INCITS: U.S. Library of Congress Uses Standardized Digital Formats for Sustainable Preservation of Archives

Background

The Library of Congress (LoC) maintains records that archive historical or cultural collections as well as collections subject to mandatory archival under U.S. copyright law. In 2000, Congress appropriated $100 million for the National Digital Information Infrastructure and Preservation Program (NDIIPP) to be led by the LoC. The legislation called for the LoC to guide a public-private partnership in the development of a national approach to digital preservation. The goal was to collect, preserve, manage, and make available significant digital content, especially information that is created in digital form only, for current and future generations.

Problem

Long-term archival of digital documents has been an ongoing challenge for academia and industry alike. Stakeholders were concerned about best practices for sustaining works as well as how to keep up with the creation of new digital formats which rapidly replace old ones. Other considerations also included the ability to encode and decode digital works, documentation of format contents, access to the tools necessary to view or edit them, and format resistance to digital degradation. The LoC was tasked with identifying the most common digital formats and evaluate their strengths and weaknesses for archiving with respect to long-term storage and sustainability.

Approach

In order to evaluate each digital format and establish a decision-making framework for support/adoption of each format, the LoC developed seven sustainability factors: disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents, and technical protection mechanisms. This framework considers policy imperatives, legal issues, lifecycle of a work in that format, and benefits of supporting multiple formats for a given type of work. When establishing the sustainability factors, the LoC recognized that the weight of each factor would vary by genre or form of expression for content. For example, significant characteristics of sound are different from those of still pictures, whether digital or not, and not all digital formats for images are appropriate for all genres of still pictures.

These factors were applied across digital formats for all categories of information to reveal the feasibility and cost of preserving the content in the face of future change in the technological environment in which users and archiving institutions operate. These factors laid the foundation for the selection of formats that offer the greatest potential for sustainability and future functionality such as migration to new formats, emulation of current software on future computers, or other challenges that may arise.

Outcome

Using the seven sustainability factors, the LoC evaluated formats that were both promising and not promising for archiving. Consequently, file formats identified by the LoC as optimal for

sustainability and long-term storage were created by standards development organizations (SDOs), such as the [InterNational Committee for Information Technology Standards](#) (INCITS) and the [International Organization for Standardization](#) (ISO). Reflecting the LoC's experience with archiving digital reproductions of historical content with its physical collections of photographs, pictorial prints, sound recordings, books and periodicals, and moving images, the formats initially adopted by the LoC fall into four basic content categories: still images (TIFF), sound (MP3), text (PDF), and moving images (MPEG). These four basic content categories also serve as building blocks for more complex digital works, e.g., serials, compilations, multimedia works, etc. Subsequent additions have included the formats appropriate for archived websites, datasets, and geospatial information. Over time, additional descriptions will be added to all categories and new categories will be added as new standards and digital formats evolve.