Should you? May you? Can you? Factors in selecting rare books and special collections for digitisation

By Janet Gertz

Selection for digitisation is about more than which items to scan. Selection is what shapes the online collections that are built by libraries, archives, historical societies, and other cultural heritage institutions. By selecting well, institutions can concentrate on the parts of their collections that are best suited to digitisation, that make the most effective use of technology, and that meet the needs of their audiences. They can create online collections that are both useful and usable, and they can create high-quality digital assets they can manage well into the future.

No institution can afford to digitise everything. Some items are not suited to digitisation. In some cases an item's level of damage or deterioration would necessitate significant repair before digitisation could be possible. Some items simply are not important enough to justify the effort.

Every institution should have a selection process in place to evaluate materials and to determine when digital conversion is most appropriate. Clearly stated goals for digitisation and careful plans to achieve them are the starting point.

Having a basic set of selection criteria to work from helps characterise materials as better or worse candidates for digitisation based on their content value and their physical features, and it provides guidance through the logistic and financial questions: Should these materials be digitised, may they be digitised, can they be digitised, and what will it cost?

Should these materials be digitised?

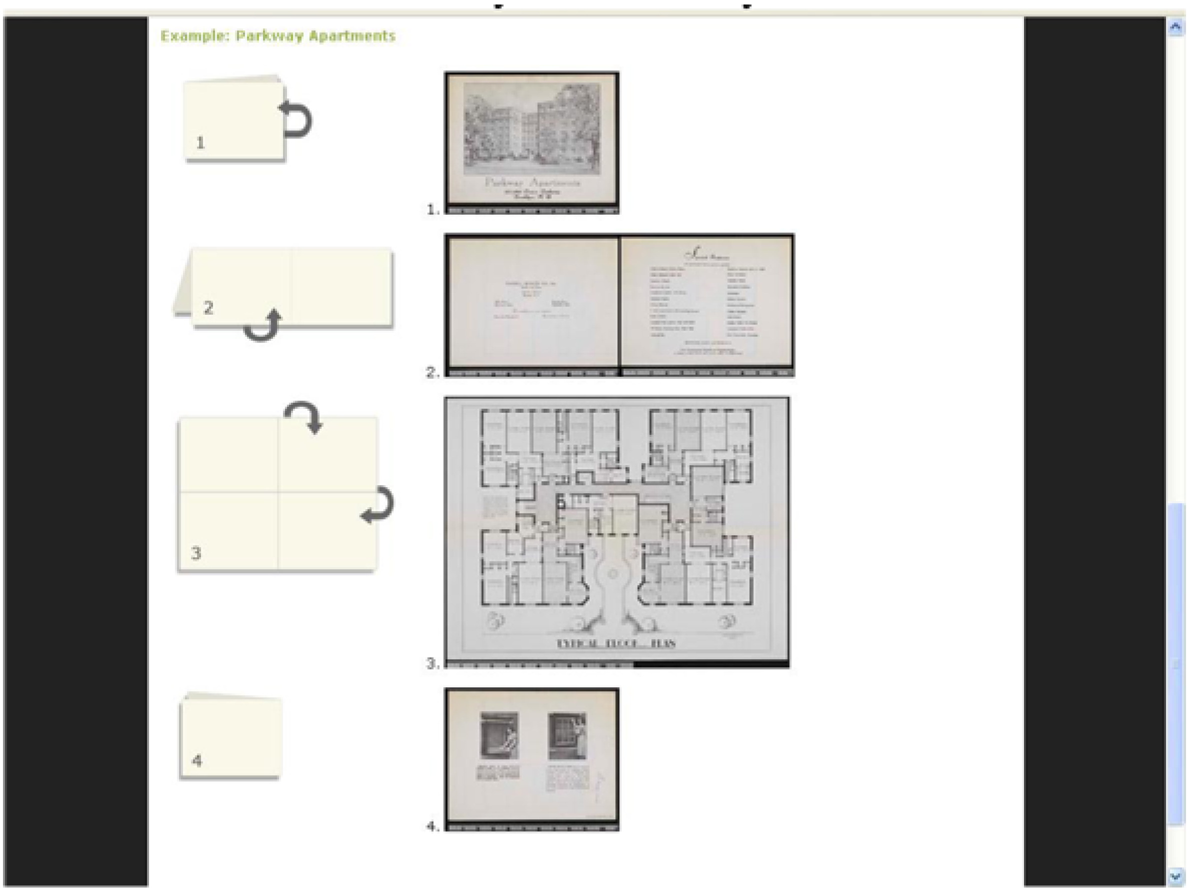
Is the collection important enough, is there enough audience demand, and can sufficient value be added through digitisation to make it worth the effort?

First, does the value of the materials merit the expenditure of effort and resources? Specific definitions of value and importance vary, but they cluster around intellectual, historic, and physical characteristics. For instance, are the materials unique or rare, aesthetically appealing, or associated with important people or events? Is the content important for scholarly or societal reasons? How do the materials relate to the institution's collecting policy, and how do they complement its other digital resources?

Value alone is not a sufficient reason for digitisation. Demand from users is a vital second factor. Digitising and mounting materials publicly is a form of publishing, and success in publishing means knowing and targeting viewers. Is there a current, active audience for these materials? Is access to the original materials inadequate, perhaps due to heavy use of popular items or because access to fragile or very costly items must be restricted? If current demand is low, will digitisation attract enough new viewers to justify the cost?

In order to satisfy demand, the institution needs to determine what audiences it hopes to serve and how it will present the digital content. Scholars, high school students, and the general public utilise online content in very different ways. How should content be presented to be most useful to audiences now and in the future? What discovery and navigation tools will be necessary? What metadata will provide adequate description and file management? What supportive and interpretive information will accompany the content? Is the plan to use a standard search interface such as Google to let people find and interpret the items on their own? Or will a special site be designed that puts the materials in context through introductory essays and related content?

How should content be presented to be most useful to audiences now and in the future?



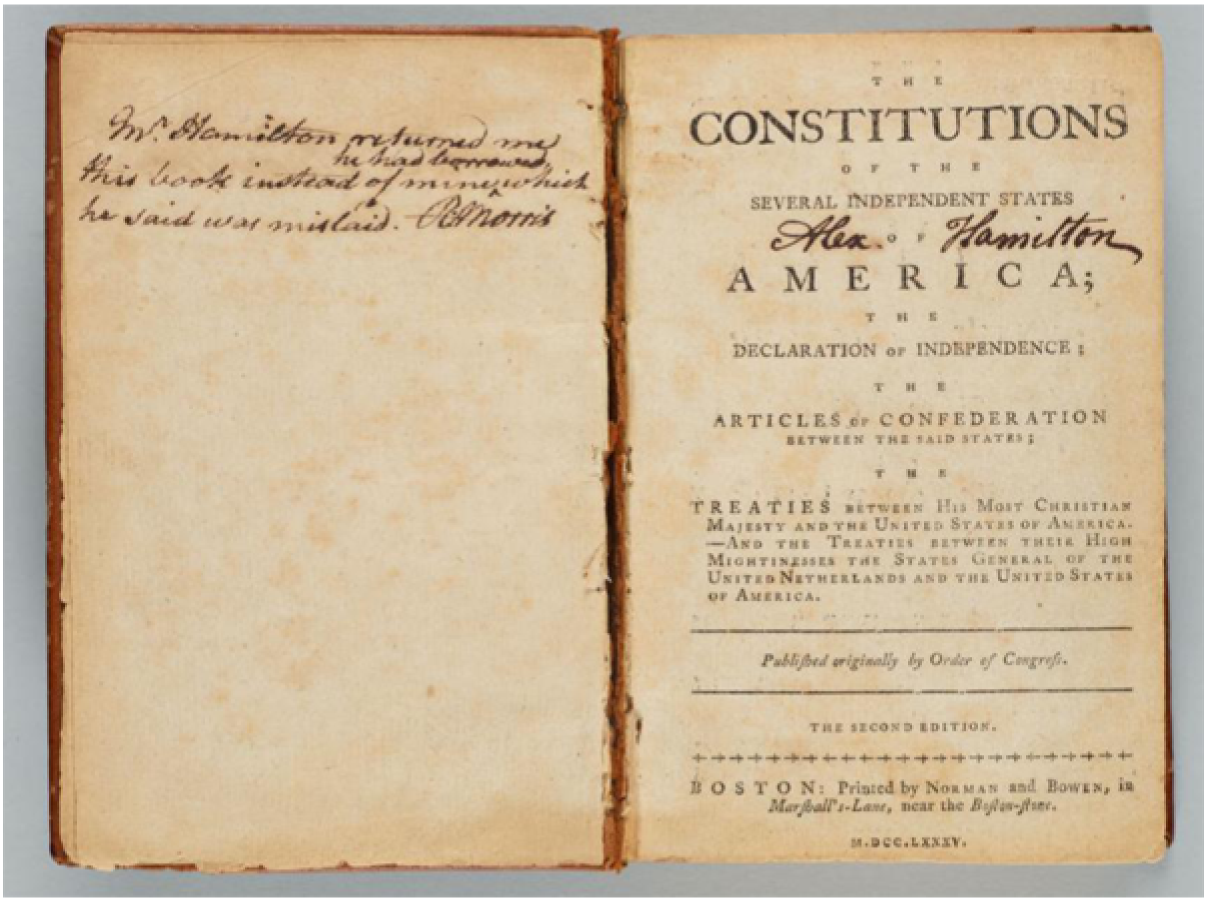
New York Real Estate Brochure Collection, About the Collection: Overview and History, screen shot (http://nyre.cul.columbia.edu/pages/about-collection, Columbia University in the City of New York). Example of information created to support a digitised collection, here explaining how to view images from complex folded objects.

Third, how will the materials' value be enhanced; what will be gained that will make them worth more in digital form? Digitisation can facilitate exposure of materials kept under restricted access due to threat of damage, theft, or vandalism or because of difficulties in handling or extreme fragility. Digital copies play an important preservation role as surrogates protecting fragile and valuable originals from handling while presenting their content to a vastly increased audience around the world. Someday, the digital version may be the only record left of an original object that deteriorates or is destroyed.

Digitisation may provide a way to improve poor legibility through technical manipulation, enabling people to work with images digitally in ways that are difficult or impossible with the original materials. An obvious source of added value is providing the ability to search and manipulate texts, most often through the application of automated optical character recognition (OCR) to bitmapped images of books or other textual materials. An institution could simply create bitmapped images, but these are not searchable – they are just pictures of the page – and they may not satisfy the needs of the readers.

Because the quality of OCR can range from excellent to poor and difficult scripts can result in plentiful errors, a decision to digitise materials and provide OCR includes planning to assure that it is reasonably correct. Providing full-text search capability for handwritten manuscripts and unusual scripts is a more complicated undertaking since OCR is not usually an option. Making these texts searchable means paying a human being to transcribe and key in the text, an expensive proposition that might provide an argument against selecting them for a project.

In thinking about digitising published materials, a fourth factor is relevant: Has anyone else already created a digital version? Redundant effort is an expensive way to use up scarce funding. Google, the Internet Archive, and other institutions are busy digitising hundreds of thousands of books and other formats. Before scanning any published item, the institution should check to see if an acceptable digital version is available. However, digitising a book, even if it has been done before, is certainly justified if the new version will be different in some way, for instance, if it contains interesting annotations or if it is more true to the original because it is at a higher resolution or in full colour.



Mr. Hamilton returned me this book instead of mine which he had borrowed, he said was mislaid. RMorris

THE

CONSTITUTIONS

OF THE

SEVERAL INDEPENDENT STATES

Alex. Hamilton

AMERICA;

THE

DECLARATION OF INDEPENDENCE;

THE

ARTICLES OF CONFEDERATION
BETWEEN THE SAID STATES;

THE

TREATIES BETWEEN HIS MOST CHRISTIAN MAJESTY AND THE UNITED STATES OF AMERICA. —AND THE TREATIES BETWEEN THEIR HIGH MIGHTINESSES THE STATES GENERAL OF THE UNITED NETHERLANDS AND THE UNITED STATES OF AMERICA.

Publiſhed originally by Order of Congreſs.

THE SECOND EDITION.

BOSTON: Printed by NORMAN and BOWEN, in
Marſhall's-Lane, near the *Boſton-ſtone*.

M.DCC.LXXXV.

The Constitutions of the Several Independent States of America, Boston, 1785, annotated by Alexander Hamilton (Rare Book & Manuscript Library, Columbia University in the City of New York).

May these materials be digitised?

Does the institution have the legal right to create and disseminate a digital version? The issues around intellectual property rights are very serious and must be addressed early in the selection process because the institution may not legally be able to disseminate digital versions. While libraries and archives do have the legal right to digitise materials that are under copyright if their purpose is preservation, those digital versions may only be accessed on the institution's premises. Obtaining permission from rights holders for public dissemination takes time, can be expensive, and is not always possible.

There is relatively straightforward guidance on how to determine whether a paper-based, published work is under copyright. But many digitisation efforts target audio-visual materials or paper formats that are unpublished. There may be complicated histories of ownership and multiple layers of authorship. Whether, and how, public access to such materials may be provided remains open to legal interpretation, and the details should be carefully investigated before starting a digitisation project. Fair use may provide options when permissions are not available. The Association of Research Libraries has recently published a set of best practices relating to fair use that offer general guidance and include examples of digitisation projects.

In general, before deciding whether to digitise, institutions need to ask the following: Is the purpose of this digitisation project purely preservation, with no intent to provide public dissemination? Or, if the purpose is indeed to put the materials online, are they in the public domain? Does the institution itself own the rights so that it can legally choose to make and display digital copies? If not, can the holder of the rights be identified and can permission be requested? If not, how much risk is the institution comfortable with in putting materials online without permission? One helpful mechanism is a click-through page where statements about rights can be displayed, including disclaimers and contact information for people who might assert rights to the digitised content.

Aside from purely copyright issues, privacy issues should also be considered.

Aside from purely copyright issues, privacy issues should also be considered. Do the materials contain personal information that should not, or legally cannot, be disseminated? On a more general level, are there issues of religious, ethnic, or community sensitivity that would make public access to the materials problematic? What sort of contextualisation would be needed in publicly mounting materials of this type to diffuse any potential upset and make it clear that the institution is presenting it from a neutral point of view?

Can these materials be digitised?

Does the institution have the technical infrastructure and expertise to create digital files and make them available to users now and into the future? And can it be done in a way that will achieve the goals of the project, given the physical nature of the materials and their organisation, arrangement, and description?

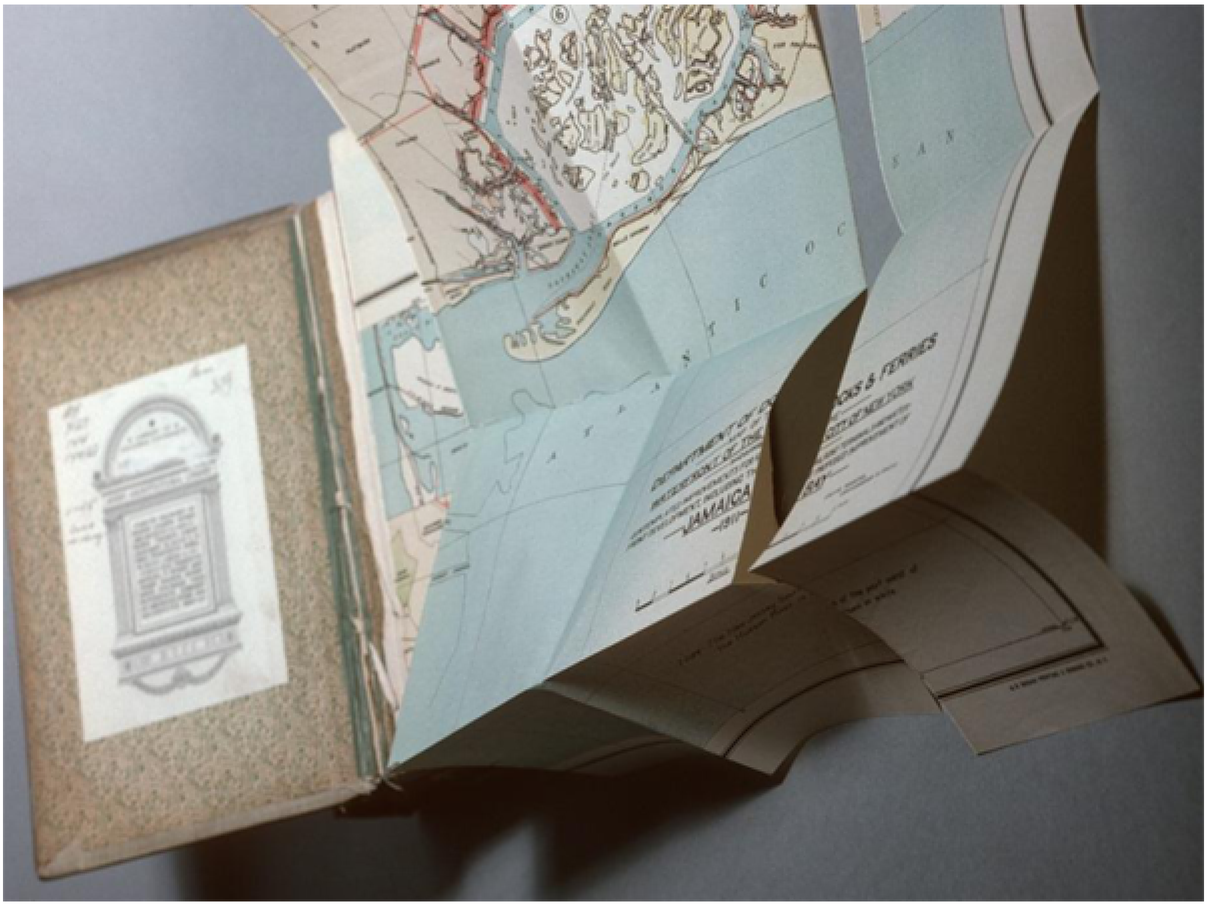
In brief, a digital conversion project requires the following:

- Preparation of materials, including physical organisation and/or collation
- Repair or conservation work if materials are in poor condition
- High-quality capture of the content according to national best practices; possible enhancements such as OCR and other functionality
- Provision of description and identification through cataloging and metadata according to best practices to record descriptive, technical, structural, and capture information.

Significant work is also required to mount the files on a website, make them accessible, and manage them over time:

- Design of the user interface, with all necessary searching and navigational tools
- Management of the website
- Planning for preservation of the files into the future.

All of this is basic to an effective product. When making the initial decision on what to digitise, success requires collaboration among the experts on the content of the materials, technical experts on preservation/conservation, digital capture, metadata creation, web design, and digital asset management. Since each set of experts has its own vocabulary, priorities, and principles, a successful digitisation project can be as much about team building as about the content of the materials. If there are no resident experts, working with consultants is strongly recommended, especially the first time.



New York City, Dept. of Marine and Aviation, Report on Jamaica Bay Improvement, 1910 (Avery Architectural & Fine Arts Library, Columbia University in the City of New York). Example of a damaged item that will require conservation before digitisation.

The technical aspects of digitisation for text, images, audio, and other formats influence choices for selection because information can be captured in many ways at many quality levels. The institution must investigate whether it can provide digital versions of the quality viewers need. A temporary online exhibition offers quite a different quality from a site meant for in-depth research. How will people use the digital images and sound, and what level of capture quality does that entail? What features of the original must be conveyed in the digital version? What features are less important? Will the digital version be of high enough quality to remain useful in the future as technology evolves?

An item's physical characteristics definitely affect what can be captured, stored, displayed, and manipulated. Issues start with the legibility of the original item. What resolution is needed to capture the smallest detail that must be viewed for the information to be useful? For instance, if there are oversize items, will the chosen format allow zooming and easy navigation within the image? There are also decisions to make about tonality, i.e., will the materials be imaged in black and white (not very satisfying, especially for illustrations), grayscale, or color? Scanning in color is not always straightforward since capture and viewing equipment must have proper calibration and color management software.

A second group of technical issues involves possible damage to the original item. For instance, tightly bound volumes do not open wide enough for capture of the entire page, which can result in images with gutter shadow and distortion or content hidden in the inner margin. Forcing a book to lie flat may cause damage. How does the value of the original object compare to the value of a better image of the content? It may be possible to deal with two problems at once, by folding in the time and resources for conservation into digital projects and by planning for digitisation of items in general as they are conserved.

Aristotle, Elenchi, France ca. 1315 (Rare Book & Manuscript Library, Columbia University in the City of New York).

A third very important issue is whether the materials are organised, arranged, and described in a way suited to online use. Once online, every image will require its own identification and description. If a book is scanned, will the institution's user interface let viewers move forward only page by page, or will it provide extra structural metadata to enable navigation at the level of chapters or other significant divisions? In preparing materials more complicated than books, for instance, sets of photographs or collections of personal papers, there can be very significant cost repercussions if the collection lacks good organisation and description. The institution will need to carefully consider what it can feasibly handle in terms of cataloging and other forms of intellectual control. The rule of thumb for archives and special materials: Don't even think about digitising until the collection is fully arranged and described.

Does the institution have a long-term commitment to preserving digital resources?

Finally, and perhaps the most difficult issue, digitisation creates new institutional assets that must be preserved if they are to remain useful over time. Choices made at the beginning of a project about capture methods and metadata directly affect the institution's ability to carry out preservation of the digital assets. Does the institution consider its digitised content to be temporary, or does it believe the content will have enduring value in digital form? Keeping digital files intact over the long term requires an infrastructure designed for the future. Does the institution have a long-term commitment to preserving digital resources? Is it willing and able to develop the necessary infrastructure or to pay for external preservation services?

What will it cost?

All of this adds up to a great deal of cost and effort. An important step in selecting a digitisation project is a cost-benefit analysis. What is the likely cost in staff time, vendor services, and equipment and supplies, from selection to metadata creation to digital capture to preserving files for the future? Does this cost match the anticipated benefits, given the value of the materials and the demand for digital access? How does the cost fit with the institution's mission and goals? How much is the institution willing or able to spend for new modes of access, wider distribution, and enhanced assets?

Everyone wants to know what it costs to digitise. The answer is, it depends. What does it cost to buy a car? It depends on the make, model, and special features; and the same is true of digitisation. It depends on the nature of the materials and all the factors discussed here. In general, based on experiences at a number of institutions, image capture of library and archival materials appears to consume, at most, one-third of the cost of a digitisation project. There is therefore no good justification for skimping on image quality. Poor images are a waste of money. Skimping on metadata is also a mistake. Metadata is expensive, but without it, people will have trouble finding and using the digitised materials, and there will be major headaches in managing and preserving them. In other words, there is presently no easy way to make digitisation really cheap if the goal is a quality product that is fully useful and usable and that will continue to be so over time.

Strategies and priorities

Digitisation makes most sense when there are materials that people want to use that are not available elsewhere online, when good metadata is available, and when copyright issues do not prevent online dissemination. The argument in favour of digitisation is especially strong when there is an opportunity to add value, whether by protecting fragile originals and documenting badly deteriorating items or by providing a new contextual framework or functionality. Developing strategies and priorities for digitisation that tie in with the institution's policies and long-range plans will allow the institution to articulate how it wants to use digitisation to support its mission and to build consensus about what criteria should guide the choice of projects and the growth of an online program.

Institutions need internal procedures for evaluating possible digitisation projects that include collection of the information relevant to understanding the true costs and benefits. Subject specialists should describe the content of the materials and the goals of the project: how it relates to the institution's mission, whether an important audience is asking for it, and how it will strengthen the institution's online collections. Information should be gathered about copyright status and whether an effort will be needed to clear copyright. Conservation staff should comment on the material's physical condition and whether any treatment will be needed before or after digitisation, and intellectual control experts should comment on what effort will be required to provide the necessary metadata. Technical staff should comment on appropriate specifications for digitisation, including any desirable enhancements, and on the work required to design the online presentation. Based on this information, institutions can make informed decisions on how well projects match up with priorities and what the workload and other costs are likely to be, in order to determine if the potential value of any project matches the resources available to carry it out.

Digitisation offers exciting options for new means of access to collections, but doing it well is not cheap or easy. By carefully thinking through decisions on what materials to digitise, institutions can produce truly successful digital assets and manage them well to the benefit of audiences now and in the future.

References

Is it under copyright? See Circular 15A, 'Duration of Copyright' Accessed at: **copyright.gov/circs/circ15a.pdf** from the U.S. Copyright Office. Charts such as Copyright Term and the Public Domain in the United States accessed at: **copyright.cornell.edu/resources/docs/copyrightterm.pdf** are also helpful.

Would it qualify as fair use? See the Association of Research Libraries' Code of Best Practices in Fair Use for Academic and Research Libraries (2012) Accessed at: **arl.org/bm~doc/code-of-best-practices-fair-use.pdf**. Best practices for digitisation: See, for instance, the Federal Agencies Digitisation Initiative, "Technical Guidelines for Digitising Cultural Heritage Materials: Creation of Raster Image," 2010 Accessed at: **digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf.**

For an example of an evaluation chart, see the Canadian Council of Archives' Digitisation Tree for Digitisation Projects, 2002 (Accessed at: **cdncouncilarchives.ca/digitization_en.pdf.**)

A version of this article first appeared in Computers in Libraries – Information Today Inc and is reprinted here with the kind permission of the author.

Janet Gertz *(gertz@columbia.edu) is director of the Preservation and Digital Conversion Division of Columbia University Libraries. She administers the libraries' preservation program for physical collections, including conservation, binding, mass deacidification, digitisation, environmental monitoring, and disaster preparedness. She manages digital conversion for both special and general collections, ranging from large-scale initiatives such as the Google book scanning project to small, specialised efforts such as imaging medieval manuscripts and Chinese oracle bones. She is also responsible for digital conversion of audio and moving image materials. In 2008, she was the recipient of the American Library Association Paul Banks and Carolyn Harris Preservation Award.*