Orphan Works as Grist for the Data Mill

Matthew Sag
Associate Professor of Law, Loyola University Chicago
matthewsag@gmail.com | Bio | SSRN

The phenomenon of library digitization in general, and the digitization of so-called 'orphan works' in particular, raises many important copyright law questions. However, as this article explains, correctly understood, *there is no orphan works problem for certain kinds of library digitization*.

The distinction between expressive and nonexpressive *works* is already well recognized in copyright law as the gatekeeper to copyright protection—novels are protected by copyright, telephone books and other uncreative compilations of data are not. The same distinction should generally be made in relation to potential *acts* of infringement. Preserving the functional force of the idea–expression distinction in the digital context requires that copying for purely nonexpressive purposes (also referred to as non-consumptive use), such as the automated extraction of data, should not be regarded as infringing.

The nonexpressive use of copyrighted works has tremendous potential social value: it makes search engines possible, it provides an important data source for research in computational linguistics, automated translation and natural language processing. And increasingly, the macro-analysis of text is being used in fields such as the study of literature itself. So long as digitization is confined to data processing applications that do not result in infringing expressive or consumptive uses of individual works, there is no orphan works problem because the exclusive rights of the copyright owner are limited to the expressive elements of their works and the expressive uses of their works.