# Chapter 5 / Capítulo 5

SAP Superintelligence Series
ISSN: 3144-5624
ISBN (complete work): 978-9915-9851-0-7

Superintelligence Series Volume 5: Intelligent Visual Arts and Audiovisual Engineering: Generative AI, Computational Creativity and Superintelligence
ISBN (this volume): 978-9915-9966-5-3
DOI: 10.62486/978-9915-9966-3-9.Ch05
Pages: 47-57

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

# Growth, use and deletion of machine-generated images on Wikimedia Commons

## Crecimiento, uso y marcadores de borrado de imágenes generadas por máquina en Wikimedia Commons

**Matilde Bolaño-García**[1]

1 Universidad del Magdalena. Facultad de Ciencias de la Educación. Santa Marta, Colombia.

**ABSTRACT**

Introduction: Wikimedia Commons, the free media repository of Wikipedia, has absorbed a rising number of machine-generated images, and their place in the project is contested. Systematic evidence on their growth, use and deletion is lacking. Methods: A registry census was conducted of the Commons category tree for machine-generated images using the MediaWiki API. All file members of the category and its subcategories were enumerated, and their upload timestamps, global usage and membership in quality-flag categories were recorded on a single snapshot date. Results: The census identified 2,548 image files across 104 subcategories. Uploads grew from 4 files in 2019 to 164 in 2022, 324 in 2024, and 1,035 in 2026. Only 2.0% of files were used in at least one wiki page. Membership in quality-flag categories reached 0.7% for low-quality tagging and 0.9% for obvious artifacts. Conclusions: Machine-generated images on Commons grew rapidly but remain almost entirely unused, and only a small share are flagged for deletion-related concerns. The repository is absorbing generated content far faster than the project is integrating it.

Keywords: Wikimedia Commons; machine-generated images; digital commons; open knowledge; generative models; content governance

**RESUMEN**

Introducción: Wikimedia Commons, el repositorio de medios libres de Wikipedia, ha absorbido un número creciente de imágenes generadas por máquina, y su lugar en el proyecto es controvertido. Falta evidencia sistemática sobre su crecimiento, uso y borrado. Métodos: Se realizó un censo del registro del árbol de categorías de imágenes generadas por máquina en Commons mediante la API de MediaWiki. Se enumeraron todos los archivos de la categoría y sus subcategorías y se registraron sus fechas de subida, su uso global y su pertenencia a categorías de marcado de calidad en una única fecha. Resultados: El censo identificó 2548 archivos en 104 subcategorías. Las subidas crecieron de 4 archivos en 2019 a 164 en 2022, 324 en 2024 y 1035 en 2026. Solo el 2,0% de los archivos se usaba en al menos una página wiki. La pertenencia a categorías de marcado de calidad alcanzó 0,7% para el etiquetado de baja calidad y 0,9% para artefactos evidentes. Conclusiones: Las imágenes generadas por máquina en Commons crecieron con rapidez pero permanecen casi sin uso, y solo una pequeña fracción está marcada para borrado. El repositorio absorbe contenido generado mucho más rápido de lo que el proyecto lo integra.

Palabras clave: Wikimedia Commons; imágenes generadas por máquina; bienes comunes digitales; conocimiento abierto; modelos generativos; gobernanza de contenidos

## INTRODUCTION

Wikimedia Commons is the shared media repository behind Wikipedia, holding tens of millions of images, videos and documents under free licenses, and its value to the open knowledge ecosystem depends on two properties that pull against each other. It must be open enough to accept contributions from absolutely anyone, and it must be curated enough that what it hosts is actually useful to the encyclopedias and projects that draw on it.[1,2] The arrival of machine-generated images has stressed both properties at once. A contributor can now produce an image in seconds and upload it in a few seconds more, which floods the upload stream with content that no one has judged, while the community has never settled the question of whether such images belong in the repository at all.[3,4]

The debate is well documented in the community's own discussions. Wikipedia's editors spent three years working out how to govern language-model-generated text, arriving at a minimalist policy that accepted some uses while restricting others, and Commons now faces the parallel question for images, with the added complication that an image is harder to verify than a paragraph.[3,5] The stakes are genuinely concrete: a repository that accepts generated images without curation risks becoming a dumping ground, while one that rejects them outright forgoes illustrations that could genuinely serve articles.[6,7] The question also sits inside a longer scholarly conversation about what a commons is and how it stays healthy. Commons is a commons-based peer production project in the sense that Benkler and later theorists defined, and the defining question for such projects is whether the collective can govern the influx of contributions faster than those contributions erode the quality that made the commons worth having.(8-10)

Two further contexts frame the problem. The first is licensing. Commons exists to hold freely licensed media, and its usefulness depends on the license attached to every file, a point that generated images complicate because the legal status of their output, and of the models that produced it, remains unsettled in the jurisdictions that matter.(11-13) A generated image is, at the very minimum, a work whose authorship and copyright status are genuinely open questions, and a repository whose entire premise is clear licensing has to decide how to treat such works. The second is the wider rise of synthetic media. Generated images are one instance of a broader phenomenon whose societal risks, from deceptive content to the erosion of trust in photographs, are being studied across the media and forensics literatures, and Commons is a case in which those risks meet a community that has to make concrete, per-file decisions.(14-16) What has been missing from all of this is measurement. Claims about how many generated images Commons hosts, whether they are used, and how often they are challenged have been made on the basis of impressions rather than counts.

The choice of method matters here, because there is no single authoritative count of machine-generated images on Commons. They are not tagged with a machine-readable flag; they are identified only by the human-curated categories into which contributors have sorted them, and those categories were themselves created and reorganized over time.[17,18] A defensible estimate therefore requires a census of the category tree, not a search, and it requires care to distinguish when a file was uploaded from when it was categorized, because the two dates can differ by years and the distinction changes the entire growth story. This study reports such a census, chosen over the alternatives of a text search or a sample precisely because the category tree is the only complete, auditable register of the phenomenon. The objective was to quantify the growth, categorization, use and deletion-flagging of machine-generated images on Wikimedia Commons.

The contribution is deliberately narrow and descriptive. Rather than argue for or against generated images on Commons, the study measures what is actually there, and it does so in a way that can be re-run as the category tree continues to grow, so that the snapshot reported here can become the first point in a continuous time series. That is the kind of evidence the community's own discussions have been missing, and it is the kind that a method-focused study is best placed to supply.

## METHODS

### DESIGN AND DATA SOURCE

A cross-sectional registry census was conducted using the MediaWiki application programming interface of the Wikimedia Commons project itself. The register under study is the

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

category tree rooted at the category for machine-generated images, whose members are files and subcategories assigned by contributors. The MediaWiki API was chosen because it is the same interface through which the repository is maintained, which makes the census reproducible by any researcher, and because it exposes both the category structure and the per-file metadata needed for the analysis.

The unit of analysis is the individual file, and the population is defined by category membership: a file is in scope if, and only if, a contributor has placed it under the machine-generated image category, whether directly or through any depth of subcategories. This is a consequential definition, because it makes the census depend on the community's own classification rather than on any independent judgment about what counts as generated. It was chosen deliberately over a content-based definition, which would have required inspecting each image, because the category tree is the operational definition the community itself uses, and a census that mirrors the community's own categories is the one that can speak to the community's own debates. The cost of this choice, that uncategorized generated images are invisible, is a limitation reported in the discussion.

## ENUMERATION PROCEDURE

The category tree was traversed breadth-first. Starting from the root category, the API was queried for its members, distinguishing files from subcategories, and each subcategory was then visited in turn until the frontier was exhausted. Pagination followed the continuation token returned by the API, so that categories with more than five hundred members were enumerated completely rather than truncated at the first page. Every file was recorded with its title and its immediate parent category. The traversal recovered 2,548 files across 104 subcategories. The upload timestamp of each file was then retrieved in batches of fifty through the imageinfo endpoint, and the global usage of each file, meaning its transclusion in any page across any Wikimedia project, was retrieved in the same batches through the globalusage endpoint. Batching was required because the API returns per-file metadata only in groups, and the batch size of fifty is the API's own documented maximum for this operation. Two timestamps are therefore distinguished throughout: the upload timestamp, which records when the file was created, and the categorization timestamp, which records when it was added to the category tree, and the growth analysis uses the former.

## VARIABLES

The primary variables were the upload year of each file, its immediate parent category, whether it was used in at least one wiki page, and whether it belonged to any of three quality-flag categories. The quality-flag categories are those that the community uses to mark machine-generated images of concern: a category for low-quality generated content colloquially tagged as generated sludge, a category for images with obvious visual artifacts, and a category for images whose generated provenance has not been identified. Membership in these categories is a proxy for the deletion debate rather than a direct count of deletion nominations, because a formal deletion nomination is a page-level action recorded in a separate log; the proxy is reported because it is the attribute that is uniformly and machine-readably visible on the file itself. The distribution of files across the immediate parent categories summarizes how the community organizes the content.

The distinction between direct and transitive category membership deserves a note. A file counted in the census may belong to the root category directly, or it may belong to a subcategory several levels deep, and the traversal records whichever immediate parent the API reports. This means the categorization counts are of the file's immediate parent, not of its position in the full hierarchy, and a file deep in the tree is attributed to its local category rather than to the root. The choice is deliberate: the immediate parent is the category that a contributor actually assigned, and it is the level at which the community's organizational decisions are visible. Aggregating everything to the root would hide exactly the structure that the categorization analysis is meant to reveal.

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

## ANALYSIS

Because the census enumerates the entire register rather than a sample, the analysis is descriptive: absolute counts and proportions, with no inferential statistics and no confidence intervals. Growth was tabulated by upload year, and the top-level organization was summarized by the frequency of each immediate parent category. The use rate was computed as the proportion of files transcluded in at least one wiki page, and the deletion-flag rate as the proportion belonging to each quality-flag category. The entire pipeline was written as a re-runnable script that queries the API directly, and it was executed with rate-limit handling that respects the API's retry guidance. The count of files was verified against the category information returned by the API where available.

The single methodological decision that deserves the most emphasis is the separation of the two timestamps. The MediaWiki API exposes, for every category member, a timestamp that records when the page was added to that category, and this is easily mistaken for the file's upload date. The two diverge substantially here, because the category tree that organizes machine-generated images is itself recent, so many files were uploaded years before they were sorted into it. Using the categorization timestamp would compress the entire growth history into the last two years and erase the early files; using the upload timestamp, which required a second, batched query per file, recovers the true trajectory. This is the kind of subtle choice on which the entire growth narrative turns, and it is reported here rather than left implicit, in keeping with the methodological focus of the study.

## RESULTS

## GROWTH

The census identified 2,548 machine-generated image files spread across 104 subcategories of the root category (Table 1). Uploads began as a trickle, with 4 files in 2019, 2 in 2020, and 5 in 2021. The first substantial increase came in 2022, when 164 files were uploaded, coinciding with the public release of the first widely used text-to-image models. The count rose to 362 in 2023 and eased to 324 in 2024, before climbing to 652 in 2025 and 1,035 in the first ten months of 2026 (Figure 1). The 2024 dip is the notable irregularity in an otherwise monotone rise, and it is discussed below. In cumulative terms the corpus grew from 164 files at the end of 2022 to over two and a half thousand by 2026, roughly a fifteen-fold increase in under four years, which is the number that any policy discussion would need to keep firmly in view.

**Table 1. Machine-generated image files by upload year.**

| Year | Files uploaded |
|---|---|
| 2019 | 4 |
| 2020 | 2 |
| 2021 | 5 |
| 2022 | 164 |
| 2023 | 362 |
| 2024 | 324 |
| 2025 | 652 |
| 2026 | 1035 |

## CATEGORIZATION

The 104 subcategories show how the community actually sorts the content, and the largest of them reveal the character of what is being uploaded in practice (Table 2). The single largest category, with 357 files, is for images that include their generation prompt, a practice the community appears to encourage for transparency. The second largest, with 344 files, is for generated coats of arms, a niche that has produced a large and sustained output. The remainder spreads across diagrams, maps, headgear, science fiction illustrations, and landscapes, a distribution that mixes clearly illustrative material with ornamental and speculative content. The

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

concentration at the top of the distribution is itself informative: two categories account for more than a quarter of the corpus, and the rest fans out into a long tail of narrowly defined subject categories, which is what one would expect from a bottom-up, contributor-driven classification rather than from a designed taxonomy.

**Figure 1. Machine-generated image files uploaded to Wikimedia Commons by year.**

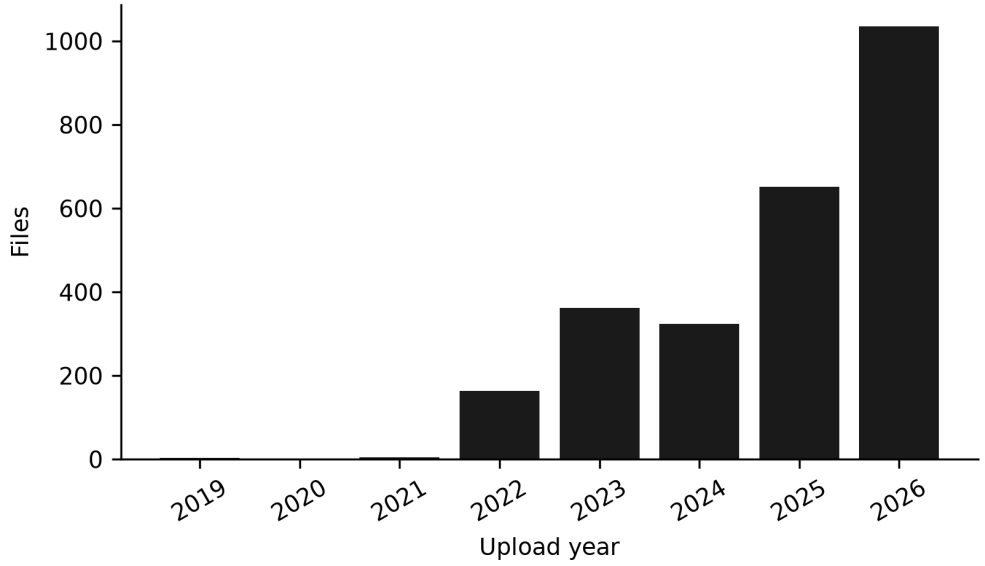



**Table 2. Largest immediate subcategories of machine-generated images by file count.**

| Category | Files |
|---|---|
| AI-generated images including prompts | 357 |
| AI-generated coats of arms | 344 |
| AI-generated diagrams | 99 |
| AI-generated images of maps | 98 |
| AI-generated images of headgear | 97 |
| AI-generated science fiction | 71 |
| AI-generated winterscapes with Wikipedia-esque globes | 71 |
| Monochrome AI-generated images | 64 |
| AI-generated computers | 59 |
| AI-generated images of objects | 59 |
| AI misgeneration | 58 |
| AI-generated landscapes | 56 |

## USE AND DELETION FLAGS

The most striking single result in the whole census is the use rate. Of the 2,548 files in the census, only 51 were used in at least one wiki page, a rate of 2.0% (Table 3). The remaining 2,497 files exist in the repository but appear in no article, gallery, or user page across any Wikimedia project. The quality-flag categories tell a quieter story. Low-quality tagging covered 17 files (0.7%), the obvious-artifacts flag covered 23 files (0.9%), and the unidentified-provenance flag covered 1 files (0.0%).

The two indicators pull in opposite directions, and their combination is precisely the point. The near-zero use rate says the corpus is not yet serving the encyclopedia, while the low flag rates say the community is not yet systematically rejecting it either. Between those two facts lies a large body of files that are neither used nor flagged, sitting in an indeterminate state that the census can name but not resolve. That indeterminate state, and its substantial size relative to the small minority that is either used or challenged, is the most accurate single description of where machine-generated images stand on Commons today.

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

**Table 3. Use and quality-flag indicators for machine-generated images.**

| Indicator | Files | % |
|---|---|---|
| Files in tree | 2548 | 100.0 |
| Used in at least one wiki page | 51 | 2.0 |
| ai_slop | 17 | 0.7 |
| artifacts | 23 | 0.9 |
| unidentified | 1 | 0.0 |

Note: Use means transclusion in at least one page of any Wikimedia project; quality flags are membership in the corresponding maintenance category.

## DISCUSSION

### MAIN FINDINGS

The census establishes three clear things about machine-generated images on Commons. Their number has grown rapidly, from a handful before 2022 to over one thousand files in a partial year, and the growth is recent enough that most of the register was created in the last two years. Their use is almost nil: two percent of files appear in a single wiki page, which means the repository is absorbing generated content far faster than the projects are consuming it. And their deletion-related flagging is small but nonzero, with a minority of files carrying the marks that the community applies when it judges a generated image to be poor or dubious. These three facts together describe a repository that is growing as an archive while not yet functioning as a supply of illustrations, and that state of affairs, more than any single number, is the headline result.

### INTERPRETATION

The growth curve matches the external history of generative image models closely enough to read as a faithful record of it. The 2022 surge follows the public release of the first open text-to-image models, and the continued rise through 2025 and 2026 tracks the spread of those tools, a pattern consistent with the broader observation that generative systems have reshaped how volunteers produce and share visual content.[18,19] The 2024 dip is the one notable deviation, and it admits two readings that the census alone cannot separate. It may reflect a genuine slowdown in uploads, or it may reflect a lag in categorization, because a file counts in this register only once a contributor has placed it under the machine-generated category tree, and a backlog of uncategorized uploads would depress the 2024 count relative to its neighbors. The distinction between upload and categorization timestamps, which the method was designed to keep separate, is therefore also the source of the one interpretive uncertainty in the growth series.

The scale of the corpus should be put in proportion. The two and a half thousand files counted here are a small fraction of Commons, which holds tens of millions of media files, so machine-generated images are not yet restructuring the repository as a whole.(1) What is notable is the direction and slope of the curve: a doubling from 2024 to 2025 and another near-doubling into 2026 means the category is growing faster than the repository's human review practices are likely to scale, and it is the slope, more than the absolute count, that carries the governance implication.

The use rate is the finding with the clearest implication. A free-media repository that accepts images which almost no project then uses has, in effect, become an archive of unused illustrations, and the near-zero use rate quantifies a concern that has otherwise been voiced only impressionistically.[2,10] The concentration of uploads in ornamental niches such as generated coats of arms, which number in the hundreds while contributing little to the encyclopedia, points the same way: a large share of the generated corpus is produced for its own sake rather than to fill an encyclopedic need.(20) This is not necessarily a defect, because Commons has always hosted material whose usefulness is prospective, but it changes what the repository is becoming.

The contrast with the human-uploaded corpus is instructive, even if the census does not measure it directly. Commons as a whole achieves a substantially higher use rate because most uploads are photographs and scans that were made to illustrate something specific, whereas a large part of the generated corpus appears to have been made to exercise a tool rather than

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

to meet a need.[1,2] If that reading is right, the near-zero use rate is not a temporary lag but a structural property of how the generated content is being produced, and it will not correct itself as the corpus matures. It would correct only if the projects began deliberately commissioning generated illustrations for their articles, which is a different activity from the one the upload stream currently reflects.(6)

The low deletion-flag rate should be read with care, and the method was chosen to make that care explicit. The three quality-flag categories capture only the files that contributors have already identified as problematic, and the flagging process itself is the bottleneck: a file can be poor without yet having been flagged, and a formal deletion nomination is a separate, heavier action that the census does not count directly.(4) The small observed proportions are therefore lower bounds on the true share of the corpus that the community would, on inspection, remove or reject. What they do show is that the visible, organized concern is concentrated in a small minority, which is consistent with a community that has not yet developed a systematic practice for reviewing generated images at the rate they arrive.(21)

The quality question also connects directly to a broader literature on how open-knowledge projects maintain verifiability and quality under sustained pressure. Wikipedia's content policies are built on a model of human review that generative tools disrupt, and the community has responded with both automated assistance and tighter norms, neither of which has yet settled the place of machine-produced media.[22-24] The near-zero use rate observed here is, in that light, a symptom rather than a cause: the projects that might consume generated images have not yet decided whether they should, and the unused corpus on Commons is the visible residue of that indecision.

The parallel with text is worth drawing out, because the two cases illustrate the same governance problem at different stages. Wikipedia settled on a restrictive-but-not-prohibitive policy for generated text only after years of debate, and its editors now treat generated prose as a thing to be used sparingly and disclosed when used.[3,5] Commons has not reached the equivalent settlement for images, and the census shows why the question is pressing: the images are already there, in the thousands, and every month of indecision adds more. The difference between the two cases is also instructive, because an image is harder to verify and easier to produce than a paragraph, which makes the image case the harder governance problem even though it has received less attention.(6)

## METHODOLOGICAL ASSESSMENT

The census method was chosen over two alternatives, and the results are conditioned on that choice. A text search over file descriptions would have recovered files whose descriptions mention generated content, but it would miss files described in other languages or not described at all, and it could not distinguish generation from mere mention. A random sample would have been cheaper but could not support the per-year counts that are the point of the growth analysis. The category-tree census is complete for what it defines, but it inherits the definition: a file counts only if a contributor has placed it, directly or transitively, under the machine-generated category. Files generated by machine but categorized only by their subject, with no provenance category, are invisible to this method, so the census is a lower bound on the true population. The strength of the method is that this boundary is stated rather than hidden, and the counts are reproducible by anyone who re-runs the same traversal.[10,25]

The robustness of the findings can be judged against the assumptions they rest on. The growth series is robust to the exact choice of root category in its broad shape, because any reasonable definition of machine-generated images would capture the same surge, but its precise per-year counts are sensitive to how the category tree has grown, which is why the categorization timestamp is reported separately rather than discarded. The use rate is robust in the opposite way: it would only fall if the census over-counted, and the census is a lower bound, so the true use rate is at most two percent and possibly lower. The deletion-flag rate is the least robust figure, because it depends on the community's flagging activity rather than on any property of the files, and it should be read as a snapshot of that activity rather than as a stable attribute of the corpus. These are the kinds of assumption-checks that a method-focused study owes its reader, and they are stated here so that the conclusions can be recalibrated by anyone who disagrees with the boundary choices.

There is also a reproducibility argument that favors the method over the alternatives. A text search depends on the query string the researcher happens to choose, and two researchers

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

with different queries will arrive at different counts; a census of a named category tree depends only on the name of the root category, which is a public, stable identifier. A sample depends on the sampling procedure and the seed. The census is the option that most cleanly separates the researcher's choices from the object being measured, which is why it is the right instrument for a question whose answer should not depend on who is asking it. This is the methodological principle behind the design, and it is stated here because the study's conclusions inherit their force from it.

## GENERALIZABILITY

The findings generalize in a qualified and carefully bounded sense. The specific counts are a snapshot of one repository at one date, and they will change, but the structural pattern, a steep upload curve meeting a flat use rate, is unlikely to be unique to Commons. Any open repository that accepts freely licensed media will face the same asymmetry once generation becomes cheap, because the cost of producing a generated image is near zero while the cost of integrating it into an article remains a human judgment. The method, a category-tree census with separated timestamps, is the more portable contribution, and it can be applied to any wiki that uses the same MediaWiki platform or to any repository with a comparable human-curated category structure. What the study cannot generalize is the community's particular response, because that response is shaped by Commons's history and norms rather than by any feature of the phenomenon.[25,26]

## IMPLICATIONS

For the Commons community, the results make the policy question quantitative. The repository is not being overwhelmed by deletion debates, but neither is it integrating what it accepts, and the gap between the upload curve and the flat use rate is where any future governance decision will have to land.(3) A policy that accepts generated images uncritically would codify an archive that no one reads, while a policy that curates them would have to scale a review process to an upload stream that has already passed a thousand files in a year.(6) The distinction between upload and categorization timestamps, which proved material here, is also a caution for any future measurement of Commons, where the two dates are easy to conflate and their conflation would distort any growth study.(18)

The finding also suggests a division of attention for the community. The deletion-related flagging, while real, is a second-order concern compared with the use gap, and a community that spends its energy adjudicating individual deletions without addressing the integration problem will be managing symptoms. The more consequential question is whether the projects that Commons serves actually want generated illustrations, and if so, under what conditions, and the near-zero use rate suggests that question has not yet been answered in the affirmative. A deliberate decision either way, grounded in the counts reported here, would be an improvement over the current state, in which the repository absorbs content by default and the projects ignore it by default.

The measurement itself is part of the answer, in a modest way. A count is not a policy, but it is a precondition for one, and a community that can cite how many files it hosts, how many it uses, and how many it has flagged is in a better position to decide than one arguing from anecdote. The census reported here provides those numbers for the present, and because it is reproducible and cheap, it can be repeated as the category grows, giving the community a running measure of a problem that is otherwise easy to misjudge by feel. That, ultimately, is the practical contribution of a descriptive study of this kind.

For the broader open knowledge ecosystem, the finding is a case study in how a commons absorbs a new, cheap, high-volume form of contribution. The same dynamics that produced a near-unused corpus of generated images are likely to recur wherever generative tools meet open repositories, and the measurement approach here, a category-tree census with separated timestamps, is portable to those settings.[26,27] The result also bears on the recurrent worry that generated content will crowd out human contribution; on Commons, so far, the generated corpus has accumulated alongside rather than replaced, but its scale relative to its use suggests the crowding question is still open.[5,28] The synthetic-media governance literature reaches the same conclusion from the other direction, arguing that the risks of generated content are best

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

managed by institutions and norms rather than by blanket prohibition, which is precisely the terrain on which Commons is now making its decisions.[29,30]

## CONCLUSIONS

A census of the Wikimedia Commons category tree identified a machine-generated image corpus whose uploads grew from a handful before 2022 to well over one thousand files in a single partial year, with a dip in 2024 that the method cannot fully resolve. The corpus was organized into over one hundred subcategories, the largest of which held generated prompts and generated coats of arms. Use was near zero, at two percent of files appearing in any wiki page, while deletion-related flagging was confined to a small minority. The repository is absorbing machine-generated images far faster than the projects are using them, a gap that any governance decision will have to address explicitly.

https://doi.org/10.62486/978-9915-9966-3-9.Ch05

## REFERENCES

Kelly EJ. Reuse of Wikimedia Commons Cultural Heritage Images on the Wider Web. EBLIP. 2019;14(3):28-51. doi: 10.18438/eblip29575
Yu Y, McDonald DW. “Why do you need 400 photographs of 400 different Lockheed Constellation?”: Value Expressions by Contributors and Users of Wikimedia Commons. Proc ACM Hum-Comput Interact. 2023;7(CSCW2):1-34. doi: 10.1145/3610094
Froneman W. Failed comprehensiveness, successful minimalism: Wikipedia’s 3-year struggle to govern AI-generated content (2022–2025). AI & Soc. 2026;41(7):7045-7061. doi: 10.1007/s00146-026-03046-1
Fisher SA, Howard JW, Kira B. Moderating Synthetic Content: the Challenge of Generative AI. Philos Technol. 2024;37(4):133. doi: 10.1007/s13347-024-00818-9
Wagner C, Jiang L. Death by AI: Will large language models diminish Wikipedia?. Asso for Info Science & Tech. 2025;76(5):743-751. doi: 10.1002/asi.24975
Reeves N, Simperl E. Machines in the Margins: A Systematic Review of Automated Content Generation for Wikipedia. Proc ACM Hum-Comput Interact. 2025;9(7):1-30. doi: 10.1145/3757626
Yu Y, McDonald DW. Computer-Aided Tagging on Wikimedia Commons: Designing for Human–AI Collaboration in Open Knowledge Work. Proc ACM Hum-Comput Interact. 2026;10(6):1-29. doi: 10.1145/3816912
Papadimitropoulos V. Commons-Based Peer Production in the Work of Yochai Benkler. tripleC. 2018;16(2):835-856. doi: 10.31269/triplec.v16i2.1009
Aryan V, Bertling J, Liedtke C. Topology, typology, and dynamics of commons-based peer production: On platforms, actors, and innovation in the maker movement. CreatInnov Manage. 2021;30(1):63-79. doi: 10.1111/caim.12392
Jullien N, Roudaut K. Are Digital Knowledge Commons Real Commons?. International Journal of the Commons. 2026;20(1):468-483. doi: 10.5334/ijc.1691
Simmonds T. Common Knowledge? The Rise of Creative Commons Licensing. Legal Information Management. 2010;10(3):162-165. doi: 10.1017/s1472669610000642
Hare C. Copyfight: Creative Commons, Open Licensing, Bringing Information to the People (and letting them use it). DJIM. 2010;3(1). doi: 10.5931/djim.v3i1.24
Quintais JP. Generative AI, copyright and the AI Act. Computer Law & Security Review. 2025;56:106107. doi: 10.1016/j.clsr.2025.106107
Amerini I, Barni M, Battiato S, Bestagini P, Boato G, Bruni V, et al. Deepfake Media Forensics: Status and Future Challenges. J Imaging. 2025;11(3):73. doi: 10.3390/jimaging11030073
Pawelec M. Decent deepfakes? Professional deepfake developers’ ethical considerations and their governance potential. AI Ethics. 2025;5(3):2641-2666. doi: 10.1007/s43681-024-00542-2
Glick J. Deepfake Satire and the Possibilities of Synthetic Media. Afterimage. 2023;50(3):81-107. doi: 10.1525/aft.2023.50.3.81
Yu Y, McDonald DW. Unpacking Stitching between Wikipedia and Wikimedia Commons: Barriers to Cross-Platform Collaboration. Proc ACM Hum-Comput Interact. 2022;6(CSCW2):1-35. doi: 10.1145/3555766
Jankowski S. Wikimedia research methodologies: a bibliometric history of peer production and AI (2005–2025). AI & Soc. 2026. doi: 10.1007/s00146-026-03359-1
Luth E. The Use of Wikipedia, Wikimedia, and Open Access Content for Artificial Intelligence and Text and Data Mining. SIPLR. 2024;7(2):37-43. doi: 10.53292/33313cc8.be33e111
Kelly EJ. Digital Cultural Heritage and Wikimedia Commons Licenses:. JCEL. 2019;3(3):1-25. doi: 10.17161/jcel.v3i3.9771
Zhao A, Naaman M, Wu L. Rule Development in Online Communities Amidst Growth and Generative AI. Proc ACM Hum-Comput Interact. 2026;10(6):1-23. doi: 10.1145/3816993
Niederer S, van Dijck J. Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. New Media & Society. 2010;12(8):1368-1387. doi: 10.1177/1461444810365297
Ferretti E, Cagnina L, Paiz V, Donne SD, Zacagnini R, Errecalde M. Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws. Information Processing & Management. 2018;54(6):1169-1181. doi: 10.1016/j.ipm.2018.08.003
Petroni F, Broscheit S, Piktus A, Lewis P, Izacard G, Hosseini L, et al. Improving Wikipedia verifiability with AI. Nat Mach Intell. 2023;5(10):1142-1148. doi: 10.1038/s42256-023-00726-1
Dulong de Rosnay M, Stalder F. Digital commons. Internet Policy Review. 2020;9(4):15 p. doi: 10.14763/2020.4.1530
Fuchs C. The Digital Commons and the Digital Public Sphere: How to Advance Digital Democracy Today. Westminster Papers in Communication and Culture. 2021;16(1). doi: 10.16997/wpcc.917
Taeihagh A. Governance of Generative AI. Policy and Society. 2025;44(1):1-22. doi: 10.1093/polsoc/puaf001
Pawelec M. Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. DISO. 2022;1(2):19. doi: 10.1007/s44206-022-00010-6
Alexis Celia, Brittany Donnie, Masaki Hasegawa, Maelle Renard. Deepfakes, Synthetic Deception, and Responsible Deployment of Generative AI. J Artif Intell Eng Consort. 2026;2(1):50-59. doi: 10.66238/jaiec58
Montague-Hellen B. Empowering knowledge through AI: open scholarship proactively supporting well trained generative AI. Insights the UKSG journal. 2024;37:10. doi: 10.1629/uksg.649