Uncovering the Past with AI – How MyHeritage Extracts Historical Records from Newspapers

By Maya Geier, Product Manager at MyHeritage

Introduction

This syllabus accompanies the Legacy Family Tree webinar led by Maya Geier, offering a deep dive into MyHeritage's advanced AI technologies for transforming historical newspapers into structured, searchable genealogical records. Attendees learn about the OldNews.com platform, the company's custom-built OCR, and entity extraction tools that can interpret context, infer missing details, and connect disparate data points. The syllabus summarizes key concepts, workflows, and practical tips to help researchers make the most of newspaper archives for family history discoveries.

---

1. Overview of OldNews.com

- Launched in March 2024 as MyHeritage's dedicated historical newspaper platform.
- Grew from 100 million to over 400 million digitized pages worldwide.
- Focuses on exclusive microfilm scanning and continuous content expansion.
- Accessible through standalone OldNews.com subscriptions or included in the MyHeritage Omni plan.

---

2. AI-Driven OCR and Entity Extraction

- Custom-built OCR pipeline designed for historical print challenges: faded ink, antique fonts, damaged pages.
- OCR accuracy of ~99.25%, outperforming major competitors.
- Entity extraction models identify genealogically relevant details—names, dates, locations, relationships—from free-text.
- System can infer missing details (e.g., calculating exact death dates from relative references).
- Structured data enables automatic record matching and integration with MyHeritage family trees.

---

3. Key Record Types and Examples

- Obituaries: Rich in genealogical value; often include relatives, residences, life events, and causes of death.
- Birth, marriage, military service, and social event announcements extracted into structured records.
- AI-generated article summaries allow quick content assessment.
- Example: Webinar host Geoff discovered a 1921 road-building donation by his great-grandfather in Mink Creek, Idaho, live during the session.

---

4. Global Expansion of AI Extraction

- Over 15 billion structured records extracted to date (a 50% increase in MyHeritage's historical record database).
- Language coverage now includes English and French; German and Spanish underway.
- Long-term goal: global coverage with multilingual extraction models tailored to each language's grammar and cultural context.

---

5. Practical Tips for Researchers

- Search by both free-text and structured records for comprehensive results.
- Provide as much detail as possible (names, places, dates) to refine searches.
- Use the article summary to assess relevance before reading full text.
- Leverage interconnected records to explore relatives and associates.
- Attach and extract data from records directly into your MyHeritage family tree.

Conclusion and Further Exploration

MyHeritage's AI technology unlocks the hidden potential of historical newspapers, turning hard-to-search clippings into rich, structured genealogical resources. By combining high-accuracy OCR, intelligent entity extraction, and contextual inference, researchers can discover connections and life events that traditional searches might miss. View the full

webinar replay to see live demonstrations—including Geoff's personal discovery—and explore the accompanying syllabus resources for practical search strategies and advanced tips.