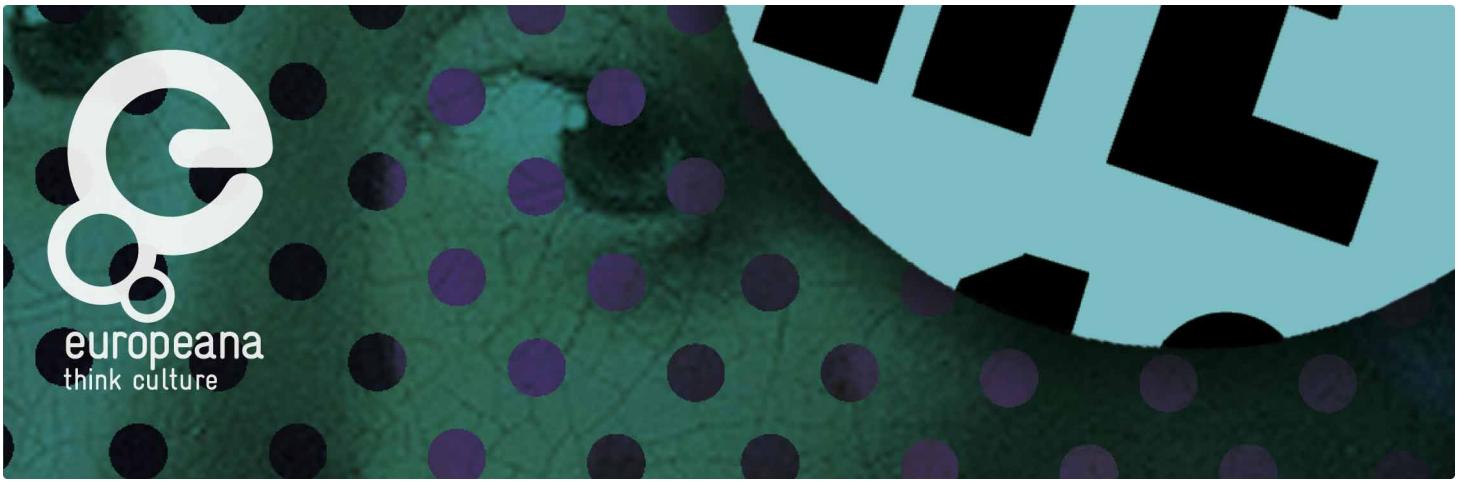


→ *Report on Requirements for Usage and Reuse Statistics for GLAM Content*

Author: Maarten Zeinstra, Kennisland

WIKIMEDIA UK

WIKIMEDIA FRANCE

WIKIMEDIA CH

WIKIMEDIA NEDERLAND

1. Table of Contents

2. GLAMTools project

The “Report on requirements for usage and reuse statistics for GLAM content” is part of the Europeana GLAMTools project, a joint Wikimedia Chapters and Europeana project. The purpose of the project is to develop a scalable and maintainable system for mass uploading (open) content from galleries, libraries, archives and museums (henceforth GLAMs) to Wikimedia Commons and to create GLAM-specific requirements for usage statistics. The GLAMTools project is initiated and made possible by financial support from Wikimedia France, Wikimedia UK, Wikimedia Netherlands, and Wikimedia Switzerland.

Maarten Zeinstra from Kennisland wrote the research behind this report and the report itself. Kennisland is a Dutch think tank active in the cultural sector, working especially with open content, dissemination of cultural heritage and copyright law and is also a partner in the project. For more information go to www.kl.nl/en/

3. Executive Summary

Wikipedia is the 6th most visited website in the world. As a collaborative encyclopaedia its mission is to share the world's knowledge. This mission is very attractive and similar to other knowledge and cultural institutions like Galleries, Libraries, Archives, and Museums (GLAMs). Cooperation between these and Wikimedia is therefore frequent and successful.

The current system of measuring the outcome of these collaborative projects between GLAMs and Wikimedia projects is based on a set of loosely connected scripts created by volunteers and hosted outside of Wikimedia projects. Although the scripts are good and useful, they lack institutional support, maintenance and development. The scripts also lack usability, a graphical user interface and easy to use export options.

Sharing cultural content on Wikimedia projects has an enormous impact on the reach of that material. Research by Kennisland has shown that sharing content can result in a 17000% increase of yearly impressions of an object which was otherwise only available on the institution's website[1]. These numbers are important for GLAMs to further institutionalise contributions to Wikimedia projects, but are very difficult to come by.

Our research shows that there is a need for a GLAM analytics system as an open dashboard that shows all contributing institutions with monthly overviews of contributed objects, their plays and their usage in Wikimedia projects. The GLAM analytics system needs to be maintained by the Wikimedia analytics team to ensure its sustainability. The Wikimedia analytics team currently develops and maintains services that are well-suited for this purpose: Kraken and Limn.

Well-designed, reliable and consistent GLAM statistics will provide a great boost of confidence for GLAMs, as well as a place of research for scholars in online dissemination of cultural works.

[1]The report of this research is available here (Dutch only): http://beeldenvoordetoekomst.nl/nl/research/effectmeting-nationaal-archief-joins-wikipedia

4. Introduction

This report on requirements for usage and reuse statistics for GLAM content is part of the GLAM-wiki Toolset Project[2] initiated by Wikimedia Netherlands, Wikimedia France and Wikimedia UK and coordinated by the Europeana Foundation.

The report gives an overview of the current state of metrics that measure the impact and use of content made available by GLAMs to Wikimedia projects within Wikimedia Commons. The report also gives an overview of requirements of GLAM tools based on interviews and surveys of three user groups: Wikipedians, GLAMs and the analytics team of the Wikimedia Foundation.

At the time of writing,[3] GLAM content amounts to 2,071,402 objects or 13.14% of the total collection of Wikimedia Commons.[4] In other words, one in every eight files on Wikimedia Commons is made available through some collaboration with GLAMs. Either a GLAM puts its public domain collection or openly licensed collections online and volunteers upload this, or an active collaboration between GLAMs and the Wikimedia community such as Wiki Loves Monuments creates GLAM content. Volunteers then place these media objects in Wikipedia-articles creating an incredible increase in their visibility.

Contributing institutions see the potential of Wikimedia as a distribution channel. However, they have very limited tools to measure the effects of the content they make available through Wikimedia Commons. The tools that are currently available show impressive statistics in the increase of visibility of the material, in one case visibility increased 17,000-fold.[5]

[2] See: http://commons.wikimedia.org/wiki/Commons:GLAMToolset_project

[3] Monday, January 21th 2013

[4] This is the total number of files in the category 'Galleries,_Libraries,_Archives_and_Museums_(GLAM)' and it's 73,822 containing categories.

[5] See research 'Nationaal Archief Joins Wikipedia Effectmeting' (Dutch) at http://beeldenvoordetoekomst.nl/nl/research/effectmeting-nationaal-archief-joins-wikipedia

5. Problem definition

Wikimedia Commons ('the Commons') is a community media repository for free and open media files. The Commons host many of the media files that are used on Wikipedia and other Wikimedia projects such as Wikisource, Wikivoyage, Wikispecies, etc. At the time of writing, over 14 million media files have been uploaded to Commons.[6] A large percentage of these files is contributed by GLAMs.

These media files share several important characteristics. They are usually more than merely decorative or illustrative media files; they are historical, primary-source documents, relating to detailed content within a specific article. Such media files typically come from repositories with authoritative metadata about the works and are made available by cultural institutions.

GLAMs now use several (proof of concept) tools created by volunteers that reside on the Toolserver[7] to measure the effects of these images. The Toolserver is a Wikimedia Germany funded project that allows developers to query Wikipedia sites and hosts small programs to perform analysis or other repetitive tasks on those sites. The life cycle of the Toolserver is currently under discussion and there is a high likelihood that it will cease to function within two years.[8]

While the existing tools for GLAM metrics and statistics have been helpful, they do not provide the breadth of necessary metrics. Also, the tools are not well maintained and could go offline at any time. Research, in the form of a questionnaire directed at GLAMS, will provide more clarity on these issues. We can conclude that there is a need for good statistical tools that measure the impact of GLAM media on Wikimedia projects. Current tools are not developed with durability and usability in mind.

---

[6] https://commons.wikimedia.org/wiki/Special:Statistics. This is, however, on Wikimedia Commons only, other Wikimedia projects hold more files.

[7] http://www.toolserver.org

[8] https://meta.wikimedia.org/wiki/Future_of_Toolserver

6. Current status

There have been several initiatives to kick-start GLAM statistics. Several independent developers like Magnus Manske, created tools such as GLAMorous and BaGLAMa on the Toolserver to make rudimentary analytics available. Also, http://stats.grok.se/ shows us the number of visits for relevant pages on Wikimedia projects. Most presentations and reports about usage of GLAM content on Wikimedia rely on these tools for their data.

As mentioned earlier, these tools were created and are maintained by single volunteers. volunteers cannot maintain these tools indefinitely, That is why several initiatives[9] have tried to kick-start more sustainable GLAM metrics programs,. Key figures in these attempts are Diederik van Liere, Dominique McDevitt-Parks, Liam Wyatt, Lori Philips, Maarten Brinkerink and Maarten Dammers[10]. All of these experts are or were involved in some aspects of GLAM collaborations, GLAM software development or statistics.

Most GLAMs that are already contributing to the Wikimedia Commons are very interested in statistics, e.g. Netherlands Institute of Sound and Vision, Rijksmuseum, Amsterdam Museum, Dutch National Archive, Brooklyn Museum, etc. International initiatives like the 'Wiki Loves…' projects also benefit from these GLAM statistics.

[9] Several initiatives have tried to kick-start GLAM metrics:

- http://www.mediawiki.org/wiki/Analytics/Pageviews/GLAM
- https://docs.google.com/document/d/1kptd2EUgZ5gSuh9PFP7SUl4wyL5JJGKp2eYxQocPro/edit
- http://etherpad.wikimedia.org/GLAMcampAmsterdamSun-stats
- http://etherpad.wikimedia.org/GLAMcampNYC-metrics-proposal
- http://outreach.wikimedia.org/wiki/GLAM/Tools_%26_Requests

[10] The writer of this document, Maarten Zeinstra, was also involved in these discussions.

7. Research overview

To answer the question 'What functional requirements are necessary for gathering metrics on impact and use of GLAM content on the Commons?' a questionnaire has been developed and distributed among several mailing lists and personal networks. A total of 32 useful (n=32) responses was received[11].

7.1. Questionnaire

The questionnaire focused on whether an institute contributed to Wikimedia, what kind of statistics they already use and what kind of statistics they require for their collection on the Commons.

7.1.1. Demography

We received a total of 32 useful responses. Most of these were people from cultural institutions (27 unique cultural institutions, 3 duplicate institutions) who hold a variety of roles from directors to Wikimedians in Residence. About one half of them already contributed to Wikimedia Commons, and their input ranged from several files to thousands of files. The other half intends to contribute in some form to Wikimedia Commons in the future. The institutions of all respondents represent over 30 million files, roughly 14 million more than Wikimedia Commons currently holds.

7.1.2. Use of analytical tools

Over half (52%) of the respondents already use Google Analytics or similar systems to monitor their collection and websites. Page views are the most mentioned interest here. Most respondents did not know how much of their traffic originated from Wikimedia projects. Those that did (n=5) indicated that this was usually lower than 5% of their total traffic (~0.87%, ~3%, 8.3%, 0,2, 1,84% for Wikipedia) and mostly unknown for Wikimedia Commons (0,1%, 1,0%). Referrals from Wikipedia usually indicate that Wikipedia visitors click on source links to these institutions in footnotes; referrals from Wikimedia Commons mean that they were interested in the institution after seeing a media item.[12]

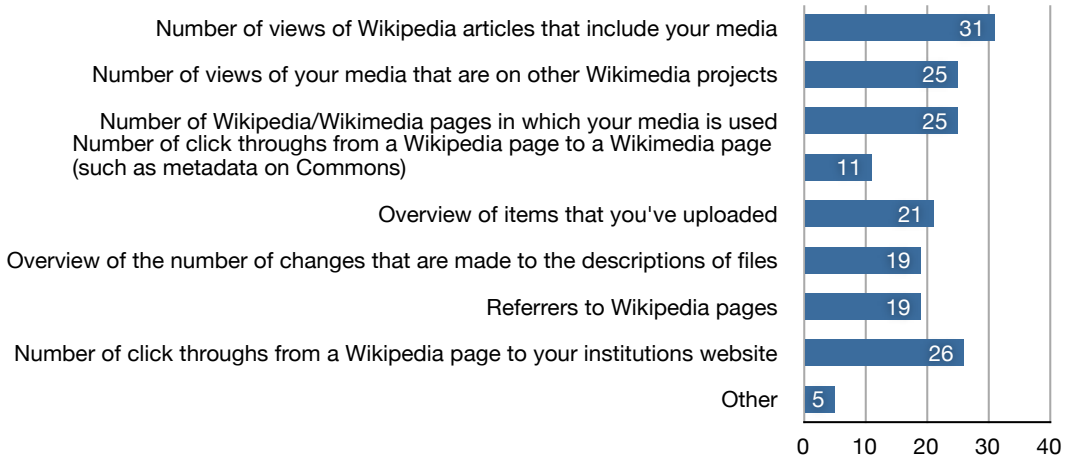
7.1.3. Requested information

Respondents were presented a list of options for the kind of information they require:

- Number of views of Wikipedia articles that include your media
- Number of views of your media that are visible on other Wikimedia projects
- Number of Wikipedia/Wikimedia pages in which your media is used
- Number of click throughs from a Wikipedia page to a Wikimedia page (such as metadata on Commons)
- Overview of items that you have uploaded
- Overview of the number of changes that are made to the descriptions of files
- Referrers to Wikipedia pages
- Number of click throughs from a Wikipedia page to your institution's website
- Other

[11] Another 11 respondents were received through a translated version of the questionnaire to French, with thanks through the efforts of Wikimedia France. These extra respondents confirmed the views represented in this document but, as the questionnaire was differently structured, it proved hard to integrate these numbers into this document.

[12] There is some controversy in the Wikipedia community about referencing media. See http://en.wikipedia.org/wiki/User:Dominic/Image_citation for more details.



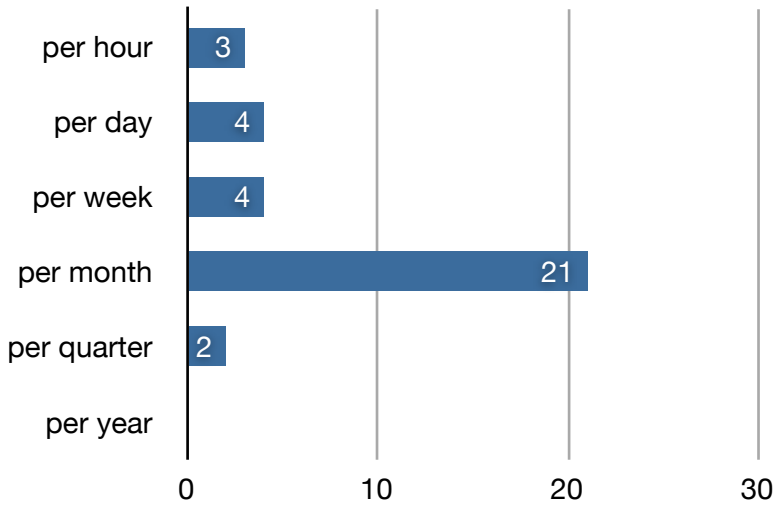
All users wanted to see the number of views and the number of click throughs from a Wikipedia page to their own institution's website 'Other' suggestions contained:

- Information on provenance / demographics of the viewers/users (country / language)
- Technology used (browser, OS)
- Number of views of the actual media files on the Commons
- Plays of content on Wikipedia pages
- Plays of content on Wikimedia pages

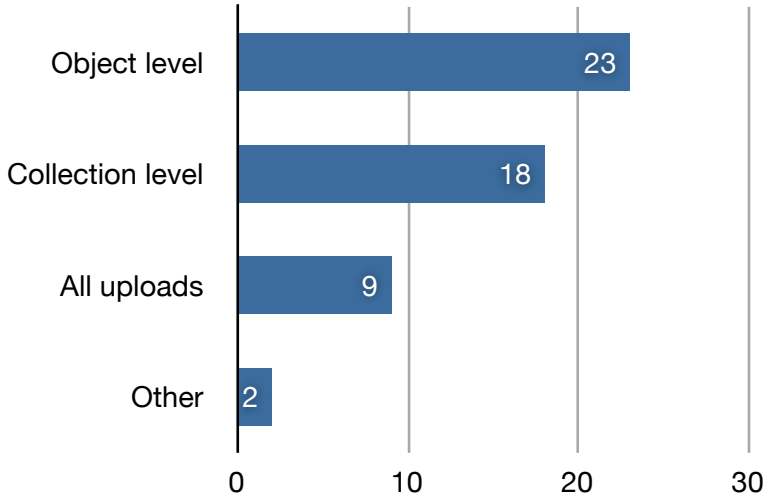
Respondents could comment on why they needed these statistics in a text box. Justifiability, relevance, and adaption were most the frequent responses. Respondents share the need for statistics to justify sharing collections to Commons to their management, funders, directors, etc. and that sharing media on Wikimedia Commons is worthwhile for successfully executing the mission of their institutions. Respondents want to keep their collections relevant by offering media from them for reuse through Wiki-projects. They want to see which media objects are popular and which media objects could best be shared next. They want to adapt to the needs of the Wiki community by seeing what works for the community and also for themselves as institutions. Some indicate that the goal of sharing media on the Commons is to achieve more views on their own website, while others are more concerned with the reach of their content as multi-channel communication and do not care about click backs.

7.1.4. Frequency and granularity

When asked how often they need these statistics (per year, quarter, month, week, day, hour) more than 50% indicated that once per month would be sufficient. Users were able to tick more than one box, so the sum of all percentages adds up to more than 100%.



When asked what level of statistics they desire (per object, per collection, all uploads), they indicated to prefer statistics on an object-level basis.



7.1.5. Other suggestions made by respondents

The following remarks were given when asked for other suggestions to add to the requirements of reuse statistics for GLAM content.

- 'Integrated with Google Analytics (i.e. provide GA tracking ID, then stats from Wikimedia Commons are also visible on GA)'
- 'Create a shared dashboard for all participating collections, in which we can compare our statistics[13]

[13]See www.museum-analytics.org or http://dashboard.imamuseum.org/ for inspiration

- 'Well-designed, visually attractive stats would be a big plus - the existing statistics dashboard of Wikimedia projects is feature-rich but looks unattractive'
- 'Create exporting functionality (PDF, CSV, images of graphs) similar to Google Analytics - so that we can send (automated or manually compiled) reports to anyone, e.g. for integration in year reports.'
- 'Measure the number of plays of audio and video material. While it is sufficient to count the display of images on articles it does not take into account whether users of wiki-projects have actually experienced the media file.'
- ‘Time spent per page visit’

7.2. Qualitative research on the technical possibilities of GLAM stats

The survey presented above gives a good overall indication of what stakeholders want in analytics stats. However, they do not take into account what is possible within the current infrastructure of Wikimedia projects.

Several discussions with the following involved experts resulted in the research below:

- Maarten Dammers (Commons volunteer; uploads most GLAM content)
- Diederik van Liere (Team leader analytics team Wikimedia Foundation)
- Dan Andreescu (Developer for analytics team Wikimedia Foundation)
- Dan Entous (Senior Developer GLAMTools project)

A discussion with Van Liere, Andreescu and Entous provided clarity on the technical restrictions of the Graphical User Interface (GUI) of the statistical tools. The Wikimedia Foundation's analytics team actively develops and maintains a tool for displaying graphs and charts called Limn.[14] Limn is a JavaScript library that creates easy-to-use charts that can be created without much programming effort. Limn is used in the statistical output of the Wikimedia Foundation: the Wikimedia report card.[15]

A discussion with Dammers and Van Liere showed the complexity of the underlying infrastructure of sharing GLAM content with Wikimedia projects. As stated before, Wikimedia Commons is the media repository of all Wikimedia projects. Most media on other Wikimedia projects embed these files from Wikimedia Commons. For most GLAMs, site statistics for Wikimedia Commons are not what they are interested in. It is the usage of these files on other Wikimedia projects, like Wikipedia, that GLAMs want to get insight in. This means that we need to measure the number of impressions[16] of the media file, regardless of which project uses the file.

Wikimedia has a complex architecture. Next to a LAMP[17] stack MediaWiki enables a complex architecture of multiple sites and nested categories. The Wikimedia analytics team uses log files on the LAMP level of the architecture to generate statistics using an open source JavaScript library to generate graphics (Limn). These statistics are easily extended with other data from these log files.

The Wikimedia Foundation is actively developing a data services platform for MediaWiki called Kraken[18] to aggregate, store, analyse, and query all incoming data that is of interest to the community. This network of 40 computers (40-node cluster) will crunch data from Wikimedia projects' log files.[19] One of the available data streams comes from Apache log files; another comes from harvesting data, such as categories, from MediaWiki APIs.

A new definition for measuring impressions that incorporates clarification of the methods of measurement needs to be created. Together with the experts mentioned above the following definitions were designed to be used to describe the technical data requirements of the statistics: FileViews, ArticleViews, and FilePlays.

---

[14] See https://github.com/wikimedia/limn/wiki

[15] See http://reportcard.wmflabs.org/

[16] An impression includes all cases where a media file from Wikimedia Commons is loaded; these include hotlinking and embedding in WikiMedia projects, as well as direct file access.

[17] Linux, Apache, MySQL, PHP

[18] http://www.mediawiki.org/wiki/Analytics

[19] http://dumps.wikimedia.org/other/pagecounts-raw/

7.2.1. FileView

A FileView is an impression of a file to be measured through Webserver logs, regardless of which Wikimedia Project or external site uses the file. The total number of impressions is the simplest statistic that can be run by the analytics team. Here we have to take into account that one 'File' can have multiple versions and/or sizes. In useful statistics, these need to be canonicalised[20] into one File. The FileViews measurement should also contain the number of referrals or hot-linked[21] views. These are not requested by GLAMs very much, perhaps because institutions are not aware that Wikimedia Commons allows hot-linking.[22]

7.2.2. ArticleView

ArticleViews are the number of views one article gets. We have to keep in mind that one article can contain multiple media files from one GLAM. An article about an artist can contain multiple media files that are measured as one ArticleView.

7.2.3. FilePlay

For audio and video, an impression is not sufficient. Methods need to be developed to measure at least the start of an audio or video file. We will call statistics of media plays 'FilePlays'. HTML5 players and other players use stills to show a preview of a movie clip. Similar to FileViews, these need to be canonicalised into one impression or play.

7.3. Research conclusions

There is an obvious need for high-quality statistics on GLAM content that the Commons hosts. This is shown both by the high number of recipients and their responses. We see that a lot of recipients show interest in the Commons but have not yet contributed content. All participants like to see statistics based on the use of their objects in Wiki projects like Wikipedia.

Statistics will be used for two main purposes, primarily to show the effect of open sharing of cultural content. Cultural institutions need these numbers for their own collections. More advanced contributors also indicated that they want to have insight in statistical information in order to adapt content for the Wiki community by selecting those kinds of objects that are most often used by the community.

The Wikimedia analytics team is actively maintaining and developing tools that could enable most of these requirements. This mainly becomes a matter of setting up a display page (Limn) and consistently gathering data (Kraken).

Mapping the requirements of end users and the possibilities and vision of the Wikimedia Foundation's analytics team, leads us to propose that a Kraken-Limn-based infrastructure will be used to measure impressions of all media files as well as the articles in which they appear.

---

[20] Canonicalising is a technical term for normalisation or aggregation of versions of an object and treating them as a single object.

[21] A file hotlink is when you embed an image from Wikimedia Commons directly on your own webpage.

[22] There does not seem to be a formal policy on hotlinking at the moment, see http://commons.wikimedia.org/wiki/Commons:Reusing_content_outside_Wikimedia/technical#Hotlinking

8. Requirements for usage and reuse statistics for GLAM content

The combination of the quantitative and qualitative research lets us describe the requirements for statistics for GLAM content as follows:

'A GLAM analytics system is an open dashboard that shows all contributing institutions with monthly overviews of objects, their plays and their usage in Wiki projects. Ideally, the data is created by Kraken and presentation is handled through Limn with well-designed graphs providing exportable datasets.'

This translates into requirements with various levels of abstraction. High-level requirements are the most abstract and deal only with the overarching questions of what the system needs to be able to do. Medium-level requirements deal with intermediary abstraction of reuse statistics. Low-level requirements are out of scope for the purpose of this document.

8.1. High-level requirements

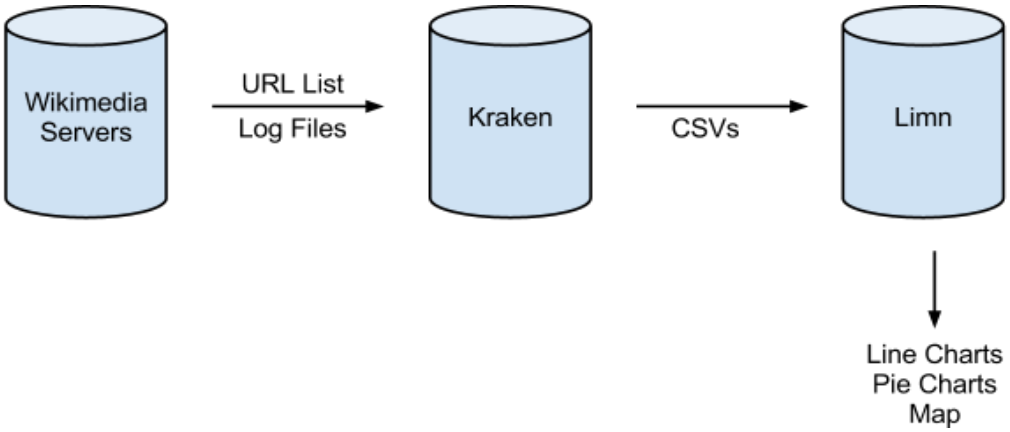
A set of high-level requirements can be defined as queries that GLAMs will ask a statistics system.

1. Give me an overview of my media files
2. Give me an overview of the usage of my media files on Wikimedia Project X
3. Give me a monthly overview of the views my media files had (FileViews)
4. Give me a monthly overview of the number of page views of Wikipedia articles that include my media files (ArticleViews)
5. Give me a monthly overview of the number of plays my audio/video had (FilePlays)
6. Give me a monthly overview of the provenance of the users that see my media files
7. Give me a monthly overview of the devices/OS of the users that see my media files

8.2. High-level technical requirements

Wikimedia Projects need to be able to give log files and File listings to Kraken. Kraken needs to be able to output CSVs that return the results of the queries in the high-level requirements

1. Limn needs to be able to provide line charts
2. Limn needs to be able to provide pie charts
3. Limn needs to be able to provide a world map



8.3. Medium-level technical requirements

8.3.1. Kraken

- Kraken needs to be able to canonicalise different versions (thumbs, stills, etc.) of files into one identity (File).
- Kraken needs to be able to cross-reference media files from specific Wikimedia Commons categories with log files of Wikimedia Commons requests

8.4. Non-functional requirements

- Kraken needs to be able to remain online even when it needs to do millions of comparisons on every hit that the Wikimedia Commons gets.

Matching canonicalised media files to the articles they are embedded in does not seem to scale very easily. It is the concern of this researcher that the analytics team does not realise the numbers of GLAM media files to be tracked.

8.5. Non-requirements / out of scope

- Being able to see how many click throughs happen from Wikimedia Commons to the institution's website.

Two things are at play here. First, this is not something to be measured by Wikimedia Projects but by the institution itself. A possible solution is to provide proper tutorials together with the analytics page. Second, the infrastructure does not promote media on Commons as source material to the end user. User Dominic has written an interesting article about this problem.[23] Because of this infrastructure, GLAMs will not see Wikipedia referrers as a result of their shared media.

8.6. Data model

Below is a draft of the data model per high-level question. The data model indicates what type of data file needs to be outputted by Kraken in order for Limn to work. The graphical examples below are based on the current version of Limn and test data gathered from the Wikimedia Scorecard and Toolserver Tools using the Openimages.eu dataset.

8.6.1. Give me an overview of my media files

On the user interface, this can simply be fixed through a link to the Wikimedia Commons GLAM category.

However, this question also refers to the canonicalisation of files. We therefore also need a dataset that has a canonicalised name and all its derivatives:

- Canonicalised name
- List of derivatives

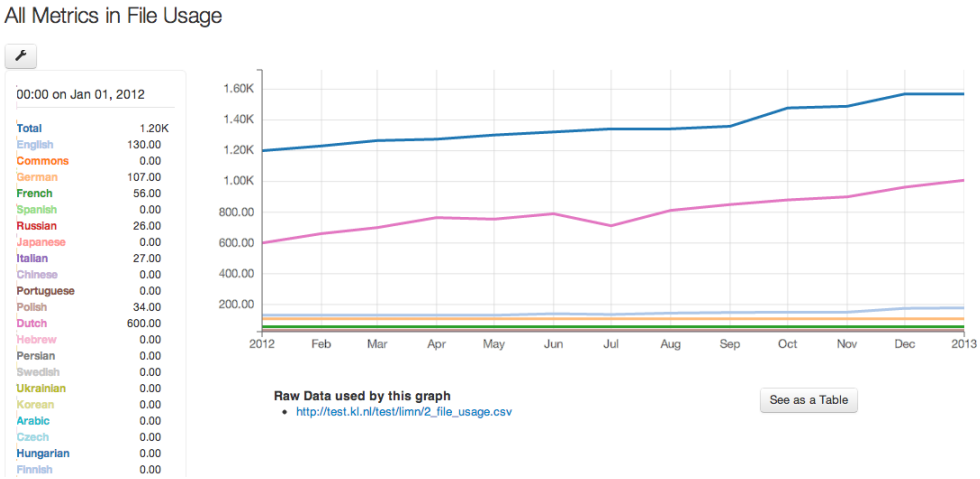
[23] http://en.wikipedia.org/wiki/User:Dominic/Image_citation

8.6.2. Give me an overview of the usage of my media files on Wikimedia Project X

Based on the GLAMorous[24] tool by Magnus Manske, this shows a live overview of objects per Wikimedia Project:

- Period (in months)
- Project
- Total objects in category
- Total used images per (pre-selected) Wiki project
- Total unique used images

This produces graphs with the period in months on the x-axis and the total objects on the y-axis, uses of objects per pre-selected Wiki project and total unique images used.

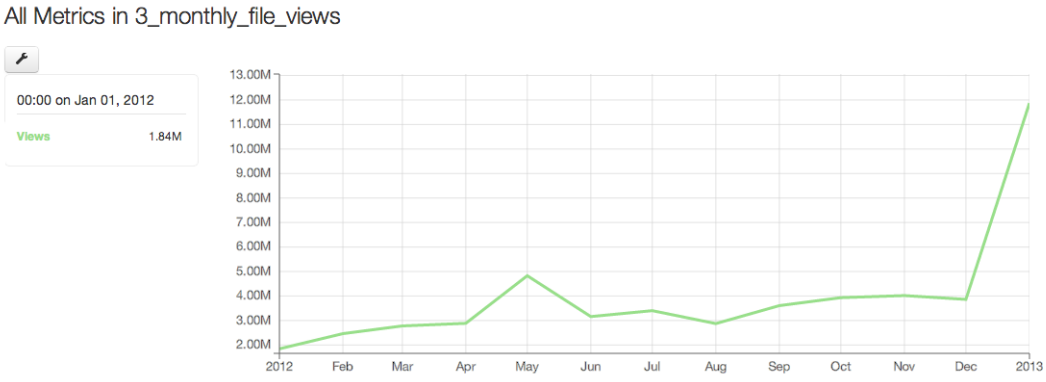


8.6.3. Give me a monthly overview of the views my media files had (FileViews)

Gives the total number of views per month, per filename:

- Period
- URL thumb
- Filename
- Views

This produces a line graph with a monthly period on the x-axis and number of views on the y-axis.



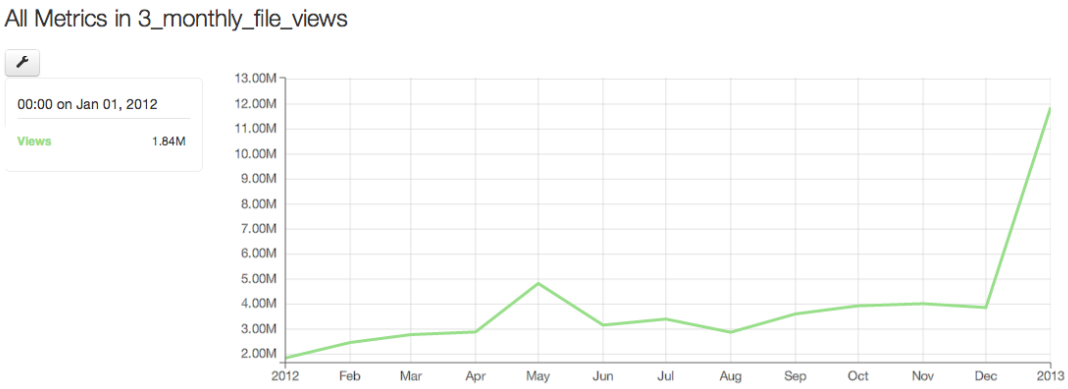
[24] See http://toolserver.org/~magnus/glamorous.php.

8.6.4. Give me a monthly overview of the number of page views of Wikipedia articles that include my media files (ArticleViews)

Shows the total number of views per month, per article:

- Period
- Article name
- ArticleURL
- Views

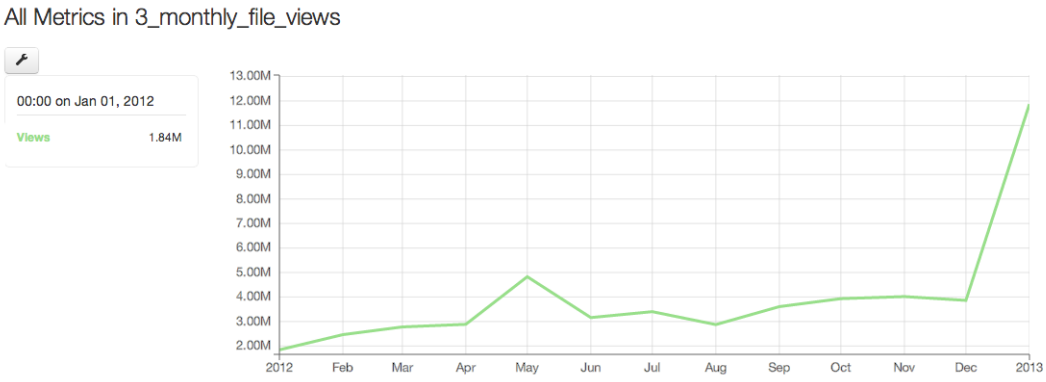
Has the ability to show which files from a category are used in an article.



8.6.5. Give me a monthly overview of the number of plays my audio/video had (FilePlays)

Gives the total number of plays per month, per filename:

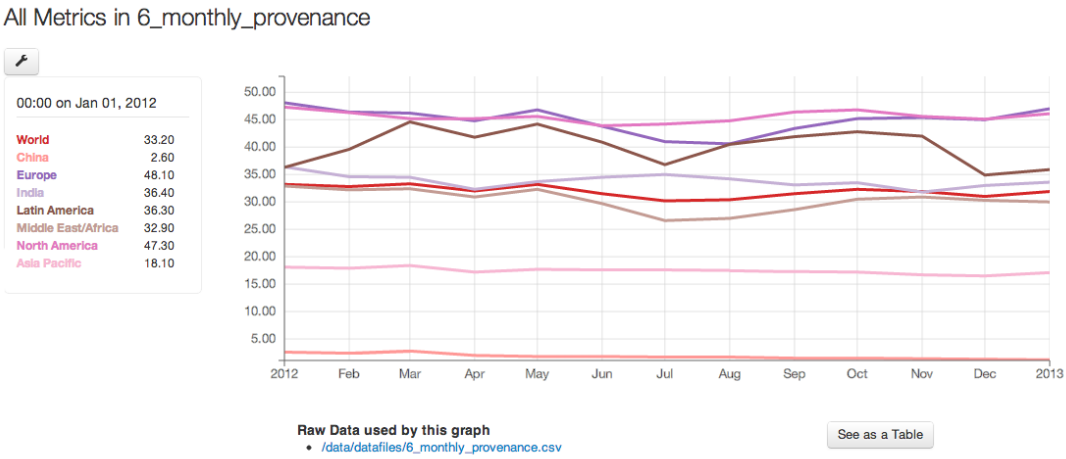
- Period
- URL thumb
- Filename
- Plays



8.6.6. Give me a monthly overview of the provenance of the users that see my media files

Gives the number of views per month, per country:

- Period
- Country
- Views



8.6.7. Give me a monthly overview of the devices/OS of the users that see my media files

Gives the number of views per month, per device:

- Period
- Device
- Views

9. Next steps

We have seen that there is clearly a need for GLAM-related statistics, for research purposes, but mainly to enable GLAMs to use Wikimedia Projects as a mature distribution channel for their collections.

We have seen that the Wikimedia Foundation is actively developing and supporting the tools necessary for the needs that are described in this document (Kraken/Limn). A data model per requirement has been proposed that can be created by Kraken and displayed by Limn.

The next steps are up to the analytics team of the Wikimedia Foundation to start gathering the information required (using Kraken) and work on a page (using Limn) to display that information. The GLAM toolset project will continually and regularly liaise with the Wikimedia Foundation to ensure this development is prioritised.