Considering WEMI and the Digital Artifacts Around Language Documentation Transcripts

Hugh Paterson III*
Collaborative Scholar
University of Oregon & University of North Texas
USA
i@hp3.me

Abstract

This paper discusses time-aligned textual transcriptions and the relationships between them and other files to which they relate. It addresses critical questions about the independent nature of transcriptions and annotations as independent works under WEMI models. Arguments are presented for an analysis where transcriptions are independent expressions if not also independent works. According to the Dublin Core 1:1 principle this means that Dublin Core based metadata schemas should support these resources with a separate description record. However, from a file based perspective these resources are not always a single file.

Keywords: Language Documentation; WEMI; Transcripts

1 Introduction

It is well known that music may be expressed in audible formats or in notated print media. Print resources have been cataloged independently from audio recordings of the performances of the musical scores. Considering WEMI entities, Riley (2008) notes that minimally music in print media and recordings of performances are separate expressions of the same work:

> It is clear from the FRBR report that Expressions have a defined form; that is, music represented in visual notation is a different Expression than music represented as an audio recording.

Additionally, musical performances may be evidenced via visual artifacts such as video or photographs. Therefore, as illustrated in Figure 1, the audio and associated textual files are part of different expressions. A remaining question exists: Are they part of the same *Work*? There are strong similarities between the score—recording and annotation–recording, along the lines of the modality of engagement, but a key differences is that one can not recreate sound in a recording from the annotation as one might do from a score.

2 Language Documentation Context

Within language scholarship, including folk studies, oral history studies, linguistics, philology, and anthropology, some of the earliest cases of audio recording include the recording of ethnolinguistic minority speech performances (Fewkes, 1890; Haddon & Myers, 1898). However, the written scholarly record has often ignored the audio artifact and given preference reductionistic print-based surrogates. Within language scholarship, textual representations are generally considered to come in two forms. The first, transcriptions, reduce the audio signal to one of three types of textual representation: phonetic, phonological, or orthographic. In addition to transcription, it is often the practice of scholars to provide grammatical or performative annotations to textual formats (Thompson, 2004). These annotations may take a variety of textual forms but always relate back to a theoretical framework of analysis and often reflect a different production processes than

*The author has a MA in linguistics and has conducted language documentation work in Nigeria and Mexico. He is currently the Thomas Intern for Research, Scholarly Communication Infrastructure at the University of Oregon Knight Library. He is also in the Information Science PhD program at the University of North Texas. https://hughandbecky.us/Hugh-CV

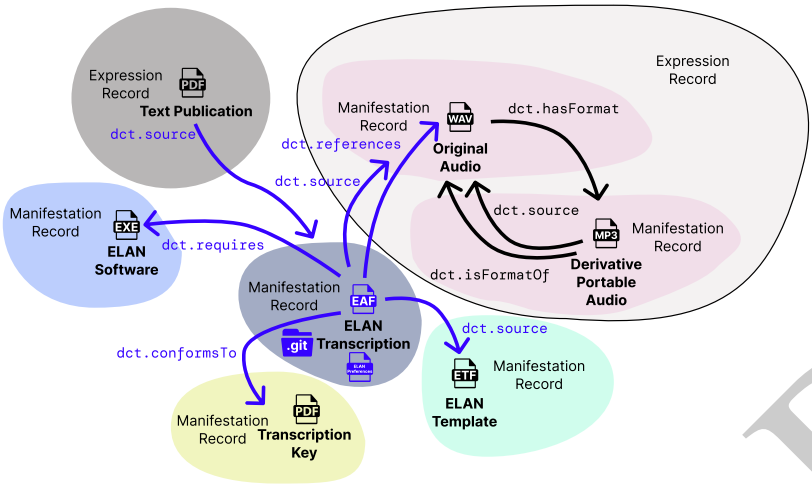

Figure 1: Relationships between transcription manifestations and other entities.
Diagram of related audio and associated entities manifestation and expression records.

the audio *Manifestation*. That is, transcripts may undergo separate revisions with additional input from those who were recorded (Good, 2016). There is a wide consensus among scholars that both annotations and transcriptions represent the application of theory and judgment apart from the source *Work* and are, therefore, independent *Works* of analysis (Bird & Liberman, 2001; Bucholtz, 2000; Jaffe, 2000; Ochs, 1979; Paterson III, 2015; Tillett, 2004). Within language scholarship Himmelmann (2012) classifies audio recordings as raw data while annotations, translations, and transcriptions are classified as primary data. Himmelmann's divide acknowledges the different contributors as well as the different types of effort which goes into making the resources. Himmelmann's views along with those of Riley (2008) and Vellucci (2007) clearly place the distinction between source audio and annotations, translations, and transcriptions as different WEMI *Expressions*. However, Tillett (2004) classifies annotated editions as new *Works*. Taken together this suggests that in cases of language scholarship that annotations and transcriptions are understood to be separate WEMI *Works* from the underlying performative act documented via the audio or video artifact. There is an important relationship between audio performances and their transcription or annotation which at times may be complex to express in digital libraries. The proposed model allows one to postulate that it might be the case that there is an underlying *Work* common to both textual and audio representations as well as being independent WEMI *Works*. This is similar to how legal scholars and copyright law in the USA treat musical works. That is, the *Manifestation* of the recorded musical performance constitutes a copyright-able work independent from the copyrights applicable to the underlying work (U.S. Copyright Office, 2020, 2021a, 2021b). Within linguistic scholarship, it is contested as to whether these underlying works actually exist for copyright purposes in all cases. For example, some speech recordings consist of lists of words. It is doubtful that such a list of audio tokens would meet the minimum artistic creation requirements of copyright applicability.[2] If there is any such underlying creativity for lists of audio tokens it is the creative effort of the list creator rather than the speaker who would likely be the copyright holder (assuming the underlying work meets thresholds for copyright). With regard to underlying works, the underlying work ought to be acknowledged. Four examples of underlying *Works*, perhaps even with different *Expressions*, in more experimental projects involving language scholarship include the use of the *Pear Story* (Chafe, 1980), the *SIL comparative African wordlist* (Snider & Roberts, 2004, 2006), the *Swadesh list* (Swadesh, 1952, 1955), and *The North Wind and the Sun* (Aesop, n.d.) as used in various Illustrations of the International Phonetic Alphabet. Other recordings in language scholarship consist of epics or myths which are common knowledge among the community and also might not qualify as a copyright-able work due to a lack of an identifiable creator. In such cases, the performance may qualify as a copyright-able work. Still yet other recordings in language scholarship do consist of authored performances similar to scripted performances as is seen in the music or movie industries. The point is that audio recordings may constitute several works. Cataloging these works is important. Textual representations of audio artifacts then have a relationship not just with an audio file (performance) but also potentially with an underlying work.

[2]Copyright applicability is not the bases of identifying underlying works, but it does point to the importance of identifying the resource at hand as well as any underlying work.

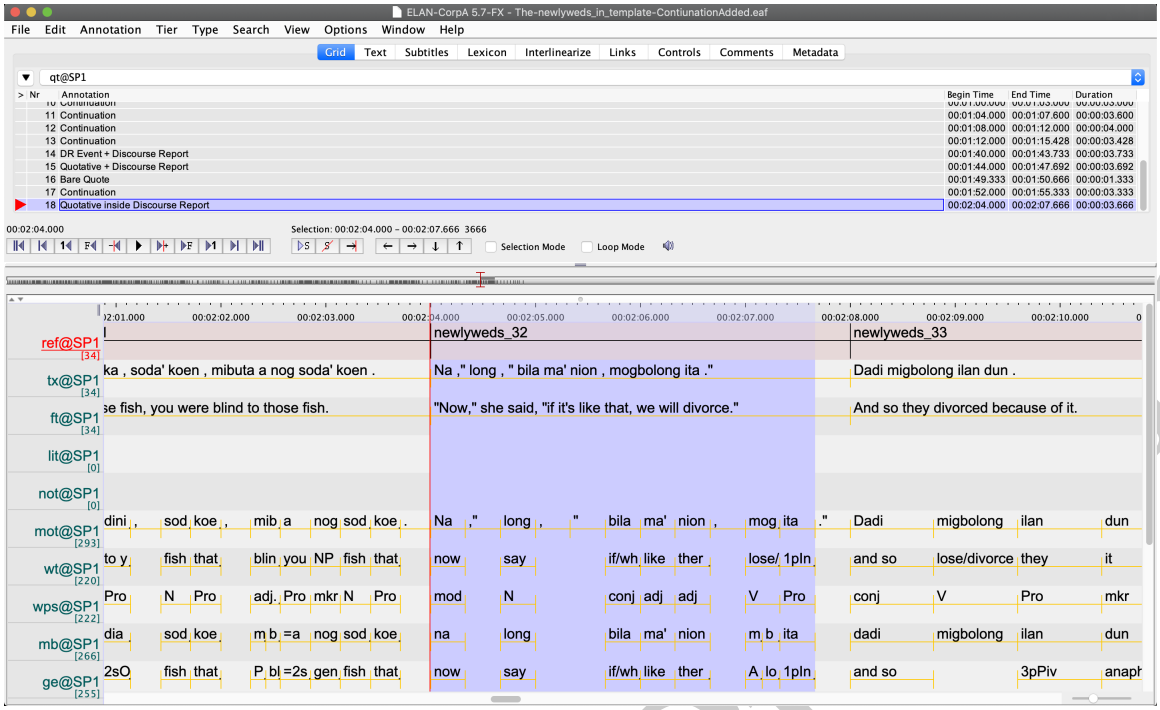

Figure 2: Multiple Annotation Tiers in ELAN. Showing the Reported Speech template Nikitina et al., 2019. Screenshot of ELAN showing multiple annotation tiers.

2.1 Technology and Expressions

The production workflows ought to inform the description and arrangement (including inter-entity relationships) of audio artifacts and their textual associated entities. With the introduction of computers, time based annotations of speech signals became more common. With tools such as ELAN (Sloetjes, 2022), PRAAT (Boersma & Weenink, 2023), and Audacity (Audacity Team, 2022), time aligned textual representations are the standard in language scholarship. Annotation and transcription of audio artifacts is expected in Language Documentation (Himmelmann, 1998, 2012), a sub-field of linguistics with strong connections to ethnolinguistic minorities, linguistics, folk studies, oral history studies, and preservation activities. Annotations and transcriptions can come in several file formats. For example, professional video workflows use `.srt` files. PRAAT workflows, often focused on word or phrase level units, use `.TextGrid` files. ELAN workflows, compatible with both audio and video artifacts, use `.eaf` files. Other file formats and software workflows also exist. However, ELAN is likely the most widely used tool. Therefore, it is given special attention. Like many textual resources, annotations and transcriptions created in ELAN may be used directly in ELAN to edit audio or video media, they may be used directly in other applications, or indirectly to derive statistics about their related media files. Finally, annotations and transcriptions are frequently used to create annotated compilations of texts or examples which are then published as evidence supporting linguistic analysis in monographs or articles. As such, transcription and annotation files may be the object of reference in a *dct:source* relationship for a variety of language resource objects. The importance of annotations and transcriptions artifacts within the knowledge ecosystem means that each *Manifestation* ought to be independently referenceable following scholarly best practices (Andreassen et al., 2018; Berez-Kroeker et al., 2018; Smith et al., 2016; Wu et al., 2018).

2.2 ELAN workflow considerations

ELAN may be the tool of a single person or of a team of scholars. Team-based workflows using ELAN `.eaf` files may contain annotations, transcriptions, or both. One should also assume that there is more than one speaker's contribution in the `.eaf` file as is common in discourse or dialogue studies and many non-laboratory based language documentation scenes (Crasborn & Sloetjes, 2010). Increasingly, multi-person teams are responsible for the investigative process. Some team members may engage in recording, while others in annotation, and still others in transcription. Additionally, some machine learning or artificial intelligence

based tools may be employed for transcription creation. In these types of situations, it is frequently the case that each team member will focus on contributing to the annotation in a specific way. ELAN's `.eaf` files are time-aligned but are also tiered so that specific speakers can be tracked through a performance. Further, within that performance, specific types of annotations can be made on specific tiers. ELAN Tiers are the horizontal rows illustrated in the bottom half of Figure 2. The nature of ELAN does not allow two team members to simultaneously annotate audio artifacts. This forces teams to share their annotations either by sequentially working on files or by simultaneously working on duplicates of the same file and then merging their files into a common file. Often research projects will set up a template file to make the process of setting up the various tiers easier and consistent across a project. Collaborators without a pre-established common appreciation for provenance retention in the workflow are prone to resort to file management practices which include the use of E-mail, USB drives, Telegram[3], and Dropbox[4] to share files across the team. As a technologist, linguist, and archivist, experience has shown that prudence includes and requires the use of distributed version control to fully support team activities and leverage the power of integrated provenance tracking including the reverse-ability of changes in scholarly outputs. One possibility is for implementing version control is Git, but depending on the project tools such as DVC[5] or Pachyderm[6] may become useful. Teams of language scholars have been noticed to use Github[7] like Dropbox—as a simple file sharing service—without regard for any of the provenance tracking features or collaboration features available via Git. Github provides more than Kanban boards and to-do lists. Technologists in software development have developed best practices when using Git which maximize the traceability of changes to a document (code base), including contributor, reason for the change, and the actual change (Lee, 2020; Tsitoara, 2020). When using Git not only is traceability of changes a possibility but also merge processes on divergent files are simplified. Reverts to a previous state of a document, which are sometimes desired based on changes in analysis, are also part of the capability of documents like `.eaf` files under Git version control. This means that in addition to the files under Git version control there is an actual Git provenance log. Finally, with team-based uses of ELAN, the screen layout preference file is considered a part of the annotation *Item* because it defines the materiality which contextualizes the transcription artifact. The model suggests that the `.git` log, the textual representation of the audio and the ELAN preference file for dealing with the single text artifact be combined in a single manifestation record.[8] This combination of files may appear to be a technical violation of the Dublin Core 1:1 Principle; however, its logical that the provenance should travel with the textual file. Therefore, the model is constructed such that the preference-file component of the viewing experience (of the `.eaf`) is part of the manifestation's material context. Other language scholarship software, such as Shoebox[9] Toolbox[10] and FLEx also have multipart issuance resources which must be used together in order to recover the full essence of the artifact. Finally, the complexities of how the contents of a `.eaf` file can be created means that transcription and annotation files may be aggregate works, containing works, expressions, and manifestations of different works. Depending on how teams are established, contributors may include the original speaker, a team of scholars with different roles, editors, analysts, and computer programs.

In contrast to the single creation event and static nature of audio recordings created in language research, their associated text files have a very different creation process. Analysis suggests that annotation and transcription files are a type of *diachronic work* (works that are embodied over time) known as continuing resources. For example, each Git *commit* is given a unique identifier lending to an analysis that each commit represents a different manifestation of a single expression. In addition to Git commits, a well managed Git project will have tags and/or release versions of the resource, these markers lend themselves to an analysis of separate expressions of the same work. Git technology also has a method for creating divergent and, sometimes, experimental instances of the content, called branching. Branches with additional commits have a derivative relationship with the main branch of a project. In contrast to new commits which instantiate new manifestations, branches seem to be new diachronic works. Best practice for Git users newly joining a project is to `fork` (clone) a project (make two WEMI *Items* from a single *Manifestation*). Then within their forked instance, they create a new branch (new diachronic work). Ensuing modifications are then tested on their own branch until the main branch of the original work assimilates any suggested changes. New branches are often merged into

[3]https://telegram.org

[4]https://dropbox.com

[5]https://dvc.org

[6]https://www.pachyderm.com

[7]https://github.com

[8]Within the Resource Description and Access (RDA) framework it appears that this manifestation then would have a *Mode of Issuance* which is *multiple unit*; fitting the definition: A mode of issuance of a manifestation that is issued as a multipart physical unit or intangible multipart logical unit.MARC/RDA Working Group (2020)

[9]https://software.sil.org/shoebox

[10]http://www.fieldlinguiststoolbox.org

the main branch or abandoned. However, sometimes they can become the main branch of a new *Work* and take on a life of their own within a new community of contributors.

Best practice for Git usage suggests that the main branch of a project ought not be the primary branch under which development activity is conducted. This behavior means that the main resource is constantly assimilating new content as the resource is developed.

Given the multi-tiered nature of `.eaf` files, it is always possible to add new layers of annotation or transcription to the resource. Hence, these resources can reasonably be expected to not have a *finite* publication state. A finite publication state is a hallmark distinction between *monographic resources* and *continuing resources*. Thus in many cases it is reasonable to treat transcription files such as `.eaf` files as continuing resources. Language annotation and transcription files may be years in the making and are often created, altered, or enhanced long after the original audio is recorded. Continuing resources may additionally be classified as indeterminate vs. determinate, and successive vs. integrating. With Git version control it seems that annotation and transcription works, such as are embodied in `.eaf` files, are indeterminate (they have no predefined time to cease their evolution) and integrating (each change is assimilated into the whole resource).

Annotation and transcription files like `.eaf` do not have separate parts such as *journal issues* are separate parts of a *journal*, and therefore, meet the qualifications of integrating resources. Many integrating resources are assimilatory and as such the identity of the integrated parts lose their identity upon assimilation. When under Git version control, `.eaf` files and their updates can be bisected so that individual updates can be reviewed. However, sans the provenance log, the distinctions would be indistinguishable.

3 Conclusion

For Dublin Core based metadata schemas, transcriptions and annotations should have their own description records. This would maintain the 1:1 Principle. However there are clear cases where several files must be considered a single unit. Transcription files, depending on their composition, may be works-of-works or expressions-of-works. This is not really any different from print bound editions of some published works. Finally, teams are frequently using git within scholarly project. The relationship between WEMI and various git processes is mostly unexplored to date. A broader understanding of how git interacts with preservation, stewardship, and scholarship would bring some clarity on how to catalog resources with multiple git controlled instances.

References

Aesop. (n.d.). The North Wind and the Sun [aesop's fables]. perry index 46.

Andreassen, H. N., Berez-Kroeker, A., & Gawne, L. (2018). The Austin Principles of Data Citation in Linguistics. Retrieved December 25, 2020, from https://munin.uit.no/handle/10037/16598 Accepted: 2019-11-05T08:07:08Z

Audacity Team. (2022). *Audacity® [computer software]. version 3.2.0*. Open Source GPL Software. https://audacityteam.org

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., & Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, *56*(1), 1–18. https://doi.org/10.1515/ling-2017-0032

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, *33*(1), 23–60. https://doi.org/10.1016/S0167-6393(00)00068-6

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer [computer software]. version 6.3.04*. Phonetic Sciences, Faculty of Humanities, University of Amsterdam. http://www.praat.org

Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics*, *32*(10), 1439–1465. https://doi.org/10.1016/S0378-2166(99)00094-6

Chafe, W. L. (Ed.). (1980). *The Pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Ablex Pub. Corp. Retrieved February 3, 2023, from http://catdir.loc.gov/catdir/enhancements/fy1511/80120122-t.html

Crasborn, O., & Sloetjes, H. (2010). Using ELAN for annotating sign language corpora in a team setting. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, & A. Schembri (Eds.), *Workshop Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* (pp. 61–64). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/workshops/W13.pdf

Fewkes, J. W. (1890). *Jesse Walter Fewkes collection of Passamaquoddy cylinder recordings*. Archive of Folk Culture, American Folklife Center (AFC 1972/003). Library of Congress.

Good, F. (2016). Voice, Ear and Text Words: Meaning and Transcription. In R. Perks & A. Thomson (Eds.), *The Oral History Reader* (3rd ed., pp. 458–468). Routledge.

Haddon, A., & Myers, C. (1898). *[voices from australia in the] ethnographic wax cylinder collection, world and traditional music*. British Library. https://sounds.bl.uk/World-and-traditional-music/Ethnographic-wax-cylinders

Himmelmann, N. P. (1998). Documentary and Descriptive Linguistics. *Linguistics*, *36*(1), 161–195. https://doi.org/10.1515/ling.1998.36.1.161

Himmelmann, N. P. (2012). Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation & Conservation*, *6*(1), 187–207. http://hdl.handle.net/10125/4503

Jaffe, A. (2000). Introduction: Non-standard orthography and non-standard speech. *Journal of Sociolinguistics*, *4*(4), 497–513. https://doi.org/10.1111/1467-9481.00127 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9481.00127

Lee, J. (2020). Using git on Small Teams: Best Practices. Retrieved January 4, 2023, from https://joshuamlee.com/using-git-on-teams-vcs-best-practices

MARC/RDA Working Group. (2020). *Recording the Mode of Issuance for Manifestations in the MARC 21 Bibliographic Format* (tech. rep. Discussion Paper No. 2020-DP16). Library of Congress. Washington, D.C. Retrieved January 5, 2023, from https://www.loc.gov/marc/mac/2020/2020-dp16.html

Nikitina, T., Hantgan, A., & Chanard, C. (2019). Reported speech annotation template for ELAN. http://discoursereporting.huma-num.fr/corpus.html

Ochs, E. (1979). Transcription as Theroy. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental Pragmatics* (pp. 43–72). Academic Press. Retrieved January 18, 2023, from http://www.sscnet.ucla.edu/anthro/faculty/ochs/articles/ochs1979.pdf

Paterson III, H. (2015). Phonetic Transcription of tone in the IPA. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Conference of Phonetic Sciences* (Paper 507). International Phonetic Association. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0507.pdf

Riley, J. (2008). Application of the Functional Requirements for Bibliographic Records (FRBR) to Music. *ISMIR 2008: Proceedings of the 9th International Conference on Music Information Retrieval, 14-18 September 2008, Philadelphia (USA)*, 439–444. https://doi.org/10.5281/zenodo.1416446

Sloetjes, H. (2022). *ELAN [computer software]. version 6.4*. Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan

Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86. https://doi.org/10.7717/peerj-cs.86

Snider, K. L., & Roberts, J. S. (2004). SIL Comparative African Word List (SILCAWL). *Journal of West African Languages*, *31*(2), 73–122. https://main.journalofwestafricanlanguages.org/index.php/downloads/summary/88-volume3102/436-sil-comparative-african-word-list-silcawl

Snider, K. L., & Roberts, J. S. (2006). SIL comparative African wordlist (SILCAWL). *SIL Electronic Working Papers*, *2006*(Article 005). Retrieved June 25, 2020, from https://www.sil.org/resources/publications/entry/7882

Swadesh, M. (1952). Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, *96*(4), 452–463. Retrieved February 3, 2023, from https://www.jstor.org/stable/3143802

Swadesh, M. (1955). Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, *21*(2), 121–137. https://doi.org/10.1086/464321

Thompson, P. (2004). Spoken Language Corpora. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (Chapter 5). Arts and Humanities Data Service: Literature, Languages, and Linguistics. https://users.ox.ac.uk/~martinw/dlc/chapter5.htm

Tillett, B. (2004). *What is FRBR? — A Conceptual Model for the Bibliographic Universe*. Library of Congress Cataloging Distribution Service. https://www.loc.gov/cds/downloads/FRBR.PDF

Tsitoara, M. (2020). Git Best Practices. In *Beginning Git and GitHub* (pp. 79–86). Apress. https://doi.org/10.1007/978-1-4842-5313-7_6

U.S. Copyright Office. (2020). *Copyright in Derivative Works and Compilations* (07/2020). https://www.copyright.gov/circs/circ14.pdf

U.S. Copyright Office. (2021a). *Copyright Registration for Sound Recordings* (03/2021). https://www.copyright.gov/circs/circ56.pdf

U.S. Copyright Office. (2021b). *Copyright Registration of Musical Compositions and Sound Recordings* (03/2021). https://www.copyright.gov/circs/circ56a.pdf

Vellucci, S. L. (2007). FRBR and Music. In A. G. Taylor (Ed.), *Understanding FRBR: What it is and how it will affect our retrieval tools* (pp. 131–151). Libraries Unlimited.

Wu, Y., Alawini, A., Davidson, S. B., & Silvello, G. (2018). Data Citation: Giving Credit Where Credit is Due. *Proceedings of the 2018 International Conference on Management of Data*, 99–114. https://doi.org/10.1145/3183713.3196910