Streamlining Data Retention Auditing and File Management Through Automation

09 June 2026 14:42

Name: Chris L Powell

Student Number: 2508868

Module: Changemakers - Creativity and Value Creation
Assignment 1: Innovation and Project Management

Task 1 - Ideation and Innovation

07 July 2026 21:56

Q1: Innovation

19 May 2026 11:56

Innovation creates value, and solves problems in creative ways, it is critical for the continued progress of society. Unit 1, ‘How innovation shapes the world’ covered of the key innovations throughout human history, as well as why innovation is important.

Some of the top Innovations from the Last Decade

1. Proliferation of contactless payments
2. Robot Vacuums
3. Generative AI
4. Proliferation of smart devices for health tracking e.g. smart watch/smart ring
5. Smart home technology such as lights, thermostats, etc

Which invention had the biggest impact on the development of society?

The invention of agriculture to replace our hunter-gatherer roots by far had the biggest impact. Technological progress, prior to agriculture, was stagnant or sporadic, with innovation becoming the norm after its introduction. [1]

Is Innovation changing things for the better?

Innovation has most definitely caused a positive impact on society. While there are still improvements to be made, health, wealth, and education are significantly improved at an accelerated rate. Not all innovations are good, but this is why we should strive to continually innovate.

We're entering the 'fourth industrial revolution', comprised of the advent of 'cyber-physical systems' to further embed technology into the everyday operation of society.

Biggest WEE Technological impact in my discipline

Big data will greatly influence the data we have available, alongside accessory industries like computing power, data storage, analytics, and machine learning/AI [2]. Data availability and accuracy can influence decisions across every single industry.

Creativity

Creativity IS innovation. We shouldn't be frightened of being wrong, as it will stifle innovation. We must try new things at the risk of being wrong, or we risk remaining stagnant. Creativity as a skill is increasingly coveted for both social and business applications [3]

Without encouragement to innovate, and my personal drive to fulfil my own creativity, I would have long since stagnated in my professional career.

Q2: Ideation

19 May 2026 11:58

Idea 1: Create bespoke training material for Python programming language [4]

Bring skills learned from this course to the rest of my team, opening opportunities to streamline processes and upskill for their own career development. Content should be customised to be relevant to the day-to-day activities of my organisation to maximise interest and knowledge retention.

Idea 2: Programmatically assist in performing Data Retention regulatory control

Manual check takes considerable time, prone to human error, and leads to vast over retention of data, potentially opening employer up to regulatory risk [5], and leading to extra cost retaining unnecessary data.

Idea 3: Create a script to streamline dynamic creation of a batch files.

Frequently need to create batch files to load csv data to cloud storage using a SQL query. Each file is unique to the data being loaded and the destination, currently must be created by hand. Very tedious and time consuming for datasets with many fields.

Chosen project = **Idea 2**

This idea would provide the maximum benefit, should it prove successful, by freeing up employee time, making material cost savings, and reducing exposure to regulatory risk [6].

Proposed Project Title: Streamlining Data Retention Auditing and File Management Through Automation

Aim: To develop a tool that will heavily reduce effort needed to audit and manage file/data retention.

Description: Shared document servers have a set disk size that semi-frequently is reached, causing processes to fail and leading to several clean-up activities each year.

Additionally, we have a regulatory responsibility to only retain data for a fixed period. Managing the data retention policy is currently manual, leading to over retention, additional storage costs, wasted employee time, and possible exposure to regulatory risk for retaining data beyond the retention period.

This project aims to streamline the process for review and deletion of data that breaches this data retention policy, theoretically alleviating all the above issues.

Objective 1: Programme will scan all directories & sub-directories from a starting location chosen by a user, returning a single output of combined results. Metadata returned should have 100% accuracy.

Objective 2: Programme should only target files according to specific criteria set by the user. e.g. files last modified before a certain date. File targeting should have 100% accuracy demonstrated during testing.

Objective 3: Programme should have the ability to delete files, also utilising the user defined criteria. 100% of eligible files should be deleted during testing, with all information recorded in a log file.

Objective 4: Programme should be simple to use so that non-technical colleagues are able to use it effectively. 80% of users should be able to complete core functionality without additional instruction.

Q3: Research

19 May 2026 11:58

The project being entirely technical means that all the key resources needed to support its progression will be learning resources of technical documentation. After some research, the following have been identified to help achieve the project aim.

1. Automate everything with Python | LinkedIn Learning

   Specifically, the sections:

- Read files with Python
- Organise directories with python
- Parse data with python
- Extract data with regular expressions
- Validate inputs with automation
- Automate error handling

2. Python from zero | LinkedIn Learning

   The sections:

- Lists: Add, insert, delete
- For Loops
- Functions
- Input

3. Accessing Files & Directories in Python - Using OS & Path Modules - isfile isdir & exist
4. Using Python to List the files in a Directory
5. os.path — Common pathname manipulations — Python 3.14.5 documentation
6. Writing CSV files in Python - GeeksforGeeks

Task 2 - Project Management

07 July 2026 21:56

Q1: Project Scope

19 May 2026 11:59

The main deliverable: Code that will have the ability to scan (loop) through directories/sub-directories, gather a suite of metadata on directory contents along the way, collate the metadata into a single data object before exporting to a user readable file.

The code should take user input to use to filter results. It should further enable the user to delete directory contents that meet the criteria.

Outcome:

1. Reduction in time required to implement data retention internal control.
2. Reduced data retention and associated cost of data storage.

Required resources:

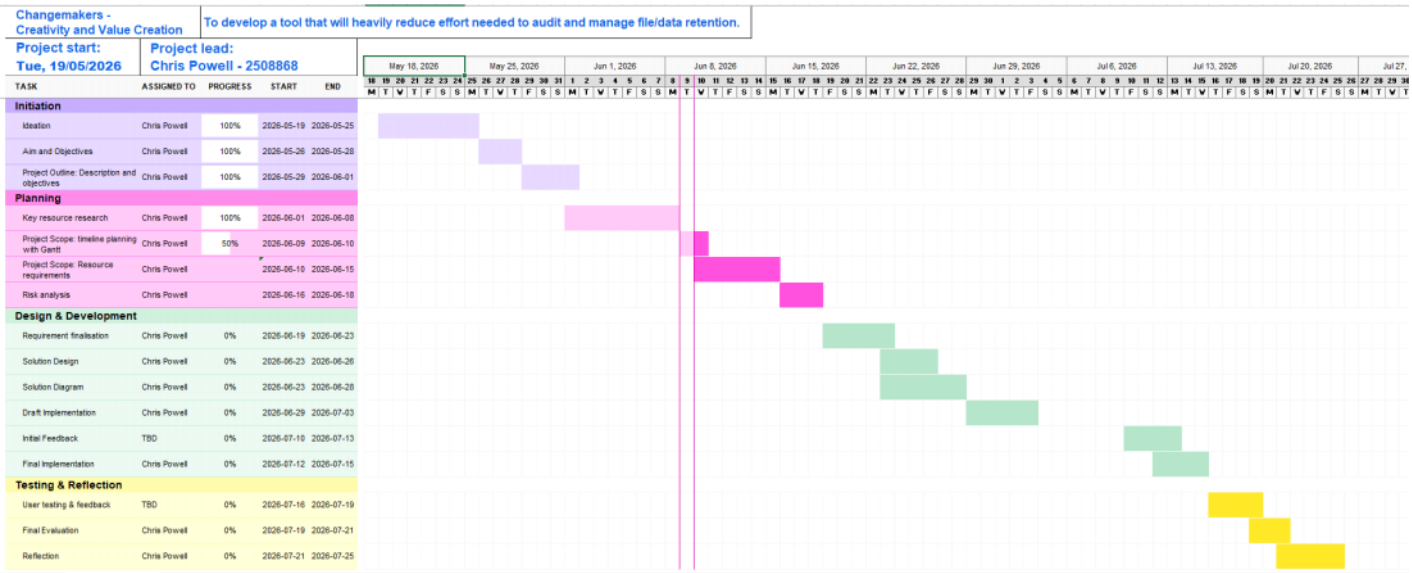
- Access to a programming language suitable for this task - Python
- Skillset enabling the developer to create robust code and guardrails to safely and efficiently meet the desired deliverable
- Varied collection of test files with different metadata, specifically age, that can be used to test functionality
- Development time sufficient to meet scope in the timeline allowed
- Testers to provide feedback on design and initial implementation

Timeframe: Design and initial implementation to be complete by 10th July 2026.

Q2: Project Gantt Chart

19 May 2026 11:59

Gantt charts are extremely useful for management of project timelines [7]. The tool of choice for my Gantt chart below was Microsoft Excel. This was chosen because of the pre-made templates available, making setup very quick and simple. It also has the additional benefit of being extremely scalable and customisable for other areas of the project.



Changemakers - Creativity and Value Creation

To develop a tool that will heavily reduce effort needed to audit and manage file/data retention.

Project start: Tue, 19/05/2026

Project lead: Chris Powell - 2508868

TASK	ASSIGNED TO	PROGRESS	START	END
Initiation				
Ideation	Chris Powell	100%	2026-05-19	2026-05-25
Aim and Objectives	Chris Powell	100%	2026-05-26	2026-05-28
Project Outline: Description and objectives	Chris Powell	100%	2026-05-29	2026-06-01
Planning				
Key resource research	Chris Powell	100%	2026-06-01	2026-06-08
Project Scope: timeline planning with Gantt	Chris Powell	50%	2026-06-09	2026-06-10
Project Scope: Resource requirements	Chris Powell		2026-06-10	2026-06-15
Risk analysis	Chris Powell		2026-06-16	2026-06-18
Design & Development				
Requirement finalisation	Chris Powell	0%	2026-06-19	2026-06-23
Solution Design	Chris Powell	0%	2026-06-23	2026-06-26
Solution Diagram	Chris Powell	0%	2026-06-23	2026-06-28
Draft Implementation	Chris Powell	0%	2026-06-29	2026-07-03
Initial Feedback	TBD	0%	2026-07-10	2026-07-13
Final Implementation	Chris Powell	0%	2026-07-12	2026-07-15
Testing & Reflection				
User testing & feedback	TBD	0%	2026-07-16	2026-07-19
Final Evaluation	Chris Powell	0%	2026-07-19	2026-07-21
Reflection	Chris Powell	0%	2026-07-21	2026-07-25

Q3: Project Risk/Challenges

19 May 2026 12:00

Anticipating, tracking, and mitigating risks to projects is often key to their success [8]. For this project, some identified risks and their contingency plans are as follows:

Risk no.	Risks	Impact	Contingencies
1.	Hardware failure & loss of data	High	All data will have appropriate backups, including a secure reliable cloud facility using OneDrive. My source code will also be appropriately version controlled and backed up using GitHub
2.	Skill set	High	Relevant resources will be used to develop the required skills to complete the project, prior-to, during and throughout the project process.
3.	Time management	Medium	Running out of time to complete the project. A Gantt chart will be used to outline key stages and tasks that need to be completed. This will be followed to ensure I complete the project on time.
4.	Scope Creep / Gold plating	Medium	A possible cause of risk #3. To avoid this risk a prioritised Requirements specification and Gantt chart will be used to help focus on the key features to be implemented. If there is time left at the end of the project, lower priority and desired / out-of-scope ideas can be added.
5.	Insufficient guardrails to prevent mis-use	High	A tool that can delete in bulk has risks to organisational data integrity, and therefore continued operation. The tool will have guardrails built in to prevent such cases, minimising the possibility of critical data loss to within risk appetite.
6.	Logical errors / code bugs	High	Robust and multi-stage testing of functionality in a controlled environment, preventing any genuine consequences of bugs until confidence is reached and the tool is ready for wider use
7.	Code performance	Low	Large directory structures could lead to poor performance and long run times, decreasing usability for some users
8.	Cross-OS compatibility	Low	Initial development scope will be Windows based. Stated clearly in documentation.
9.	Third party code library dependencies	Medium	If used, third party libraries can be built on or changed, breaking functionality. This project will aim to use only built-in libraries to mitigate risk.
10.	User usability	Low	Use of the tool will need to be tested with real users to make sure it's understandable and usable without too much instruction

Task 3 - Project Development method and Design

07 July 2026 21:55

Q1: Development Method

19 May 2026 12:00

After considering the various different development methodologies, the project seems to initially suit a ‘Software Development Life Cycle’ (SDLC) approach, however when also considering the project scale and available resources (one person), a combination approach of SDLC with Agile Feature-driven development (FDD)/Kanban board would likely be the best way to implement this project.

Given the project aims to produce software, the iterative nature of the combined approach [9] suits extremely well. The implementation of the various features of the software will be tracked via a kanban board, with tasks split by the different features.

This combination approach will help to split development into manageable segments, easily track status, and maintain a robust but flexible approach to each feature from requirement to deployment.

Q2: Design Work

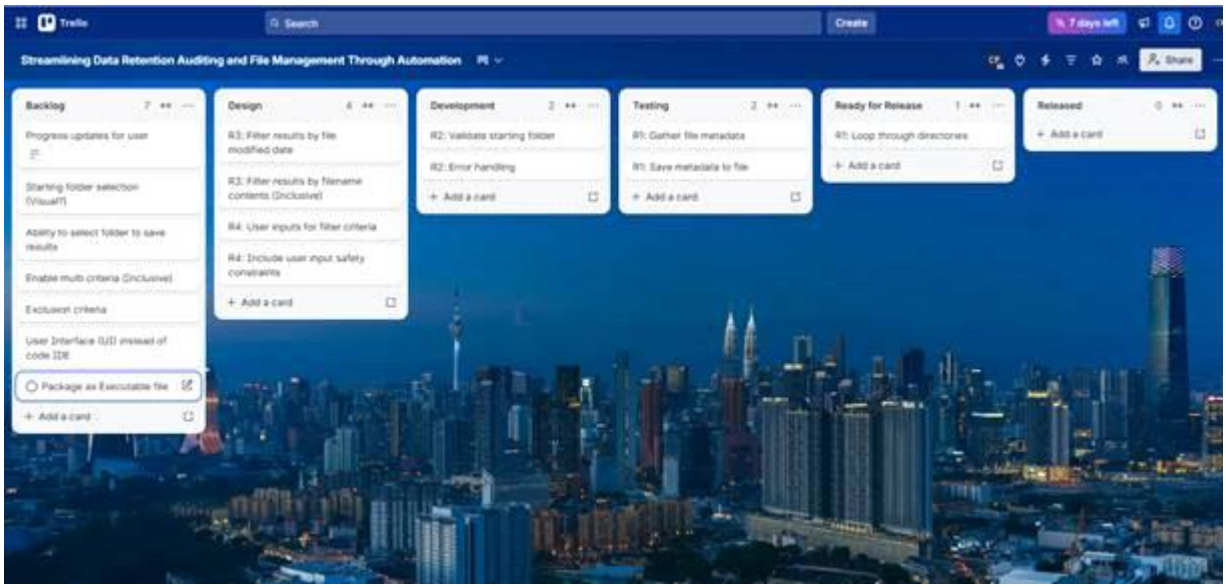
19 May 2026 12:00

As with most projects, a specification of requirements was one of the first steps to help design the work needed to achieve the desired outcome [10]. This was developed in the form of user stories that were further split into functional or non-function technical requirements. Each main requirement and it's lower-level technical requirements have been given unique IDs for easy tracking. The specification has been collated in a table format as seen below.

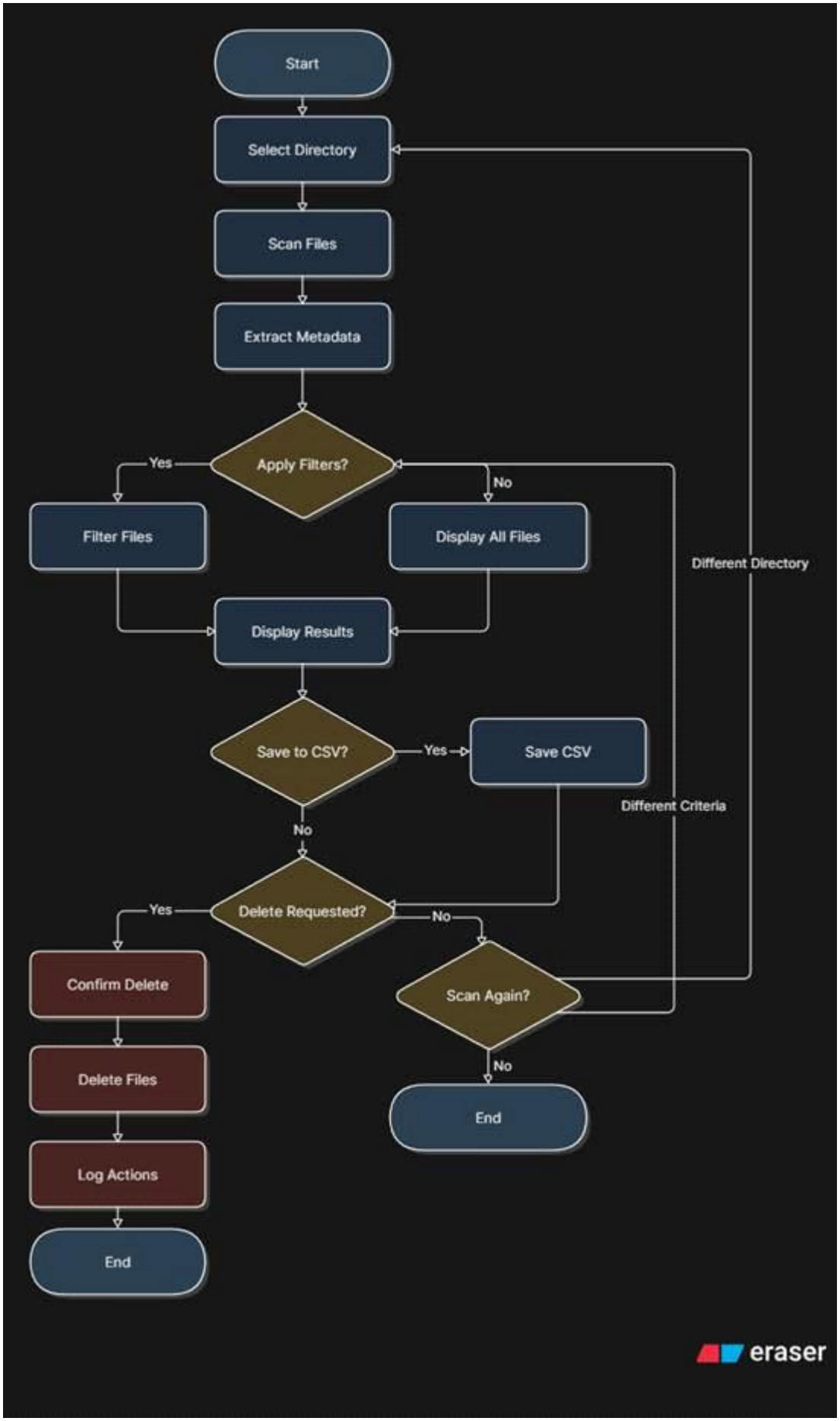
Requirement ID	User Story	Requirement Sub-ID	Type	Detail
PR-01	As a USER, I need the ability to scan all directory contents, returning information on each item	PR-01.01	Functional	Scan a selected directory and subdirectories
		PR-01.02	Functional	Gather metadata for each file: 1. Location 2. File name 3. File type 4. File size 5. Last modified date
		PR-01.03	Functional	Export collated results to a csv file for maximum compatibility
		PR-01.04	Non-Functional	Be able to operate on large directory trees with a reasonable run time
PR-02	As a USER, I need the ability to specify criteria by which the results will be filtered	PR-02.01	Functional	Enable string filtering based on filename
		PR-02.02	Functional	Enable date filtering based on last modified date
		PR-02.03	Functional	Only files matching all criteria are to be selected
PR-03	As a USER, I need the ability to delete files in bulk, maintaining to option to filter the files in scope as per PR-02	PR-03.01	Functional	Add in ability to delete files according to filter criteria
		PR-03.02	Non-Functional	Selectable time range for deletions. Limit to a last modified date <2 years in the past as they're far more likely to still be required
		PR-03.03	Non-Functional	Add in extra confirmation screen before commencing deletion operation.
PR-04	As a USER, I would like the user experience to be simple and streamlined	PR-04.01	Non-Functional	Implement input validation for incorrect inputs. Examples: 1. selecting an invalid/non-existent directory path 2. Invalid input for date filter
		PR-04.02	Non-Functional	Include exception handling during tool operation. e.g. if tool tries to read filename containing invalid characters. These should not halt the operation of the tool
PR-05	As a USER, I would like relevant information presented to me at	PR-05.01	Functional	After file deletion operation, CSV output should be saved containing a log of all

	appropriate times of the user journey			the files that were deleted, along with a date/time of the action
		PR-05.02	Non-Functional	Provide progress messages for the user to see that the tool is still running and the scope of results found so far.
		PR-05.03	Non-Functional	Error log should be collated and provided to user after activity completion should any errors occur during operation

Building on the chosen development method, to assist in the development cycle design the below feature driven kanban board was created, splitting the development work up into individual features, and then grouping those features together into ‘releases’ that will be subject to the Software Development Life Cycle.



To then further design the logical flow of the various interconnected features, a flow diagram was created as shown below:

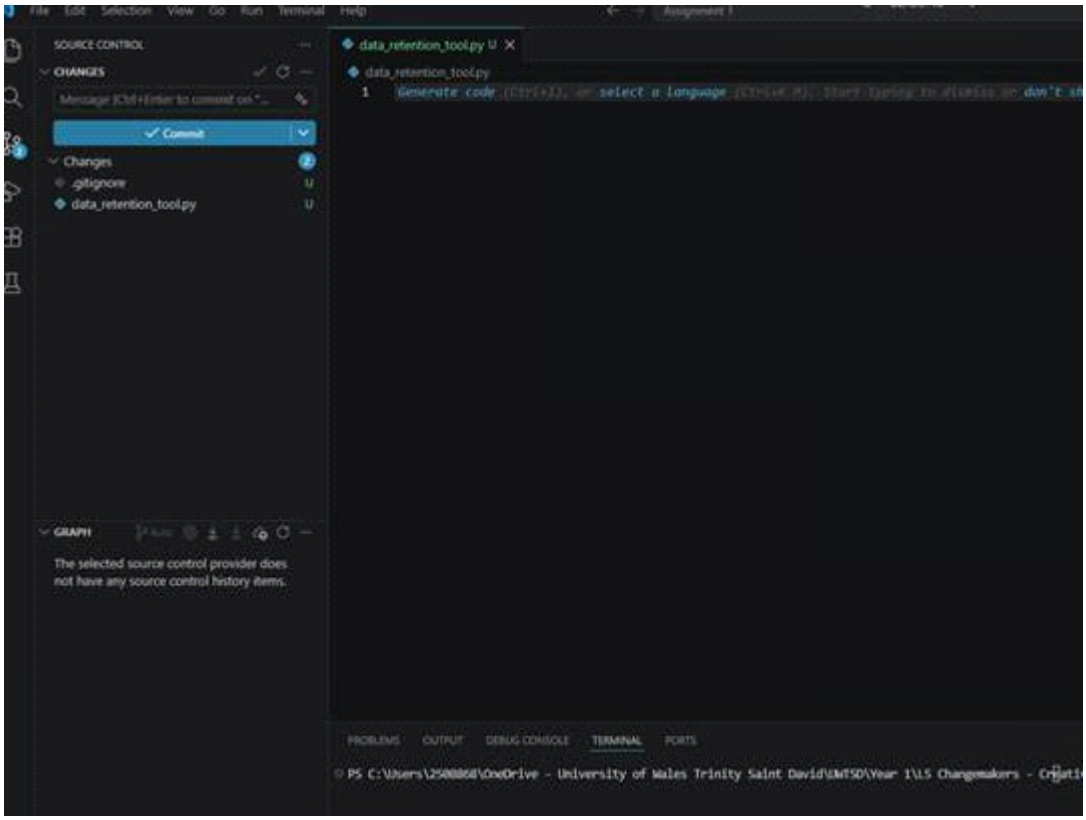

Start
Select Directory
Scan Files
Extract Metadata
Apply Filters?
Yes
No
Filter Files
Display All Files
Display Results
Save to CSV?
Yes
Save CSV
No
Delete Requested?
Yes
No
Confirm Delete
Delete Files
Log Actions
End
Scan Again?
No
End
Different Directory
Different Criteria
eraser

Q3: Initial Practical Implementation

19 May 2026 12:00

Some documented evidence of development from start to first basic prototype can be seen in the images below:

Initialising a development environment, and activating appropriate code version control



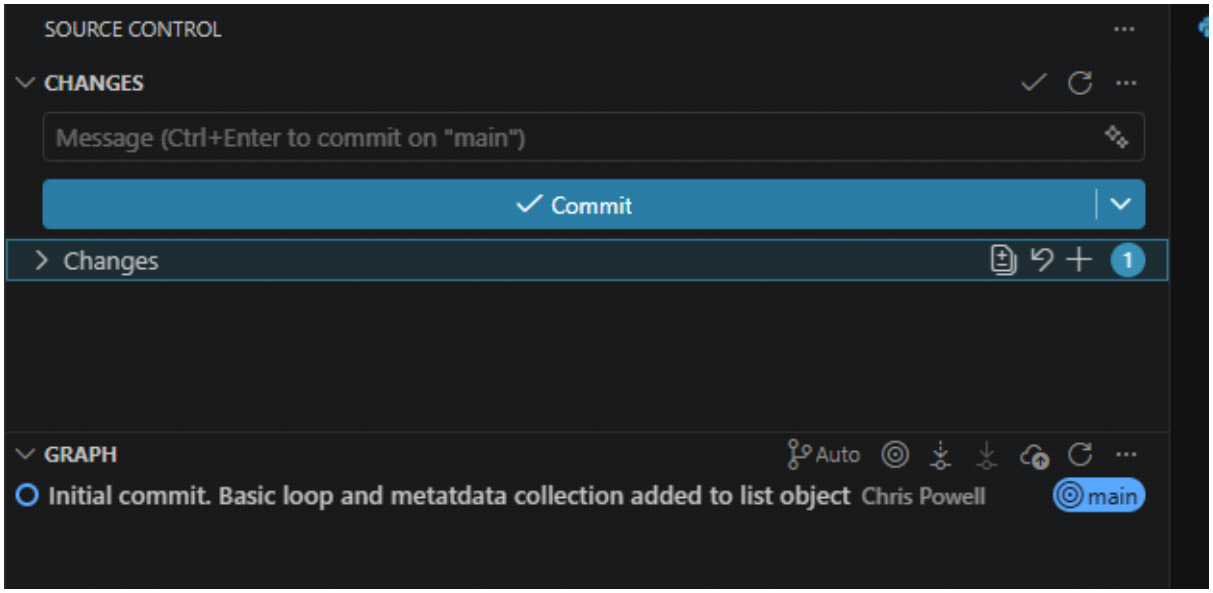
Initial code development

```python
from pathlib import Path
from datetime import datetime

def collect_metadata(starting_dir):
    """
    Loop through a directory and collects metadata for all files.

    Returns:
        list: List of objects containing file metadata.
    """
    file_list = []

    root = Path(starting_dir)

    for file_path in root.rglob("*"):
        if file_path.is_file():
            stat = file_path.stat()

            file_info = {
                "file_location": str(file_path.parent),
                "file_name": file_path.name,
                "file_type": file_path.suffix.lower() if file_path.suffix else "No Extension",
                "last_modified_date": datetime.fromtimestamp(
                    stat.st_mtime
                ).strftime("%Y-%m-%d %H:%M:%S"),
                "file_size_mb": round(stat.st_size / (1024 ** 2), 2),
```

First set of results during local testing:

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL    PORTS

(.venv) PS C:\Users\2508868\OneDrive - University of Wales Trinity Saint David\UWTSD\Year 1\L5 Changemakers - Creativit
                 'Saint David\\UWTSD\\Year 1\\L4 Intro to Web & Database '
                 'Concepts',
 'file_name': 'Scheme of Work.xlsx',
 'file_size_gb': 0.0009,
 'file_size_mb': 0.88,
 'file_type': '.xlsx',
 'last_modified_date': '2025-09-29 12:05:06'}
{'file_location': 'C:\\Users\\2508868\\OneDrive - University of Wales Trinity '
                 'Saint David\\UWTSD\\Year 1\\L4 Learning in the Digital Era',
 'file_name': 'Images UWTSD amp AC Logos-20251006.zip',
 'file_size_gb': 0.0005,
 'file_size_mb': 0.55,
 'file_type': '.zip',
 'last_modified_date': '2025-10-06 14:53:26'}
{'file_location': 'C:\\Users\\2508868\\OneDrive - University of Wales Trinity '
                 'Saint David\\UWTSD\\Year 1\\L4 Software Development',
 'file_name': 'Installing Python and PyCharm IDE Jan 2026.docx',
 'file_size_gb': 0.0011,
 'file_size_mb': 1.12,
```

Committing initial development to code version controller ‘git’


SOURCE CONTROL
CHANGES
Message (Ctrl+Enter to commit on "main")
Commit
Changes
1
GRAPH
Auto
Initial commit. Basic loop and metatdata collection added to list object
Chris Powell
main

Reference List

07 July 2026 21:46

[1] G. K. Dow and C. G. Reed, “Stagnation and innovation before agriculture,” *Journal of Economic Behavior & Organization,* vol. 77, no. 3, pp. 339-350, 2011.

[2] A. Zharovskikh, “Impact of Big data on society,” InData Labs, 13 Feb 2025. [Online]. Available: https://indatalabs.com/blog/big-data-impact-on-society. [Accessed 30 Jun 2026].

[3] D. Henriksen and R. K. Sawyer, Explaining creativity : the science of human innovation., New York: Oxford University Press, 2023.

[4] G. Strategies, “Why Your Organization Needs a Continuous Technical Training Program,” GP Strategies, 27 Dec 2023. [Online]. Available: https://www.gpstrategies.com/blog/why-your-organization-needs-a-continuous-technical-training-program/. [Accessed 05 Jul 2026].

[5] I. C. O. (ICO), “Principle (e): Storage limitation,” 30 Jun 2026. [Online]. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/storage-limitation/.

[6] L. F. Associate, “The pitfalls of data retention,” Legal Future, 05 Oct 2022. [Online]. Available: https://www.legalfutures.co.uk/associate-news/the-pitfalls-of-data-retention. [Accessed 05 Jul 2026].

[7] J. M. Wilson, “Gantt charts: A centenary appreciation,” *European journal of operational research,* vol. 149, pp. 430-437, 2003.

[8] C.-N. Bodea, A. Purnus, M. Huemann and M. Hajdu, Managing project risks for competitive advantage in changing business environments, Hershey: Business Science Reference, 2016.

[9] B. Nancholas, “The software development life cycle: an in-depth study,” Sheffield Hallam University, 20 Dec 2024. [Online]. Available: https://online.shu.ac.uk/the-software-development-life-cycle-an-in-depth-study/. [Accessed 07 Jul 2026].

[10] J. Knutson, “That first step can be the most important,” Project Management Institute, Sep 1999. [Online]. Available: https://www.pmi.org/learning/library/first-step-important-initiating-project-3597. [Accessed 07 Jul 2026].

Ideation Support Questions Lecture Activity

02 June 2026 15:09

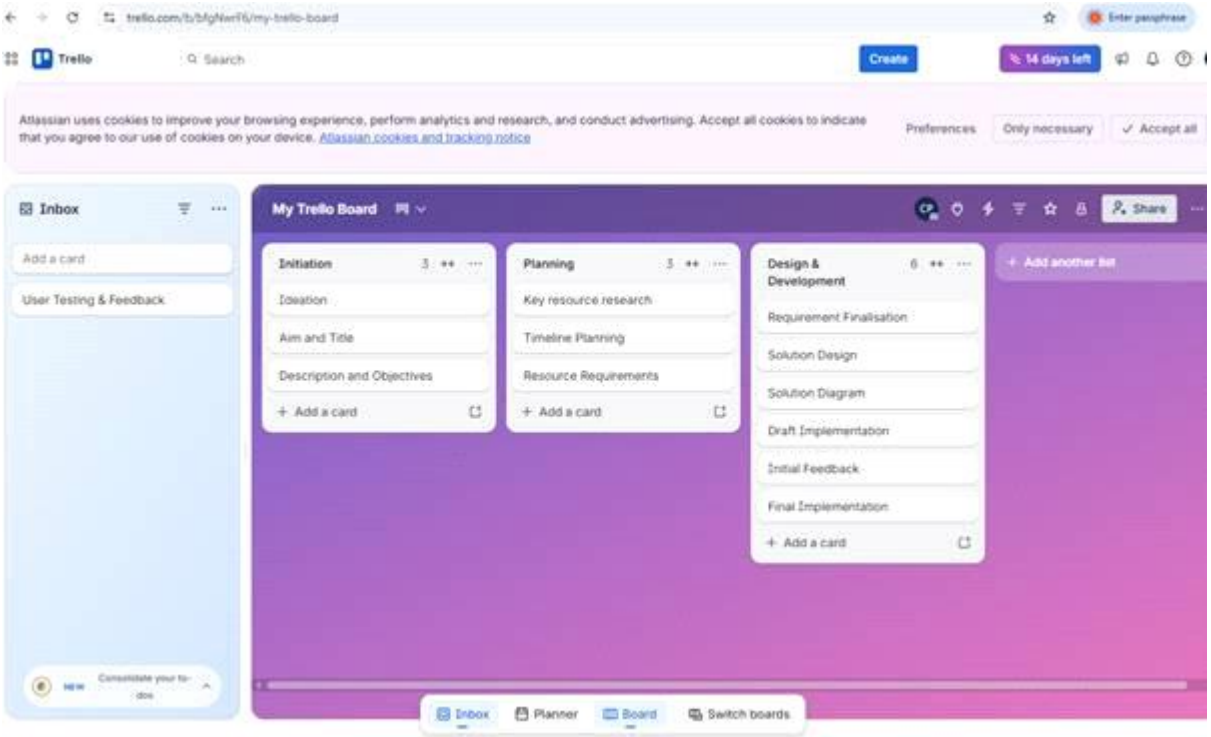
What do I want to do?

- Skill Development: Q. What new skill or technology would I like to learn or enhance?
  A. Programming to support task automation / simplification
- Positive Social Impact: Q. If I could create something to make the world a better place, what would it be?
  Automation of repeatable tasks to free up time for more value adding activity
- Employability: Q. What technology or skill would I like to develop that would make me more attractive to potential employers?
  Programming is highly sought after in many data related fields, and even in non-data specific industries
- Project Idea: Q. What would I like to do for my L5 Computing Project?

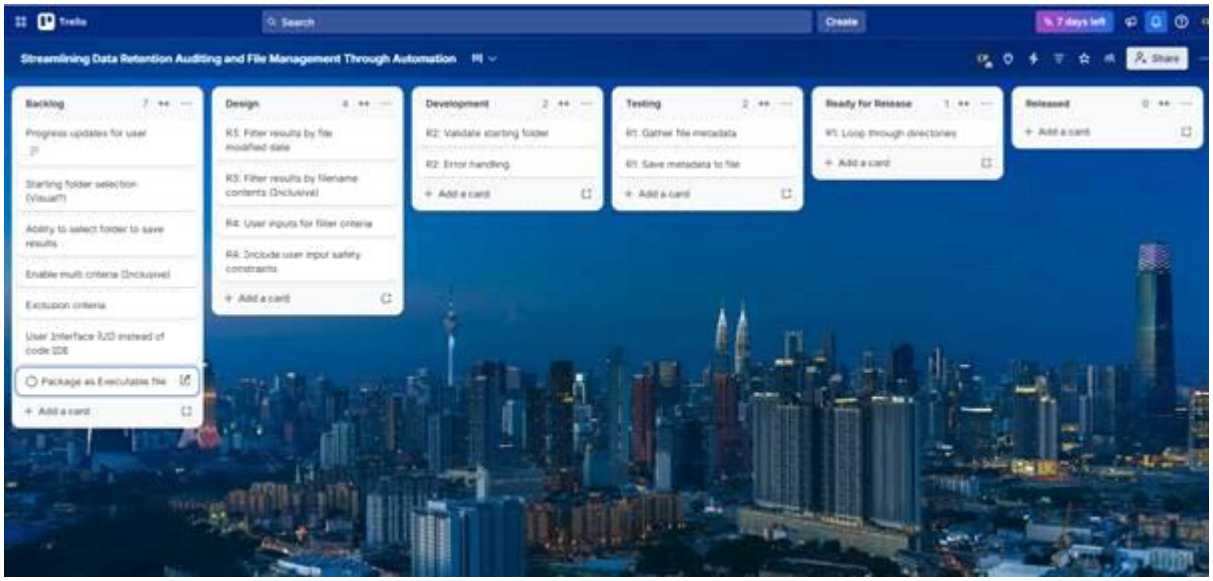
Project Management Tool Use

02 June 2026 16:39

Initial Testing



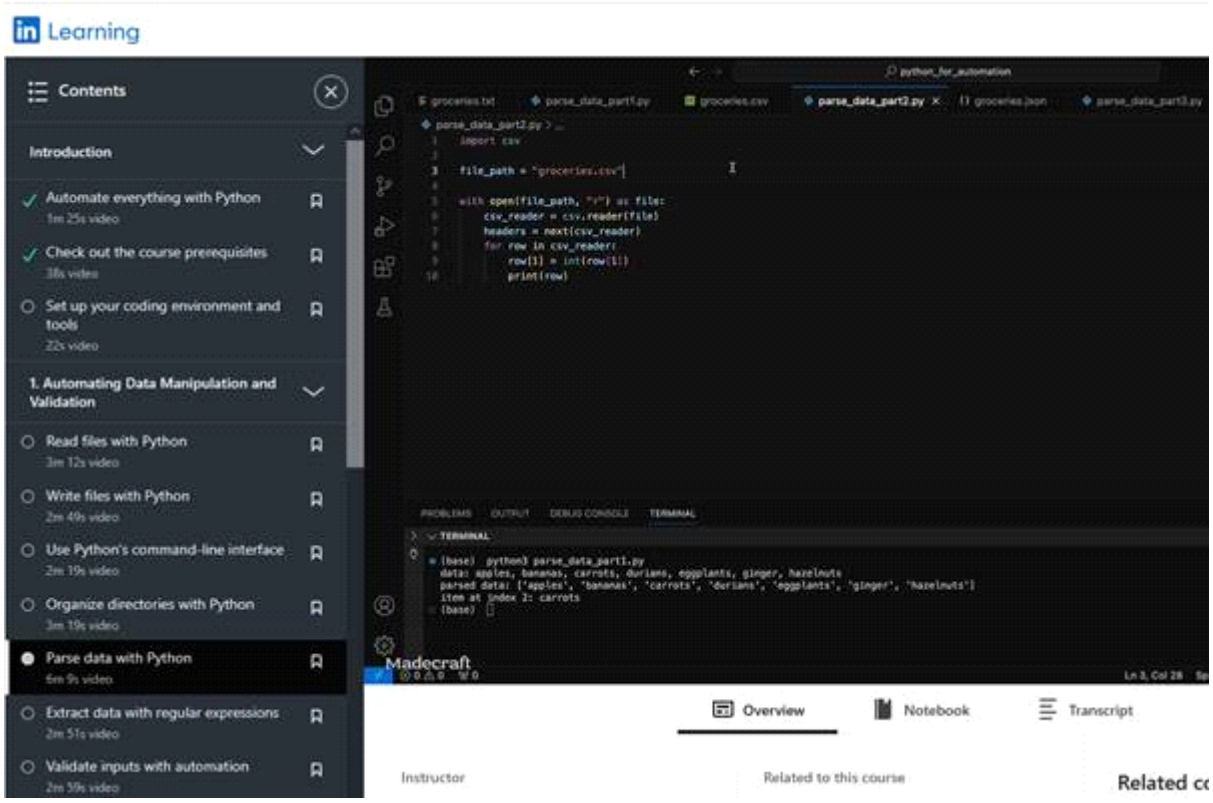
Feature Driven Kanban Board



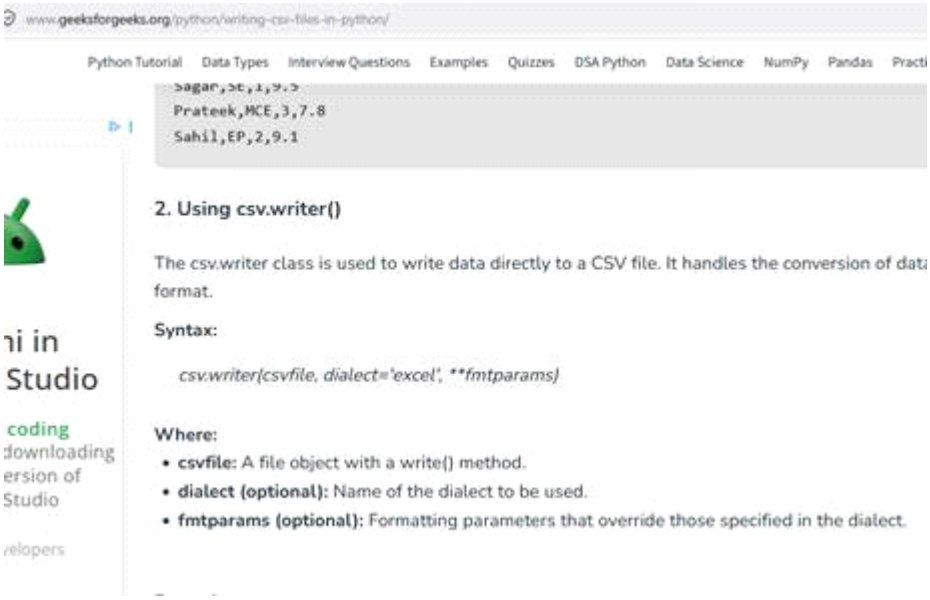
CPD Evidence

07 July 2026 21:48

Python Basics - LinkedIn Learning



Exporting data into a CSV - Independent Online Documentation



Parsing directories YouTube Video

```
import pathlib

desktop = pathlib.Path("C:\Users\alexa\OneDrive\Desktop")
  Cell In[2], line 1
    desktop = pathlib.Path("C:\Users\alexa\OneDrive\Desktop")
                                                           ^
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes in position 2-3:
XXXXXXXX escape


desktop = pathlib.Path(r"C:\Users\alexa\OneDrive\Desktop")

desktop.iterdir()
<generator object Path.iterdir at 0x000001AF4B9CBCA0>

list(desktop.iterdir())
[WindowsPath('C:/Users/alexa/OneDrive/Desktop/desktop.ini'),
 WindowsPath('C:/Users/alexa/OneDrive/Desktop/Notes'),
 WindowsPath('C:/Users/alexa/OneDrive/Desktop/realpython'),
 WindowsPath('C:/Users/alexa/OneDrive/Desktop/scripts'),
 WindowsPath('C:/Users/alexa/OneDrive/Desktop/todo.txt')]
```

Using Python to List the files in a Directory

Real Python

Subscribe