Charleston Library Conference

# Library of Congress Recommended Format Specifications: Encouraging Preservation Without Discouraging Creation

Ted Westervelt
*Library of Congress*, thwe@loc.gov

Donna Scanlon
*Library of Congress*, dscanlon@loc.gov

# Library of Congress Recommended Format Specifications: Encouraging Preservation Without Discouraging Creation

*Ted Westervelt, Library of Congress (Author)*
*Donna Scanlon, Library of Congress (Presenter)*

## Abstract

The Library of Congress has a fundamental commitment to acquiring, preserving and making accessible in the long term the creative output of the nation and the world. The Library has devised the Recommended Format Specifications to enable it to identify what formats will most easily lend themselves to preservation and long-term access, especially with regard to digital formats. The Library is doing this to provide guidance to its staff in their work of acquiring content for its collection, but we also seek to share this with other stakeholders, from the creative community to vendors to other libraries, each of which has a need and interest in preservation and access. To ensure ongoing accuracy and relevancy, the Library of Congress will be reviewing and revising the specifications on an annual basis and welcomes feedback and input from all interested parties.

## Why the Library of Congress Developed the Recommended Format Specifications

Throughout its history, the Library of Congress has been committed to a goal best described in its mission statement "to further the progress of knowledge and creativity for the benefit of the American people." At its core, the Library's ability to advance the nation's progress has depended upon its collection, which in turn embodies the knowledge and creativity of the many authors, composers, journalists, artists, and scientists whose work is contained there. The quality of the collection reflects the Library's care in selecting materials and the effort it invests in preserving them and making them accessible to the American people for the long term.

To build such a substantial and wide-ranging collection and to ensure that it will be available for successive generations, the Library relies upon a wealth of expertise. In order to maximize the scope and scale of the content in the collection, the Library calls upon the knowledge in languages, subject matter and trends in publishing and content creation provided by the specialists who identify and acquire material for the Library's collection.

But knowledge of the technical characteristics of the production of creative works is required as well. In the past, the lasting power of the collections depended exclusively upon the endurance of such materials as the paper, ink, and binding of a book; the acetate or paper coated with gelatin in a photograph; or the shellac, vinyl, and coated polyester that comprise a sound recording. Although these materials remain in use today, creators and publishers have also begun to employ a wide array of intangible digital formats, as well as continuing to change and adapt the physical formats in which they work. The Library needs to be able to identify the formats that are suitable for large-scale acquisition and preservation for long-term access if it is to continue to build its collection and ensure that it lasts into the future.

To do this in the past, the Library of Congress has relied upon the specifications included in the copyright regulation known as the Best Edition Statement.[1] This has offered clear guidance to Library of Congress staff on the hierarchy of preference between certain physical characteristics in creative works. For example, it states clearly that when it comes to printed textual matter, "hard cover rather than soft cover." The detail in the Best Edition Statement

---

[1] United States Copyright Office. (2012). Best edition of published copyrighted works for the collections of the Library of Congress. Retrieved from http://www.copyright.gov/circs/circ07b.pdf

has been extremely useful for the Library for decades; however, it has some serious drawbacks. Since it is a regulation, the Best Edition Statement is not revised or updated frequently and there are preferences within it that no longer keep up with changes which have taken place in the creation of tangible media, such as the decline of the use of diskettes. Even more importantly, the Best Edition Statement does not address digital content at all, with the sole exception of online serials. For an institution with the broad goals and remit of the Library of Congress, having guidance that fails to address at least half of all formats in use will not work. Specifications are required that cover the whole range of content it intends to collect and that means digital content at least as much as analog.

In response to this need, in 2011 the Library began a process that would lead to the development of the Recommended Format Specifications (www.loc.gov/preservation /resources/rfs/index.html). The Library began its work by examining the Best Edition Statement, which enabled it to work closely and collaboratively with its colleagues in the Copyright Office (http://www.copyright.gov/) and take advantage of their input and unique expertise. Yet it was not merely the Best Edition Statement that provided a base from which to carry out the group's work. For digital formats, the working group took full advantage of the work done by Library of Congress staff with regard to its work on digital format sustainability (http://www .digitalpreservation.gov/formats/index.shtml) to provide it with a starting point.[2]

Between these two established fields of endeavor and sources of expertise, the Library had a strong basis on which to build the Recommended Format Specifications.

## Parameters of the Recommended Format Specifications

Before discussing the specific aims the Recommended Format Specifications attempt to address, it is best to make clear what they do not

attempt to do. The specifications which the Library is now publishing do not replace or supersede the Best Edition Statement, which provides guidance to publishers and creators in fulfilling their obligations with regard to the registration or deposit of their works under the terms of the Copyright Law. It seeks to complement that work, building upon the knowledge gained from working with the Best Edition Statement and providing a broader set of recommendations, aimed at providing guidance and clarity in a creative world, which is rich with both potential and problems and which affords numerous competing options for content format or container.

Likewise, the creation and publication of the Recommended Format Specifications is not intended to serve as an answer to all the questions raised in preserving and providing long-term access to creative content. They do not provide instructions for receiving this material into repositories, managing that content or undertaking the many ongoing tasks which will be necessary to maintain this content so that it may be used well into the future. Tackling each of those aspects is a project in and of itself as each form of content has a unique set of facets and nuances. These specifications provide guidance on identifying sets of formats which are not drawn so narrowly as to discourage creators from working within them, but will instead encourage creators to use them to produce works in formats which will make preserving them and making them accessible simpler. Following these specifications helps make it realistic to build, grow and save creative output for our individual and collective benefit for generations to come.

## Developing the Recommended Format Specifications

In 2011, a working group comprised of stakeholders from across the Library was established to examine the existing Best Edition Statement and determine a structure upon which the Library could model its own specifications. The

---

[2] Library of Congress. (2014). Sustainability of digital formats: Planning for Library of Congress collections. Retrieved from http://www.digitalpreservation.gov/formats/index.shtml

Library identified six basic categories of creative output, which represent significant parts of the publishing, information, and media industries, especially those that are rapidly adopting digital production and are central to building the Library's collections: Textual Works and Musical Compositions; Still Image Works; Audio Works; Moving Image Works; Software and Electronic Gaming and Learning; and Datasets/Databases. Technical teams (http://www.loc.gov /preservation/resources/rfs/contacts.html), made up of experts from across the institution bringing specialized knowledge in technical aspects of preservation, ongoing access needs and developments in the marketplace and in the publishing world, were established to identify recommended formats for each of these categories. These technical teams also engaged other subject matter experts throughout the Library, and where appropriate, at other organizations (though not for public comment). The teams also reviewed the currently available formats of published materials—both print/tangible and digital—in their categories, as well as the Library's other guides to selecting collection materials (such as the Collections Policy Statements (http://www.loc.gov/acq/devpol /cpsstate.html) and Sustainability of Digital Formats guidelines).[3] The results of their work are the core of the specifications, which seek to provide a framework within which creative works should have the flexibility to grow and develop, and also help ensure that these creative works be accessible and authentic into the future.

The Recommended Format Specifications seek to provide structure without being enslaved to it. Like the Best Edition Statement, the Recommended Format Specifications use hierarchies for the physical and technical characteristics of creative formats that will maximize the chances for survival and continued accessibility, though in the case of the specifications they are comprehensively digital as well as analog. Yet the hierarchies are not so rigid as to make them unworkable. Each basic category is broken down in logical ways—print text and

digital text (http://www.loc.gov/preservation /resources/rfs/textmus.html), print photographs and digital photographs (http://www.loc.gov /preservation/resources/rfs/stillimg.html)—but these divisions are determined by the specifics of the category and subdivision, not by a forced attempt to fit them in identical boxes. While for text and photographs, it makes sense to have sections for print text and digital text, print photographs and digital photographs, for audio works the key subdivisions are "On Tangible Medium (digital and analog)" and "Media-independent (digital)" (http://www.loc.gov /preservation/resources/rfs/audio.html). This carries through when identifying the specific characteristics of types of works, for they are not the same for print text and for digital photographs and the particulars in the specifications reflect that.

This need for a level of flexibility, especially with regard to digital formats, is also apparent in the arrangement of the technical characteristics of a given type of work into two groups, preferred and acceptable. In many situations, there is a long list of file formats that could be or are included in the Recommended Format Specifications. Arranging them in a numbered order is visually useful and makes them more apparent to the user and therefore more easily accessible, but has the potential drawback of leading to unproductive debates over the placement of a given file format sixth in line as opposed to third. The Library is more concerned with whether a file format is "preferred" or "acceptable" and less whether it is number four or six in a list of file formats within those groupings. If a file format or a technical characteristic is listed as preferred, the Library has identified its use as promoting preservation and long-term access. If it is acceptable, then that file format or technical characteristic may or may not promote preservation, but at the very least is not an impediment to it. In dealing with digital content, it is important to avoid being too dogmatic and this is one attempt to keep a necessary flexibility within an equally necessary structure.

---

[3] Library of Congress. (2014). Collection policy statements and supplementary guidelines. Retrieved from http://www.loc.gov/acq/devpol/cpsstate.html

## Recommended Format Specifications Goals and Uses

The Recommended Format Specifications seek to fill two key needs. One is to provide internal guidance within the Library to help inform acquisitions of collections materials (other than materials received through the Copyright Office). The Recommended Format Specifications do not allow Library of Congress staff to start recommending, selecting and acquiring any content that comes in a file format listed in the document. In some instances, there are digital acquisition workflows, tools and processes in place, and Library of Congress staff working with them should be able simply to integrate the specifications into their work. But in many cases, the necessary workflows and infrastructure have not yet been created to handle these formats. However, Library of Congress recommending officers and acquisitions librarians are often made aware of digital content that is available via gift, exchange or purchase. In all such cases where the intellectual content would be of benefit to the Library, the staff member must be cognizant of the technical characteristics of that content. By tracking the formats of digital material available for acquisition by the Library, Library staff can identify the range of content that is both intellectually suitable for the Library's collection and available in one of the preferred formats listed in the specifications. They can also gain insight into the types of format in use and provide feedback on this for use in future revisions of the Recommended Format Specifications.

It is also expected that Library of Congress staff will be able to inform potential acquisitions sources of the parameters within which the Library sees itself collecting in the future. This does not mean the Library will refuse to acquire any content that is not in one of the formats listed. The scope of the Library of Congress's acquisitions is broad and governed by the terms of its Collection Policy Statements. If the value of an item, for its intellectual content or for other reasons, is great enough, the Library will acquire the content even if it is not in one of the formats listed in the Recommended Format Specifications. However, these are expected to be exceptional

cases. For the Library to build the digital content in its collection on the scale it does with analog content, the Recommended Format Specifications will have to be used by staff as a guide to help identify content for acquisition into the collection.

The Recommended Format Specifications also fill a second, broader need. The work that the Library has undertaken in developing them has definitely been from its own particular perspective. Nonetheless, the fundamental purpose of the specifications, to identify the characteristics of creative works which best enable them to last and to be accessible in the long-term, is not one specific to the Library alone. The Library of Congress recognizes that the broader communities look to America's foremost library for guidance; and one of the Library's fundamental goals is to provide the benefit of its expertise and knowledge to support and assist those other communities and institutions. The Library intends to disseminate the Recommended Format Specifications as broadly as possible so that others might benefit from them and also that the specifications might benefit from the feedback the Library receives from those other stakeholders.

## Future Work

The Library's commitment to the long-term survival of the creative output of the nation and the world means that this set of specifications must be a living document. The creative world by its very nature is a dynamic one and so the framework must live, adapt and grow alongside it. As such, the Library will be revisiting these specifications on an annual basis. It is not expected that this will result in root-and-branch changes in the course of any one of these revisions. It is in fact hoped that, by engaging with the specifications on an annual basis, revisions will be smaller and more manageable as there be less room for the Library's specifications to slip out of sync with developments in the creative world.

Over the months preceding the annual review, the Library will seek out and request input from stakeholders to ensure that all parties who could use and will benefit from this set of specifications are fully aware and engaged in any and all

revisions. Not only will this provide for the best-informed decision-making when it comes to the revision of the Recommended Format Specifications, it also offers a chance to engage with other groups, organizations and institutions that have a vested interest in the goals of the specifications and, hopefully, move us all towards greater clarity and precision.

For example, with regard to the digital formats, there is a lot of fluidity and uncertainty in identifying the best way forward in meeting the needs of all concerned in encouraging and rewarding creativity, but also making it easy to share, preserve and access. This can be seen directly in the category dealing with datasets and databases (http://www.loc.gov/preservation /resources/rfs/data.html) in the Recommended Format Specifications. Instead of being able to identify precise file formats and technical specifications, the Library was forced to describe the desired attributes in more open-ended terms, for the simple reason that there is no clear

consensus on any precise specifications with regard to them.

## Conclusion

At a time of such great growth in the production of creative output, when not only are the frontiers expanding, but new ones seem to crop up faster than we can grasp them, there is a definite need for some expert guidance, so that this amazing creative content is not lost to us. The Library of Congress appreciates that it is uniquely positioned to provide that guidance and, in fact, that its position has given it that responsibility. The Library is the nation's premier institution instructed to further the progress of knowledge and creativity for the benefit of the American people. In producing and publishing the Recommended Format Specifications, it seeks to meet that charge, and to provide the benefit of its expertise for creators, vendors, and archivists, so that they might succeed in their goals to share and disseminate their creative output and benefit the nation generally.