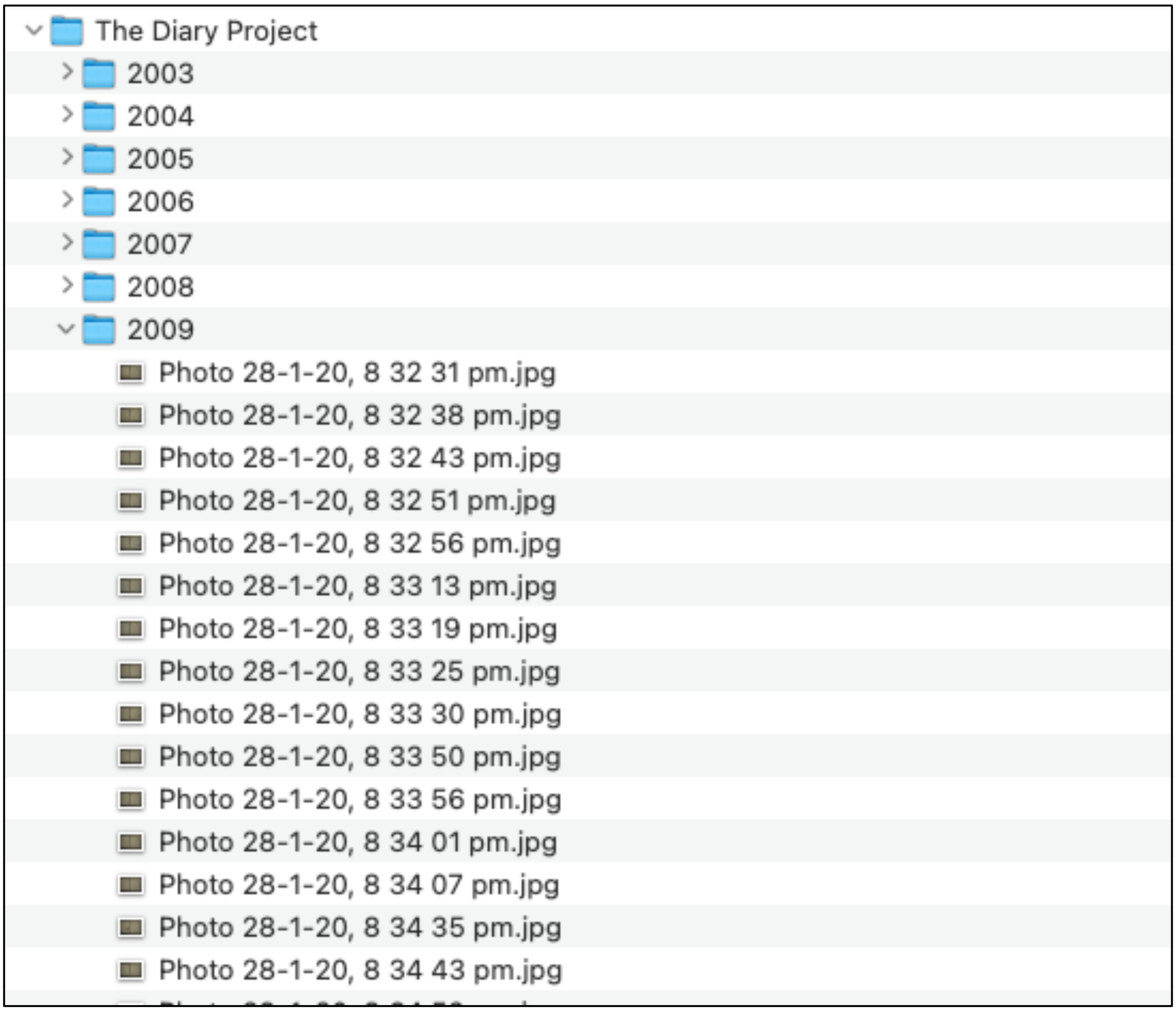
G7: Improve the existing information structure and format
Tara Hulcome

Project summary: Convert handwritten diaries to machine-readable format ('The Diary Project').

Sample 1: Current file structure



```
The Diary Project
  2003
  2004
  2005
  2006
  2007
  2008
  2009
    Photo 28-1-20, 8 32 31 pm.jpg
    Photo 28-1-20, 8 32 38 pm.jpg
    Photo 28-1-20, 8 32 43 pm.jpg
    Photo 28-1-20, 8 32 51 pm.jpg
    Photo 28-1-20, 8 32 56 pm.jpg
    Photo 28-1-20, 8 33 13 pm.jpg
    Photo 28-1-20, 8 33 19 pm.jpg
    Photo 28-1-20, 8 33 25 pm.jpg
    Photo 28-1-20, 8 33 30 pm.jpg
    Photo 28-1-20, 8 33 50 pm.jpg
    Photo 28-1-20, 8 33 56 pm.jpg
    Photo 28-1-20, 8 34 01 pm.jpg
    Photo 28-1-20, 8 34 07 pm.jpg
    Photo 28-1-20, 8 34 35 pm.jpg
    Photo 28-1-20, 8 34 43 pm.jpg
```

Figure 1: *Current file structure (.jpg files containing scanned diary entries, saved in cloud-based Dropbox folder)*

Sample 2: New structure (JSON)

```
{
  "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"
    }
  ]
}
```

Question 1: How will your structure promote maintaining and measuring quality easier?

In terms of the seven steps to ensure and sustain data quality (Shen, 2019), my new structure will have a particular focus on promoting **rigorous data profiling and control of incoming data.**

I will use optical character recognition (OCR) tools to extract text from the scanned .jpg images and then convert to JSON format *(as shown in Sample 2 above)*. The use of a standardized JSON format promotes quality control by organizing the extracted data into pre-defined fields. This means that, for example, missing or distorted dates are easier to detect compared to free-form text.

I will then experiment with writing Python code that automatically checks for the OCR confidence level and includes the result in a metadata field *(see JSON extract below)*. According to Microsoft (2024), 'the estimated confidence level of each word is calculated as a range of 0 to 1. The confidence score represents the certainty in the accuracy of the result. For example, an 82% certainty is represented as an 0.82 score.'

OCR confidence level	**Interpretation**
"ocr_confidence": 0.94,	We have 94% certainty in the accuracy of this diary entry. This is an **excellent** confidence score.
"ocr_confidence": 0.54,	We have 54% certainty in the accuracy of this diary entry. This is a **poor** confidence score.

Evaluating the OCR tool's accuracy in interpreting my (messy!) handwriting is an effective method of data profiling. This promotes easier quality measurement and maintenance because, if the confidence level score for a particular diary entry is much lower than others, that will flag that there are possible data anomalies or inconsistencies requiring manual corrective action.

It will also be useful to conduct regular testing of the OCR tool for different diary entries, to monitor whether there are any discernible changes in the confidence level over time (for example, due to changes in handwriting style or paper quality).

Question 2: How will your structure enable any information security issues to be identified and managed down the road?

Of the three information security objectives (confidentiality, integrity and availability), **confidentiality** is my major priority as the diary entries contain highly sensitive, personally identifiable information. Confidentiality means that ‘only authenticated and authorized individuals can access data and information assets’ (Exabeam, n.d.). With a structured dataset and clearly defined API (Application Programming Interface), I aim to manage data confidentiality by implementing a highly secure access protocol that supports multi-factor authentication. Ideally, this protocol will also support audit trails and breach detection, logging attempts by unauthorized users to gain access.

As a second strategy for promoting confidentiality, the JSON entries will include a metadata field that classifies sensitivity levels *(see Sample 2 extract below)*. This makes it easier to apply appropriate security controls to higher-risk content.

```
"sensitivity": "confidential_personal",
```

References

Microsoft. (2025, October 9). *Get optical character recognition (OCR) insights.* https://learn.microsoft.com/en-us/azure/azure-video-indexer/ocr-insight

Shen, S. (2019, July 28). *7 steps to ensure and sustain data quality.* TechTarget. https://www.datasciencecentral.com/7-steps-to-ensure-and-sustain-data-quality/

Exabeam. (n.d). *The 12 Elements of an Information Security Policy.* https://www.exabeam.com/explainers/information-security/the-12-elements-of-an-information-security-policy/