The Diary Project

Making My Memories Machine-Readable

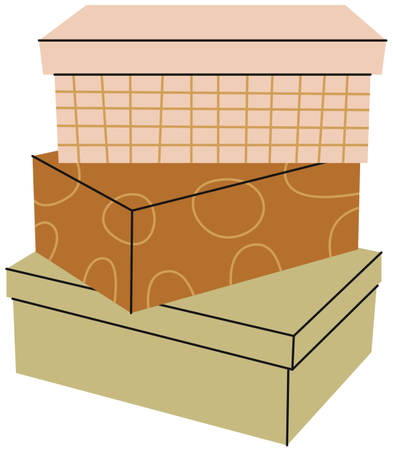
Tara Hulcome

IMT 542 A: Portable Information Structures

3 June 2025

TABLE OF CONTENTS

INTRODUCING: MY PERSONAL INFORMATION STORIES



The Shoebox Problem

People often store meaningful personal information in handwritten, messy, unstructured formats that are hard to access, share, or analyze.

Key users

Ink-based diarists... like me!

My goal

Develop strategy to convert my diary entries into portable, accessible digital structure that enables search, analysis & long-term preservation.

PROCESS > PERFECTION

Figure 1: Personal collection of hard copy diaries

DESIGNING A NEW DIGITAL INFORMATION PRODUCT

In Scope

- Use of dummy diary data (300 entries).
- Test end-to-end process.
- Create sample JSON structure, taxonomy & metadata.
- Build information system to host diary entries through web-based API.
- Include interactive topic filters.
- Design clean, simple user interface.

Out of Scope

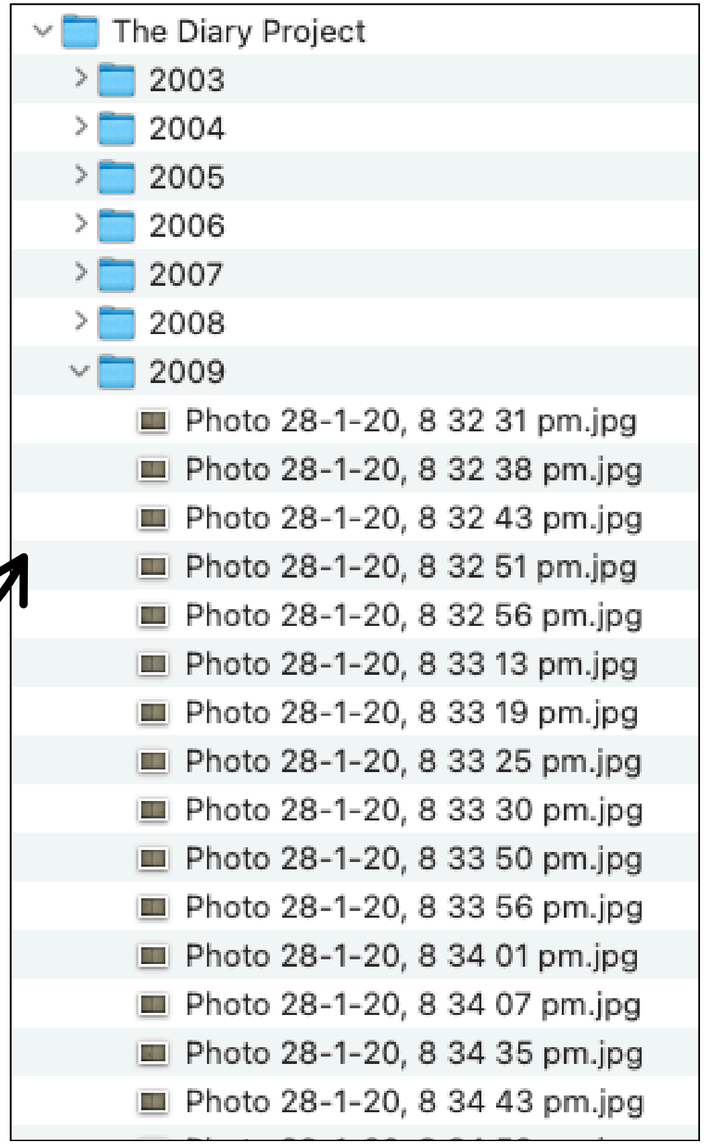
- Use of real diary entries.
- JPEG-to-JSON conversion with Optical Character Recognition (OCR) tools.
- AI-powered sentiment analysis.

Key Requirements

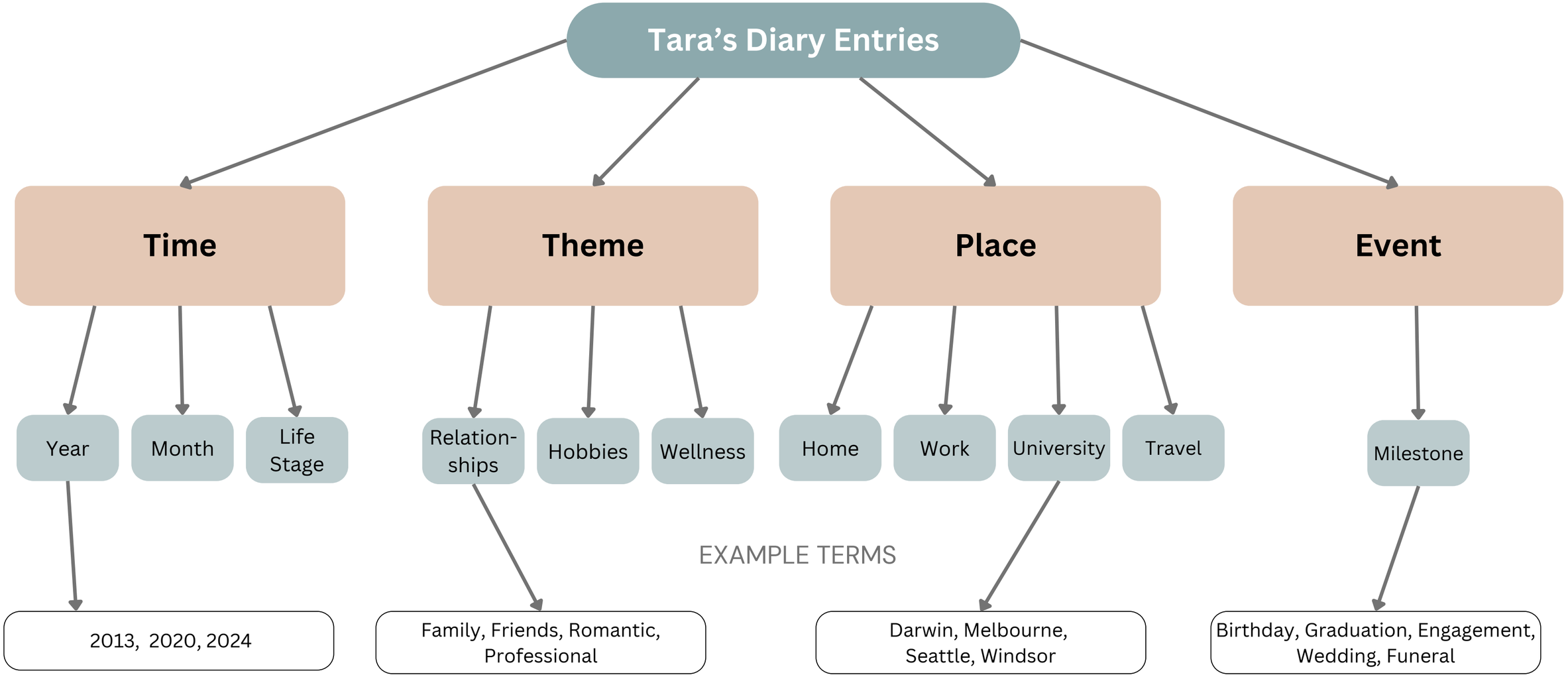
- Security emphasis.
- Easy & cost effective to maintain.
- An enjoyable user experience.

EXISTING STRUCTURE: WHAT I'VE GOT TO WORK WITH

Existing Information	**Hard copy files**	**Digital files**
Structure / Quality	• 22 diaries (~8,030 entries) • Poor quality: inconsistent layout & handwriting.	• ~1,144 JPEG files. • High quality: large file size & high resolution images.
Storage Solution	• Hard copies stored in shoe boxes.	• Stored in cloud (DropBox). • Password-protected.
Access / Analysis	• Very hard to retrieve. • Search & analysis cumbersome / impossible.	• Poor folder structure hinders retrieval & analysis.
Portability	• Very low portability - difficult to move.	• Better portability - but room for improvement.
Risks	• Vulnerable to physical damage, loss, or exposure.	• Potential for data breaches.



GIVING STRUCTURE TO THE STORIES: A NEW TAXONOMY



GIVING STRUCTURE TO THE STORIES: CONTROLLED VOCAB

- Developed controlled vocab list *(to align with taxonomy)*
- Four top-level categories
  - Time
  - Theme
  - Event
  - Place
- Defined sub-categories & specific terms
  - *See example →*
- An iterative process!

EXAMPLE CONTROLLED VOCABULARY LIST

Top-level category	Sub-category	Term	Example tags
Time	Month	September	
Time	Month	October	
Time	Month	November	
Time	Month	December	
Time	Life Stage	High School	
Time	Life Stage	University - Undergraduate	
Time	Life Stage	University - Graduate	
Time	Life Stage	Cadetship	#time-life_stage-cadetship
Time	Life Stage	Career	
Time	Life Stage	Van Life	
Theme	Hobbies	Music	
Theme	Hobbies	Camping	#theme-hobbies-camping
Theme	Hobbies	Baking	
Theme	Hobbies	Cooking	
Theme	Wellness	Hiking	
Theme	Wellness	Yoga	#theme-wellness-yoga
Theme	Wellness	Pilates	
Theme	Wellness	Running	
Theme	Wellness	Strength	
Theme	Wellness	Aerial	
Theme	Wellness	Kayaking	
Theme	Wellness	Skiing	
Theme	Relationships	Family	
Theme	Relationships	Friends	
Theme	Relationships	Romantic	#theme-relationships-romantic
Theme	Relationships	Professional	
Event	Milestone	Birthday	#event-milestone-birthday
Event	Milestone	Graduation	
Event	Milestone	Engagement	
Event	Milestone	Wedding	
Event	Milestone	Death	

Top-level category	Sub-category	Term	Sub-Term	Example tags
Place	Home	Australia	Darwin	#place-home-australia
Place	Work	Australia	Canberra	
Place	Work	Australia	Gold Coast	
Place	Work	China	Beijing	
Place	Work	Fiji	Suva	
Place	Work	India	New Delhi	
Place	Work	Indonesia	Jakarta	
Place	Work	Japan	Tokyo	
Place	Work	PNG	Port Moresby	
Place	Work	Russia	Vladivostok	
Place	University	Australia	Darwin	
Place	University	Australia	Melbourne	
Place	University	Canada	Windsor	#place-university-canada-windsor
Place	University	United States	Seattle	
Place	Leisure Travel	Australia	Adelaide	
Place	Leisure Travel	Australia	Brisbane	
Place	Leisure Travel	Australia	Coffs Harbour	
Place	Leisure Travel	Australia	Gold Coast	
Place	Leisure Travel	Australia	Melbourne	
Place	Leisure Travel	Australia	Sydney	
Place	Leisure Travel	Australia	Hobart	
Place	Leisure Travel	New Zealand	Auckland	
Place	Leisure Travel	New Zealand	Queenstown	
Place	Leisure Travel	New Zealand	Christchurch	
Place	Leisure Travel	New Zealand	Taupo	
Place	Leisure Travel	United Kingdom	London	

FROM PAPER TO PLATFORM: THE TRANSFORMATION PROCESS

- **Transformation process**: Convert handwritten diary entries into structured, machine-readable format.
- **Goal**: Enhanced accessibility & findability, enabling user (me!) to interact with data in more enjoyable way.

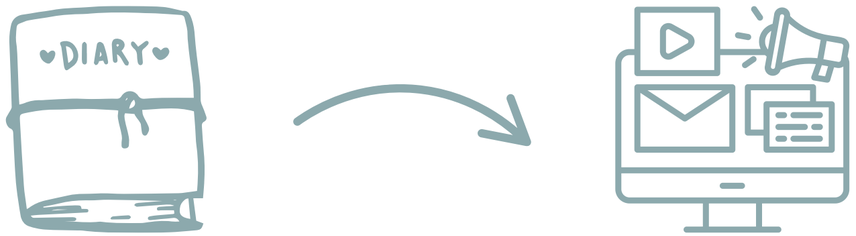
Old vs New: Key differences

Information remains the same.

Format: Hard copy → JPEG → JSON

Structure: Unstructured, messy → Highly structured schema with defined fields.

Access: Manual reading → Digital access (personal DropBox) → Digital access (web-based API)



FROM PAPER TO PLATFORM: THE TRANSFORMATION PROCESS

Key steps

1 Extract text from JPEG file & convert to standard JSON format.
(...or use AI to create a dummy dataset to play with!)

2 JSON structured in line with new taxonomy & defined metadata fields.

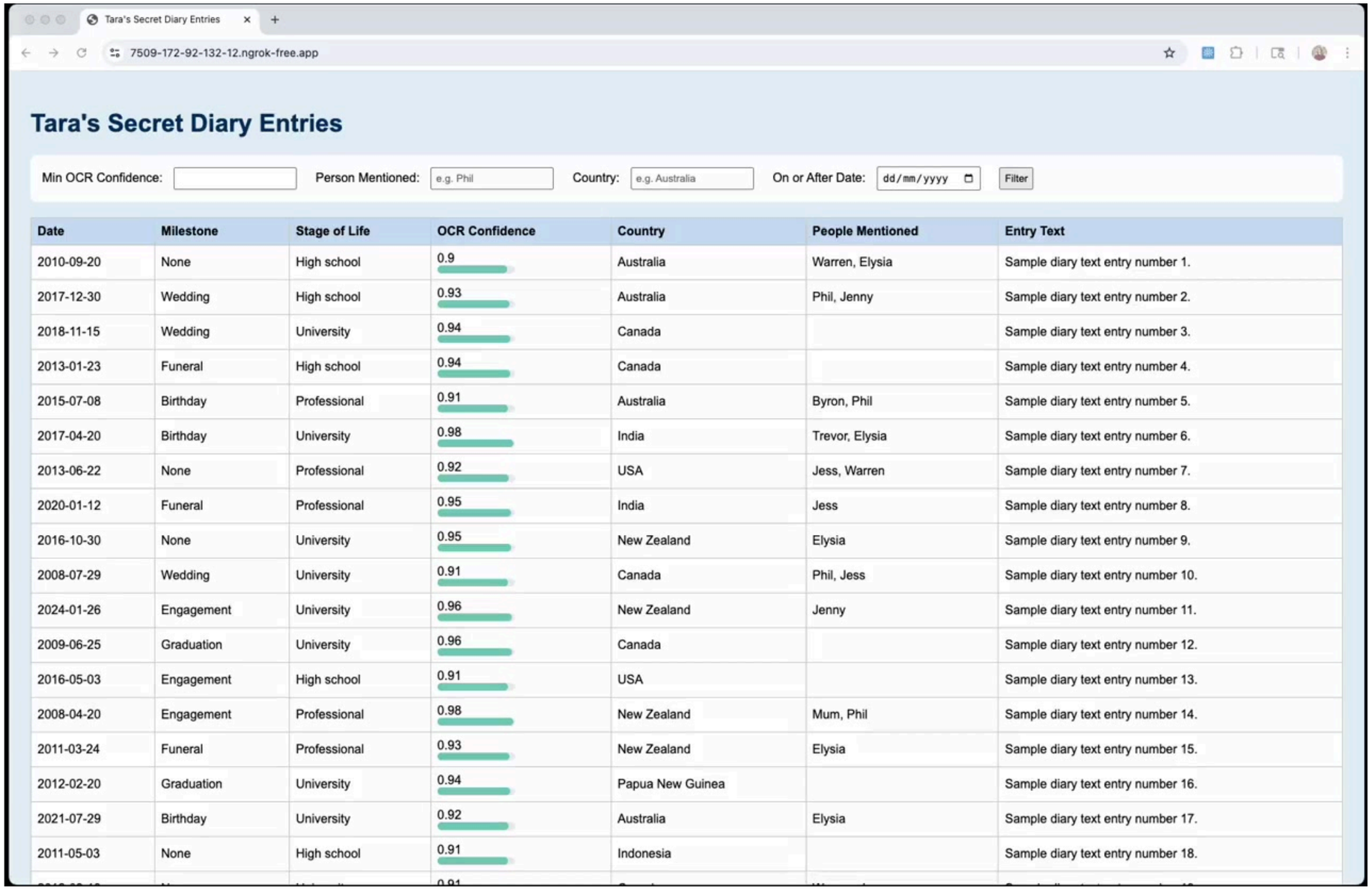
3 Build & run Flask-based API with multiple endpoints, allowing user to access, filter & retrieve specific data.

4 Develop front-end HTML interface with user-friendly features.

5 Use ngrok to expose app to web, enabling access from any internet-connected device.


```
{
  "entry_id": "2025-05-14_01",
  "date": "2025-05-14",
  "time": "20:45",
  "location": "Seattle, WA",
  "entry_text": "Today, I went to a dentist appointment in the morning and then went to campus for several hours of Information Architecture study. I can't believe we only have 3 weeks' left of class before we graduate! Later on, I did a HIIT class with Phil at Community Fitness in Roosevelt.",
  "topics": ["hobbies", "university", "fitness"],
  "people_mentioned": ["Phil"],
  "tags": ["study", " graduation"],
  "ocr_confidence": 0.94,
  "sensitivity": "confidential_personal",
  "audit_log": [
    {
      "action": "ocr_extracted",
      "timestamp": "2025-05-15T14:33:00",
      "performed_by": "system"
    },
    {
      "action": "manual_edit",
      "timestamp": "2025-05-15T09:15:22",
      "performed_by": "tara_hulcome"
    }
  ]
}
```

A SNEAK PEEK OF TARA'S DIARY ENTRIES



Date	Milestone	Stage of Life	OCR Confidence	Country	People Mentioned	Entry Text
2010-09-20	None	High school	0.9	Australia	Warren, Elysia	Sample diary text entry number 1.
2017-12-30	Wedding	High school	0.93	Australia	Phil, Jenny	Sample diary text entry number 2.
2018-11-15	Wedding	University	0.94	Canada		Sample diary text entry number 3.
2013-01-23	Funeral	High school	0.94	Canada		Sample diary text entry number 4.
2015-07-08	Birthday	Professional	0.91	Australia	Byron, Phil	Sample diary text entry number 5.
2017-04-20	Birthday	University	0.98	India	Trevor, Elysia	Sample diary text entry number 6.
2013-06-22	None	Professional	0.92	USA	Jess, Warren	Sample diary text entry number 7.
2020-01-12	Funeral	Professional	0.95	India	Jess	Sample diary text entry number 8.
2016-10-30	None	University	0.95	New Zealand	Elysia	Sample diary text entry number 9.
2008-07-29	Wedding	University	0.91	Canada	Phil, Jess	Sample diary text entry number 10.
2024-01-26	Engagement	University	0.96	New Zealand	Jenny	Sample diary text entry number 11.
2009-06-25	Graduation	University	0.96	Canada		Sample diary text entry number 12.
2016-05-03	Engagement	High school	0.91	USA		Sample diary text entry number 13.
2008-04-20	Engagement	Professional	0.98	New Zealand	Mum, Phil	Sample diary text entry number 14.
2011-03-24	Funeral	Professional	0.93	New Zealand	Elysia	Sample diary text entry number 15.
2012-02-20	Graduation	University	0.94	Papua New Guinea		Sample diary text entry number 16.
2021-07-29	Birthday	University	0.92	Australia	Elysia	Sample diary text entry number 17.
2011-05-03	None	High school	0.91	Indonesia		Sample diary text entry number 18.

...BUT ARE THEY CORRECT AND COMPLETE?

What I'm looking for

Completeness:

- GET request returns full list of entries in JSON format.
- Initial GET request returns >300 entries.
- Application of filters returns entries including expected metadata fields (date, year, text, people, topics, etc).

Accuracy:

- OCR confidence score ideally >0.95.
- Manual scan of text looks good.

Red flags

Completeness:

- GET request returns fewer entries than expected / not in structured format.
- Application of single filters does not return data as per expected logic.
- Combination of filters does not return data as per expected logic.

Accuracy:

- OCR confidence level <0.90.
- Spot check: text just looks wrong.

QUALITY GOALS AND PERFORMANCE TESTING

Quality Goals

- JSON schema validated.
- Metadata fields populated consistently.
- Returned data is as expected.
- OCR level > 0.95

Identifying + Remediating Issues

- Flag low-confidence OCR for review.
- Manual QA / spot-checking.
- Conduct functional & performance tests to identify specific improvements.

Functional Testing

- **End point accessibility**
  - GET requests
- **End point filtering**
  - Empty, single & combined filters
  - Abnormal inputs

Performance Testing

- **Response time**
  - <1 second for 100+ entries.
- **Filter efficiency**
  - Combination of filters ≠ delay.

Security Testing

- Critical due to use of personal data.

NEXT STEPS: STILL SOME WORK TO DO

Next steps

- Address filtering issues.
- Implement robust security plan.
- Explore cloud hosting options.
- More sophisticated UI.
- Transition to use of real data.
- Longer term: AI sentiment analysis?