UM Library/Google Digitization Partnership FAQ, August 2005

The University of Michigan and Google, Inc. have entered into a partnership to digitize the entire print collection of the University Library. The digitized collection will be searchable by Google, and the University Library will receive and own a copy of all images to integrate into new and existing UM Library user services.   This FAQ addresses questions in the following areas:

The UM-Google Project (aka MDP)

Q. 1: What is the UM–Google project?
A: The UM-Google project is a partnership between UM and Google that will make the seven million UM University Library volumes searchable via the Google search engine, and open the way to universal access to information. Google will digitize our library collection and make the items accessible through the Google site. The University Library will also receive and own a high quality digital copy of the collection to use for its own purposes.

Q. 2: Why does UM sometimes refer to the project as "MDP"?
A: Prior to the announcement of the project, Michigan and Google decided to refer to the work as the Michigan Digitization Project, or MDP. Because the University of Michigan's work with Google encompasses a number of activities and Google products (e.g., Google Scholar), we often refer to the digitization project as MDP.

Q. 3: How long will the project take?
A: Estimating how long the project will take is difficult, but we are currently planning for approximately six years of scanning.

Q. 4: How does this fit in with Google Print and Google Scholar?
A: Our project is part of the Google Print program to make books and other offline material searchable online. Google Print also includes books from publishers. Google Scholar is a parallel, yet separate project. (see http://scholar.google.com/)

Michigan's decision to work with Google
Q. 5: Why did Michigan decide to do enter into this partnership with Google?

A: The project with Google is core to our mission as a great public university to advance knowledge—on campus and beyond. By joining this partnership that makes our library holdings searchable through Google, UM serves as an agent in an initiative that radically increases the availability of information to the public. The University embraces this project as a means to make information available as broadly and conveniently as possible. Moreover, the UM Library embarked on this ground-breaking partnership for a number of very compelling reasons:

- We believe that, beyond providing basic access to library collections, this activity is critically transformative, enabling the University Library to build on and reconceive vital library services for the new millennium.
- This work will create new ways for users to search and access library content, opening up our collections to our own users and to users throughout the world.
- Although we have engaged in large-scale, preservation-based conversion of materials in the Library's collection for several years, and have been a leader in digital preservation efforts among research libraries, we know that only through partnerships of this sort can conversion of this scale be achieved. Our program is strong, and we have been able to digitize approximately 5,000 volumes/year; nevertheless, at this rate, it would take us more than a thousand years to digitize our entire collection.

Q. 6: How is this project advancing the mission of the University of Michigan as a great public university?
A: It is the very essence of the mission of a great public university to capture and preserve the sum of human knowledge, and to make that knowledge widely available for the purposes of teaching, learning and scholarship. We believe this project is perfectly aligned with the mission and values of the University of Michigan.

Collections to be converted, and materials handling

Q. 7: What collections in the library will be digitized?
A: Most of the University Library's bound print collections will be digitized (see **Question 10** below for exceptions), beginning with all volumes in the Buhr Shelving Facility.

Q. 8: How many volumes will be digitized?
A: Nearly seven million volumes will be scanned by the time the project is finished.

Q. 9: In what order will the different libraries be scanned, and will the project include new acquisitions?
A: A timetable and strategy for digitizing volumes in locations other than Buhr will be developed over time. We are currently focusing on the 2.5 million volumes in Buhr; consequently, newly acquired materials are not factored into the conversion process. As we move into other libraries, we will formulate strategies for taking new acquisitions into account.

Q. 10: What materials have been excluded?

A: Some materials are excluded or their conversion is currently on hold:
1. Special Collections is not included in the project at this time, but we hope to include materials from Special Collections later in the project.
2. Because this is a conversion project, electronic resources (e.g., those for which we pay a subscription access fee) are excluded.
3. Extremely large materials (e.g., folio format) and unbound materials are not included at this time.

Q. 11: Will the books and journals be disbound?
A: No, Google has developed technology that makes large-scale scanning without disbinding library materials a practical reality. Volumes will be returned to the shelf as soon as they are scanned.

Q. 12: How long will materials be unavailable while they are being scanned?
A: Books and journals will be unavailable for a very brief time, just a few days (rather than weeks or months).

Q. 13: Will the books still be part of the UM's collection?
A: Yes. After they are scanned, books are returned to the shelf, intact.

Q. 14: Will we be digitizing only books in the public domain, or in-copyright materials also?
A: All books will be scanned, but the way that they are made available to users will differ depending on their copyright status. For information on how these two different types of materials can be accessed, see **Question 23**.

Technology issues

Q. 15: Can we see the scanning facility?
A: Google does not permit access to the scanning facility. Because of the proprietary hardware and software in use, Google prefers that the processes and equipment not be made public.

Q. 16: What file format does Google use or provide to Michigan?
A: The capture methods used by Google produce extremely high-quality raw image files. Although Google is likely to use these files to create JPEG files for their online service, Michigan has chosen to receive files that that conform to library community standards (e.g., see http://www.diglib.org/standards/bmarkfin.pdf) and are consistent in format with those used in our preservation reformatting activities. Specifically:
- Most pages (i.e., those that consist of print without illustrations) are delivered to Michigan as 600dpi TIFF images using ITU G4 compression.
- Occasionally, pages include significant illustrations; these are provided to Michigan as 300dpi JPEG2000 images.
- OCR (performed by Google) is provided with each page.

Legal issues

Q. 17: Does our agreement with Google comply with copyright law?
A: Yes, to the best of our understanding, our agreement is in compliance with copyright law. Copyright is not a completely predictable area of the law, however. Because the law is designed to balance the needs of owners, publishers, and users, it is often difficult to predict how a court will decide any given case. Thus, new technology and business models (e.g., player pianos, copy machines, VCRs, computers, cached web pages, peer-to-peer filesharing, thumbnail images, etc.) that affect many copyrightable works are often challenged at first.

Q. 18: Is scanning a violation of copyright law?
A: No. The use Google makes is fully consistent with both the history of fair use under copyright law and all of the principles underlying copyright law itself. Copyright law strikes a balance between rewarding creators of intellectual property for their creations and facilitating public access to these works in ways that do not create a business harm. For books, this balance means the law needs to ensure authors write books, publishers sell them, libraries lend them, and people get to read them, either by purchasing them or borrowing them. The project is working to ensure that people can find books so that they can read them. By making books more discoverable, Google is enhancing the ability of authors and publishers to sell books to a global audience, well beyond the relatively small world of the traditional book market.

Q. 19: But publishers say you are violating copyright law. What do you say about that?
A: We disagree. Copyright law is a two-way street. It talks both about rights granted to copyright holders and fair use rights to the public. Our project falls squarely under fair use. And, indeed, the proposed online access (both at Google and at the University of Michigan) is designed so that it is helpful *and not hurtful* to publishers and authors.

Q. 20: What does UM get from the deal?
A: We get a copy of the digital files with no significant constraints on our ability to use them in ways that are consistent with copyright law. The value of these files is significant, as it would have taken us 1600 years and hundreds of millions of dollars to convert these materials on our own.

Q. 21: Why did UM make its contract available publicly?
A: As a public institution, Michigan feels strongly that openness in these types of partnerships contributes significantly to public understanding of the purpose and benefits of its initiatives. When Michigan began its negotiations with Google, we also worked to convey the value of this type of transparency to Google, and together worked to plan an appropriate release date for the contract. The contract is online at http://www.lib.umich.edu/mdp/um-google-cooperative-agreement.pdf.

Access to the content online

Q. 22: How and where will the digitized books be available?

A: As always, we want to emphasize that we will only make materials available within the bounds of copyright law. Initially, the books will be accessible through Google by entering a normal Google search. Specifics on how the Library makes is own copy available are being developed, but we do know that the Library will make its own copy available in a variety of ways, including direct links in Mirlyn records, searches through the Digital Library Production Service (DLPS) system, as well as other methods.

Q. 23: Can I see the full text of an out-of-copyright work? How does this differ from the display of in-copyright works?
A: The full text (all pages and covers) will display for titles out of copyright; users will be able to page back and forth within the volume, and retrieve all text through searches. For titles in copyright, a Google search result will display three "snippets" of text from throughout the volume, in addition to the bibliographic data associated with the book; a count of how many times total that search term appears in the book; and information on where to buy the book or find it at a local library. See also **Question 24**.

Q. 24: What is a "snippet"?
A: A "snippet" consists of approximately 3 sentences. The example below, from http://print.google.com/googleprint/about.html, describes Google's use of snippets and provides an illustration.

Q. 25: How much of the UM collection is available through the new Google service today?

A: Users can find a small number of sample titles in Google Print now. We will start to see a more significant body of material over the next six months or more, as we move from pilot to production. It will be approximately six years until the collections at the University Library are converted in their entirety. (see also **Question 3**).

Q. 26: Will people searching on Google know that the volume they've found is from the UM library?
A: There will be no explicit identification of the source. However, a Google search returning results from a book out-of-copyright will allow users to access every page of a given volume, including both front and back covers. The UM Library stamp is imprinted inside the front of books. A UM barcode is also included in the back of most UM Library books. These are the ways that a person might identify where the book is housed.

Q. 27: Will the Library put the digitized materials online also?
A: Yes, we are planning for that eventuality.

UM's Digital Archive

Q. 28: Why does UM want its own digital copy?
A: There are many reasons that Michigan wants its own digital copy of these materials. Some include:

1. Having a copy allows us to archive and curate the printed record in digital form. Only libraries do this today and if is not done by libraries, there is substantial risk that entire bodies of work will be lost. Most publishers, for example, do not have copies of the majority of their out-of-print books. It is not in their mission to preserve these materials for future generations. Similarly, Google's mission is subject to market pressures, whereas libraries have a perpetual responsibility for their materials.
2. With this rich body of work Michigan can cooperate in the development of an appropriate system that manages access to intellectual property. We now conceive of this system being governed by a "rights matrix." We plan to start with a "dark archive" and develop a rights matrix into which each of the individual works fits. We can then "light" the various parts of the archive appropriately. Public domain works can be brought fully into the light. Works that are still in print and in copyright can be left dark, but preserved. In many cases, we have found that rights holders are pleased to allow us to provide access to their work, and this decision can be documented and registered in the "rights matrix."
3. With a dark archive ensuring preservation, libraries around the country can begin to collaborate and consolidate library storage. Importantly, this should not change the purchasing behavior of libraries. That is, we do not see the Google project as having impact on library purchases (e.g., we are currently negotiating the purchase of a number of back files from journals even as Google moves forward because the purchased back files come with a variety of advantages). Where it will have impact is on how and what we choose to *store*.

Q. 29: What is the importance of maintaining a digital archive of library materials?
A: Universities are the only entity that has the interest in and the ability to preserve the

vast spectrum of written works, and specific editions of written works, that otherwise will eventually be lost to society. Many of the works in question are out of print, and will not be produced by any publisher in the future. Some of these are books that now exist only in libraries. Others are in brittle condition and will eventually be lost if they are not preserved through digital means. The role of archiving and preserving written works is an essential function of university libraries.

Q. 30: Why would Michigan provide access to these files when Google is also providing access?
A: Michigan believes there is considerable value in mounting a copy of the content. Some of the reasons are outlined elsewhere (e.g., **Q: Why does UM want its own digital copy?**). Additionally, we believe that our community of researchers will have specialized needs that are not met by the Google version. These needs may be as simple as supporting mechanisms for citation (to support research and scholarship), but may also include more specialized mechanisms such as advanced searching. In any case, we do not believe that this parallel version of the materials will be in competition with Google; rather, it should complement the sort of implementation that Google builds.

Q. 31: Will journals be searchable at the level of the individual article?
A: Although the full text of journal volumes will be searchable online, and all pages of out-of-copyright journals will be shown, it is unlikely that this content will be broken out into individual articles with corrected and clearly identified bibliographic information. Your search results will take you to specific pages, and you will be able to page forward and backward through an article.

Impact on existing library services

Q. 32: How does the project affect the Library's preservation and digitization program?
A: The Preservation program at UM Library remains strong. It will be a number of years before we are able to digitize the entire collection, so current selection and digitization work is and continues to be very valuable in identifying items that will benefit from immediate attention. Our project with Google fits into our use of digitization as our preferred method for preservation-related reformatting. Over the years, we are likely to see our preservation reformatting program concentrate on areas that are complementary to Google's efforts, including digitizing rare and other materials that need special treatment, quick turnaround conversion for materials that are part of a special project, and digitizing materials that are not well-suited to the mechanisms that Google has developed.

Q. 33: How will the project affect the library's conservation of its print collection?
A: The University's commitment to the conservation and preservation of books as physical objects will not diminish. We will continue to actively acquire material in all formats and we will continue to conserve them.

Q. 34: What effect will this have on Interlibrary Loan lending?

A: As the project progresses and material begins to show up in Google's index, the library will monitor ILL requests and take appropriate action. We will also be able to satisfy interlibrary loan requests for out-of-copyright materials by putting these items online. We will not be able to fill interlibrary loan requests for in-copyright materials with the online copy digitized by Google.

Q. 35: What are Michigan's financial burdens in converting its collection?
A: The University of Michigan does not bear significant costs in its cooperative efforts with Google. In fact, all scanning is conducted by Google, Google pays for necessary costs related to conversion and transmission of data, and all costs related to pulling and reshelving materials are borne by Google.

Q. 36: Do Library staff perform the scanning?
A: No, Google staff manage the scanning, and all Google hiring (e.g., for scanner operators) is conducted outside of the University of Michigan.

For questions, contact mdp-info@umich.edu.