Data Transparency and Online Privacy on Dropbox Files

Project Overview and Plan

1. Introduction

This project aims to improve the understanding of the relationship between online privacy and data transparency with the use of metadata.

The primary objective of the project is to raise awareness on the user's end on how to share their online content, in this case, the files upload on Dropbox cloud that contains metadata associated that could be exploited or misused by other users or Dropbox itself.

With the constant increase of content uploaded or shared on cloud storage networks, having a direct effect on the amount of personal and business information shared online. The potential of information breaches increments exponentially since no one can guarantee 100% of data governance over the internet.

As established before users must beware of their sensitive data and metadata, these ventures focus primarily on how to manage the metadata related to the archives. The intent is to develop a web application that functions as a metadata preview filter on files uploaded to Dropbox.

This metadata preview filter will allow users to see their files related information, and this can be illustrated briefly by uploading an image to the screen showing the associated metadata like the camera specifications, the coordinates, creation date and so forth.

One of the goals of technology is to provide tools and services that are widely used by everyone. At first glance the approach is harmless but having a good understanding of how it works behind the scenes is a must on taking informed decisions on how to safely share business and personal info.

Drawing a fine line between public and private is a tall task, everything that goes online can be tracked or targeted to different groups or users. For instance, nowadays most people use clouds to store their data and acquire accessibility mimicking a virtual USB.

However, this approach has a subtlety, the owners of the global storage network storing the data have the choice to use that information to enhance their algorithms to sort out data, create predictors or to target a user's location.

Discussing the relationship between the data ethics [1] for computer scientists and new legislation GDPR [2] to establish the new regulations on how to handle consumer information to explain why is not on the best interest of those companies to reveal user info.

2. Background

The formal definition of metadata is data that describes other data [3] a few examples are: author, file size, date created date modified and so forth. Metadata is inside document files, images, videos, spreadsheets and web pages.

Metadata put things in perspective because it gives context instead of content. Hence is commonly used to answer questions like Who? When? What? How? Where?. Making it very interesting for computer scientists to generate knowledge from it. The collection, aggregation and analysis of metadata can provide very detailed information regarding individual belief's, preferences and behaviour [5].

The University of Manchester (UoM) establishes compliance with General Data Protection Regulation (GDPR) on top of their agenda, all about personally identifying information (PII) and the how is collected, stored and used [4]. They are raising awareness on how to protect the personal data entrusted to their staff as part of business processes to perform their jobs.

Moreover, UoM enhances their document lifecycle especially of documents containing any personal info. UoM has a Dropbox Business service to provide a secure, cloud-based file sharing and synchronisation tool for internal and external collaboration [5]. Providing access to network drives to members of the staff on the unit (P:) [6] mapped to each computer to share files.

However, this strategy does not mention metadata as part of enhancing document lifecycle. This omission leaves a mark on the efforts to the transparent person identification information on the lifecycle. Here is where this project takes relevance, to mend the lifecycle while uploading documents to Dropbox stripping out metadata.

Every effort to personalise services is consistently valuable to the consumers. However, it may come at the cost of consumer privacy [7]. Fostering new consumer-driven personalised services is significant nowadays since it adds value but the price is to find consumers who are willing to provide information.

Moreover, the privacy paradox comes into play, illustrated as where people seem to be concerned about privacy, but they do exactly the opposite [8]. Hypocrisy is not new conduct for human beings, actually is part of a denial of a defect or uncomfortable situation. Can be exemplified, when someone expresses discomfort with his or her social lives but dismiss going out with friends, to play video games.

The lack of privacy sense commonly associates with indifference or ignorance. People value their online privacy differently exchanging info for a short-term benefit or entirely cluelessly about it. Privacy is a human right that is getting harder to protect over the internet making it more illusory than real.

Nevertheless, that can change establishing a robust data ethics framework to address the ethical impact and implications of data science and its applications [1]. Another extended look where metadata has the upper hand shifting the paradigm from being information-centric to being data-centric this shift brings into focus the different moral dimensions of all kinds of data, even the data that never translate directly into information but can be used to generate behaviours [1].

As discussed above, fine-grained metadata can be a game changer for applications that are popular to become trendy. Granting them a technical advantage over others such as Netflix recommendations, The wow effect induced by this strong prediction/recommendation of a rarity, triggered a spur-of-the-moment discussion which may serve to illuminate different aspects of metadata analytics [9].

However, indiscriminate mass metadata surveillance of entire populations by governments generates abundant amounts of personal data and consequently represents a substantial threat to the privacy, equality and liberty of individuals [10]. Initially, this project was going to focus on Facebook data and metadata management until they changed their user agreements thus their manage metadata.

Is well known that politicians hire private companies to perform dirty jobs against the citizens they represent to take advantage of them. The most recent security data breach happened to Facebook stock loses $25 BN by the Cambridge Analytica case [11].

To put in perspective how this internet juggernaut lost that much money by allowing third-party enterprises to access users data/metadata and misuse that information to turn the political behaviour around and give votes to a specific politician in the 2016 US Presidential Election [12].

3.   Research Methodology

The research methodology for this project is to develop a web application that will act as a metadata remover or filter on files and upload them to Dropbox.

The technologies involved to achieve this project are Java and JavaScript. Using an MVC (Model – Controller – View) framework based on JSP's and Servlets to create the structure of the application. Also communicating through the Dropbox API libraries technologies, as part of the back end strategy utilising Oracle Glassfish to host the tool.

As part of the technical details, Java is going to process the files gathered by the screen and clone them, so the original data stored on the computer are not modified at all. Accordingly, the newly created dossiers are going to match with the preferences template selected by the user and insert metadata upon that.

The way in which the application is going to interact with the user is as follows:

I. The web application is going to open on the home page, which has the functionality to upload one or multiple files from different extensions.

II. Once the server has the files, it will show a menu that allows filtering the metadata attached to each of them.

III. The selected filter is going into settings as a template saved on the temporary files as an XML, so the next time the user opens the web application, it has them as default values for the mime type.

IV. Then the upload button will enable, and on click, it will ask for the Dropbox credentials.

V. The authenticated user will be able to select the files in the carousel to confirm the submission into the repository.

Test-driven development is the implementation technique that fits better the project objectives and goals. An agile mindset is suitable providing the flexibility, adjustability and speed needed to execute this agile development considering that there is just one developer.

4.   Ethics and professional considerations

The approach of this project takes into account a framework of data ethics and professionalism that has to find a balance between professional secrecy and public interest [8]. Moreover, the project is based on codes and rules to keep personal info safe while taking the most advantages of technology.

Like on most of the social-oriented applications, insight is to keep a safe way of coding that permits privacy and transparency on the use of user's data. Moreover, the design revolves around the idea of collecting, processing and storing information that holds to high accountability standards.

On the technical aspect of the application, users are going to type in their Dropbox credentials after selecting their filters for their files. Details are going to travel around the network encapsulated, wrapped and encrypted on an object provided by Dropbox API which will return a connection-token to manage and browse files on the cloud for the given account.

However, the application is going to be client side oriented to provide transparency to the filter process as follows:

a. Files are going to be uploaded and cloned from the originals. Therefore, the original archives are intact.

b. The program would not copy the original or cloned files on any local server, so they are just accessible at execution time.

c. The process is going to match data and templates and create an XML file stored on the temporary internet files to use it as a cache for the subsequent executions and would work as defaults.

d. The user is going to provide their Dropbox credentials to navigate through their cloud storage network and save the selected metadata filtered files.

5.   Risk consideration

Let's classify the risks can fall into three categories technical, usability and tuning and explained as follows:

Technical Risks have more to do with the code cohesion and interaction between classes. How the design takes knowledgeable consideration using libraries and frameworks that are compatibles and most adequate to perform the task. Nonetheless, experience tells us that every development has their shortcomings and challenges, rely upon a third-party library to connect to Dropbox could lead into a knowledge curve while having his limitations or bugs.

So one of the most critical decisions is picking the right third-party library to avoid any unnecessary risks. Since one of the primary focus should be to obtain reliability, portability, flexibility, standardisation and support from the community that had already implemented something alike.

The best option would be the one that they advertise; it is called Java SDK for API v2 helps to integrate Dropbox [13] with Java applications. As expected is highly well regarded on the developer's reviews and

have a GitHub repository [14] full of examples displaying different functionalities. For instance basic operations like write a file, read a folder or share a link to a file to a group of users.

In the programming scene, everything has to have a foundation based on well taken, and executed decisions that fit the narrative, attached to web development industry standards to avoid the intrinsic complexity. Initial observations suggest that MVC is the most commonly used framework by the industry [15], using Java technologies based on Servlets is the selected framework on the back end and JavaScript with JSP's on the front end hosted by Glassfish web server.

Another technical problem could be that Dropbox itself implements the functionality of striping out the metadata on all the files based on the new legislation and his business model.

Usability Risks, this is one of the most common causes of failure for new applications in the industry. Moreover, one of the primary reasons that every successful app is continuously updating or developing new functionality suggested by the clients or the way they adapted it to their needs.

There is a phenomenon that commonly happens on new applications that sometimes developers manage to build functionality to fulfil a requirement. However, users misuse that functionality and devise another utility for it that proves to be more useful than the original thought of it. Reshaping the primary purpose and changing the innovative approach to this that proves more useful.

To mitigate these risks, the strategy to deploy must provide the best user experience regarding usability and flexibility furthermore consider a handy user interface design that maximises those qualities to offer assurance.

That is why having implemented a simple user interface that provides multiple options to the same functionality. A useful example in our case is the requirement to upload files, have an implementation that handles it in two different ways by drag and drops and a file browser to select the data.

Tuning Risks, how many times a user has a complaint about a function taking ages to deliver a result or to complete a basic operation. Nowadays user patience is a limited resource, and that could be the breaking point of using an app or discarding it.

 Another scenario to account for is when the application has bugs, deliver inaccurate information or the lack of adequately considered situations. While most of these inconveniences come from a poorly design phase most of the times but sometimes those are inevitable from a lack of knowledge or background with the problem at hand.

The most commonly accepted and used solution is to invest heavily in testing the app from the beginning to the end and at every development stage unit testing, functionality testing and system testing. To increase the internal software qualities and provide solid building blocks.

The best technique to alleviate these risks is test-driven development (TDD), generating the synergy required on the components and structure. Also creating a set of system tests that challenge the performance, accurateness and correctness of the system. Another goal of this approach is to maximise test coverage to expose bugs in earlier phases of coding building long-term sustainability and saving developer time.

6. Project evaluation

The project is going to use an agile approach to management and evaluation that focuses on speed and flexibility, creating sprints per week taking advantage of the high interaction between user and developer (supervisor and supervised). Allowing for quick adjustments, identify issues promptly and reducing complexity.

We can divide the constant evaluation into two main phases

- Scrum methodology, benefitting from the scheduled meeting every week to present work done, during sprint planning to create the new detailed tasks for next week. While getting the energised with the three fundamental questions:

  - ✓ What did you do yesterday?

  - ✓ What do you plan on doing today?

  - ✓ Are there any blocks or impediments that keep you from doing your work?

  Creating the backlog of the specified tasks and inserting them into the activity report [16].

- User's evaluation, at the last part of process development and to increase view and perspective on the project the users are going to be massively involved. This app is not for a specific group of users, on the contrary, to broad and diverse groups with a different preference, usage and knowledge of technology. So this part of the process is going evolve as follows;

  - o Firstly, users are going to fill a google form answering simple questions about their use of technology. How do they get involved in social networks, cloud storage networks, share content online, if they own a Dropbox account, the privacy of their data and if they are aware of the difference between data and metadata?

  - o Secondly, If they own a Dropbox account then invite them to test the application, showing the functionality and what the concepts meant in practice.

  - o Thirdly, surveys with questions to receive feedback on the functionality of the tool, the interface and utility in their experience of sharing files on the cloud storage network.

7.   Planning

As the course of action of this project is to identify, evaluate and test the technical milestones and list them as follows:
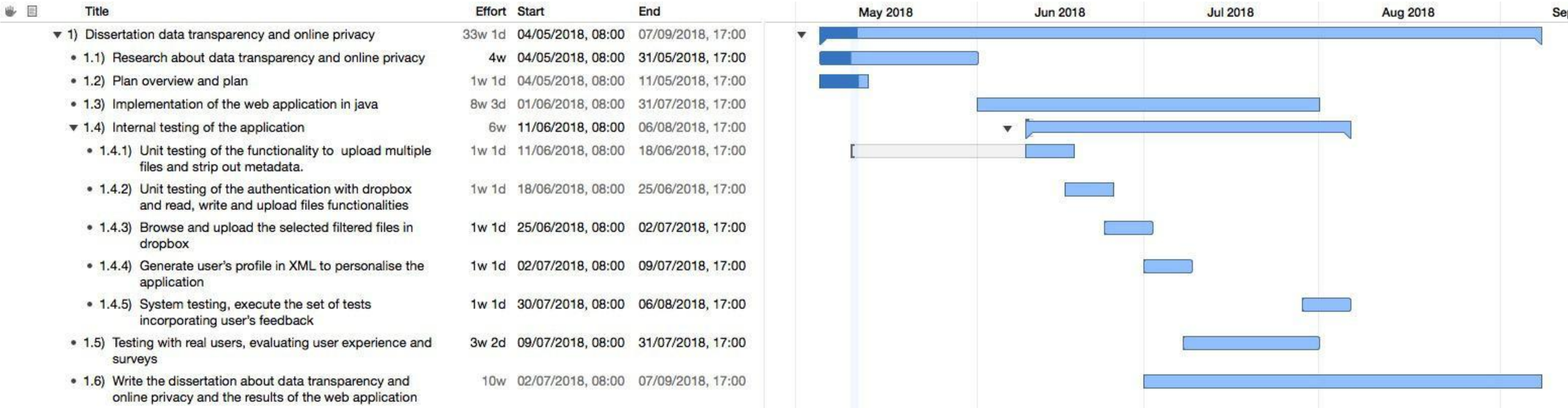
- ✓ Identifying the pillars of the plan and lined them up with objectives and aims. Set deadlines for each taking into consideration that there must be a deliverable that must be measurable in amounts of work, these can be listed as follows:

    I.   Upload multiple files in a webpage and able to strip out the metadata of each of them using the metadata filter template assigned for each of them (11.06.2018-18.06.2018).

    II.   Authenticate the user's account to Dropbox using the Java SDK for API v2 (18.06.2018-25.06.2018).

    III.   Browse and upload the selected filtered files to the user's account (25.06.2018-02.07.2018).

    IV.   Generate user's profile in XML to personalise the application saving the templates in temporary internet files (02.07.2018-02.09.2018).

    V.   System testing set, after the last release that includes changes led by users feedback (30.07.2018-06.08.2018).

- ✓ Evaluation of the milestones and review of the status, to continually seek the feedback on how the project is progressing by asking the following questions:

    - What is going right on the project?

    - What is going wrong?

    - What problems are emerging?

    - Where is the project concerning the schedule and milestones?

    The answers to these questions indicate the progress on the project. If developing phases are on time and form accordingly to the plan or a change of direction is needed and urge of applying corrective actions to get the project back on track.

- ✓ Testing as a development technique, the project is going to be developed using TDD, so the core functionality is frequently tested regardless, although there must be unit test sets and metric test sets per each milestone.

Moreover, a set of tests that need to comply with software metrics like performance and accurateness tests. At the end, of the development, phase there must be a set of system tests.

Data transparency and online privacy: Gantt Chart

Title	Effort	Start	End
▼ 1) Dissertation data transparency and online privacy	33w 1d	04/05/2018, 08:00	07/09/2018, 17:00
• 1.1) Research about data transparency and online privacy	4w	04/05/2018, 08:00	31/05/2018, 17:00
• 1.2) Plan overview and plan	1w 1d	04/05/2018, 08:00	11/05/2018, 17:00
• 1.3) Implementation of the web application in java	8w 3d	01/06/2018, 08:00	31/07/2018, 17:00
▼ 1.4) Internal testing of the application	6w	11/06/2018, 08:00	06/08/2018, 17:00
• 1.4.1) Unit testing of the functionality to upload multiple files and strip out metadata.	1w 1d	11/06/2018, 08:00	18/06/2018, 17:00
• 1.4.2) Unit testing of the authentication with dropbox and read, write and upload files functionalities	1w 1d	18/06/2018, 08:00	25/06/2018, 17:00
• 1.4.3) Browse and upload the selected filtered files in dropbox	1w 1d	25/06/2018, 08:00	02/07/2018, 17:00
• 1.4.4) Generate user's profile in XML to personalise the application	1w 1d	02/07/2018, 08:00	09/07/2018, 17:00
• 1.4.5) System testing, execute the set of tests incorporating user's feedback	1w 1d	30/07/2018, 08:00	06/08/2018, 17:00
• 1.5) Testing with real users, evaluating user experience and surveys	3w 2d	09/07/2018, 08:00	31/07/2018, 17:00
• 1.6) Write the dissertation about data transparency and online privacy and the results of the web application	10w	02/07/2018, 08:00	07/09/2018, 17:00

References

[1] L. Floridi, "What is data Ethics," University of Oxford, Oxford, 2016.

[2] S. Saltis, "Coredna," 3 April 2018. [Online]. Available: https://www.coredna.com/blogs/general-data-protection-regulation. [Accessed 11 May 2018].

[3] M. Rouse, "whatis?.com," [Online]. Available: https://whatis.techtarget.com/definition/metadata. [Accessed 11 May 2018].

[4] Manchester, University of, "GDPR 5 minute guide," Manchester, 2018.

[5] Manchester, University of, "IT Services," [Online]. Available: http://www.itservices.manchester.ac.uk/ourservices/catalogue/commscollab/sec/.

[6] Manchester, University of, "Staff Services," 28 May 2018. [Online]. Available: http://www.staffnet.manchester.ac.uk/news/university-news/display/?id=19803.

[7] N. F. A. &. M. S. Krishnan, "The personalization privacy paradox," University of Michigan Business School, Michigan, 2006.

[8] D. J. K. Andrea Bellinger, Network Public Governance, Transcript Verlag, 2018.

[9] G. Anadiotis, "ZDNET," 20 January 2017. [Online]. Available: https://www.zdnet.com/article/mix-and-match-analytics-data-metadata-and-machine-learning-for-the-win-2/. [Accessed 11 May 2018].

[10] N. N. loideain, "EU Law and Mass internet metadata surveillance in the post-Snowden era," Faculty of Law, University of Cambridge, Cambridge, 2015.

[11] A. Griffin, "The independent," Independent, 19 March 2018. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-cambridge-analytica-personal-data-breach-stock-shares-slump-latest-updates-a8263386.html. [Accessed 11 May 2018].

[12] M. Hindman, "The independent," 13 4 2018. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/how-cambridge-analytica-s-facebook-targeting-model-really-worked-according-to-the-person-who-built-a8289901.html. [Accessed 1 5 2018].

[13] Dropbox, "Dropbox for developers," Dropbox, [Online]. Available: https://www.dropbox.com/developers/documentation/java#overview. [Accessed 11 May 2018].

[14] GitHub, "Dropbox SDK," Dropbox on GitHub, [Online]. Available: https://github.com/dropbox/dropbox-sdk-java. [Accessed 10 May 2018].

[15] V. Sidorenko, "Gearheart Web development," Gearheart, [Online]. Available: https://gearheart.io/blog/7-best-frameworks-for-web-development-in-2017/. [Accessed 11 May 2018].

[16] University of Manchester, "Projects 2017-2018," University of Manchester, [Online]. Available: https://studentnet.cs.manchester.ac.uk/pgt/2017/COMP66090/project/assessment/projectprogress.php. [Accessed 2018 11 5].