AI for the Media Assignment 2 Datasheet

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created as input for an image generation model. It has a creative focus on how these types of models break down images using a dataset that interests me. I think there are potential uses that the outcomes of this project could be used, in creative settings.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

This dataset was created by scraping several blogs, searches from Pinterest and the internet archive, as part of my AI for the Media Module assignments and mini project.

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

This dataset is made up of only images (PNGs) of illuminated manuscripts, a type of medieval western art that was the main method of transcribing text, in a western context, onto paper prior to the printing press.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

This data is self-contained.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

This dataset doesn't contain any currently confidential information

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

There is the potential that this dataset contains imagery that might be offensive to some. This dataset is a reflection of historical ideas and is bound to contain imagery related to sexism, due to the historical period the work is from there is less likely to contain other forms of offensive imagery like racism and homophobia, though it cannot be ruled out. The dataset does also contain sexual content as well as nudity, that while could be considered disturbing to some, the style of the manuscripts leads to the visuals appearing more comedic than anxiety-inducing.

- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

This dataset doesn't relate to any living people. It should be mentioned that because of the inherent religiosity of the texts I struggled with the ethics of using what could be considered religious material. Whilst I understand that these are religious, some of the imagery included is so far away from what modern Christianity espouses I do believe there is room for using this dataset in a way that is not offensive.

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

This data is directly observable.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?

This data was collected using webscrapers including pin down and tumblr-crawler, (https://github.com/dixudx/tumblr-crawler) as well as internet-archive-downloader (https://github.com/terrybroad/internet-archive-downloader).

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

This dataset was collected by one Masters's student (that student being me). The data that was scrapped from the Penn Libraries Manuscripts Tumblr blog which makes up the vast majority of my dataset and was curated by Dot Porter.

- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

This dataset was collected over a day, though is not time sensitive as the work is historical and archived.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms

It is best practice to relay some of the potentially sensitive content that might be included in this dataset, as it relates to nudity, sex, violence and the threat of damnation. It would be responsible at the very least to caution any potential user of the dataset of potentially sensitive outcomes from an image generator. As previously mentioned, the dataset does contain this sort of potentially sensitive imagery but due the imageries aburdity if a model was able to create images with any detail, they would reflect this stylistic absurdity.

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

I will not be distributing this dataset as I have been unable to find clear information on the use of some of these images. Whilst the manuscripts themselves belong to their respective museum or holder the photos are not perfect representations of the entire work. I am unsure of the ramifications of this dataset outside of a research or personal capacity.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

As stated above there could be potential licensing issues that I am unsure on in regards to museum copyright ownership.

- **Who will be supporting/hosting/maintaining the dataset?**

The dataset doesn't need to be maintained as it is mainly for a single objective rather than a long running project, and as the content in it is historical it doesn't need to be kept up to date in the same way other datasets might require. If the dataset were to be released then there could be conversation for how to make the dataset more representative of other historical examples as I was limited in the scope of what I could collect.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

The curator of the dataset can be contacted through her email address (s.nathan062022@arts.ac.uk)

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

So mentioned above, there is little reason to update this dataset, though there are ways it could become more robust by widening the dataset in regards to image generation