

Copyright & The Distributed Lexicon

Hugh Paterson III

Talk context

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 are extensible to the 176 countries who have signed the Berne Convention.

Talk context

This is not legal advice. This is social discussion on legal issues which impact language resource production.

The scope is law as it is practiced in the USA.

Principals are extensible to the 176 countries who have signed the Berne Convention.

Laws represent social contracts.

Laws only come into existence if somebody feels they can make money off of it (or not lose more money 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: A tribal council and United States federal law might both claim authority over fishing rights, or watershed protection.

Human rights and minority peoples literature contain many arguments for the codification of minority peoples social contracts with relation to *Intellectual Property*.

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.

Terms - What is Copyright?

Copyright is one type/category of *Intellectual Property Rights*.

Intellectual Property Rights and what they cover vary by jurisdiction.

- Sui generis database rights
- Moral rights

Copyright is now automatically ascribed, whereas it used to be required to be applied for.

Creative works which are not under copyright are legally known as being in the “public domain”. *Public domain* does not mean *publicly accessible*.

Terms - What is Copyright?

Copyright is one type/category of *Intellectual Property Rights*.

Intellectual Property Rights and what they cover vary by jurisdiction.

- Sui generis database rights
- Moral rights

Copyright is now automatically ascribed, whereas it used to be required to be applied for.

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.

US code 17.1§106

Terms - What is a Lexicon?

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

Lexicon

Corpus like with a data structure

May include **facts** and **creative works**

Usually focused around Words, Roots & Morphemes and their contexts

Terms - What is a Lexicon?

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

Lexicon

Corpus like with a data structure

May include **facts** and **creative works**

Usually focused around Words, Roots & Morphemes and their contexts



Dictionary

Output format

Includes formatting and typography and intended usage as a reference work

Terms - What is a Lexicon?

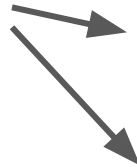
The foundational data set upon which derivative works like dictionaries are created.

Lexicon

Corpus like with a data structure

May include **facts** and **creative works**

Usually focused around Words, Roots & Morphemes and their contexts



Dictionary

Output format

Includes formatting and typography and intended usage as a reference work

Spell Check

Database, or rule based output format

Grammar Check

Database, or rule based output format

Terms - What do I mean by *Distributed*?

How it is built.

One person



Flex & WeSay

Give us the power to organize our work and collaborators in cross-entry ways.

The screenshot shows a dictionary entry for the Hausa word **ɓɓɓɓɓɓ** [ɓɓ:ɓɓɓɓ] (be) amusing; funny. The entry is displayed in a table-like format with various linguistic fields. The 'Lexeme Form' section is highlighted with a green background and shows the IPA transcription **ɓɓ:ɓɓɓɓ**. The 'Morph Type' is **stem**. The 'Citation Form' is **ɓɓ:ɓɓɓɓ**. The 'Pronunciation' is **ɓɓ:ɓɓɓɓ**. The 'Publish Pronunciation In' is **Main Dictionary**. The 'Note' section is expanded, showing the Hausa character **ɓ**. The 'Blench Nigeria List Number' is **0**. The 'Tone Study 2013' is empty. The 'Researcher' is empty. The 'New Custom Field' is empty. The 'Hausa Elicitation blarne' is empty. The 'Grammatical Category of Analysis' is empty. The bottom right corner shows 'Sorted by: Glosses' and '20/2198'.

Entry		Show Hidden Fields
ɓɓɓɓɓɓ [ɓɓ:ɓɓɓɓ] (be) amusing; funny		
Lexeme Form	geI_IPA ɓɓ:ɓɓɓɓ	
	geI_ModOrth ɓɓ-ɓɓɓɓ	
	geI_TONE	
Morph Type	stem	
Citation Form	geI_IPA ɓɓ:ɓɓɓɓ	
	geI_ModOrth	
	geI_TONE	
Complex Forms		
Components		
Pronunciation	geI_IPA ɓɓ:ɓɓɓɓ	
Publish Pronunciation In	Main Dictionary	
Note	Eng ɓ	
	Hau	
	Hau_CL	
	geI_orth	
Blench Nigeria List Number	0	
Tone Study 2013		
Researcher		
New Custom Field	geI_IPA	
	geI_ModOrth	
	geI_TONE	
Hausa Elicitation blarne		
Grammatical Category of Analysis		

Sorted by: Glosses 20/2198

Why do we acknowledge the presence of © & ®?

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

- 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 rights of the researcher.

Why do we acknowledge the presence of © & ®?

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

Art. 89 of the GDPR provides 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 vary from country to country.

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

Outbound restrictions vs. Inbound restrictions

Licenses focus on what can happen relative to the reserved rights of copyright with a dataset after it is created.

Outbound restrictions vs. Inbound restrictions

Licenses focus on what can happen relative to the reserved rights of copyright with a dataset after it is created.

CLA — Contributor License Agreements focus on gaining consent prior to data inclusion in a data set.

A CLA is much more in line with the linguistics industry standard of IRBs and Informed consent.

Contributor license agreements

Typically CLAs assign copyright to one of two entities.

- *The sponsor of a project.*
 - *The contributor but licensed to the project.*
 - *Projects themselves can't hold copyright.*
-
- *A CLA which assigns the content to the public domain gives the data the most freedom.*

If the copyright is assigned to the sponsor of a project they 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.

*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.*

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.

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.

Thank you