatism of the Spine, and for three months I was confined to my bed, and for part of the time was helpless. I was under the care of two physicians, but got no relief. I was induced to try your SULPHUR BITTERS, and before I had taken one bottle I was able to get out of bed, and before I had taken three bottles I was able to walk without crutches very well, which I have used for the last eight months. I am better than I have been for 15 years. I feel like a new man. They are the greatest medicine I ever saw. Respectfully, RANDALL N. WILSON.

Do not delay, but try this great Spring Remedy. There is no calomel mercury in Sulphur Bitters and is absolutely free from Alcohol.

A RUNNING SORE

For 14 Years Cured.

An Extraordinary Cure.

LEWISTON, Me., April 6, 1880.

DEAR SIRS:—I have suffered with a running sore on my leg for the past 14 years, caused by exposure in the army during the rebellion. I have had the best of the most discouraging constantly. I have tried every remedy that I heard of, but never received any benefit until I commenced using SULPHUR BITTERS. I commenced taking them 2 weeks ago, and the sore is entirely well. I can say that SULPHUR BITTERS HAVE SAVED MY LIFE. Yours truly and thankful, HENRY HANCOCK.

Before we purchase goods, we ask you to call at

DR. KAUFFMAN'S

SULPHUR BITTERS

ARE PREPARED ONLY BY

A. P. ORDWAY & CO., CHEMISTS, Sole Proprietors for U. S. and Canada, Lawrence, Mass.

LIEBIG CO'S COCA BEEF TONIC.

SOLD BY ALL DRUGGISTS.

The Bleachery is open with the Spring 81 Ins from Boston and New York.

Straw Chip, Leghorn, Neapolitan, and Fancy Braids cleansed and made into the latest styles.

Oak Street, back of the Post Office.

S. N. WOOD.

BASS! BASS! BASS!

Come One! Come All!

—TO—

"HAMMOND'S GROVE,"

IN MANCHESTER.

and catch the shabby bass. Only 15 miles from Augusta. Take the Kennebec & Portland R. R. to Manchester Station. There you will find teams in waiting to meet those for picnic parties, a good camping ground for picnic parties, a good camping ground for those who wish to spend a few days or weeks. Ice and pure cooked water will be furnished every day.

C. H. HAMMOND, Proprietor.

EXECUTORS' SALE.—Pursuant to the order of the subscribers, as the Probate Court for the County of Kennebec on Monday, May 3, 1880, we the petition of Benjamin F. Allen & Co., praying for leave to sell at public auction or private sale, all the right, title and interest of the said Benjamin F. Allen & Co. in and to the following described property, viz: The entire stock of the Clothing and Furnishing Goods, and other personal property belonging to the said Benjamin F. Allen & Co. at the store of said petitioners, situated on Water street, in Augusta, and there sell the same, at such time and place as we may appoint, after giving notice of the time and place of such sale by publishing the same three weeks successively in the Kennebec Journal, a newspaper printed in Augusta, that all persons interested may attend thereat, and that the sale be made at Augusta, and there make return to said Court. Now, therefore, we give notice that the sale of said property will be made at the store of said petitioners, on Water street, in Augusta, on Saturday, the 15th day of May, A. D. 1880, at ten o'clock in the forenoon, and that the same be sold at public auction.

H. L. BAKER, Judge.

Attest: CHARLES HEWINS, Register.

may6 3w

Tales and Sketches.

HOW CLAY WHITTLE'S MANHOOD WAS CROWNED.

M. L. STONE.

Ethel Bruce sat on a low rocker in the recess of the bay windows, apparently in deep thought. A very pretty picture she made, so her Brother Hal thought as he passed along the street in company with Clay Whittle. Tufts were creeping above and around her head reaching after the sunlight, while lilies, geraniums, and lobelias were blooming at her feet.

The gentlemen passed on, as also did the hands of the clock, but still Ethel sat in evident study. The room was handsomely furnished with cultured taste, and skillful hands kept it looking very inviting to the family and their friends. It was library and sitting room combined. Ethel's parents had always striven to have their home attractive, even if other expenses had to be curtailed very closely.

Ethel was an only child, and had been for several weeks, and the willing hands and thoughtful heart of the mother were sadly missed by the daughter. The rest of the family said "that Ethel took mother's place to them all so well, that for the time they were not great losers," and the mother confined to her room for lack of strength to be elsewhere, rather than from actual pain said "that she was having a rest in mind and body, for Ethel did not neglect anything" and she kept the little mother fresh and dainty, and the room so bright and cozy, that all the family would gather around her in social glee.

Ethel would first notice mother's tired look if they stayed too long, and pleasantly chiding them would drive them off to the sitting room. Clay Whittle used often to come in with Hal to the mother's room, and these little visits were seasons of delight to him. His mother had passed the portals over into the other side. How he missed her room and her counsel, and the time spent over at the Bruces, were choice minutes and great helps to him.

His father was a leading politician, his sister a fashionable woman, while he himself was a late student.

The evening before Clay and Ethel had attended an entertainment together and on passing out from it, she expressed a moment to speak to a friend and Clay was joined by his sister.

Clay she asked "Why didn't you bring Eva Dryton this evening."

"I prefer the company I have," he answered.

"Company good enough I dare say, but I like to see a girl with some style of dress. Everything she wears is a year old."

"Which of itself is a crime of the first degree in the fashionable world" he answered, joined Ethel and passed on.

So on this the following morning, Ethel was trying to study out some plans for new clothes. Of course she was indignant over the remark, but it was true that all she had, was a year old, but she had felt that they were good enough, until now. She knew too, that all they had to live upon was her father's salary, which might have seemed large to many, but their expenses were correspondingly large, and with mother's sickness and Tess's college bills, or a part of them, was absolutely had to economize and really deny themselves, and she had felt willing, and even anxious to do her share; but after overhearing the remark of the night before, denying herself had not been so sweet.

She examined the closets to see if there was anything available to be made over, but returned to her low rocker disheartened. Her father came in, and seeing something amiss, asked, "What are your thoughts like, daughter, and what are your wishes?" Speak, and if half of the kingdom is desired, it shall be yours!"

"I thought I was queen of the realm, and mistress of all I survey. I desire no more than the home kingdom, but a robe of royal purple."

"Or some other fashionable shade," he said. "Frankly, Ethel, what were you thinking about?"

"I was wondering if there was any surplus money in the family treasury, that I could use for myself," she said.

"I think about four dollars. Would it help you?"

"Yes, indeed. Just the sum I want."

"Well father seemed to hesitate. Ethel spoke quickly, "do not mind it if it is needed elsewhere."

"I'll let you know at ten-time," he said. "Will balance accounts for the month, and rest assured, you shall have it, daughter, if father can spare it," and seated on one knee with her arm twined around his neck, she stroked his manly beard, caressingly.

"We are not rich, Ethel," he said, "but this home is ours, and is a very pleasant, and an attractive one, and I have a family of whom I am both proud and thankful. There is one thing upon which I congratulate myself every week of my life, and it is this—that there are no unexpected bills coming in for payment contracted by my family. I know that they never buy without my advice and counsel, and then under the full extent. At the end of the month I come out even, at least. Here is an incident which struck me forcibly, but is only one among the many of my experience. As I passed through the store one day, I saw Mrs. Drayton and Eva buying goods. The bill was brought to my desk, copied and 'please remit' added, as is our custom, and placed in an envelope to be handed to parties or mailed. Opportunity for giving it to Mr. Drayton occurred on my way to dinner. He opened it, glanced at the total amount, and looked very much disturbed."

"Was it large?" asked Ethel.

"Something over an hundred dollars" "Inside of thirty days will answer" he asked.

"Yes, I answered and passed on. The time has expired and no payment yet."

"He is very able to pay it?" exclaimed Ethel.

"Perhaps no more so than your father, Ethel; and several such bills coming unexpectedly would seriously embarrass

[Continued on fourth page.

Ducrot, *menuisier*, r. de la Mortellerie, 150. 408
Ducrot, *parfumeur*, r. S.-Merri, 18. 441
Ducrot, *potier de terre*, r. Montorgueil, 62. 456
Dudognon, *march. de plâtre*, r. des Marais-S.-Germain, 7. 403
Duboit, *maçon*, r. de la Bucherie, 14. 402
Ducz, *menuisier*, r. de la Tixeranderie, 51. 408
Dufaillet, *chaudronnier*, r. Vieille-du-Temple, 107. 314
Dufart (Pierre), *libraire*, q. Voltaire, 19. 392
Dufau et Clémendot, *pharmaciens*, r. de la Chaussée-d'Antin, 34. (Elig.) 449
Dufaut (Victor), *libraire*, r. du Gros-Chenet, 2. 392
Dufay, *essayeur du commerce*, place Dauphine, 5. 359
Dufay, *chaudronnier*, r. du Pont-aux-Choux, 15. 314
Dufay (V.e), *grenetière*, r. du Faub.-S.-Denis, 20. 372
Dufay, *libraire*, r. Montmartre, 82, passage du Saumon, 38. 392
Dufay, *papetier*, r. S.-Martin, 20. 437
Dufay, *march. de toiles*, r. des Mauvaises-Paroles, 22. 487
Dufay-Coqueret, *quincaillier*, r. Phélipeaux, 28. 458
Dufeu, *charcutier*, r. Montmartre, 89. 318
Dufey fils, *bijoutier*, passage de la Réunion, 7. 284
Dufey, *boucher*, r. S.-Louis-au-Marais, 3. 297
Dufey, *march. de vins*, r. S.-Honoré, 216. 498
Duffaud, *bijoutier*, r. S.-Martin, 107. 284
Duffaud, *maçon*, r. S.-Honoré, 353. 402
Duffaut, *chaudronnier*, r. de la Sourdière, 1. 314
Dufils, *limonadier*, r. des Colonnes, 6. 395
Duflocq, *architecte*, r. Basse-Porte-S.-Denis, 28. 277
Duflocq, *march. de bois à brûler*, r. de l'Arcade, 13. 291
Duflos, *horloger*, r. du Faub.-S.-Martin, 60. 376
Duflos, *papetier*, r. Poissonnière, 10. 437
Duflos, *rouennerie*, r. Bourg-l'Abbé, pass. de l'Ancre, 34. 412
Duflot, *quincaillier*, quai de l'Ecole, 19. 458
Duflot, *loueur de carrosses*, r. de Paradis-Poissonnière, 22. 505
Dufort aîné, *bottier*, r. Pagevin, 7. 295
Dufort, *bottier*, Palais-R., gal. vitrée, 215. 295
Dufort fils, *bottier* (*exp.* 1819.), r. J.-J.-Rousseau, 18. 295
Dufort, *tabletier*, r. Jean-Pain-Mollet, 10. 476
Dufossé, *teinturier*, r. du Faub.-Montmartre, 63. 485
Dufour, *ag. d'affaires*, r. du Chaume, 15. 275
Dufour, *march. de planches*, r. Chapon, 28. 290
Dufour et Besnard, *march. de bois à brûler*, quai de la Tournelle, 17, et r. des Fossés-S.-Bernard, 11. 291
Dufour jeune, *march. de bois à brûler*, r. de la Pépinière, 53. 291
Dufour (Mad.), *march. de chapeaux de paille*, passage du Panorama, 11. 307
Dufour, *chapelier*, r. S.-Honoré, 385. 309
Dufour, *cirier*, r. de l'Arbre-Sec, 66. 316
Dufour et co., *commiss.*, r. des Contres-S.-Gervais, 18. 320
Dufour, *corroyeur*, r. des Vertus, 11. 327
Dufour (Charles), *épicier*, r. S.-Denis, 307. 352
Dufour, *fab. d'éventails* (*exp.* 1819), r. Beaubourg, 48. 360
Dufour, *faiencier*, r. S.-Anne, 50. 362
Dufour, *h. garni*, r. Ventadour, 7. 378
Dufour (Gabriel), *libraire*, r. de Vaugirard, 34. 392
Dufour, *maçon*, r. S.-Jean-Baptiste, 4. 402
Dufour, *architecte*, r. de Chartres-S.-Honoré, 12. (Elig.) 278, 402
Dufour (Mad.), *nouveautés*, r. Neuve-des-Petits-Champs, 4. 412
Dufour Chabrol, *papetier*, r. S.-Martin, 132. 437
Dufour (J.) et co., *fab. de papiers peints*, (A 1819), r. Beauveau-S.-Antoine, 10. 439
Dufour et co., *fab. de schals*, r. Ste.-Apolline, 5. 371, 495
Dufour, *march. de vins*, r. Ste.-Marguerite-S.-Germain, 6. 498
Dufour, *march. de vins*, r. des Boucheries-S.-Germain, 67. 498
Dufour, *march. de vins*, r. des Fossés-S.-Germain-l'Auxerrois, 31. 498
Dufourmantelle (L.), *peaussier*, r. Tiquetonne, 3. 446
Dufourmantelle (Louis), *peaussier*, r. du Petit-Lion-S.-Sauveur, 19. 446
Dufourmantelle aîné, *march. de peaux*, r. S.-Denis, 62. 446
Dufourmantelle jeune, *peaussier*, r. de la Vieille-Monnaie, 19. 446
Dufourmantelle (Jacques), *peaussier*, r. Bourg-l'Abbé, 50. 446
Dufraisne, *tailleur*, r. Richelieu, 4. 479
Dufrenel, *plumassier*, r. S.-Denis, 307. 453
Dufrénoy, *march. de vins*, r. Guérin-Boisseau, 48. 498
Dufresne (J.-G.), *courtier de commerce*, r. S.-Denis, 97. 333
Dufresne, *fab. de crayons* (*exp.* 1819), r. de la Contellerie, 19. 336
Dufresne, *épicier*, r. du Marché-aux-Poirées, 12. 352
Dufresne (Mad.), *lingère*, r. S.-Denis. 67. 398

Figure 6: Sample page from French Trade Directories.

99282. Jamain, P. E. M., Dijon; Vertr.: Dr W. Häberlein, Berlin N.W., Karlstr. 7. Maschine zum Pressen von Pastillen. 17/8 97. — A 713.

99283. Kugel, Dr L., Wien; Vertr.: A. du Bois-Reymond u. M. Wagner, Berlin N.W., Schiffbauerdamm 29a. Hörrohr. 23/11 97. — A 730.

99307. Katz, H., Berlin C., Landsbergerstr. 71. Cuvettenhalter. 28/8 97. — A 748.

99308. Büdingen, Dr Th., Mainz, Rheinstr. 71. Einrichtung für Heißluft-, Dampf- und Wasserbäder. 19/9 97. — A 748.

99363. La Société Générale des Verseurs hermétiques „L'Egide", Paris; Vertr.: O. Lenz, Berlin N.W., Luisenstr. 31b. Druckluft-Vorrichtung für Spritzen u. dgl. 23/9 97. — A 764.

99383. Neumeister, Dr R., Leipzig u. Dr M. Matthes, Jena. Verfahren zur Herstellung eines diagnostisch verwerthbaren Eiweißpräparates. 1/2 98. — A 781.

99427. Cournand, J., Paris; Vertr.: M. Mintz, Berlin W., Unter den Linden 11. Vorrichtung zum Reinigen der Zähne. 12/9 97. — A 764.

99428. Hegedüs, St., u. K. Krug, Szentes, Ung.; Vertr.: G. Dedreux, München. Bett mit Walzen zum Heben und Auswechseln des Bett-Tuches. 18/9 97. — A 764.

99617. Farbenfabriken vorm. Friedr. Bayer & Co., Elberfeld. Verfahren zur Darstellung einer gegen Säuren beständigen und in Alkalien schwer löslichen Tanninalbuminat-Verbindung. 13/10 97. — A 835.

99765. Helmers, Dr O., Hamburg, Bohnenstr. 21. Verfahren zur Herstellung fast geruchloser Präparate aus den durch Einwirkenlassen von Schwefelsäure auf Mineralöle und ähnliche Kohlenwasserstoffe gewonnenen, sulfidartig gebundenen, Schwefel enthaltenden Substanzen. 15/12 97. — A 835.

99798. Völkl, G., Ingolstadt. Verband zum Strecken und Einrichten gebrochener oder verrenkter Beine oder Füße seitens des Arztes ohne fremde Beihilfe. 11/11 97. — A 931.

99799. Kumpf & Rethwisch, Stuttgart, Heusteigerstr. 108. Saugpfropfen mit Luftzuführungs-Vorrichtung. 10/12 97. — A 911.

99800. Grünewald, F., Berlin W., Grunewaldstr. 11. Selbstfeuchtende Kopfwaschbürste. 22/1 98. — A 931.

100083. Estermann, L., Berlin W., Friedrichstr. 171. Leibbinde. 21/10 97. — A 931.

100084. Gardiner, S. H., Brooklyn; Vertr.: C. Fehlert u. G. Loubier, Berlin N.W., Dorotheenstr. 32. Verbandschiene. 23/11 97. — A 912.

100085. The French Hygienic Society Limited, London; Vertr.: H. E. Schmidt, Berlin S.W., Friedrichstr. 234. Zerstäuber mit Vorrichtung zum Regeln der zu zerstäubenden Flüssigkeitsmenge. 5/2 98.

100241. Chemische Fabrik auf Aktien (vorm. E. Schering), Berlin N. Apparat zur Desinfektion mit Formaldehyd; Zus. z. Pat. 96671. 13/6 97. — A 912.

100249. Krämer, W., Charlottenburg, Pestalozzistr. 31. Formtisch-Schaltwerk von Stempelpressen zur Herstellung von Pastillen u. dgl. 12/6 97.

100256. Volland, P., Berlin S., Prinzenstr. 87. Elektrische Bohr- und Fräsmaschine für zahnärztliche Zwecke. 19/11 97.

100376. Fütterer, H., Düsseldorf, Städt. Tonhalle. Haartrocken-Vorrichtung mit elektrischer Heizung. 19/2 97.

100377. Wittkowski, J., Berlin N., Friedrichstr. 125. Zahnklammer mit Gestell zum Ueberspannen einer Gummihaut (Kofferdam). 25/12 97. — A 912.

100378. Adam, H., Breitenstein b/Stolberg, Harz. Saugflasche. 23/2 98.

100424. Bruns, W., Brandenburg a/H. Verfahren zur Herstellung einer für Zahnfleisch-Nachahmung geeigneten Kautschukmasse. 3/4 98. — A 932.

100467. Paschen, R., Dessau. Verbandschiene für Knochenbrüche, Verrenkungen und andere Verletzungen. 30/3 97.

100603. Gerson, Dr K., Berlin S.W., Friedrichstr. 9. Elastische Pflasterbinde, hauptsächlich als Ersatz des Suspensoriums. 11/11 97.

100604. Tresenreiter, M., München, Neuhauserstr. 40. Befestigung künstlicher Zahnkronen an natürlichen Wurzeln. 23/11 97.

100605. Timar, J., Berlin C., Gr. Hamburgerstr. 20. Athmungsvorrichtung, besonders zur ärztlichen Behandlung der Athmungsorgane. 8/4 98.

100617. Lahr, E., Würzburg, Räsburgstr. 5. Verfahren zur Herstellung einer dichten, haltbaren Trennungswand in Glasröhren. 18/3 98.

100707. Knoll & Co., Ludwigshafen a/Rh. Verfahren zur Darstellung einer fast geruch- und geschmacklosen Ichthyoleiweißverbindung. 4/4 97.

100850. Kröning, Frau H., Magdeburg, Olvenstedterstr. 28d. Vorrichtung zum Auffpreizen natürlicher Körperkanäle, hauptsächlich zwecks Einführung von Arzneimitteln. 31/12 97.

100851. Averbeck, F., Bonn-Poppelsdorf, Friedrichstr. 24. In einem durchbohrten Flaschenpfropfen verschiebbarer Tropfstöpsel. 1/1 98.

100852. Schmidt, J. A., Barmen, Alleestr. 116. Zerlegbares Heft für Werkzeuge, besonders zu chirurgischen Zwecken. 8/3 98.

100853. Nesselbosch, R., Vaihingen a/Enz, Württ. Instrument zum Untersuchen von Kehlkopf und Rachen. 20/3 98.

100854. Schille, J., Wien; Vertr.: Dr W. Häberlein, Berlin N.W., Karlstr. 7. Federnder, sich durch Druck der Hand einem Körpertheil dicht anschließender Trichter für ärztliche Zwecke. 26/3 98.

100888. Ferguson, R. G., Lakewood, Ocean Cty, N.-J., u. A. M. Elkus, New-York; Vertr.: F. C. Glaser u. L. Glaser, Berlin S.W., Lindenstr. 80. Apparat zum Mischen von Flüssigkeiten mit Gasen, im Besondern zur Herstellung von Brausebädern. 19/10 97.

100909. Ferracciu, R., Rom; Vertr.: R. Deißler, J. Maemecke u. Fr. Deißler, Berlin N.W., Luisenstr. 31a. Suspensorium. 2/11 97.
Der Patentinhaber nimmt für dieses Patent die Rechte aus § 3 des Uebereinkommens mit Italien vom 18/1 92 auf Grund einer Anmeldung in Italien vom 2/7 97 in Anspruch.

101021. Kaeding, A., Halberstadt. Einrichtung zum Auswechseln des Bett-Tuches von Krankenbetten. 18/8 97.

101022. Bogren, A., Malmö, Schweden; Vertr.: J. Tunelo u. H. Hallencreutz, Berlin N.W., Lessingstr. 48. Vorrichtung, um Zahnhöhlen mit erwärmtem Wasser auszuspülen und mit erwärmter Luft auszutrocknen. 3/3 98.

101185. Giese, Gebr., & Co., Offenbach a/M. Fahrbare Tragbahre. 21/8 97.

101186 Haacke, J., geb. Bornschein, Apolda, Goldgasse 4. Halter beim Schleifen künstlicher Zähne. 1/12 97.

101187. Hentschel, C., Kattowitz, Mühlstr. 17/19. Zerstäuber für Flüssigkeiten. 11/2 98.

101188. Herrmann, Dr, Alt-Küstrinchen, N/M. Aufblasbares kugelförmiges Pessar. 27/3 98.

Figure 7: Sample pages from the German Imperial Patent Gazette dataset.