Clifford A. Lynch, "Open Computation: Beyond Human-Reader-Centric Views of Scholarly Literatures," *Open Access: Key Strategic, Technical and Economic Aspects*, Neil Jacobs (Ed.), (Oxford: Chandos Publishing, 2006), pp. 185-193.

Introduction

This chapter is probably the book's most speculative, in that it discusses broad-based computational access to scholarly literatures — a collection of developments that are likely to happen largely as a consequence of increasing open access. Traditional open access is, in my view, a probable (but not certain) prerequisite for the emergence of fully developed large-scale computational approaches to the scholarly literature. It may not be a sufficient prerequisite, particularly if the legal and systems architecture frameworks currently being developed and deployed to support traditional open access are not quickly adjusted to accommodate the needs of open computational access. Indeed, even if such accommodations are made, and if appropriate open access provisions were to be universally established for all scholarly works going forward, there is still an enormous, long-lasting problem with the established historical base of scholarly literature. While scholars tend to focus largely on new contributions to the literature, computational technologies value and demand scale and comprehensiveness in the literature base that they address; constraints on the use of the historical literature will continue to represent a massive barrier to such computational uses. A move to open access may not help much with this retrospective material.

I am confident that the other chapters of this volume have done a fine job of describing the various access models and practices that are being characterized by the term "open access" in different settings, and the virtues and benefits that they share in terms of democratizing access to varying degrees and in varying dimensions. Indeed, we are seeing some of these benefits — for example, access by readers in developing countries — today not just as a result of author and publisher choices about open access, but sometimes even as a result of publisher practices that could only be termed any kind of real "open access" by the most imaginative and dedicated public relations functionary. Similarly, we are seeing some developments in computational access to literature —most prominently for indexing (think of Google and similar search engines, and their explicit arrangements with publishers, or their efforts to implicitly compromise with publishers within the framework of copyright law's fair use provisions through the indexing of copyrighted text but the presentation only of brief "snippets" of copyrighted material) — outside

of the open access framework. Some publishers are also making explicit provisions for experimental text mining, or allowing rehosting under license agreements which opens the door to arbitrary computational exploitation or representation of their material within closed organizational contexts.

The case for the benefits of open computational access to the scholarly literature is also much more complex than the arguments usually marshaled for traditional open access — in part because these benefits are indirect, and in part because they are still considered largely speculative and unproven. They are indirect in that they merely open the way for various players with good ideas to advance the progress of research and scholarship in perhaps new and perhaps more accelerated ways; presumably, in the long run, such research progress is of value to everyone. (Note that, paradoxically, computational access to a scholarly literature for the purposes of indexing may also make that literature more economically valuable in the non-open access case, in that it may increase demand: witness the interest of commercial journal publishers in having their material indexed in search engines.)

The benefits are speculative in the sense that we are just beginning to understand and demonstrate what we can accomplish, computationally, with large scholarly literature corpora. A number of inter-related technologies such as text mining and analysis are very active, vibrant and well-funded research areas, attracting extensive participation and investment from government and industry as well as academia. And, more recently, we are seeing experiments not only in computing on literatures to derive insights, but in the actual *rehosting* of literatures within new analysis, usage and curation environments: here a scholarly literature is actually imported into a new usage environment that adds value through computation and perhaps also through social interaction — leading examples of this might include the work of the US National Center for Biotechnology Information at the National Library of Medicine for the molecular biology literature, or the fascinating experiments carried out by Greg Crane and his colleagues at Tufts University in the Perseus Project. But it is important to recognize that while researchers focusing specifically on computational manipulation of scholarly literatures are reporting great advances in their work, I think that the broad community of working scholars remain to be convinced of the critical future contributions of such technologies.

This brief chapter begins an exploration of both the technical and the legal issues involved in enabling widespread application of computational

techniques and technologies to the research literature. There are many more questions than answers at this stage.

Technological Opportunities

Let's perform a thought experiment. Let us suppose, for the moment, that the only copyright encumbrance on the scholarly literature was that of attribution; articles could be freely replicated, and arbitrary computations could be performed upon these articles. The results of these computations could be freely and widely employed and shared. In such a world, what do current technology trends suggest might be done with the collection of articles that constitute the vast majority of the scholarly literature in so many fields?

Clearly we would see the widespread creation of copies of the scholarly literature, or very sizeable subsets of this literature; these copies would reside in a great range of personal, workgroup, and disciplinary settings for convenience of access and searching. Storage is getting very cheap, and students and researchers cannot always count on the ubiquitous availability of very inexpensive broadband connectivity. We would see these copies of the published literature federated in various ways with unpublished, preliminary, and proprietary materials, forming knowledge bases that were unique to specific researchers, research groups, corporations and other entities. These federations would be facilitated by the ability to computationally re-arrange and re-structure the literature.

We would also see an explosion in services that provided access to this literature in new and creative ways. Such services would also incorporate specialized vocabulary databases, gazetteers, factual databases, ontologies, and other auxiliary tools to enhance indexing and retrieval. They would rapidly transcend access to address navigation and analysis. One path here leads towards more-customized rehosting of scholarly literatures and underlying evidence into new usage and analysis environments attuned to the specific scholarly practices of various disciplines.

We would also see a move beyond federation and indexing to actual text mining and analysis, to the extraction of hypotheses and correlations that would help to drive ongoing scholarly inquiry. Indeed, the literature would be embedded in a computational context that reorganized and re-evaluated the existing body of knowledge as new literature became available. Initially, we would likely see a series of leap-frog breakthroughs as these technologies rapidly advanced, but I think it is likely that, over time, the state of the art in text mining and analysis would stabilize or

converge to a point where new computations over the common literature base using the best state-of-the-art tools would only produce, at best, modest incremental advances. At this point the key leverage for wringing new discoveries from the literature would pivot on two points of competitive advantage. The first would be early access to and rapid integration of new contributions — including, most likely, preprints that had not, at least yet, been peer reviewed, and perhaps segments of the historical literature base newly entering the digital domain. The second would be the ability to quickly and successfully integrate and exploit unreleased or non-public information — not just unreleased preprints, but data, including negative data that had never seen publication, in conjunction with the common shared public literature base and ancillary public data and knowledge bases.

It's also near certain that these innovations would not apply to all scholarly disciplines uniformly. Areas such as biomedicine or chemistry, where much of the literature is relatively well-structured and where a base of investment in the development of auxiliary knowledge structures such as factual databases, ontologies, specialized vocabularies and vocabulary mappings and similar tools has been extensive, would likely be fertile ground for early advances. Indeed, in these fields we are already seeing the beginning of a re-evaluation of authorial practices that propose the incorporation of markup to facilitate exactly such computational processing of the literature — consider the work of scholars such as Peter Murray-Rust in chemistry, or the various proposals for specialized markup languages in areas as diverse as history and molecular biology. (In other web settings, these efforts are being characterized as "micro-formats".) Other "hard" sciences, and certainly many branches of the social sciences, would yield results more slowly. Many of the humanities would remain recondite. And, of course, changes in disciplinary practices of scholarly authoring would have a great influence: to the extent that new articles in the public literature base are routinely structured to facilitate computational verification, integration or correlation, these disciplines would presumably see greater payoffs for the applications of textual mining and analysis. One can even imagine, in certain highly competitive and commercially significant fields, deliberate release of what is in effect *disinformation* to divert the attention of research driven by text mining and literature analysis in deliberately unproductive directions.

Finally, in an environment largely unencumbered by intellectual property issues, it's likely that the tension between distributed and centralized computation will be resolved primarily according to the mandates of technical simplicity and universality rather than being shaped by the contortions enforced by licensing agreements and the services that individual publishers choose to make available. While in theory there's a

performance tradeoff between the choice of moving an interoperable, transportable network based representation of the computation to the servers where the data resides, and doing remote execution of procedural computational code on this remote database — the concepts implicit in the seminal work of Kahn and Cerf in their classic report "The World of Knowbots" for example — and the infinitely simpler model that just *copies* all relevant data to a local store upon which computation occurs, it seems to me most probable that in the absence of intellectual property concerns and licensing constraints that the obvious and universally understood framework of creating local copies will triumph. The practical will dominate the theoretically optimal. The local replication model is so much simpler and more reliable and predictable than the alternatives, where it seems likely that every remote execution environment will have its local idiosyncrasies and constraints, and where large-scale literature analysis will have to adapt to the variety of interfaces offered by different publishers. These interfaces will inevitably incorporate a series of tradeoffs that publishers design to prevent computational access from allowing actual copying of the literature base (consider, for example, the as yet nebulous Open Text Mining Interface proposal — see http://blogs.nature.com/wp/nascent/2006/04/open_text_mining_interface_1.html).

And it also avoids the very real additional complexities of correlating and consolidating results from multiple remote computations executing in a range of remote, most likely publisher-based, literature silos. So it seems absent proprietary content ownership constraints, the dominant paradigm and the fastest path to the payoffs of textual mining and analysis, of the application of new digital library technologies designed to import and host literatures in ways that add value to that literature, will be to accumulate a local representation of the relevant literature, and then to perform ongoing computations on that literature locally.

Real-World Conundrums

Let's move on from our idealized thought experiment.

We are very unclear today about whether even the systems that claim to offer "open access" to collections of scholarly literature are being — or should be — designed to permit simple, large-scale *replication* of these collections in order to facilitate the creation of local resources that can be computed upon. This is both a technical question (is it easy to make a copy of the full collection?) and a legal one (concerning what uses are allowed under the implicit or explicit licenses). So one set of questions is about whether we will provide the enabling technical infrastructure and

legal permission that facilitate computational access to scholarly literatures even in the context of the various definitions of open access.

For the proprietary scholarly literature, today's license agreements generally preclude the creation of large literature subsets external to the publisher's site, and, indeed, user attempts to perform large-scale downloading have raised alarms and led to difficult and awkward discussions involving publishers or aggregators, licensing institutions (universities) and end users about the appropriateness and legality of creating such local mirror databases. At least in theory, if the creation of local copies of literature databases derived from large-scale downloads from various publishers becomes a standard and accepted practice for faculty at licensing universities, one might presume — or at least hope — that most publishers (though there would undoubtedly be holdouts) would revise and adapt their license agreements to recognize and permit such practice.

For open access materials, the creation of large-scale collections of copies is often ambiguous in the absence of specific permissions; we are moving towards a legal understanding that suggests public-access content is available for reading, but the ability to re-host long lived copies is less clear. Open access content offered under terms such as the Creative Commons license agreements reduces the uncertainty here — but not necessarily for downstream *use*, as I will shortly discuss.

Clear legal rights to make large-scale copies of the literature are just the beginning of the legal conundrums that will create barriers to open literature computation. What is the legal status of the results of computations upon such copies? What is the legal status of a re-hosting of these materials within a new computational context that facilitates linkages, re-presentation, exploration and analysis of a literature corpus? As far as I can determine these questions are largely unexplored and unresolved in law – both case law and legislation. We have the well-established concept of a derivative work — for example, a translation or a work; creating a derivative work requires permission from the rights holder of the original work. At least when the process of creating the derivative incorporates substantial new human intellectual effort, new rights are overlaid upon those of the original author in the ownership of the derivative. It is completely unclear whether an algorithmic computation produces a true derivative work or whether it is just considered a re-presentation of the original, but in either case, rights in the algorithmic product certainly seem to include claims from the source work. In cases where the computation process takes as input an entire literature base, consisting of perhaps hundreds of thousands of individual works the authors of *each and every one* of these input works might have

a claim on the output. It is not at all clear that we can make the case that only a small and selected subset of the input works made a material contribution to the output and thus have claims upon that output. Is it the case, for example, that if we rerun the algorithm on a copy of the literature base excluding a single article and get the same result as if we had not excluded that article that we could argue this proved the result was independent of the source article in question.

The sheer volume of rights that need to be cleared may effectively preclude the application of computational technologies to large literature bases. If the literature base is offered by a publisher operating within a framework where authors transfer copyright to the publisher, then presumably the publisher could grant the necessary rights to allow meaningful text mining of the corpus, or the importation of the corpus into a new analysis and presentation environment. (Whether publishers will actually be willing to do so is another, and doubtful, proposition.) In cases where the corpus is produced through open access type arrangements, unless the transfer of (most likely nonexclusive) permissions to the host of the corpus are crafted with great care and specific focus on the computational opportunities, text miners and those wanting to import materials into new use environments will have to engage in completely impractical and unrealistic author-by-author clearing of permissions.

The Creative Commons (CC) license is a good case study here. It is a very valuable tool in reducing ambiguity about the permitted uses of scholarly works, but it also illustrates how little thought has been given to computational applications. The CC license offers authors options about whether to permit the creation of derivative works, and also options about whether they can insist on author attribution in downstream uses of their works. Permission to create derivative works seems to be a clear prerequisite for computational use of articles; yet this is rather different that the way that this choice is presented to authors creating a CC license to their works today. Even the attribution requirement may be a source of problems — will we have to list author attributions for every work in a literature corpus as part of the attribution for any computational result from this literature corpus? And, if so, how will we practically meet this mandate? Is there a need for a new Creative Commons provision that specifically deals with authorizing and enabling the potential to text-mine, re-host or otherwise compute upon works offered under CC licenses?

Creative Commons is beginning to examine some of these issues through its Neurocommons initiative within the Science Commons program.

Preliminary Conclusions

As the scholarly literature moves to digital form, what is actually needed to move beyond a system that just replicates all of our assumptions that the this literature is only read, and read only by human beings, one article at a time? What is needed to permit the creation of digital libraries hosting these materials that moves beyond the "incunabular" view of the literature, to use Greg Crane's very provocative recent characterization. What is needed to allow the application of computational technologies to extract new knowledge, correlations and hypotheses from collections of scholarly literature?

Part of the answer is legal. Clearly we need freedom to copy, rehost, repurpose and compute upon the components of this literature. (Note that while I have not explicitly discussed large-scale retrospective digitization projects here, this is equally applicable to these efforts, not just to new contributions to the scholarly literature.) We need license terms that minimize or render moot the uncertainties surrounding the creation of derivative works and possibly even the requirements of attribution for source materials that have contributed to the production of these derivative works. The Creative Commons licensing framework offers a particularly urgent and compelling environment for exploring these requirements.

The other part of the requirement is technical. We need to see provisions in hosting systems for large-scale replication as well as item-by-item downloads of occasional copies of parts of the scholarly literature. While in theory this need might be mitigated by the availability of interfaces that allow us to export computations to repositories, I suspect that these will not fully satisfy the needs for literature analysis and for new content analysis and synthesis environments that assume the ability to rehost materials.

The opportunities are truly stunning. They point towards entirely new ways to think about the scholarly literature (and the underlying evidence that supports scholarship) as an active, computationally enabled representation of knowledge that lives, grows and interacts with its contributors rather than as a passive archive or record. They suggest ways in which information technology can accelerate the rate of scientific discovery and the growth of scholarship. It would be a disgrace if we allowed the inertia of historic scholarly publishing practices and the intellectual property arrangements that underlie these patterns to foreclose such opportunities. Open access offers an important simplification and reduction of the barriers if its development is shaped in

a way that is responsive to these opportunities, although it is certainly not a panacea in its current form.

What is ultimately at stake here is a fundamental reconceptualization of the roles and uses of scholarly literatures and the evidence that supports scholarship. The traditional intellectual property framework of scholarly publishing is not hospitable to this reconceptualition. The implications of resolving this incompatibility will ultimately have far more extensive ramifications than what we might today characterize as the “traditional” open access movement; but they will be crucial to the future of science and scholarship.