Entities that collect data from websites through an automated device, including but not limited to “scraping,” “harvesting,” and “crawling” (such entities, “**Data Collectors**”) should be mindful of potential copyright interests in the content collected. As a general matter, the Copyright Act gives the copyright owner, here, the website owner, the exclusive rights to reproduce, perform, display, distribute, and create derivative works of an expressive work of authorship.

**Copyright does not protect facts**

In most jurisdictions, copyright does not protect facts. However, copyright may exist in a compilation of factual data or database insofar as there is original selection, coordination, or arrangement of data in that compilation, although that copyright has been described by US courts as “thin” and the copying of facts is otherwise not only permitted, but encouraged under copyright law. Note that EU laws in relation to database rights are more prohibitive, as detailed below. Data Collectors often extract purely factual data from online sources and import the data into their own pre-existing, differently structured databases.

**Collection of copyrighted content may be permissible as “fair use” (US)**

Copyright’s fair use doctrine provides an additional layer of legal protection by permitting the use of copyright-protected works without a license in some circumstances. The Copyright Act provides for at least four factors for a court to consider in determining whether a given use of a copyrighted work constitutes fair use: (1) the purpose and character of the use, (2) the nature of the copyrighted work, (3) the “substantiality” of the portion used in relation to the copyrighted work as a whole, and (4) the effect on the potential market for the copyrighted work.

In many cases facing Data Collectors, the most immediate of the four factors for consideration will be the purpose and character of the use, and in particular whether the use is “transformative.” A use of copyrighted content is transformative where, rather than “superseding” the object of the original work, the use is for a new purpose or adds a different character. Specific examples of transformative and fair uses recognized by the courts have included copying student research papers into a database to determine whether a student has committed plagiarism, copying books to enable the creation of a searchable index, and copying for the purpose of extracting unprotected elements of a computer program, as well as more traditional uses involving parody and criticism. The more a given use of a copyrighted work is considered “transformative,” the less weight is given to those of the other three factors that weigh against a finding of fair use.

With that said, the fair use doctrine is applied on a case-by-case basis and is highly dependent on the particular facts and circumstances of each case.[[1]](#footnote-2) However, it is common for Data Collectors in the Alt Data space to use copyrighted works for purposes that are markedly different than originally intended. For example, where an expressive work such as a blog post is copied for training machine learning algorithms to assess the sentiment of the post, the purpose is not to express the content’s sentiment, but to measure it and improve the assessment of sentiment in other works.

Additionally, hewing to industry standards for web data collection (such as those discussed in the FISD Web Data Collection Considerations) will increase the likelihood of a finding of fair use or the application of other defenses to copyright infringement, such as estoppel or the finding of an implied license.

**Collection of copyrighted content (Europe)**

Permissible collection of data differs under EU copyright law. Where the focus of US copyright is to promote the progress of science and useful arts, EU copyright focuses on the protections of the publisher.

There are many pieces of EU legislation that affect copyright but the main Directives are:

* Directive 2001/29/EC on copyright and related rights in the information society (Copyright Directive);
* Directive 2009/24/EC on the legal protection of computer programs (Computer Programs Directive);
* Directive 96/9/EC on the legal protection of databases (Database Directive); and
* Directive 2011/77/EU on the term of protection of copyright and certain related rights.

Exceptions for copyright-protected works are laid out in the Copyright Directive and EU member states outline copyright laws in domestic legislation based on the Copyright Directive. While the Copyright Directive itself does not have anything akin to the fair use doctrine under US law, some EU member states have a similar concept of fair dealing under common law. The concept of fair dealing is not as broad as the fair use doctrine and only applies in certain countries.

The Court of Justice of the European Union (CJEU) has heard two prominent web scraping cases in relation to copyright law. In both cases the CJEU decision hinged on whether: (i) intellectual property rights subsisted in the data which was mined, (ii) whether the scraping was an infringement of those rights and (iii) whether it is possible for the website owner to limit the re-use of the data through the use of T&Cs.[[2]](#footnote-3)

The EU Database Directive limits copyright protection of databases to the ‘selection or arrangement of their contents’ and which ‘constitute the author’s own intellectual creation’ (Article 3(2), Database Directive). This does not extend to the contents of the database.

1. [↑](#footnote-ref-2)
2. [↑](#footnote-ref-3)