1. Introduction

Exhibitions, physical and digital, make art relevant and accessible to artists, scholars, and the public in a managed and directed environment. Exhibitions are a catalyst point that creates a much wider set of content that can highlight the wider contextual impact of the exhibition and the objects included. This content can be used to connect and enhance data about collections across institutions. This allows academics, artists, and the public to connect objects in museum collections around the world, enabling new insights and ways to engage with shared histories. It allows us to investigate the potential of integrating historic and contemporary exhibition data in differing formats and from multiple sources. We have connected social media, exhibition-related labels, catalogue essays, critic reviews, and press coverage to structured data (Linked Open Data) using applied computational techniques to align socially based textual and structured data, making reconciliation easier and more effective. In this work we outline the process for one of the more challenging aspects of this process: enhancement and alignment of records using social media data.

2. Structured Data

Structured data standards, and specifically Linked Data, can be used to enable consistent access to information across institutions, and in turn as a foundation for Digital Humanities scholarship including the work reported here [1]. It enables the combination of data, and the sharing and reuse of downstream applications that use this data. The Linked Art collaboration [2] has developed a Linked Open Usable Data (LOUD) [3] set of specifications to describe and publish art-related knowledge, including which objects were shown during which exhibitions. This is a data representation model and set of concepts and categories that describes art and cultural heritage objects following standardised design patterns. This allows for a common structure of data describing art objects, artists, exhibitions, collections, events, and places in a way that can be useful to scholars working with data across collections, while retaining the ability of institution-specific cataloguing practices to be reflected accurately.

The Yale Center for British Art lent the Ashmolean Museum nineteen artworks for an exhibition in 2012. This included the painting “British Gentlemen in Rome”, by Katharine Read. That object was lent to the Yale University Art Gallery in 1965 and in 2015 to the University of Arizona Museum of Art. Newspaper articles, such as in The Evening Standard and The Oxford Mail describe the exhibition and works present.

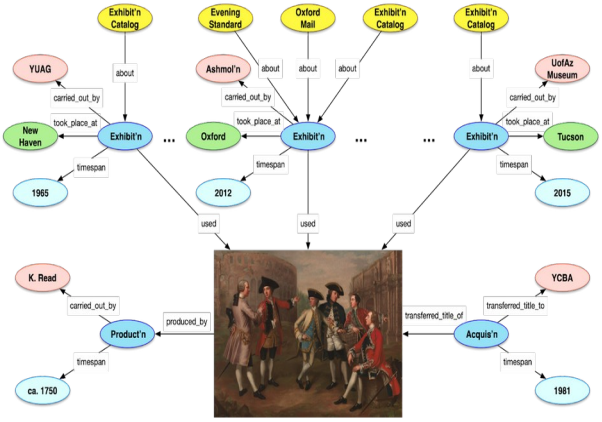


Fig. 1 The production, acquisition, and exhibitions of British Gentlemen in Rome, by K. Read

By mining these social texts (in yellow, fig. 1), and connecting with the structured data from across institutions, we can bring together both quantitative and qualitative information about objects presented and audience responses.

Exhibitions discussed in contemporary social texts provide the context of these exhibitions and their works as described at the time they were held. In recent times the volume of available public reaction to live exhibitions has drastically increased through social media responses. In the work described here we will study information posted on social media using a variety of hashtags promoted by the Ashmolean Museum and consider whether the provision of specific hashtags provokes a reaction to specific objects.

3. Labyrinth, Knossos Myth & Reality

The Ashmolean Museum held the exhibition *Labyrinth, Knossos Myth & Reality*, from February to July 2023. The first international exhibition to focus on the archaeological site of Knossos on Crete, it included over 100 loan objects from Greece and archival material from the Ashmolean telling the story of the exploration and excavation of the site associated with the mythical Labyrinth. In total there were 235 exhibits including two digital exhibits: an experience of the video game Assassin's Creed and a film by artist Elizabeth Price. This was a well-attended exhibition, with 62,000 visitors, and a high level of social media engagement.

The Ashmolean actively promoted the exhibition on various channels and visitors posted about their own experience. This research aimed to stimulate the production of content rather than investigating social media communication. Exhibitions can be experimental places; their short time-frame provides the ability to try to affect visitor behaviour in a directed space in a defined time, by prompting visitors to participate, take photos and post on social media. In the exhibition, we invited visitors to use the hashtags #LabyrinthAshmolean, #LabyrinthLabel and #LostInTheLabyrinth. A panel introduced the project and encouraged the use of #LabyrinthLabel to write their own label for a museum object. #LabyrinthAshmolean, the overall exhibition hashtag included in Ashmolean social posts, was also on a panel at the entrance to the exhibition. #LostInTheLabyrinth was hidden around the exhibition on six exhibition labels with no further explanation.

4. Methodology

4.1 Data Gathering

Gathering useful social media content is currently a challenging task [4]. Content is available through Application Programming Interfaces (APIs) directly from social media sites or through aggregators. Access to data has become limited for example, Facebook withdrew free access in 2018 and Twitter in 2023, vastly limiting this kind of research. There are options to pay for data through the aggregators from the technology companies directly although this limits the research to those that can pay for data, luckily for this project Yale University has institutional access to Brandwatch (https://www.brandwatch.com) which we

used to access data, a paid-for data aggregator service that gives access to social networks, including Twitter, Facebook, and Instagram, blogs, forums, reviews, and news sites. We gathered data using the historical search which provides a sample of the data, and the search terms were the hashtags described above. Brandwatch returned data from 1st January until 30th September 2023. We know the set provided is an incomplete data set; for example, it does not contain any Instagram data although we are aware that there were posts. The data is aggregated by Brandwatch who observe the terms of service for each site, these vary by site, and often only give access to public accounts, which accounts for the lack of Instagram data. This work offers a limited view of social media as it uses a very limited sub-sample of the data produced in reaction to the exhibition, although this limited data does allow use to test our approaches to align records.

We gathered 2,323 pieces of content. This was mainly from the social media site Twitter (2,065) but also included data from news sites (150), blogs (55), Tumblr (25), forums (16), Facebook (5), review sites (5), and Reddit (4). As most of the data was from Twitter it contained many retweets, the number of unique pieces of content is 577. Interestingly, although the content returned was generally on topic, not all the social media contained the keywords searched for.

4.2 Data Annotation

The texts of the pieces of content were compiled in a spreadsheet and annotated to identify if objects from the museum were discussed. Annotator 1 was the exhibition curator, who had a detailed knowledge of each object in the exhibition. When an exhibition object was present a code was recorded, according to the internal Ashmolean identifier, using the annotation codes in table 1. He first looked at the text to see if particular objects were unambiguously described (e.g. the sculpture of a Minotaur, Assassin's Creed) and then checked the image on the source website to confirm this identification [t]. If he could not identify objects from the texts because the description could refer to several different objects (e.g bulls, Minoan pottery), he looked at the image on the source website to identify them [ti]. If there was no clear reference to objects in the text he looked at the source website to see if there were images of objects [i] or no reference to objects [n]. The main ambiguity occurred where users posted images of the exhibition in general, where the subject was the exhibition rather than particular objects. Surprisingly there were almost no selfies in the sample, which would also have posed problems of identification because the subject is primarily the selfie-taker. The proportion of each category is given in figure 2. The nature of the task (to identify objects) and the use of supporting images where necessary meant that there was little ambiguity in object identification but the task was labour intensive and required specialist knowledge. A second annotator familiar with the exhibition, coded a 10% sample, to evaluate the difficulty of the task. The inter-annotator agreement score for this data is 0.40, which is fair agreement, indicating that this is a difficult task to automate. The main source of disagreement was not the identified objects but the method of identification: Annotator 1 identified twice as many objects from the text as Annotator 2. There was also some disagreement about the subject of the text [n/x] and whether the subject of the image was the exhibition or the particular objects [n/i]. These are subjective judgments unrelated to the identification of particular objects.

t	I could identify the object from the text
ti	The text referred to objects, but it was only possible to identify them from the images
i	The image showed one or more objects, but the text didn't mention them
n	There was no reference to specific objects in the text or image
x	The post was not about the exhibition/unreadable

Table 1 Annotation guidelines

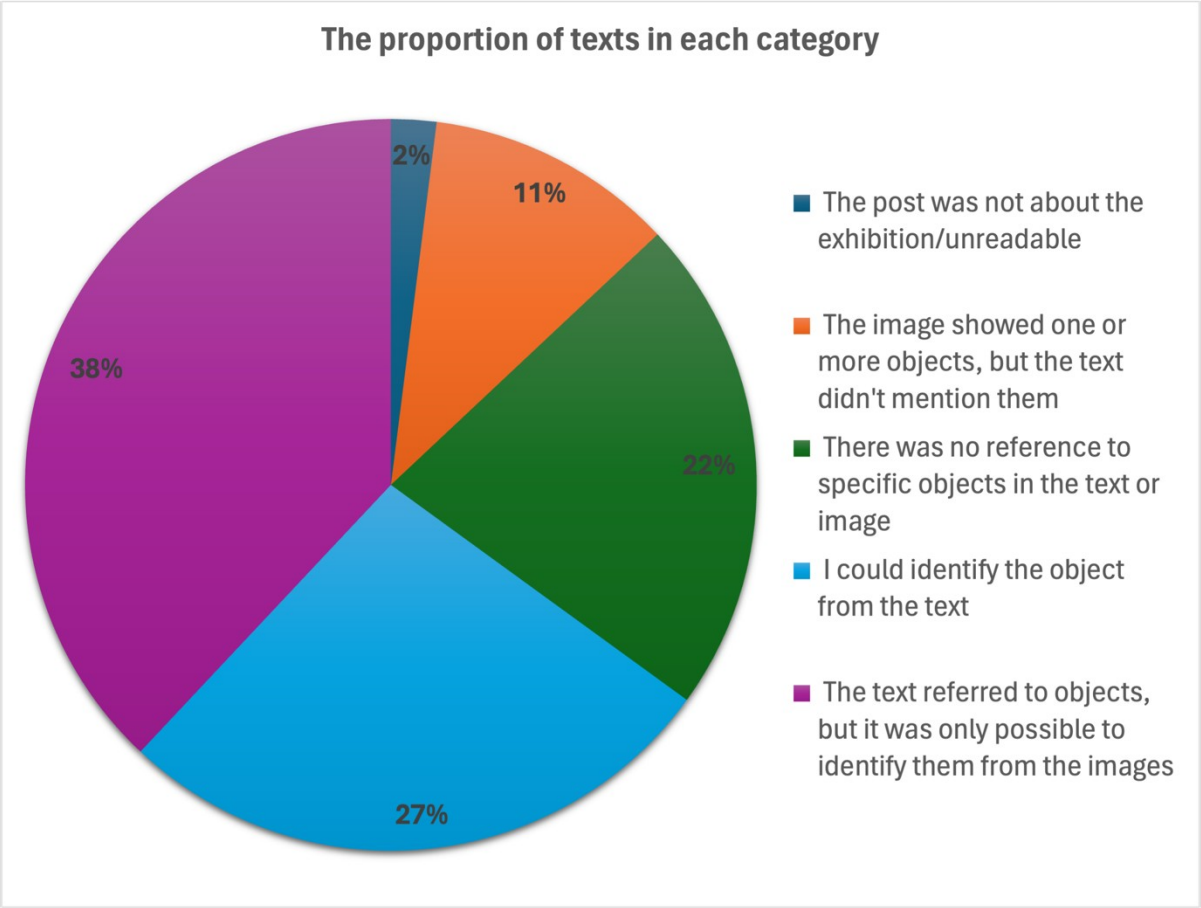


Fig. 2 The proportion of texts in each category (annotator 1)

5. Findings and Discussion

The aim is to link the social media and web data to structured data on objects from the exhibition. We have done this by human annotation and automatically.

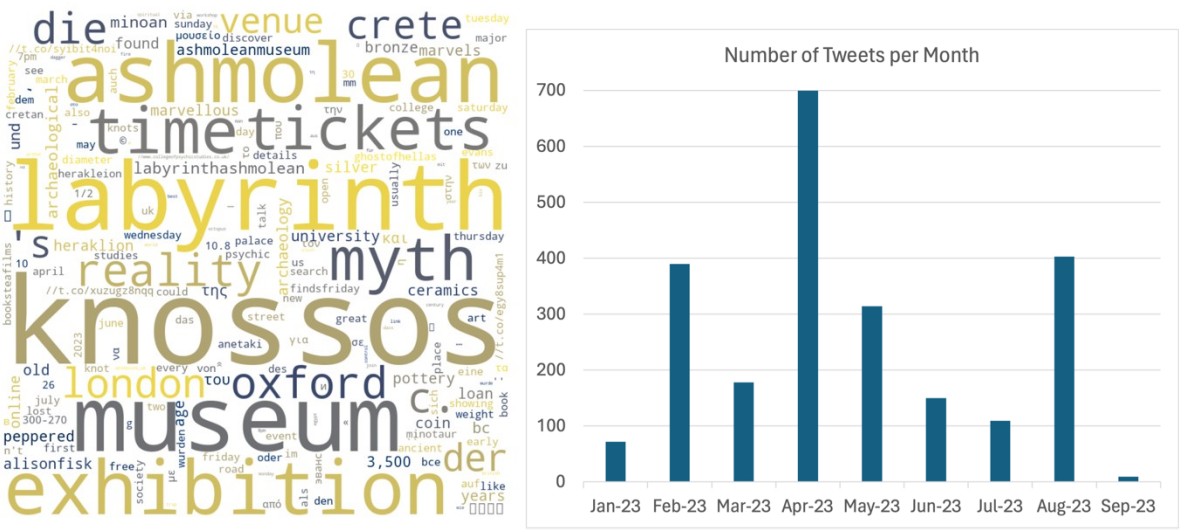


Fig.3 Overview of data gathered

The data gathered was generally on topic with only 38 posts identified as off topic. As shown in figure 2 many of the posts referred to a specific object either in the text directly (627) in the image(s) (249) or were identifiable through a combination of both (890). The content was released in a burst pattern with clear peaks in February, April, and August (Figure 3). This coincides with the peak visitor numbers for the exhibition, at the beginning, end and during the Easter holiday. The exhibition closed on the 30th July, in August we see content that is made up of retweets of earlier posts, reflections on earlier visits made and interactions with wider content such as press references to the exhibition.

There were 141 different hashtags used in the dataset (table 2), the most popular was the official exhibition hashtag #labyrinthashmolean. The other hashtags promoted in the exhibition were not used widely, with #lostinthelabyrinth six times and #labyrinthlabel once. The other frequent hashtags were either related to the exhibition such as #minoan, #knossos, #ashmolean, #oxford, #assassinscreed, or other widely used archaeology or museum hashtags #findsfriday, #internationalmuseumday, #imd2023. This suggests that tasks and special hashtags were not required, but we are aware through observation that these were used quite frequently in Instagram data which is not gathered as it is not publicly available.

labyrinthashmolean	604
archaeology	469
findsfriday	267
minoan	65
internationalmuseumday	54
imd2023	54
drthehistories	49
labyrinth	31
knossos	29
minotaur	23
ashmolean	20
oxford	17

ashmoleanlabyrinth	15
tilesontuesday	15
assassinscreed	15
museum	11
greece	11
assassinscreedodyssey	10
crete	8
ubisoft	8

Table 2 Top 20 hashtags in the data gathered.

5.1 Human Evaluation

The exhibition displayed 238 objects, and of these objects 89 were referenced in tweets. Figure 4 shows the object with the highest impact in this data set was The Poros Ewer (EX24.105). This had 10 tweets and 494 retweets. Highly followed accounts influence the reach of an object within social media, one popular tweet skews data towards four items, all pots (EX24.105, EX24.065, EX24.018 and EX24.072) this user has many followers (62k), a Labyrinth coin (EX24.018) was highlighted by the Ashmolean Museum account (105k followers), a bull vessel (EX24.077) was tweeted by a user with 26K followers. The object with the most original tweets (22) was a prominent object in the exhibition, the Minotaur sculpture (EX24.002).

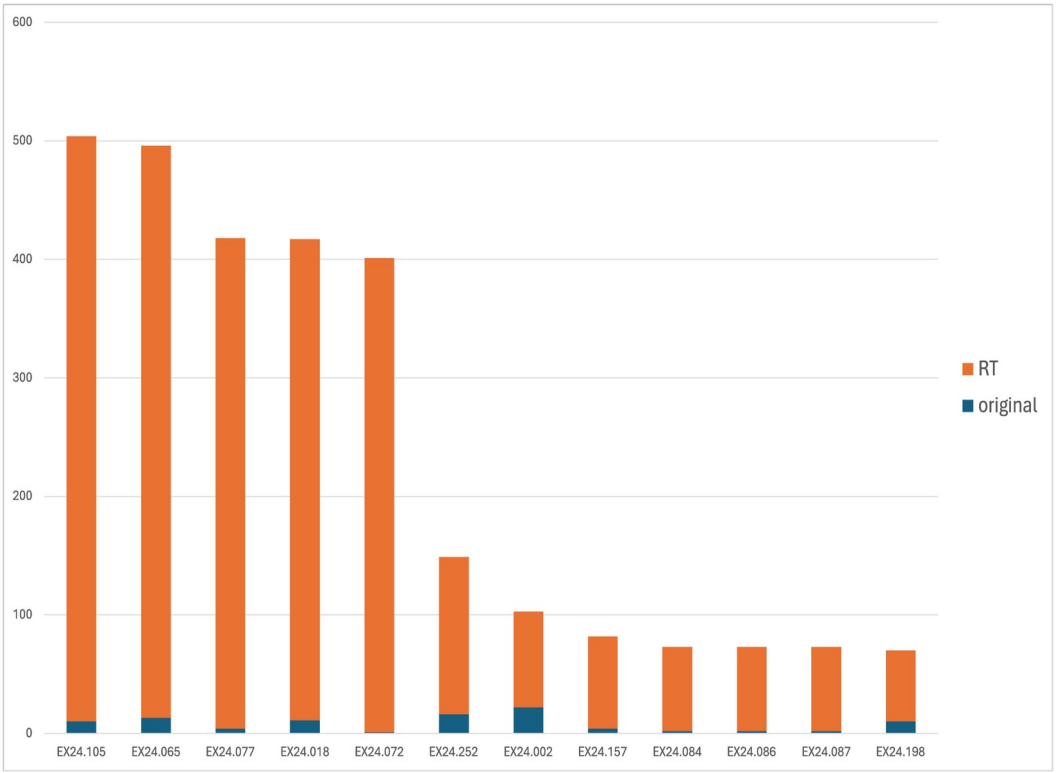


Fig.4 Tweet frequency of museum objects

5.1 Automatic Evaluation

This is a difficult task for a computational model. The human annotators could understand the variety of language used to identify the objects and the often bad spelling and typos found in social media text. We consider two classes in automatic evaluation, the positive class, those which had the human annotations t (I could identify the object from the text) or ti (the text referred to objects, but it was only possible to identify them from the images) and everything else as the negative class. The inter-annotator agreement for this slightly different task was 0.45, indicating an easier task to agree on.

5.1.1 Information Retrieval

We model this as an information retrieval task where the information taken from the structured data on the exhibition objects (names and descriptions) was used as queries to search for matches in the social media posts. We used BM25, an information retrieval model that ranks texts into an ordered list as to how likely they are relevant to a query, to retrieve tweets that were relevant to museum objects. We considered how far down the rank we would have to search to find a reasonable proportion of the tweets identified by the annotators as relevant. We experimented with expanding the queries, using word embeddings to find and add words that were semantically similar to the query terms.

5.1.2 BM25 General Performance

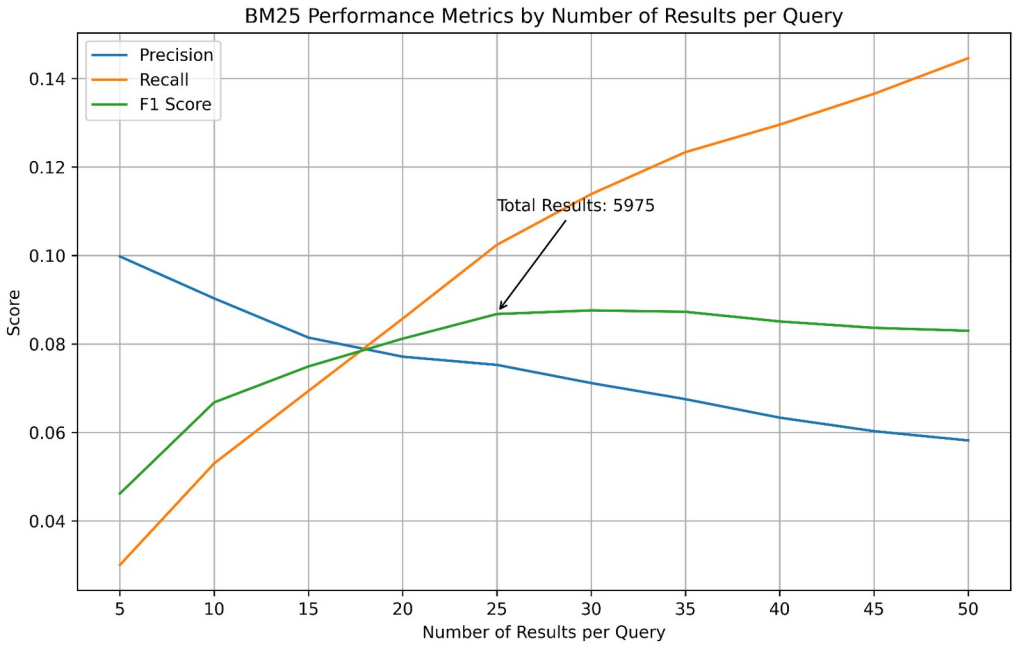


Fig. 5 BM25 performance metrics for all queries

Figure 5 shows performance is generally poor, with the F1 score plateauing at 0.09. In order to achieve this we must consider results to rank 25 for each query, which is not practical for this task.

5.1.3 BM25 Query Expansion

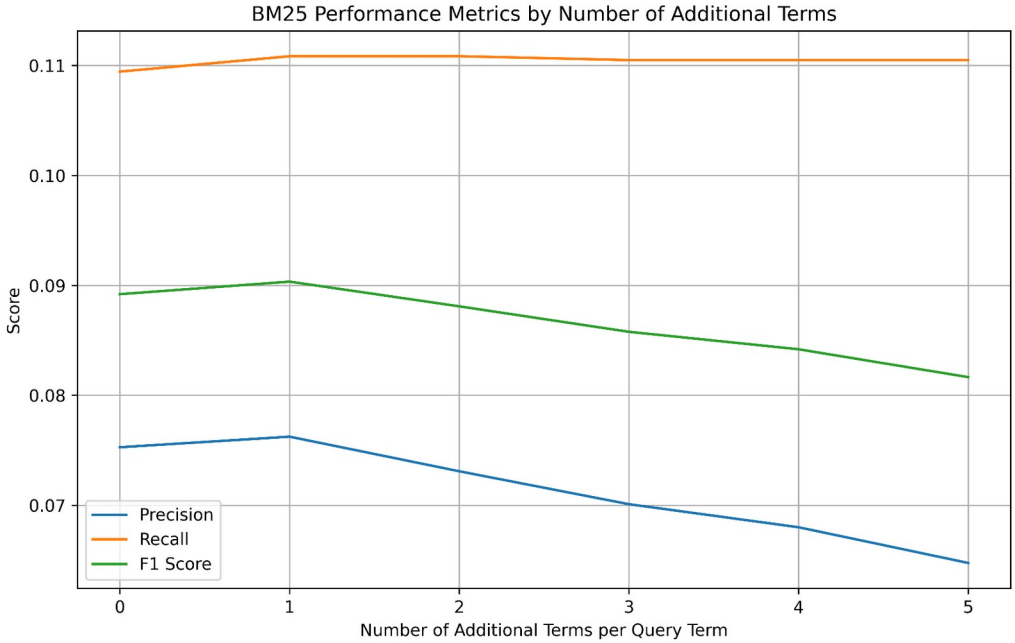


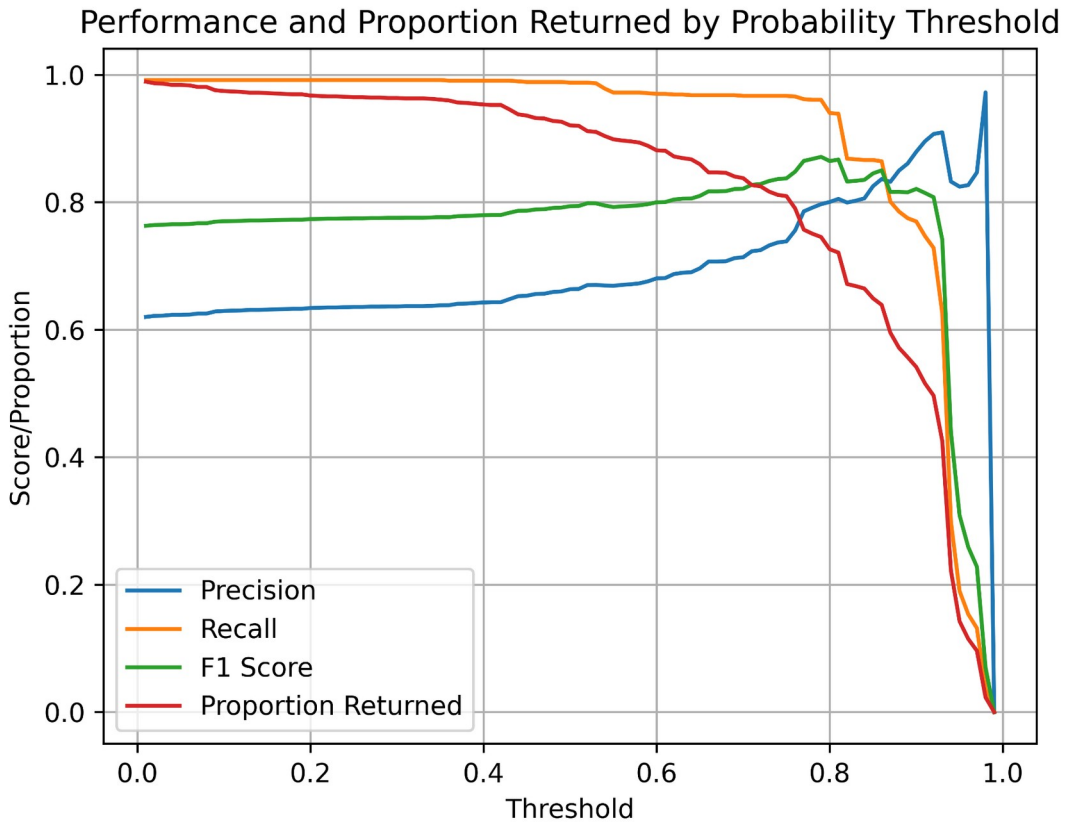
Fig. 7. BM25 with query expansion tweets performance metrics

Query expansion was performed by computing BERT embeddings for each query term. Cosine similarity between term embeddings and precomputed embeddings. The top N most similar words added to the original query, expanding it with semantically related terms. Figure 7 shows query expansion results in a very small increase in performance at one additional term per query term before reducing performance.

5.2.5 Large Language Model Classification

We used an open-source Large Language Model (LLM) FLAN-T5 [https://huggingface.co/google/flan-t5-large], in the few-shot setting asking the model to classify the texts, we provided the model with three positive and three negative tweets for training. We used the few-shot classification module from Skorch, which provides an interface for training a model with a small number of labelled examples. The module allows the pre-trained Flan-T5-Large model to be fine-tuned on a subset of data, enabling it to make predictions on unseen text with minimal labelled input.

The model generates probability scores for each prediction, which are then analysed across a range of probability thresholds from 0.01 to 0.99. By varying these thresholds, the method explores the trade-offs between false positives and false negatives. Adjusting the threshold allows for a better understanding of how the classifier's recall and precision change, ultimately helping to identify a suitable threshold. For each threshold, performance is evaluated using metrics such as precision, recall, F1 score, and the proportion of positive predictions.



Threshold	Precision	Recall	F1	Proportion Returned
0.5	0.663	0.988	0.794	0.920
0.6	0.680	0.970	0.800	0.881
0.7	0.714	0.967	0.821	0.838
0.8	0.800	0.940	0.865	0.726
0.9	0.879	0.770	0.821	0.541

Figure 8. Flan-T5 performance metrics

Figure 8 shows that the proportion of texts classified as relevant is consistently lower than recall.

Discussion

The lack of freely available social media datasets has severely limited this research. Historically, data could have been acquired freely from social media companies but this is no longer possible. The data used here also does not also provide the full picture of social media activity referencing the Ashmolean exhibition. The data that we have does allow us to see what was discussed, although the bias in this data is unknown and this will also affect the reliability of findings. We can be confident that using a subsample of data is useful for evaluating automated approaches, but we must also consider that the inter-annotator

agreement for this task is fair (0.4), which will introduce some error into the automation evaluation process. We found that information from the tasks and special hashtags we experimented with were not obviously present in the data we gathered. We also found that highly-followed accounts influence the reach of an object within social media,–leading to significant data on those objects within this set.

This work aimed to match museum objects to social media texts that could be used to enhance records. This is a difficult task even for humans, and was unlikely to be completely automatable. When evaluating whether using applied computational techniques makes the reconciliation task easier, we found that traditional information retrieval processes did not work well, even when enhanced by work embeddings. We found that the performance of the LLM Flat T5 with 3 shot learning did perform better. To achieve a high F1 score (0.9) the proportion of texts returned was 54.1%. This means that in conjunction with human curation (required to ensure the accuracy of data) 54.1% of the texts would need to be curated to find 77% of the relevant texts thereby reducing the human workload.

Conclusion

Temporary exhibitions are good places to experiment with new approaches because they are high-profile and time-limited. Social media offers a set of diverse responses to exhibition content. This project shows how social media data can be used to complement a museum's existing marketing and evaluation strategies. Objects which had been identified as 'highlights' prior to the exhibition also appeared most frequently in social media. The project also showed that the Museum's promoted hashtags were widely used although the influence of certain hashtags such as #findsfriday or particular users (i.e. 'influencers') was also notable. This shows the potential for exploiting external social media trends and trend-setters in museum marketing. Gathering this fine-grained data about exhibition content, however, is currently labour-intensive and requires specialist input. This project has shown that automating this process, at least in part, provides a way to gather exhibition data more efficiently and with fewer resources, it also illustrates the potential for incorporation of social media data analysis in future exhibition design and evaluation.

7. Funding

This work was supported by the Arts and Humanities Research Council [AH/Y006011/1] and the National Endowment for the Humanities [HND-284978-22].

8. References

[1] Nurmikko-Fuller, Terhi. Linked Data for Digital Humanities. Taylor & Francis, 2023.

[2] Page, Kevin; Delmas-Glass, Emmanuelle; Beaudet, David; Norling, Samantha; Rother, Lynn; Hänsli, Thomas. Linked Art: Networking Digital Collections and Scholarship. Digital Humanities 2020 (DH2020) Book of Abstracts, pp.504-509.

[3] Sanderson, Rob. Standards and Communities: Connected People, Consistent Data, Usable Applications (Keynote). 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. xxviii-xxix.

[4] Avalle,M., Di Marco N., Etta, G. *et al.* Persistent interaction patterns across social media platforms over time. 2024 *Nature* 628, 582–589 https://doi.org/10.1038/s41586-024-07229-y