Copyright & The Distributed Lexicon

Hugh Paterson III

Grimes (2006) describes a distributed lexicography project. Technological and sociological advances in the last twelve years have normalized digital collaboration. Workflows for distributed teams with tools such as FLEx and WeSay have been optimized to allow for many community members to collaborate on lexicon or dictionary making projects. Built-in sync tools allow teams to work in distributed environments for both local network syncing and synchronization through languagedepot.org. For the researcher or project manager, often working with a distributed team means dividing the workload so that team members can contribute in their areas of strength. Many times this means that a single "dictionary entry" might have edits or partial contributions from multiple people.

Many Institutional Review Boards (IRB) proposals never address the issue of copyright, rather they focus on the ethical treatment of participants in linguistic projects, and issues such as informed consent. Ethical issues related to IRB policies are discussed in the literature (e.g., van Driem 2016). How projects should deal with licensing intellectual property developed within the context of a project is discussed less frequently. Newman (2011:454-456) provides some templates for the transfer of copyright claims from project participants to a researcher. But to some this seems unethical. For instance those who suggest that all rights should be retained by a language community. Newman (2007) addresses some of the concerns that researchers can have towards the legal use of their own (and others') field data. Undiscussed are highly collaborative project and the tangible creations from these projects.

When left unaddressed intellectual property claims can become catastrophic for a project like a dictionary, if a contributor (or former contributor) wanted all of their contributions removed. Frawley et al. (2002) describe several such cases. On the technical implementation side, tools like FLEx do not provide managerial interfaces to the underlying data to track changes by specific project contributors.

I present and discuss two options for overtly addressing copyright interests for teams and groups who want to contribute to a project. I discuss some best practices which projects can implement in their operational strategies. I believe the presented solution to be viable in at least 175 countries based on their shared commitment to copyright law via the Bern Convention (1979).

I take the offered solution and present it as applied to the distributed team context – such as the context of team-based lexicon/dictionary building.

Hugh Paterson III. 2019. Copyright and the Distributed Lexicon. Presentation at the 6th International Conference on Language Documentation & Conservation (ICLDC), Connecting Communities, Languages & Technology, February 28-March 3, 2019, at the Hawaiʻi Imin International Conference Center on the campus of the University of Hawaiʻi at Mānoa. http://icldc6.icldc-hawaii.org/

Introduction: Context

Scope of discussion

- This is not legal advice. This is social discussion on legal issues which impact language resource production.
- The scope of this presentation is law as it is practiced in the USA.
- Principals in this presentation are extensible to the 176 countries who have signed the Berne Convention which directly relates to copyright.

Some things to remember when talking about law

- Laws represent social contracts.
- Laws only come into existence if somebody feels they can make money off of it (or not lose more money than they would otherwise make by enacting the law).
- Social contracts evolve more quickly than law does.
- Enshrining a social contract in written law is not always the healthiest social option.
- We do not know what a law means, (how it should or shouldn't be applied) until a court acts on it.
- It costs money to enforce law.

Legal Pluralism

Legal pluralism exists when two separate governing bodies assert legal authority over a jurisdiction or situation. The common situation is to have a state actor and a non-state actor.

> Example 1: A tribal council and United States federal law might both claim authority over fishing rights, or watershed protection.
>
> Example 2: Sharia law and National law in Nigeria.

Human rights and minority peoples literature contain many arguments for the codification of minority peoples social contracts within the majority culture's legal frameworks these arguments often include the issue of Intellectual Property.[1]

[1] When reading Academic literature on this topic which is written in English, one must always carefully consider the legal context to which the author(s) are referring. Though the issues may sound similar, legal contexts may actually frame the issues very differently.

Some facts

Copyright already exists for minority peoples if they are citizens of countries which have signed the Berne Convention.

- They may not be aware of these these legal rights.
- There may be social structures which prevent the efficient assertion of these rights through legal channels.

Copyright can only be held by legal entities — individuals and legally registered corporations.

What is *Copyright*?

Copyright is one type/category of Intellectual Property Rights, in US code it is defined in Title 17 chapter 1 section 106.
Intellectual Property Rights and what they cover vary by jurisdiction.
In the USA Copyright does not include:

- Sui generis database rights
- Moral rights

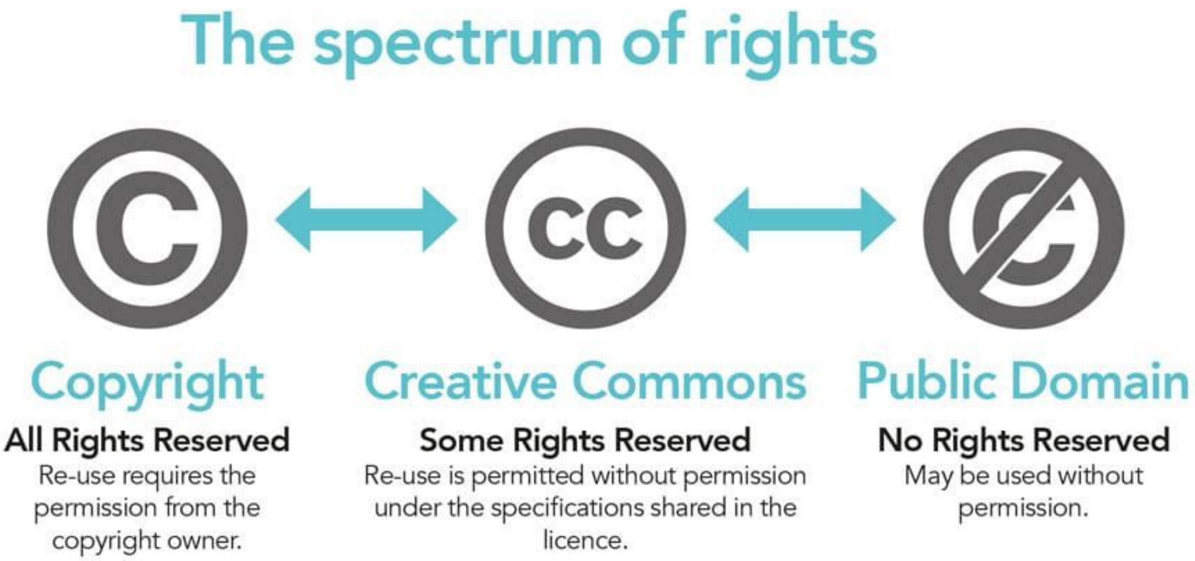
These concepts do exist in other legal frameworks.

Since 1978, copyright is now automatically ascribed, whereas prior to 1978, it used to be required to be applied for, or published with notice.
Creative works which are not under copyright are legally known as being in the "public domain". Public domain does not mean publicly accessible. That means that the following rights are not reserved for the creator.

Copyright is the right to:

- to reproduce the copyrighted work in copies or phonorecords
- to prepare derivative works based upon the copyrighted work
- to distribute copies or phonorecords of the copyrighted work to the public by sale or other transfer of ownership, or by rental, lease, or lending
- in the case of literary, musical, dramatic, and choreographic works, pantomimes, and motion pictures and other audiovisual works, to perform the copyrighted work publicly
- in the case of literary, musical, dramatic, and choreographic works, pantomimes, and pictorial, graphic, or sculptural works, including the individual images of a motion picture or other audiovisual work, to display the copyrighted work publicly
- in the case of sound recordings, to perform the copyrighted work publicly by means of a digital

audio transmission.



The spectrum of rights

Copyright

All Rights Reserved

Re-use requires the permission from the copyright owner.

Creative Commons

Some Rights Reserved

Re-use is permitted without permission under the specifications shared in the licence.

Public Domain

No Rights Reserved

May be used without permission.

Image source: Steel 2018

What is a *Lexicon*?

Lexicons are the foundational data set upon which derivative works like dictionaries are created.

Characterized by these attributes

- *Corpus like* with a data structure
- May include ***facts*** and ***creative works***
- Usually focused around *Constructions*, *Words*, *Roots* & *Morphemes* their contexts, articulations, and meanings.

Lexicons can be used to generate dictionaries. But can also be used to generate other useful tools.

Dictionaries: One output of a lexicon

Dictionaries are reference works which include the formatting, typography, of the lexical elements of a language.

- Different kinds of dictionaries can be generated from a single lexicon
- Different modalities of a dictionary can be generated from a single lexicon

Spell Checkers: One output of a lexicon

Database, or rule based, or combination of both as an output format

Grammar Check: one output of a lexicon

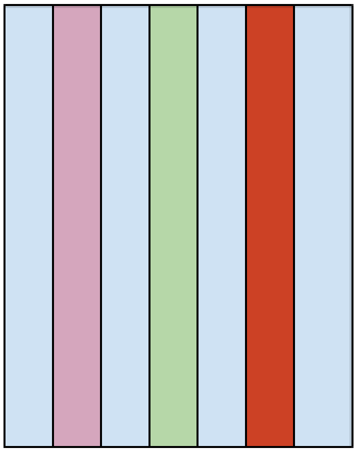
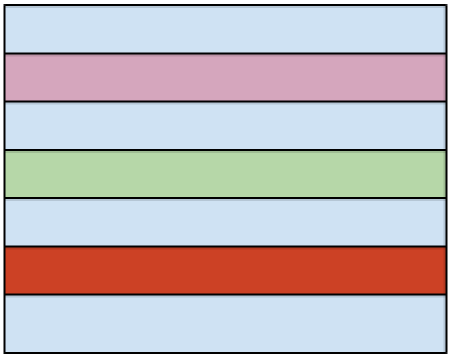
Database, or rule based, or combination of both as an output format

What do I mean by *Distributed*?

How it is built.

There are different relationships between the contributors of a lexicon.

- One person — single author.
- Several co-Authors each focusing on a specific domain of the lexicon, but responsible for the whole entry. I.e. fish, plants, humans, etc.
- Several or Many Contributors with many people touching many different portions of many different entries.

Co-authors	
Onion Model	

Why do we acknowledge the presence of Copyright?

The Austin Principles #2

> *Data citations should facilitate giving scholarly credit and normative and **legal attribution** to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.*
>
> In linguistics, citations should facilitate readers retrieving information about who contributed to the data, and how they contributed, when it is appropriate to do so. One way to do this is through citations that list individual contributors and their roles. Another way is by using citations that link to metadata about contributors and their roles.

Ownership

- Un-ownable
- Individual
- Corporate
- Work-for-hire

Just as much as we have to be aware of minority language communities' perspectives on intellectual property, we need to be aware of the legal frameworks and rights assignments of the national laws impacting our work — including work-for-hire claims or the assignment of rights by law to the organization of the researcher. That is, Universities and NGO's may in many cases have silent or unacknowledged intellectual property rights in the works which linguists and collaborating communities create, via the work-for-hire doctrine. These rights need to be addressed just as much as the rights of the minority language contributors need to be addressed.

Redaction

When left unaddressed intellectual property claims can become catastrophic for a project like a dictionary, if a contributor (or former contributor) wanted all of their contributions removed. Frawley et al. (2002) describe several such cases. On the technical implementation side, tools like FLEx do not provide managerial interfaces to the underlying data to track changes by specific project contributors.

GDPR

While not technically a right granted via copyright, The EU General Data Protection Regulation impacts what we can and cannot do with linguistic data. The GDPR applies to organizations who process data. So while copyright might be an expiring expiring protection for data at rest or in a static form, GDPR is a protection of data in transit, and in a dynamic form. Transforming lexicons to another form would qualify as "processing data".
Art. 89 of the GDPR

> "Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes"

provides nation-states the option to provide exemptions. But we don't know what these are, who they apply to, or in what contexts they apply, or how they will very from country to country.

Outbound restrictions vs. Inbound restrictions

When it comes to freeing data, or making sure that the data we store together is treated with the same freedoms as it is used by others, we can have one of two approaches. The first approach is to license the data after it has been created to restrict or "free" the data users to be able to use the data as the terms of use dictate. However, a second method exists, which says: "'This data will be available in these ways:____'. 'Who would like to contribute to this collection of data?'" And then only allowing contributors to contribute if they agree with the terms of contribution.
Licenses focus on what can happen relative to the reserved rights of copyright with a dataset after it is created. The authority to license data derives from the authority of ownership. In contrast to licenses which focus on downstream data users, the CLA — Contributor License Agreement, focus on gaining consent prior to data inclusion in a data set. The CLA is a common tool used in open source software development, where a company or organization allows contributions from outside of its set of direct employees. Some organizations such as the W3C require contributors to sign a CLA to be on their mailing lists. A CLA is much more inline with the linguistics industry standard of IRBs and Informed consent.

Copyright assignment

CLAs generally assign copyright to one of two types of entities: The sponsor of a project or the contributor of a project. Projects in and of themselves can not hold copyright as they are not corporations and legally recognized entities. For the purpose of copyright only the individual and the corporation are recognized as legal possibilities.

When the copyright is assigned to the contributor a license is then used to declare how the new creative work can be used. An analogous process happens in OpenAccess edited volumes where authors retain copyright.
In contrast, when the sponsor of a project is transferred copyright, there is generally an agreement in the project for how the information will be shared. Wikipedia uses this model.

If the copyright is assigned to the sponsor of a project the sponsor can change the license at any time. If the contributor retains copyright but the content is licensed to the sponsor, if later sponsor needs to change the license then it is extremely difficult and costly to do.

Two web services for implementing CLAs on github

- https://cla-assistant.io
- https://www.clahub.com

If the data is in the public domain it might still be subject to GDPR regulations. Copyright is about data at rest, GDPR is about data in motion.

OpenData Licenses

The choice is between public domain and a license, not between a unlicensed data and licensed data.

Why not use an “open license”?

- Most licenses restrict the flow and or use of data, thereby limiting the data’s utility and value.
- Data is like monetary currency.
  If it is hoarded it loses value.
- We shouldn't over license data. Our products should strive to build communities, licenses by their nature strive to exclude certain interests.
- No Open Data license currently has a data processing release clause.

Problematic licenses

- CC-BY-**ND** - As data we want it to remain relevant. This restriction means ***no*** updates to the data set.
- CC-BY-**NC** - ‘Non-commercial’ does not have a legal definition. Universities are engaged in commercial activities.
- CC-BY-**SA** - Can prevent commercial adoption
- ODbL - is a **SA** equivalent data license.
- CC-BY is not designed for code. Version 4.0 does address *Sui generis* database rights. CC-BY may not be as liberal as CC0, for inclusion in some digital tools.

Solution

My preference would be for a CLA which puts content in the public domain, and also contains a clause which permits unlimited processing of the data.

Such a solution will not be amenable to every context. And results will vary based on how the various kinds of situations are presented to communities.

Referenced Items

Citeded

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. doi:10.1515/ling-2017-0032.

Bern Convention Treaty. 1979. Berne Convention for the Protection of Literary and Artistic Works. World Intellectual Property Organization. Accessed: 27 August 2018. http://www.wipo.int/wipolex/en/treaties/text.jsp?file_id=283698

van Driem, George. 2016. Endangered language research and the moral depravity of ethics protocols . Language Documentation & Conservation. Vol. 10: 243-252. http://hdl.handle.net/10125/24693

European Union. General Data Protection Regulation (GDPR) Article 89: Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 ; Regulation (EU) 2016/679 (General Data Protection Regulation) version of the OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018 https://gdpr-info.eu/art-89-gdpr

Frawley, William, Kenneth C. Hill & Pamela Munro. 2002. *Making dictionaries: preserving indigenous languages of the Americas*. Berkeley, Calif: Univ. of California Press.

Grimes, Charles E. 2006. One dictionary, one language, one team, but different locations? *Paper presented at Tenth International Conference on Austronesian Linguistics. 17 - 20 January 2006. Puerto Princesa City, Palawan, Philippines.*

Newman, Paul. 2007. Copyright Essentials for Linguists. Language Documentation & Conservation. Vol. 1(1): 28–43.

Newman, Paul. 2011. Copyright and other legal concerns. In *The Handbook of Linguistic Fieldwork*, ed. by Nicholas Thieberger, pp. 430-56. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199571888.013.0020

Steel, Graham. 2018. License to share: How the Creative Commons licensing system encourages the remixing and reuse of published materials. Research Outreach (105). 06–09. doi:10.32907/RO-105-0609. https://researchoutreach.org/articles/license-to-share-how-the-creative-commons-licensing-system-encourages-the-remixing-and-reuse-of-published-materials/ (2 March, 2019).

United States Code Title 17. Chapter 1 Section 106. https://www.govinfo.gov/app/details/USCODE-2010-title17/USCODE-2010-title17-chap1-sec106/summary; https://www.law.cornell.edu/uscode/text/17/106

United States Copyright Office. 2017. Works Made for Hire. (Copyright Concepts). Circular 30 Revised: 09/2017. https://www.copyright.gov/circs/circ30.pdf

Consulted

Austin, Peter K. 2010. Communities, ethics and rights in language documentation. (Language Documentation and Description 7) 34–54.

Anderson, Jane. 2005. Indigenous Knowledge, Intellectual Property, Libraries and Archives: Crises of Access, Control and Future Utility. *Australian Academic & Research Libraries* 36(2). 83–94. doi:10.1080/00048623.2005.10721250.

Ezeanya, Chika A. 2013. Contending Issues of Intellectual Property Rights Protection and Indigenous Knowledge of Pharmacology in Africa South of the Sahara. 19.

Finetti, Claudia. 2011. Traditional knowledge and the patent system: Two worlds apart? *World Patent Information* 33(1). 58–66. doi:10.1016/j.wpi.2010.03.005.

Graham, Lorie & Stephen McJohn. 2005. Indigenous Peoples and Intellectual Property. *Washington University Journal of Law & Policy* 19(1). 313–337. http://openscholarship.wustl.edu/law_journal_law_policy/vol19/iss1/16.

Maskus, Keith E. 2004. Integrating Intellectual Property Rights and Development Policy. *Journal of International Economics* 62(1). 237–239. doi:10.1016/S0022-1996(03)00084-9.

Robinson, Daniel F. 2013. Legal Geographies of Intellectual Property, ‘Traditional’ Knowledge and Biodiversity: Experiencing Conventions, Laws, Customary Law, and

Karma in Thailand: Customary Law and Traditional Knowledge. *Geographical Research* n/a-n/a. doi:10.1111/1745-5871.12022.

Timmermans, Karin. 2003. Intellectual property rights and traditional medicine: policy dilemmas at the interface. *Social Science & Medicine* 57(4). 745–756. doi:10.1016/S0277-9536(02)00425-2.

Traynor, Cath. 2017. Data Management Plan: Empowering Indigenous Peoples and Knowledge Systems Related to Climate Change and Intellectual Property Rights. *Research Ideas and Outcomes* 3. e15111. doi:10.3897/rio.3.e15111.

Turin, Mark, Claire Wheeler & Eleanor Wilkinson. 2013. *Oral Literature in the Digital Age: Archiving Orality and Connecting with Communities* (World Oral Literature Series 2). Cambridge, England: Open Book Publishers.

Williams, Temple C. 2017. The Underlying Matrices that Frame Divergent Views in the Debate on Intellectual Property and Indigenous Knowledge Protection. *Trans-Humanities Journal* 10(1). 81–105. http://ejournals.ebsco.com/direct.asp?ArticleID=4FB7B7EDC7834939737C.

Woo, Seokkyun, Pilseong Jang & Yeonbae Kim. 2015. Effects of intellectual property rights and patented knowledge in innovation and industry value added: A multinational empirical analysis of different industries. *Technovation* 43–44. 49–63. doi:10.1016/j.technovation.2015.03.003.

Zerbe, Noah. 2005. Biodiversity, ownership, and indigenous knowledge: Exploring legal frameworks for community, farmers, and intellectual property rights in Africa. *Ecological Economics* 53(4). 493–506. doi:10.1016/j.ecolecon.2004.10.015.