Data and the law: beyond the sweat of the brow

Who owns published data? And what is data?

You gather data for research, you publish – and who then owns your data? That begs the philosophical question: what is data? The US Supreme Court has ruled that "data" are facts and cannot be copyrighted. So can everyone use your data? And why do authors find themselves paying to use data that others have gathered? In the new age of supposedly open access to digital and big data, **Gerald van Belle** and **Leslie Ruiter** discuss the background, current status and future challenges to who legally owns data.

Background

In 2012 the first author of this article, Gerald van Belle, published a textbook on the design and analysis of experiments. One of the key objectives of the book was to use real data, so that students could get used to the vagaries, gaps and inconsistencies of reality, as opposed to the problem-free perfection of invented examples. So we found real data, from research papers and the like. We hit a snag: our publisher insisted that permission be obtained for every data set used, and insisted on this even in the case where only a portion of a table was used, or the data were rearranged. Any permission fees were to be borne by the authors (i.e. ourselves). In addition, the publisher wanted, for each data set, a broad permission for unlimited numbers of copies and editions. This last makes a lot of sense because it would be hard to keep track of different requirements of copyright holders – one allowing print but not digital, one allowing digital for a limited period, and so on. The requirements seemed rather strong, but in order to get the book published the first author complied. This resulted in total fees of $2800, with one fee of $2000 for permission. In one case, where only eight data points from a table were used, the publisher would only give permission for 1000 copies, demanded additional fees for electronic use, and no permission for a second edition, if it should appear.

We discussed our woes with another statistical editor who works with a different publisher. He suggested that permission to use data from a table was not in fact needed. Unfortunately, this editor could not give documentation that could be used because the issue had come up with a previous employer. However, the editor stated that his company would not require an author to get the permissions that had cost us so much effort (and money) .

> Apedionsequia nobitatur, ium aruptatur maion nulloritat erumquidia consequate nis ullant aut aperum quis modi

This led to consultation with the second author: Leslie Ruiter is an attorney specialising in intellectual property. The conflicting opinions and practices led us to write this article in the hope that it will help others. It sets out conditions under which permission is not needed, and needed. The results are based primarily on US law, but we do give opinions about the applicability to European and Canadian law. The second author provided an initial, more technical legal opinion for those who might find it useful; this document can be accessed at xxxx and is referenced below[1].

What is copyright?

Copyright is a form of protection given to authors of "original works of authorship". US copyright law does not protect everything reduced to writing (or drawing). The Copyright Act denies protection to ideas, methods, systems, mathematical principles, formulas, equations, and devices based on these. One of the key guiding principles of the courts has been that data are facts, and facts cannot be copyrighted. This is the point that the law starts from; lots of consequences follow.

Furthermore, the courts have ruled that facts are discovered, not invented, which is why they cannot be copyrighted. Data in any form consist of bare, naked facts. Indeed, the word *data*, the plural of *datum* from Latin *dare*, means "something given". The bedrock of statistical inquiry is that the facts existed before, but have now been "found". Facts, by definition, are not created.

In law, as in statistics, definitions can make all the difference. The terms *data*, *information*, and *knowledge* are frequently used for concepts that in fact overlap. The main difference is in the level of abstraction being considered. Data are at the lowest level of abstraction, information

is the next level, and finally, knowledge is the highest level among all three. Data on their own carry no meaning. For data to become information, they must be interpreted and take on a meaning. For example, the height of Mt Everest is generally considered a "datum", a book on the geological characteristics of Mt Everest may be considered as "information", and a report containing practical information on the best way to reach the top of Mt Everest's may be considered as "knowledge". When I tell you that the height of Everest is 8848 metres above sea level I am infringing no one's copyright, even though it may have taken a dedicated team much effort, much time, much scientific skill and imagination to establish that figure. Simple enough so far? It does not remain so.

US Supreme Court opinion

In 1991 the US Supreme Court gave its opinion in a landmark case concerning telephone directories. A company called Rural Telephone Services (RTS) had compiled a telephone directory of all its customers. It covered a small area of Kansas. A company called Feist Publications wanted to publish a directory covering a larger area. They asked if they could use the listings in the RTS directory rather than gather all that information again; RTS said no, but Feist went ahead anyway. RTS sued Feist for infringement of copyright. *Feist Publications, Inc.* v. *Rural Telephone Service Co.* has become the starting point for analysing copyright protection for facts, data and databases.

A line of cases prior to *Feist* had granted protection to "sweat of the brow" or "industrious compilation", that is, protection simply because it took much effort to gather the database of facts; "the underlying notion was that copyright was a reward for the hard work that went into compiling facts"[2].

The US Supreme Court in *Feist* changed all that. It held that telephone book white page facts are in the public domain and are constitutionally beyond Congress's power to include within copyright protection. The Court rejected RTS's argument that Feist's employees should have to re-collect the same data door-to-door to construct its own directory; it noted that raw facts may be copied at will. The Court soundly rejected the "sweat of the brow" doctrine.

But sweat of the brow is not the same thing as creativity. *Feist* does not bar copyright for original or creative selection, coordination or arrangement of facts. "The *sine qua non* of copyright is originality."[2] But one would not expect a statistician or a scientist compiling his facts or statistics to take the position that the selection or arrangement of data had a subjective

A modicum of originality

Table 1 contains data on the number of active health professionals in the USA in 1980. The data – and the table – come from a government publication and are in the public domain. Thus, there is no question of copyright or protection. Table 2 comes from van Belle's book *Statistical Rules of Thumb*[3]. The data in it comes from Table 1, but van Belle noted the non-informative ordering by alphabet in the original, and the curious pattern of rounding in some cases and not in others. For example, the number of physical therapists is given as 50 000, the number of physicians as 427 122 – probably reflecting the sources of the data. In addition, the number of zeros suggested expressing the numbers in units of 1000. This led to Table 2.

The virtues of this table are fewer numbers, a more meaningful ordering of occupations, and a grouping according to approximate size. Now this table is protected by copyright (it is owned by the book's publisher) and permission is needed (and has been obtained) to reproduce it here. Thus, the original data and table are in the public domain but the revised table is protectable. The numbers in the table are not protected, but their compilation is.

Table 1. Number of active health professional according to occupation in 1980: United States[a]

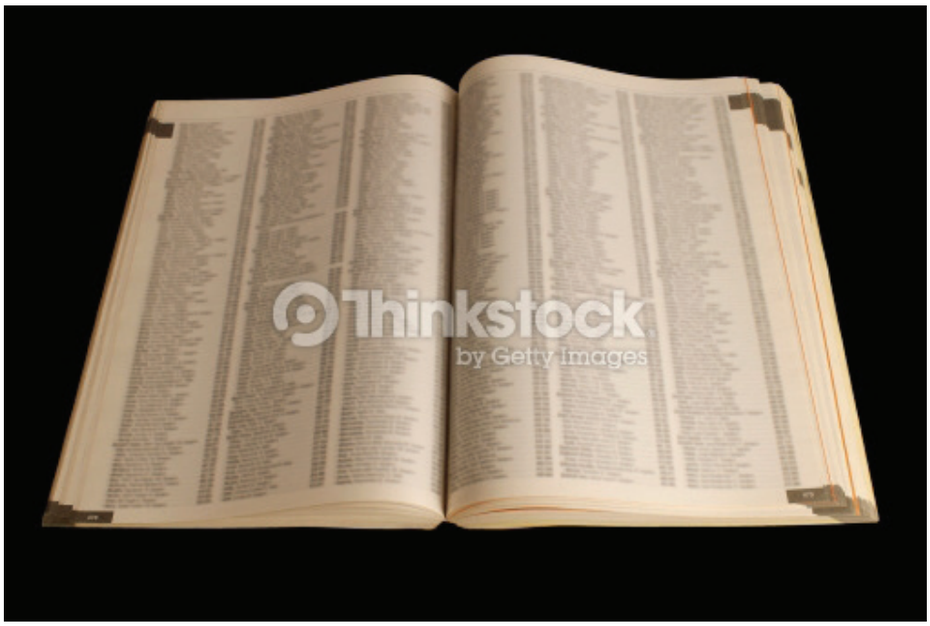
Occupation	*1980*
Chiropractors	25,600
Dentists	121,240
Nutritionist/dieticians	32,000
Nurses, registered	1,272,900
Occupational therapists	25,000
Optometrists	22,330
Pharmacists	142,780
Physical therapists	50,000
Physicians	427,122
Podiatrists	7,000
Speech therapists	50,000

[a]From the US National Center for Health Statistics, 2000; Table 104.

Table 2. Table 1 re-arranged by number in occupational category and rounded to the nearest 1000

Occupation	*1980 (1000s)*
Nurses, registered	1,273
Physicians	427
Pharmacists	143
Dentists	121
Physical therapists	50
Speech therapists	50
Nutritionists/Dieticians	32
Chiropractors	26
Occupational therapists	25
Optometrists	22
Podiatrists	7

Reprinted with permission from John Wiley & Sons.



VukasS/iStock/Thinkstock

or creative component. Statistics, and science, strive for objectivity: the assumption, or at least the hope, is that the data would be the same whoever had gathered them. In reality, of course, a degree of personal human input is almost inevitable. So how much originality do you need before a set of facts have become a work of your own creativity?

The *Feist* case is often referred to in arguments over the required "modicum of originality": whether a particular change in the selection and arrangement of the material is enough to be protected. A change in font is not original enough, but what about a change in column placements and headers? The two tables in Box 1 illustrate the issue. One is protected by copyright. One is not.

The Court's opinions in the *Feist* case contains much language helpful to scientists and generators of data. The Court noted that refusal to use copyright law to protect fact-compilers is "neither unfair nor unfortunate. It is the means by which copyright advances the progress of science and art." Further, "[c]opyright law intends to make available to all the fruits of previous research". And finally, "[t]he 1909 [Copyright] Act did not require, as 'sweat of the brow' courts mistakenly assumed, that each subsequent compiler must start from scratch and is precluded from relying on research undertaken by another. Rather, the facts contained in existing works may be freely copied because copyright protects only the elements that owe their origin to the compiler – the selection, coordination, and arrangement of the facts."[2] In Table 2 the facts in Table 1 have indeed been selected, coordinated and rearranged – which means that Table 2 is protected by copyright.

Pudit quis maxim ut hite paruptatest verferchicae prem reritaque re ipienimusda volores tionem qui ute corro blaborum

Noting the tension between two established principles of copyright law – facts are never copyrightable, but compilations of facts are generally copyrightable – the Court reached its compromise position: originality in selection, coordination or arrangement of facts is protectable and the scope of protection is limited to those original contributions.

After *Feist*, courts struggled to find the line between facts and their selection and arrangement. For example, courts have denied protection for facts explained in a scientific model which mimicked certain behaviours of millions or particles in a photonic device – the model was an attempt to "represent and describe reality for scientific purposes" and the "scientific reality was not created by the plaintiffs"[1].

On the other side of the equation, some compilations of facts were still protectable after *Feist*. For example, courts granted protection for a quick reference pocket guide for nurses based on the argument that the pocket guide involved creative choices from a universe of potentially relevant facts[1].

The leading treatise on US copyright law asserts that statistics reported as the results of various tests or surveys are predestined, that is, they are not selected at the discretion of the scientist but are determined by the process and method chosen by the scientist, and therefore unprotectable under US law.[4] Perhaps most important for scientists, copyrights do not subsist in facts *per se*. In the scientific community, publishers, authors, scientists, schools, and other owners of scholarly works may claim federal protection only in the particular expression of facts or in the selection and arrangement of those facts[4].

At first blush this may appear to be a significant body of material not protected by US copyright law, but to conclude such would be an error. Since lawsuits and the resulting case law are put forward by those claiming ownership and wanting to expand their rights, protections, and profits, the pressure on the boundary line is constant and severe. Also, since the copyright owners are typically well funded and the users of non-protectable material typically less so, the boundary is often pushed in favour of more material being proprietary and less material being public domain or unprotectable. The well funded can hire better lawyers.

The US Copyright Office (`http://www.copyright.gov/help/faq/`) is a good source for determining US government policy – as well as references to international conventions. For example, a search on *Feist* produces a detailed discussion of the results of this decision.

The Copyright Clearance Center

The Copyright Clearance Center, a US company acting primarily on behalf of publishers, provides a mechanism for getting permissions (`http://www.copyright.com`). In the process it also determines the fee to be charged. The algorithms used for determining whether a fee will be charged are not nuanced to incorporate the distinctions between data and their compilation. The website does not point to the *Feist* decision, nor to any circulars of the US Copyright Office. It will usually be preferable to identify an editor within a publishing firm for requesting permission. Networking can be very useful in this case. For an interesting history of the Copyright Clearance Center see the Wikipedia article (`http://en.wikipedia.org/wiki/Copyright_Clearance_Center`).

International copyright

Looking outside the US, the US Copyright office says: "There is no such thing as an 'international copyright' that will automatically protect an author's writings throughout the entire world. Protection against unauthorized use in a particular country depends, basically, on the national laws of that country. However, most countries do offer protection to foreign works under certain conditions, and these conditions have been greatly simplified by international copyright treaties and conventions."[5]

European Union

In 2007, the then 27 member states of the European Union agreed to adhere to various directives and regulations on the use of data. Article 1.2 of Directive 96/9/EC on the Legal Protection of Databases defines a database as a "collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". This directive offers copyright protection to databases which, by reason of the selection or arrangement of their contents, constitute the author's own original intellectual creation. With this protection, the author has the exclusive right to reproduce, alter and distribute the work. In stark contrast to US law, it provides an exclusive right to protection *sui generis* for databases, regardless of the degree of originality. With this protection of investment, the makers of databases can prevent unauthorised extraction and reutilisation.

There may be some flexibility and variation between the member states with respect to copyright protection of scientific works. Under Copyright Directive 2001/29/EC, Article 5.3(a), EU member states have the freedom to support non-commercial science by making copyright less restrictive for academic use of copyrighted work. However, note that Directive 96 does not rely on copyright law to prevent extraction and reutilisation, so it is questionable whether such flexibility for scientists would have any benefit. And if the data were protected by copyright by the EU directive prior to its inclusion in the

database, then it remains individually protected as a work separate from the copyright itself under the *Berne Convention*[1]. A general conclusion from which might be that European lawyers are in little danger of unemployment.

Canada

Canada and the US have a "fortunate similarity in matters of compilation of data"[1]. Both countries require that a work be "original" within the definition of respective laws. However, the Canadian legal concept of originality allows for copyright protection where discernment, skill, and judgement are involved in compiling data, while US law emphasises protection for the creative, novel, and unique. This distinction may result in subtle differences on the legality of extraction of data, in that data can never be "creative" by definition. But we can imagine a court convinced in some circumstances that data extraction required discernment, skill and judgement to obtain the data. At this point it appears that EU copyright law is more restrictive than the US law, with Canadian law somewhere between the two.

Catquidem faciisque ant, teceprat occum doluptat fugit, quo dolupta tiusci cor minita ne delitae

Storm of stars in the Trifid nebula. Data in astronomy are freely copied (and usually acknowledged). Image courtesy NASA/JPL-Caltech

Conclusions and recommendations

What are the implications for "big data", or even for small data? No permission is required for analyses of published data or the creative use of a subset of data, for an example in a textbook. (Collegial courtesy would require acknowledgement of the source, of course.) Data from a table used to make a graphical presentation can be used freely without permission. Similarly, data read off a graph do not need permission.

Beyond those reasonably clear guides, there is tremendous variation in determining whether permission is needed: the variation is by scientific area, by publishers and by nations. Data ownership in astronomy is a non-starter: data are freely copied (and usually acknowledged). It would seem that data generated in the observational sciences are considered less protected than data generated in the experimental sciences. This is consistent with the interpretation of data being discovered rather than created.

As we indicated at the start, there is wide difference of opinion among publishers about who owns the data. Even within publishing firms there is variation. The very large publishing firms have distinct divisions with different editors and different perspectives. Finally, among nations there is variability with attempts to attain a minimum standard of international acceptability.

There may be several challenges to the current court opinions. First, the argument that data are facts, discovered not invented, is not very strong when data are in some sense created. For example, the score on an IQ test is an example of a latent trait which is characterised by a number. Is it reasonable to think of this number being discovered or created? The deviser of an IQ test has in some senses also devised the scores it will yield. Another instance arises in simulation studies; it is difficult to think of such numbers as being discovered. The next legal challenge to *Feist* will have to wrestle with these questions. Challenges are likely to arise when a publisher senses a trend of data usage that threatens its view of copyright. This will initiate an investigation about the extent of the practice, resulting perhaps in a conclusion that it is financially worthwhile to begin a court challenge. When this happens everyone thought to have violated its view will be named. It might make for a large and interesting court case.

References

1. Placeholder for Ruiter Web material
2. *Feist Publications, Inc.* v. *Rural Telephone Service Co.* 499 U.S. 340, 111 S. Ct. 1282, 113 L. Ed. 2d 358 (1991).
3. van Belle, G. (2008) *Statistical Rules of Thumb*, 2nd edn. Hoboken, NJ: John Wiley & Sons.
4. Nimmer, M. B. and Nimmer, D. (2012) *Nimmer on Copyright*. New York: Matthew Bender.
5. US Copyright Office (2012) *Copyright Basics*, Circular 1. Washington, DC: US Copyright Office, Library of Congress. `http://www.copyright.gov`

Gerald van Belle is a professor in the Department of Biostatistics and the Department of Environmental and Occupational Health Sciences at the University of Washington in Seattle. Leslie Ruiter is a practising attorney at Stokes Lawrence in Seattle; she specialises in intellectual property.