Preliminary proposal of a metric for assessing and improving the impact of open-access heritage collections on Wikipedia

Irene Iriarte Carretero; UCL Institute for Sustainable Heritage; London, United Kingdom
Jason Evans; National Library of Wales; Aberystwyth, United Kingdom
Scott Allan Orr; UCL Institute for Sustainable Heritage; London, United Kingdom

Abstract

Institutions are increasingly moving toward providing access to collections digitally through open-access initiatives. Measuring this impact and providing metrics to benchmark progress remains an open challenge. This project aims to tackle part of this challenge, focusing on digitized image collections made available via Wikimedia. This work in progress is attempting to create a generalizable and consistent metric which can be used to maximize the potential impact of featuring an image from a collection within a certain Wikipedia entry. The metric considers several quantitative elements that represent specific aspects of impact and is demonstrated for a collection of landscape painting. We focus on impact through a public engagement lens, optimizing our metric to drive an increase in the number of people interacting with items in a collection. Future work will explore more advanced NLP methods and different collection types.

Background and motivation

In order to keep up with current trends and demands, libraries must now also consider the electronic resources that they offer as part of their collections, with a key aspect of managing these digital resources being the consideration of how to provide the means of access for users [1]. Open access is therefore becoming an increasingly important topic within management strategies, with institutions taking into account partnerships or strategies that align with open-access platforms such as Wikipedia [2]. Arguably, this trend has only increased during the COVID pandemic, during which online access to news, entertainment and learning has played a key role.

It is acknowledged that there are many benefits from making collections widely available on platforms such as Wikipedia. Part of the benefits are brought about through an increase in public engagement, manifested through connecting a collection to a wider audience and therefore increasing its reach. Uploading collections to open access platforms also allows institutions to make use of engaged and collaborative communities of volunteers [2]. Finally, it makes it easier to enhance the value of the collections by connecting items to Wikidata, thereby making use of linked open data [3]. However, this move to open access does not come without its challenges, including having a consistent way of measuring the benefits that making collections open access brings.

There is therefore a need to create tailored metrics that can help institutions manage their open access collections and report on their progress, which is harder to do with only traditional metrics such as licensing income and number of hits on the institution's website [4].

Problem

This project aims to tackle part of this challenge, focusing on digitized image collections made available via Wikimedia. The level and type of impact such a collection can bring to an institution will in part depend on how effectively the content of the collection is then leveraged within Wikipedia (and its subsidiaries). This work in progress is attempting to create a generalizable and consistent metric to measure the potential impact of featuring an image from a collection within a certain Wikipedia entry. We initially focus on impact from a public engagement perspective and therefore place emphasis on driving an increase in the number of people viewing and interacting with the items within a collection.

Approach

The metric which we are trying to define is made up of several different components, which we believe are representative of different aspects of the impact the linking of an image may bring to an institution. These components include the average number of views that the Wikipedia entry has had for a recent time period, which represents the wider interest of a topic. Another component used is the length of the entry, which is used as a proxy for the entry's completeness and therefore the chance of users engaging with the content. Including these two components in the algorithm maximize the chance of the image having a large impact as defined above.

However, to make the impact meaningful, the relevance of the content that the images are linked to is also crucial. We therefore propose a third component of the metric, focusing on measuring the relevance of an image to the Wikipedia entry. This is calculated through the image title, which is required to be meaningful and closely related to the image itself, and the text within the Wikipedia entry. This has initially been approached using common Natural Language Processing (NLP) techniques such as term frequency-inverse document frequency [5]. As a last component, the impact metric considers other images which are already present within the entry and how similar they may be to the image being considered.

Component	Description	Relevance
Number of views	Average number of views on Wikipedia entry in past 6 months	Component represents a measure of the wider interest in the topic

Length of entry	Number of words in Wikipedia entry	Component is a proxy for the completeness and usefulness of the entry
Image relevance	Numerical measure of image relevance to Wikipedia entry based on image title	Component represents how meaningful adding the image to the entry is
Image uniqueness	Measure of how many other images are similar to image in question	Component represents the uniqueness that the image will bring to the Wikipedia entry

Table 1. Table containing a summary and justification of all the components in the metric

The metrics are normalised so they can be combined into a single combined metric incorporating weights, which also enables comparison to iterations of the model based on different parameters.

Using the impact metric described above, the model is then able to suggest relevant Wikipedia entries for the images in the collection, which either the institution or volunteers can then action. The information is displayed in an intuitive user interface, allowing users to explore different images in the collection and the potential Wikipedia entries they could be placed within, as well as the values for all the different components of the impact metric.

Preliminary results of work in progress

The pilot study is focused on the Wales Landscape Collection from the National Library of Wales [6] and the model has therefore been optimised for use with landscape-related images, which typically contain pictures of things such as castles, rivers, and churches. Initial results show that the model can suggest relevant Wikipedia entries to link to the images in the collection, demonstrating the potential of the model to increase the overall impact of an open access collection.

One example which showcases the potential increase of reach of items in the collection is a picture of Monmouth castle displayed below. This image is currently exclusively in use in the *Monmouth Castle* Wikipedia entry, which has accrued 12506 views in the six months prior to May 2021. However, the algorithm has also identified the entry *Henry V of England* as highly relevant, given that he was in fact born in Monmouth castle. This entry has had over 1.3 million views in the same time period which represents a significant increase of the number of people who can view and interact with the image.

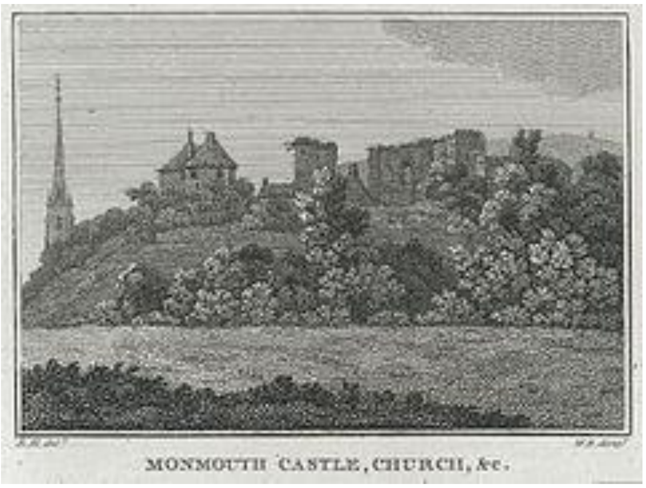

Figure 1. Image showing the castle and church at Monmouth, available on Wikimedia through the National Library of Wales (Image in the public domain)

Conclusion

To improve this work further, more sophisticated NLP methods of calculating the image relevance within the Wikipedia entry text will be explored. More quantitative ways of measuring the success of the model will also need to be developed. The model will be extended to support other types of collection of images as well as landscape related ones. Finally, the impact measure can be extended to include other relevant factors.

We believe this work could have many applications for the management of open access collections in Wikimedia. It could be used by institutions as a success measure and to improve the impact of collections over time. The model could also be used to optimize time and resources available to link images to Wikipedia entries, ensuring that the highest impact combinations are tackled first. Finally, the model could be used to determine the potential impact of different collections, based on the image titles, at the point of determining which collection should be made open access first.

References

[1] V. L. Gregory, Collection development and management for 21st century library collections: an introduction (American Library Association, 2019), pg. 1-2.

[2] J. Lubbock, “Wikipedia and libraries”, Alexandria, 28.1, 55-68 (2018).

[3] S. Allison-Cassin, D. Scott, “Wikidata: a platform for your library’s linked open data”, Code4Lib Journal, 40 (2018).

[4] E.D. Marsh., R.L. Punzalan, “Beyond Clicks, Likes, and Downloads: Identifying Meaningful Impacts for Digitized Ethnographic Archives”, Archivaria, 84 (2017).

[5] S. Qaiser, A. Ramsha, “Text mining: use of TF-IDF to examine the relevance of words to documents”, International Journal of Computer Applications, 181.1, 25-29 (2018).

[6] Wales Landscape Collection from the National Library of Wales can be accessed at https://commons.wikimedia.org/wiki/Category:Welsh_Landscape_Collection

Author Biography

Dr Irene Iriarte Carretero is an MSc student on the MSc Data Science for Cultural Heritage at UCL, having previously completed a PhD in Computational Chemistry and moved into a Data Scientist position in industry. Her current research focuses on measuring the impact of open data collections.

Jason Evans is the National Wikimedian at the National Library of Wales. He has managed a number of projects to improve content on the Welsh language Wikipedia. He works to advocate for open access within the culture sector by openly sharing library data and demonstrating the benefits to the organisation and the public. Jason is a regular contributor to digital heritage conferences with a particular interest in linked open data.

Dr Scott Allan Orr is a Lecturer in Heritage Data Science at the UCL Institute for Sustainable Heritage. An engineer with broad interests, his research within heritage science primarily uses data-driven approaches to assess environmental impacts on the historic built environment. He is the Deputy Programme Director of the MSc Data Science for Cultural Heritage, on which he teaches modules on heritage science, heritage data visualisation and digital technologies and approaches for built heritage.