DATA QUALITY & MULTILINGUAL ANNOTATION

High-precision data engineering for international archival and speech datasets.

Role: Data Processor

Company: LifeWood Data Technology Ltd.

Languages: English, Italian, French, Spanish

OVERVIEW

This work demonstrates my capability to transform complex, low-quality archival and audio data into validated, structured datasets used in international research and language technology workflows. My focus was on maintaining absolute data accuracy, consistency, and operational efficiency across diverse multilingual sources.

01 THE CHALLENGE: UNSTRUCTURED DATA AT SCALE

LifeWood Data Technology Ltd. required the processing of raw, high-variance genealogical and audio data. The project was hampered by three primary hurdles:

ILLEGIBILITY

Historical archives featured archaic, handwritten text that was difficult to decipher.

INCONSISTENCY

Original documents lacked uniform structures across different regions and time periods.

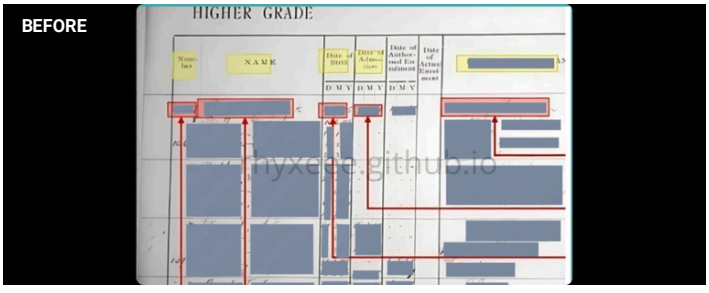
LANGUAGE VARIABILITY

Records spanned four languages (French, Italian, Spanish, and English), requiring multilingual technical proficiency.

02 CORE CONTRIBUTIONS: ARCHIVAL RESTORATION

A. Archival Data Restoration & Digitization

I extracted and digitized vital data from scanned birth, marriage, and death certificates.



Methodology: Standardized formats across diverse records to ensure uniform schema and field integrity.

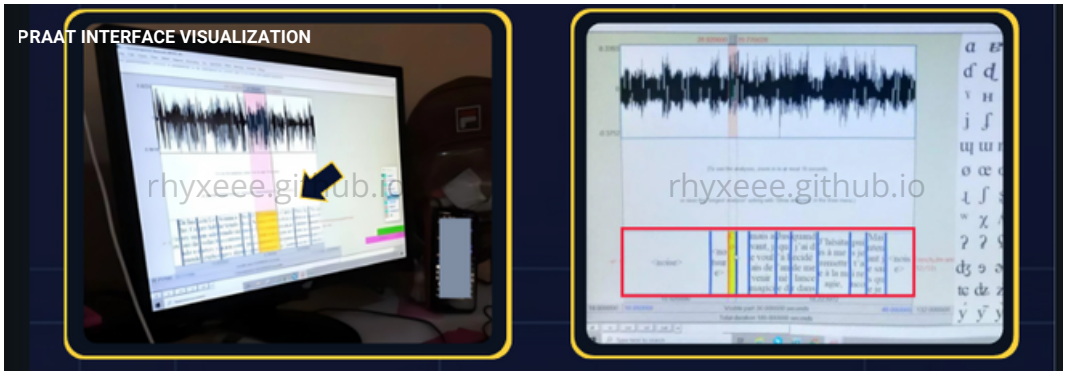
Result: Delivered structured genealogical datasets with over 98% validated accuracy.

TECHNICAL SOPHISTICATION & IMPACT

02 CONTINUED: ACOUSTIC SEGMENTATION

B. Acoustic Segmentation & Phonetic Annotation

I performed signal-level audio processing to create high-fidelity datasets for speech analysis.



Methodology: Utilized Praat to align waveforms, spectrograms, and IPA phonetic annotations for Italian and English audio.
Result: Achieved 99% accuracy in language identification and resolved all transcription ambiguities.

03 QUANTIFIABLE IMPACT

70– 80%
Increasing in Data Usability

25– 30%
Reduction in Processing Time

25,000+
Records Digitized

RESUME SUMMARY

- **Data Quality Engineering:** Digitized and structured 25,000+ high-variance genealogical records across four languages, achieving a validated 98%+ accuracy rate.
- **Acoustic Data Processing:** Performed speech segmentation and phonetic annotation using Praat, resolving ambiguities to deliver fully validated training datasets for international clients.
- **Process Optimization:** Improved dataset usability by ~80% and reduced archival processing cycles by 30% through standardized data-validation protocols.

Technical Stack Applied

Data Validation: Spreadsheet-based normalization and schema integrity checks.
Speech Analysis: Praat (Segmentation, Spectrogram Analysis, IPA Annotation).
Multilingual Support: Proficient processing of French, Italian, Spanish, and English datasets.