The legal status of event data

Philip A. Schrodt, Ph.D.
Parus Analytics LLC
Charlottesville, VA 22901

schrodt735@gmail.com

Full URL

https://asecondmouse.wordpress.com/2014/02/14/the-legal-status-of-event-data/

Posted on February 14, 2014

The legal status of event data

Posted on February 14, 2014

So, for once, only a minimal level of snark. We're still in the process of recovering from the recent series of unfortunate events in the world of event data but, indeed, we appear to be recovering and there will be some interesting announcements forthcoming,[1] probably around the beginning of March, and certainly before ISA, as otherwise I could only attend in disguise. [update: done—see this link and now this link] One of the many rabbit holes I've been traveling through has involved a fair amount of research on the legal issues surrounding the creation and dissemination of event data, and while this is still fresh in my mind, figured I might as well share this.

Let us start with the obligatory caveat that I AM NOT A LAWYER. I repeat:

I AM NOT A LAWYER!

So this is not legal advice, not should be construed as such, and if you wind up doing 20 years to life of hard labor in solitary in Leavenworth for coding the 2006 military coup in Fiji using Agence France Press [2], that is not my responsibility. But from the perspective of an academic researcher reviewing the available material, I find that rather unlikely.

[But for those of you who are lawyers, and even more so, those of you who are teaching courses on intellectual property, this would make a pretty good little project, right? Or maybe it has already been done: I would be delighted to publicize any such efforts. Similarly, if those who are lawyers see error/omissions in what I'm saying here, I will be happy to post additions/corrections, or simply approve your comments on the blog.]

Here's where I *think* we stand on this issue.

Event data are a codified collection of facts, and in the 1991 decision Feist Publications, Inc., v. Rural Telephone Service Co., the U.S. Supreme Court ruled "that information alone without a minimum of original creativity cannot be protected by copyright." [3] Evidence that event data are a codification of facts rather than dependent on the expression of those facts is provided by at least the following

- The original expressive content cannot be recovered from the event data coding: in fact per linguistic theory, there are an infinite number of sentences that could map to a given event code.[4]

- The de-duplication algorithms we use—the ubiquitous "One-A-Day" filters based on source-target-event tuples—are completely indifferent to the expressive content;
- The fact that de-duplication is required on many of our events indicates that these are reporting factual information in the natural world, not some original creative effort (to the contrary, we do everything we can to eliminate fictional references).In fact, while it would be inconvenient, I have little doubt that a perfectly good event data set could be assembled *solely* from events that are multiply-sourced, and that might in fact be a less noisy set than the ones we currently work with. [5]
- While we do not code direct quotations out of scientific concerns, that further reduces the originality of the material involved in the coding;
- The mere collection of material—which in any case we are aggregating from a variety of sources—is insufficient to support a claim of copyright, per the 1996 decision inPublications International, Ltd. v. Meredith Corporation

The noun phrases which are provided for actors which are not in dictionaries (or rather will be in PETRARCH, our new Python-based event coder) are for the most part simply factual; in the rare cases where they have expressive content, for example a particularly flowery description of some hero or villain—which are generally not found in expository reports anyway—these are a very tiny proportion of the text and probably qualify as "fair use" under at least the research component of that doctrine.

And indeed, the second big protection is "fair use," though this varies with who is generating the data and why. It would probably be strongest when used in research and instructional uses in a state-run educational institution (which are, in fact, the circumstances of a large number of event data collection efforts) and weakest in a commercial, for-profit enterprise that has no pretension of research (which was the situation in *Feist*: the defendant simply produced phone books).

Based on the 1994 *Texaco* decision, the mere fact that an enterprise engaged in research is for-profit does not, by itself, mean that this is not "research" under "fair use." *Texaco* did, however, place an emphasis on whether the action deprives a copyright holder of revenue, and in my reading of subsequent cases, this has become an important factor. Keep in mind, the actions in *Texaco*—the large-scale copying of materials—go far beyond anything involved in event data production, but the issue of commercial competition could, conceivably, be relevant at some point, for example were one of the major news services to start providing their own event data. [6]

Source texts

Here the situation seems completely unambiguous: don't share source texts unless you have a clear license or other intellectual property right to do so (which, for example, ICEWS has within the US government, a major advantage of ICEWS for those users). This has implications for replicability, and lots of people get unhappy when you say you can't share the source texts, but this is about as close to a black line as we've got, and I've got a long mini-sermon on this that I inflict whenever I'm lecturing on event data. I repeat: do not share copyrighted source texts.

Bummer. And I've long wondered why the news services—whose interest, one would think, would be in making historical text series more widely available so that more people would subscribe to their *current* news services—did not make available some sequences at a reasonable cost: the marginal costs for doing this would be very low as they already have the data.

Well, it turns out they have: the Linguistics Data Consortium has produced the GigaWord corpus which contains all of the major international news sources for roughly 2000-2010, and this can be licensed for research purposes at very reasonable rates (and is free if your institution is already a member of the Consortium). Potentially a game-changer, and more on this at a later date.

URLs are not subject to copyright—they are information rather than expressions of creative content—so there is no problem with the events reported in a data set containing URLs.

Ontologies (WEIS, CAMEO, IDEA, etc)

The copyright status of an ontology or coding framework is a complete morass, though the CAMEO ontology has an open source license, and to the best of my knowledge VRA has not made copyright claims to IDEA. Were such claims made—and we have absolutely no intentions of doing so with CAMEO—they would be difficult to support given that both systems are based on earlier work: CAMEO was based predominantly on WEIS; IDEA deliberately on a series of ontologies, including WEIS, COPDAB, World Handbook, and CAMEO. These coding systems were widely available and uncontested in the academic world, and, as they were created before the days of intellectual property trolls, did not originally have explicit licenses. [7]

All of this is to say that while it is conceivable that an ontology could be copyrighted, it is not obvious, but more importantly, it is irrelevant so long as you are content with CAMEO and open-sourced extensions of CAMEO. And probably the same for IDEA, though I can't speak for

them.

Patents

While past is not necessarily precedence, the event data field has thus far been almost completely immune from patent claims, egregious or otherwise. The core concepts have been around since the 1960s, virtually all of the core work has been publicly funded (mostly by DARPA and NSF), and the automated coding technology has been around, without challenge, since the early 1990s and open source since the 2000s. Most of the ancillary software used in current systems is also open source. The intellectual property departments of both the University of Kansas and Penn State examined the KEDS and TABARI/CAMEO systems at various times and decided not to pursue patent claims despite the option to assert such rights under the Bayh-Dole Act.

The amount of prior art accumulated in the half-century or so of open work in this field is huge, and in our experience, people just getting into the field with what they believe to be brilliant new ideas are almost invariably merely pursuing plausible and obvious approaches that were either proven to be dead-ends decades earlier, or are standard procedures: a month of coding can save an hour in the library. Granted, the U.S. patent process is monumentally screwed up at the moment and would probably seriously consider a patent on entertaining kittens using a ball of yarn, and seemingly anything *could* happen, but it seems highly unlikely that claims against the core concepts and technologies would survive even a cursory examination of prior art.

In the absence of patent reform, of course, trolls will wantonly pillage and plunder, as that is the nature of North American trolls [8] and there is no guarantee this will not happen in this field. But in terms of the core concepts and technologies, that's all they are—trolls with no defensible claims—and we can only hope trolls will focus on more important issues, like the multi-billion-dollar life-altering issue of whether it is permissible to use rounded-corners on a smart phone interface. Undoubtedly precisely the sort of thing the Founders had in mind with the phrase "the advancement of useful knowledge and discoveries."

That whirring sound you hear northeast of Charlottesville is James Madison spinning in his grave. I digress…

A Few Additional Thoughts

1. The norm that facts derived from copyrighted material do not inherit that copyright is actually

implicit in all secondary data sets used in the study of political conflict. Were this*not* the case, not only would event data be affected, but so would every other data set assembled by reading copyrighted material—COW, MID, Polity, ACLED, everything. We've simply assumed all along that this is okay. As it happens, it is.

2. A curious side-effect, as it were, of this appears that the *intellectual property rights* to data derived from a set of texts that were obtained illegally—let us say WikiLeaks—would not be affected by the origin. While there was a great deal of *strum und drang* from various U.S. government sources on the initial release of WikiLeaks [9], this seems to have subsided. Possibly on this issue, possibly on that fact that there wasn't much in the WikiLeaks that hadn't long been suspected: I believe it was Michael Gerson who observed that the big surprise of WikiLeaks was that there were no big surprises. [10]

This does not, of course, suggest that data coded from questionable sources are without issues, in particular if one of the questions is the identity of the source texts being coded. But that is an issue of scientific integrity, not intellectual property.

3. If your web crawlers are well-behaved and don't go places they aren't supposed to go, you are far less likely to end up in any trouble. Duh… The courts have definitely been more sympathetic with plaintiffs who had internal sites essentially hacked than with those where the information is readily available. This, by the way, is why you aren't going to be seeing a lot of Agence France Press material in near-real-time event data sets: it is the only one of the four major international sources (Reuters, BBC, Xinhua and AFP) that does not makes its reports easily available (e.g. through RSS feeds).

4. There are still more legal issues than one would like [11] that are in flux, with potentially critical court cases still pending, contradictory rulings at lower levels, and warnings about subjective judgements. Furthermore, the event data community are very, very small fish in this pond: the critical open issues involve the web-based data collection practices of billion-dollar enterprises such as Kayak and other price comparison sites, and this is where the "facts" emphasized in *Feist* and the "commercial interests" emphasized in*Texaco* are coming head-to-head. There are definitely some unresolved issues here, though unless these rulings wildly change the status quo—or if event data analysis becomes a wildly successful commercial venture [12]—they probably will not affect event data collection.

That said, as in outsider to the legal world, it is interesting to see the courts gradually coming to consensus positions on these issues which are reasonably coherent: they aren't there yet but they definitely no longer put up with some of the more outrageous claims that were circulating

in the early days of the Web. Patent reform is at least on the agenda now—including no less than Obama's 2014 State of the Union address—though when and if "the best Congress money can buy" will do anything about it remains to be seen: they sure don't seem in much of a hurry. But things are getting clearer.

Once again, I AM NOT A LAWYER, and I would be delighted to share further insights on these matters from those who are.

Footnotes

1. And I haven't forgotten "Feral + 7", though at the rate things are going, it will be "Feral + 8."

2. I said "minimal" snark, not "zero"

3. http://en.wikipedia.org/wiki/Feist_v._Rural

4. Okay, realistically, it is not "infinite"—and in any case, it's only countably infinite even in theory—but it is really large. That's the big problem in automated coding: if there were only a small number of ways to describe a category such as "threat of military intervention", or any other code, we'd have really small dictionaries and really accurate data coding. There isn't, and we don't.

5. This, of course, runs against the "needle in a haystack" theory of event data analysis—if we had only had more localized data, we could have predicted that the self-immolation of an obscure fruit vendor in a marginal town in Tunisia would trigger the Arab Spring. But there are plenty of reasons to think this is an illusion arising from hindsight bias, and in fact big changes come from big indicators, properly analyzed. Not everyone agrees.

6. But why would they: the derived data cannot be copyrighted, so there is little incentive to do so. Such data could still be licensed in a fashion that it could not be provided to the public, and some of this is available, though not, to my knowledge, at any significant scale: We hear far more frequently complaints that such data is *not* available even though people want it.

7. CAMEO does: originally GPL and now switching to the MIT License. We are doing the same with all of our coding dictionaries.

8. Contemporary Nordic trolls, in contrast, appeared to have mellowed. The real trolls, not the

ones wearing expensive suits.

9. Though these threats seemed remarkably selective: at the same time I was hearing from some individuals that they were all but threatened with an indefinite sojourn in Guantanamo if they even *thought* about using the WikiLeaks data, I was seeing extended analyses of it other places. Particularly certain elite universities. Some animals are more equal than others. Or maybe some institutions have better lawyers than others.

10. The same cannot be said for the next big set of leaks, from Edward Snowden.

11. Unless one is an intellectual property lawyer, or troll, in which case this is called "job security."

12. Hey, a guy can dream… But seriously, there is a great deal of commercial potential here, but it is going to have to look like political polling looks like now (not like it looked in the days of *Literary Digest):* multiple sources, some public, some private, supported by a community of professional analysts with a sophisticated scientific understanding of the strengths and weaknesses of the approach, as well as with sufficient open-access data to provide a reference point for determining when something doesn't look right. Flim-flam artists with one-size-fits-all proprietary black-boxes have no place in this model, but plenty of other approaches—academic, NGO, government and commercial—will.

Posted in Methodology | 9 Comments | Edit