October 9, 2015

United States Copyright Office
Library of Congress
Docket No. 2015-3

Re: Mass Digitization Pilot Program; Request for Comments

The Wikimedia Foundation respectfully submits these comments in response to the Copyright Office's notice of inquiry, published June 9, 2015, regarding a mass digitization and extended collective licensing (ECL) legal framework and pilot program.[1] The Wikimedia Foundation is a non-profit and charitable organization that operates Wikipedia, the Internet's largest and most popular general reference work, and its sister projects, including the mass digitization projects Wikimedia Commons and Wikisource.

Executive Summary

In creating a pilot program for mass digitization, the Copyright Office has an admirable goal: facilitating the creation of and access to collections of digital works for education and research purposes. However, aspects of the pilot program and the Office's general approach to mass digitization are misguided. As a result, the pilot program and any legal framework for mass digitization that comes out of it will not achieve the Office's goal as effectively as a mass digitization initiative could and should.

The pilot program addresses only a narrow subset of mass digitization projects. It builds off the example of Google Books, using Google Books as a model for both who is doing the digitizing (a single institution) and what is being digitized (literary works). There is a universe of mass digitization projects beyond the Google Books model, including projects with digitizers spread all over the world and projects that are digitizing sound and video recordings. The Copyright Office's mass digitization

[1] Mass Digitization Pilot Program; Request for Comments, 80 FR 32614.

framework should facilitate these sorts of projects by broadening its conception of mass digitization and providing greater protection for digitizers. Only in so doing can the Office plan for the mass digitization of the future.

The proposed pilot program places burdensome restrictions on mass digitization projects. The ECL regime does nothing to alleviate the legal risks to digitizing institutions whose work is authorized by fair use. The pilot program anticipates imposing severe limitations on which users can access digitized collections (based on affiliation with or physical presence at the digitizing institution), how they may access the collections (with "security measures"), and how they may use works in the collections (with vague and unnecessary "noncommercial" limitations). In addition, the Copyright Office does not seem to have considered the potential for a mass digitization framework to help or impede mass digitization of public domain works.

If the Copyright Office plans to go forward with its mass digitization pilot program, we encourage it to reformulate the program so that it encompasses the broader landscape of mass digitization projects and better provides for all education and research uses for mass digitization.

Introduction

The Wikimedia Foundation's mission is to "empower and engage people around the world to collect and develop educational content under a free license[2] or in the public domain, and to disseminate it effectively and globally." As the largest part of that mission, we operate over a dozen free knowledge projects, including Wikipedia, Wikisource, and Wikimedia Commons.

In short, Wikipedia and its sister projects aim to allow users all over the world to access the sum of all knowledge at the click of a button. Wikipedia gives access to knowledge in the form of an encyclopedia, with over 35 million articles across its approximately 291 language versions. Wikimedia Commons gives access to knowledge in the form of a central repository of media content (including images, sound recordings, and video) for all content consumers. Most images that illustrate

[2] Free licenses makes works available to the public for use, reuse, and modification for free in perpetuity. The most common free license used on the Wikimedia projects is the Creative Commons Attribution-Sharealike license (usually version 3.0). That license allows anyone to use, reuse, and modify the works available on the projects (including the text of Wikipedia) in any way they like for any purpose they like as long as they (1) provide attribution to the author of the original work and (2) license any derivative work they make under the same or a compatible free license. For more information about what we mean by "free," see Definition of Free Cultural Works.

Wikipedia articles are hosted on Wikimedia Commons. Wikimedia Commons is a mass digitization project: works are scanned, copied, or created; uploaded; and collected in one place. Wikisource gives access to knowledge in the form of "a free library of source texts which are in the public domain or legally available for free redistribution." Wikisource contributors upload books and other writings then translate and annotate those books and writings to make them accessible to a wider audience.

Because the Wikimedia projects represent already one of the world's largest collaborative mass digitization projects, we are interested and invested in the Copyright Office's pilot program. A mass digitization and ECL framework has the potential to affect the Wikimedia projects. Our projects consist almost entirely of public domain and freely licensed works, and, for that reason, we do not anticipate that any version of ECL would apply to our projects directly. That said, a mass digitization framework could—depending on its implementation—either help or harm the projects. Accordingly, we are submitting these views on the proposed pilot program.

I. The Copyright Office should ensure it has a robust understanding of "mass digitization."

The Copyright Office must be fully aware of the purposes of mass digitization projects and develop the legal framework for mass digitization in light of those purposes. The Office indicates that the pilot program is meant "to facilitate the work of those who wish to digitize and provide full access to certain collections of books, photographs, or other materials for nonprofit educational or research purposes."[3] The design of the pilot program does not reflect the range of mass digitization projects that are consistent with that purpose, and the purpose itself does not seem to take into account the broad possibilities for mass digitization.

The Copyright Office has planned its pilot program with Google Books as the model mass digitization project. That narrow focus fails to consider mass digitization projects with other structures, such as projects where the digitizers are not the same as the hosts of the digital collections. A mass digitization framework that works for Google Books will not necessarily work for these differently-structured projects.

The Copyright Office acknowledges the possibilities for mass digitization to benefit education and research, but its limitation of the mass digitization framework to published books and photographs severely restricts those possibilities. Mass

[3] Mass Digitization Pilot Program; Request for Comments, 80 FR 32615.

digitization of sound recordings, video recordings, and some unpublished works is both possible and necessary for the realization of the full range of benefits of mass digitization to education and research.

If the Copyright Office truly wants to plan for the mass digitization of the future, it should design a legal framework for mass digitization that facilitates making a collection of all kinds of digital works available to the world, no matter how the collection is assembled.

A. Mass digitization is more than just Google Books.

The Copyright Office's June 2015 report, *Orphan Works and Mass Digitization* (*OWMD* report), devotes considerable attention to Google Books as an example of a mass digitization project and relatively little attention to other example projects.[4] Excessive focus on the example and precedent of Google Books could lead to the creation of a legal framework that fails to account for the breadth of potential for mass digitization. The Google Books project consisted of one company scanning a collection of print materials and making the digital versions available online. The Copyright Office's mass digitization framework attempts to make such a program easier with ECL, which would allow a digitizing institution following the Google Books model to pay a licensing fee for the whole collection instead of getting a license for each work in the collection. However, not all mass digitization projects follow the Google Books model, and the Copyright Office's mass digitization framework fails to address these differently-structured projects.

Distributed mass digitization projects like Wikisource and Wikimedia Commons provide examples of a mass digitization structure that differs from Google Books and is left out of the Copyright Office's framework. Rather than a project where Wikimedia is digitizing a library of books it possesses, Wikisource is an example of a mass digitization project where the digitizers and the materials they are digitizing are spread all over the world. Wikisource contributors, which include individuals as well as libraries and other organizations, digitize books and other writings independently then upload them to Wikisource.[5] Wikimedia Commons operates on a similar model, with contributors uploading images , sound recordings, and video that they have independently digitized, created, or found elsewhere on the Internet.

[4] *See* United States Copyright Office, Orphan Works and Mass Digitization 72–75 (2015).
[5] For an example, see *infra* Part II.D.

The structure of these projects allows for sharing, collaboration, and greater democratization of knowledge. However the Copyright Office's mass digitization framework does not account for these projects. The framework assumes that there is a single digitizer and that the digitizer is the same as the institution that is making the digital collection available. Neither assumption holds true for projects like Wikisource and Wikimedia Commons, and such projects do not benefit from the mass digitization framework as a result.

If a project with a distributed digitization structure like Wikisource and Wikimedia Commons were to use ECL,[6] it is not clear that the license would protect the individual digitizers and uploaders throughout the country and the world from copyright liability. If that project chose not to use ECL and instead rely on fair use as a protection from copyright liability, it risks attracting an expensive lawsuit.[7] If the project were to use ECL, but the ECL regime included licensing terms that would restrict access through geographic limitations or "security measures", the project may have difficulty attracting contributors—people are less likely to want to contribute if their contributions are only available and useful to a few (possibly not even including themselves). These problems are magnified if the mass digitization framework ever *requires* projects to use ECL.

We understand that the Copyright Office is proposing only a pilot program, and such a program cannot cover all possible mass digitization projects. However, having an overly narrow scope now will ultimately limit the program's usefulness and could lead to unintended consequences if the program were to apply to projects that were not considered in the program's design.

B. The proposal leaves out important categories of work.

The *OWMD* report and notice of inquiry's limitation of consideration for mass digitization to narrow categories of works leaves out vast swaths of valuable cultural and historical artifacts. The Copyright Office is proposing ECL only for literary works, embedded pictorial and graphic works, and photographs. Significantly, those categories exclude sound and video recordings (such as musical works and motion pictures). Ignoring such works in discussions of mass digitization is a mistake. As the Library of Congress is aware, the preservation of these works is a significant problem. Mass digitization efforts could be an important step in the preservation process. The promise that collections of digitized (and already digital) works could be made available

[6] Not that Wikisource or Wikimedia Commons would use ECL themselves, as they are collections of public domain and freely licensed works.
[7] For more on fair use, see *infra* Part II.A.

to the public would encourage their digitization. Sound and video recordings are an important part of our cultural and technological history, and the people of the world are entitled to document and access that history for educational and research purposes.

The *OWMD* report also recommends its mass digitization framework apply only to published works, and not unpublished ones. While the rationale of not wanting to interfere with an author's "right of first publication"[8] is sound, there are categories of unpublished works that are still protected by copyright but whose digitization would have great value, at least to scholars. Some examples are early versions of creative works, such as drafts, sketches, and demos, and letters and other correspondence. A mass digitization framework could respect the "right of first publication" while still allowing for some of unpublished works to be made available in digital collections, such as by limiting mass digitization of unpublished works to ones by deceased authors.

C. The proposal should be more forward-thinking.

The Copyright Office's proposed mass digitization framework does not fully embrace the mass digitization of the future. The Office is focused on individual institutions acting independently to digitize discrete collections, and it is entertaining the possibility of restricting access to digitized collections to "affiliates... of the digitizing institution," and "only through onsite computer terminals".[9] The notion that a person would need to be in a specific physical location in order to access a set of digital files is archaic in the modern age of constant, high-bandwidth, global Internet connectivity. The reason for collections to be digitized is so people can *access* them; that access combined with human ingenuity will create new ways for old works to be used and reused. A siloed, institution-based system of mass digitization and ECL makes assumptions about its users that do not allow for change and innovation. A siloed, institution-based system of mass digitization is not the future.

II. Responses to specific questions from the notice of inquiry.

Below are answers to the questions from the notice of inquiry that we believe are particularly relevant to the Wikimedia Foundation and the Wikimedia projects.

---

[8] Orphan Works and Mass Digitization at 85. The "right of first publication" allows authors to decide if and when their work is published. It is not a right explicitly enshrined in the Copyright Act, but it is "reflected in several provisions" as well as "incorporated in the Berne Convention." *Id.* at 85 n.346.
[9] Mass Digitization Pilot Program; Request for Comments, 80 FR 32615.

A. (Question 1) What are some examples of projects for which ECL would be appropriate?

Here are some examples of mass digitization projects that any legal framework for mass digitization, though not necessarily ECL, should facilitate:

- An archive, library, gallery, museum, or educational institution making available online a collection it has curated.
  - Note that several categories of works—commercially-available, not commercially-available, orphaned, or in the public domain—could all be present in one such collection.
  - A more specific example is "Picturing Canada," a collection of public domain photographs from Canada that The British Library and an organization of Wikimedians in the UK digitized and uploaded to Wikimedia Commons.
- The collection of digitized books and other materials to create open educational courses.
  - Existing such projects include Wikiversity, edX, and OpenCourseWare initiatives (such as the one at MIT).

We are unconvinced that ECL is an appropriate framework for these mass digitization projects. Even though many of these projects would qualify as fair uses and may involve a significant amount of public domain or freely licensed works, we are concerned that the projects will obtain ECL licenses unnecessarily simply to avoid being sued for copyright infringement. The Google Books cases demonstrate how expensive copyright litigation can be, even over a fair use, and the recent district court case over the copyright to the "Happy Birthday" lyrics demonstrates that many would rather pay licensing fees than risk litigation, even if the entity they are paying is not actually the copyright holder.[10]

If the Copyright Office is to propose statutory changes that would benefit education- and research-motivated mass digitization projects, clarification and strengthening of fair use protection would be a better fit than ECL. If the Copyright Act were to specify that education- and research-motivated mass digitization projects qualify as fair use, it would reassure digitizing institutions (and individual digitizers) that their projects will not get them sued. Explicit statutory authorization of mass digitization could also be

[10] *See* Marya v. Warner/Chappell Music, Inc., No. CV 13-4460-GHK, 2015 WL 5568497 (C.D. Cal. Sept. 22, 2015).

added to the DMCA safe harbor provisions,[11] as an alternative or complement to fair use[12] clarification.

The examples in this section are the ideal sort of mass digitization projects: ones that produce a clear benefit to society. Any mass digitization framework should encourage projects like these and make them easier to do by reducing their legal risks.

B. (Question 1(b)) Which people should be able to access the collections?

Mass digitization projects have the potential to be powerful tools for increasing access to knowledge. Wikimedia believes strongly in making knowledge available for free to everyone, for any purpose. Access to knowledge is a key economic driver that leads to new creation and entrepreneurial innovation. For this reason, we oppose any proposal that would limit access to digital collections to scholars, students, library users, or other limited groups. Digitizing institutions should be allowed to provide access to the general public, not merely their affiliates. Additionally, it is critical that digitizing institutions be permitted to offer remote access to a collection, and not be restricted to providing access only through on-site computer terminals.

By not limiting access, curators of digitized collections can serve those who are most in need of educational resources. In the United States, people living in rural areas do not have the same access to print materials as their urban counterparts and would not be able to effectively access knowledge if it were limited to on-site computer terminals. The digitization of books, images, and sound and video recordings has the potential to democratize access to knowledge across the United States and the world; access to digitized materials should not be limited to scholars, employees of digitizing institutions, or people who are able to travel to use the digitized resources at an on-site computer terminal.

With the Internet, mass digitization projects can also help spread access to knowledge around the world—notably in the Global South, where people may not have access to libraries, archives, or printed materials and would not be able to use on-site resources at digitizing institutions. Mass digitization projects stand to have the greatest impact on populations that are underserved by more traditional educational models, but only if they are not burdened with restrictions that prohibit access to digital collections for anyone but a limited group of users.

[11] 17 U.S.C. § 512.
[12] 17 U.S.C. § 107.

C. (Question 1(c)) What technical security measures should be required?

When deciding what digital security measures to require ECL licensees to implement it is particularly important to remember that the purpose of mass digitization projects is to provide access to the digitized works. The more people can access and use a collection, the more valuable it is. We believe strongly in open access and the creative use of existing content. Wikipedia and its sister projects are collaborative platforms with pages that anyone can edit or augment with their own knowledge and work. Millions of people have added to and edited Wikipedia over the past nearly fifteen years, working together to make a resource far greater than anything a single individual could create. True access to works entails the ability not just to view them but to use and reuse them—that is how new, transformative creative projects are made possible.

No "security" requirement, also known as DRM, should prevent legally authorized uses of works in digital collections. That includes fair use, any use of public domain materials, and uses of freely licensed works that comply with their licenses. The notice of inquiry specifies that the ECL-required "digital security measures" would only "prevent *unauthorized* reproduction, distribution, or display of the licensed works."[13] The security measures should not go beyond that limitation by putting up barriers to fair use of copyrighted works. No security measures should apply to freely licensed or public domain works, even if they are in a collection that also includes copyrighted works. These works belong to the commons, and it is not for the Copyright Office or digitizing institutions to restrict their availability, use, or reuse.

As part of allowing for open access and reuse, ECL should not be imposed on collections that do not wish to use it. Existing mass digitization projects that choose to freely license their content, like Wikimedia Commons and Wikisource, should not be forced to adopt ECL. No new mass digitization project that likewise wants to make its collection available under Creative Commons or similar licenses should be forced to use ECL instead.

The point of mass digitization is for people to have access to the works in the digitized collections. ECL should not undermine that access by imposing overly restrictive security measures on those collections, and open access collections should not be hamstrung by overzealous application of ECL.

[13] Mass Digitization Pilot Program; Request for Comments, 80 FR 32615 (emphasis added).

D. (Question 5) Other Issues

Mass digitization policy does not only affect copyrighted works. The expansion of mass digitization initiatives may also make available works in the public domain that would otherwise be more difficult to learn of or access. Any licensing scheme should account for the fact that mass digitization has the potential to play a socially useful role of preserving and widely distributing works that are not under copyright. Wikisource is an example of a public domain-dependent mass digitization project. It currently receives tens of millions of pageviews per month and has hundreds of active contributors. Those contributors have formed partnerships to digitize institutions' collections of print materials and make them available on Wikisource. Recently, for example, the Bibliothèque et Archives nationales du Québec worked with Wikisource contributors to add dozens of documents and books to the project. An expansion of mass digitization efforts has the potential to encourage these sorts of partnerships, broadening Wikisource's scope and increasing its value as an open resource.

Digitized collections' licenses should not include restrictive reuse provisions, particularly for public domain content. Only allowing "noncommercial" reuse is one such restrictive provision: it prevents many forms of reuse that are beneficial for education and research, including use on the Wikimedia projects. Because Wikimedians support the freedom to use, share, and remix content as broadly as possible, uninhibited by the opaque limitations imposed by a "noncommercial" restriction, there is no such a restriction in the licensing of the Wikimedia projects. Because the projects license their content without a "noncommercial" restriction, works distributed with "noncommercial" license restrictions generally cannot be used on Wikimedia projects.

In addition, the traditional commercial/noncommercial distinction does not make sense in a digital context. The notice of inquiry specifies that the pilot program would contain restrictions prohibiting uses that involve "direct or indirect commercial advantage", but the scope of what could qualify as an "indirect commercial advantage" is vague.[14] As a result, an ECL scheme that places restrictions on commercial uses makes works unavailable for many educational and research purposes. For example, sharing a work on social media networks could be considered commercial reuse depending on the sites' policies, or a professor whose reputation benefits from a particular use of digitized works could be understood to reap an indirect commercial advantage. If the

[14] *Id.*

pilot program contains a “noncommercial” restriction on the use of digitized content in addition to its “educational and research purposes” restriction, it will only cause confusion and eliminate many beneficial uses for mass digitization.[15]

Conclusion

We urge the Copyright Office to reflect on its proposed mass digitization framework in light of the above points. The framework’s current narrow focus limits its ability to facilitate mass digitization projects that have educational and research purposes. If it is to truly accommodate current and future mass digitization projects, the framework should allow for projects with distributed digitizers by providing those digitizers with greater legal protection. It should apply to sound recordings, video recordings, and some unpublished works as well as published literary works and photographs.

If the framework is to truly facilitate education and research it should strengthen fair use protections for mass digitization projects, it should not restrict access to small subsets of potential users, it should not impose overly restrictive DRM “security measures”, and it should encourage public domain mass digitization projects. With the incorporation of these considerations, a mass digitization framework has the potential to increase access to knowledge in a transformative way.

Sincerely,
Wikimedia Foundation

[15] *Id.*