

The Bridge Informatics Report

Hugh Paterson III

i@hp3.me

Introduction

This is a report of work conducted by Hugh Paterson III in 2023 as part of the LEADING fellowship program supported by IMLS grant RE-246450-OLS-20. It contains information science analysis in support of Haverford's *The Bridge* project: A language learning tool to facilitate text use in curriculum.

Statement of the Problem

The Bridge project has over 1500 source documents in Latin in various stages of aggregation. Latin is only one of the languages of antiquity that The Bridge infrastructure seeks to provide services around. The Bridge Data is sourced from at least seven pipelines of data. These sources all use different terminologies for the categories of data they provide. Some of these categories may be congruent across data sources; others may have the same category title but represent different data. This confusion of terms and categories creates a data management challenge within the Bridge Project, as well as presents challenges for transparently presenting aggregated bibliographic and lexicographical data to potential downstream data users.

This final report contains the agreed upon components for the final deliverable. Namely:

1. Terms used in the project data arrays and their definitions;
2. Data Providers and the terms used by those data providers;
3. Preferred terms to use within The Bridge Project to increase transparency around the data;
4. An explanation of the chosen terms and why other investigated models are not a good fit; and,
5. A recommendation for the data structure of the project database.

These issues are addressed in the various sections of the report.

Table of Contents

Introduction.....	1
Statement of the Problem.....	1
Table of Contents.....	2
Technology Review.....	3
Legal Considerations.....	16
Future Publishing Opportunities.....	17
Import Workflows.....	18
Perseids.....	18
Simple List.....	21
Textbooks.....	23
Simple locally generated lemmatized text.....	25
Text from LASLA.....	25
Text from a Concordance.....	26
Text from PROIEL Treebank.....	27
Data Spreadsheet Header Summary Table.....	29
Data File Organization Suggestion.....	30
Bibliographic Record Management.....	31
New Metadata Elements.....	35
Titles.....	38
Contributors.....	39
Relationships.....	40
Language.....	41
Dates.....	43
Description.....	45
Divisions.....	51
Provenance.....	52
External Link.....	53
Extent.....	53
Distribution Status.....	55
Role Appendix.....	56
OLAC Roles.....	56
MARC Relator Roles.....	57

Technology Review

The Bridge Current Status

The Bridge is currently a Python application with an operational web interface. The web interface is actively used by those who study the languages of antiquity at Haverford. It is currently impacting classroom activities by supporting students and instructors. Current approaches to data management and engagement are further discussed in the section [Data Organization](#).

Open Data

The Bridge aims to be a collaborative citizen within the Linguistic Linked Data environment. Linked data comes in two varieties: *Open Linked Data*, and *Proprietary Linked Data*. The main factor differentiating these two ecosystems of linked data is the impact on business models: specifically, what it is that The Bridge can or cannot do with the data developed by data providing partners. This is further discussed under the section [Legal Considerations](#).

RDF & JSON-LD

Technologically, Linked Data approaches could implement one of several technical structural encoding formats. The first being RDF¹ and the second being JSON-LD.^{2,3} The technical requirements to support either of these syntactic varieties is different. General web-technologists will recommend JSON-LD, usually without understanding the “LD” part, because JSON is something with which they are familiar. In contrast, most implementations of Linguistic Linked Data use RDF.

RDF in turn can be implemented in several varieties, that is, with different ontologies or metadata models (metadata categories and their interrelationships) and metadata elements (metadata fields). Applying linked data to use-cases and scenarios of linguistic and/or digital humanities analysis has taken different approaches. Bosque-Gill and colleagues review and categorize over 100 models.⁴ Most models are influenced by project specific goals and are, therefore, incompatible with other projects. Of course reuse is possible, but the underlying engineering is focused on solutions with a limited scope. This scope occurs up and down the linguistic analysis hierarchy: Phonetics & Phonology, Morphology, Syntax, Semantics, Discourse & Pragmatics. In contrast to models which are application specific, some models such as the

¹ <https://www.w3.org/RDF>

² <https://json-ld.org>

³ <https://www.w3.org/TR/json-ld11>

⁴ Bosque-Gil, J., J. Gracia, E. Montiel-Ponsoda, and A. Gómez-Pérez. 2018. “Models to Represent Linguistic Linked Data.” *Natural Language Engineering* 24 (6): 811–59. <https://doi.org/10.1017/S1351324918000347> .

Natural Language Processing Interchange Format (NIF)^{5,6} and POWLA^{7,8} seek to encode text resources such that they may be used in a broad range of contexts. NIF and POWLA are further discussed in the section [Data Organization](#).

Servers

Three types of servers are needed to facilitate linked data, if Haverford were to choose to go this way. First, they would need to publish their controlled vocabularies and ontology; second, they would need to host the data in a linked data accessible way; third, they would need to host the web-based user interface.

Ontology Server

An ontology server defines the linked data URIs for linked data represented categories within a model. For example, Oregon Digital uses a domain name⁹ and deploys an application thereon to manage controlled vocabularies and make them reusable in other systems. They do this using the software called *Controlled Vocabulary Manager*.¹⁰ UNT has a Django app which does mostly the same thing.¹¹ The advantage of the Django app is that it is written in Python and

⁵ <https://persistence.uni-leipzig.org/nlp2rdf/specification/core.html>

⁶ Cimiano, Philipp, Sebastian Walter, and Matthias Hartung. 2015. Guidelines for Developing NIF-Based NLP Services. Wakefield, MA: W3C.
<https://www.w3.org/2015/09/bpmlod-reports/nif-based-nlp-webservices>.

Hellmann, Sebastian. 2015. "Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data." Ph.D Dissertation, Leipzig, Germany: University of Leipzig.
<https://nbn-resolving.org/urn:nbn:de:bsz:15-gucosa-157932>.

Hellmann, Sebastian, Jens Lehmann, and Sören Auer. 2012. "NIF: An Ontology-Based and Linked-Data-Aware NLP Interchange Format." Working Draft.
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fef24baddf1a3cbd811fc5ebc8b4c47eb3608f87>.

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. "Integrating NLP Using Linked Data." In *The Semantic Web – ISWC 2013*, edited by Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, 98–113. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer.
https://doi.org/10.1007/978-3-642-41338-4_7.

Menke, Peter, Basil Ell, and Philipp Cimiano. 2017. "On the Origin of Annotations: A Module-Based Approach to Representing Annotations in the Natural Language Processing Interchange Format (NIF)." *Applied Ontology* 12 (2): 131–55. <https://doi.org/10.3233/AO-170180>.

⁷ Chiarcos, Christian. 2012a. "Interoperability of Corpora and Annotations." In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, edited by Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, 161–79. Berlin, Heidelberg: Springer.
https://doi.org/10.1007/978-3-642-28249-2_16.

Chiarcos, Christian. 2012b. "POWLA: Modeling Linguistic Corpora in OWL/DL." In *The Semantic Web: Research and Applications*, edited by Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, 225–39. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer.
https://doi.org/10.1007/978-3-642-30284-8_22.

⁸ <https://purl.archive.org/powla/powla.owl>

⁹ <http://opaquenamespace.org>

¹⁰ <http://github.com/OregonDigital/ControlledVocabularyManager>

¹¹ <https://github.com/unt-libraries/django-controlled-vocabularies>

closer to the core dev requirements for sustainability (the rest of *The Bridge* project uses Python).

Language Data Server

Two options exist here. First is to encode texts as a single string and then annotate portions of that string. Second is to, in some way, create identifiers for information created in a SQL database. Native RDF based files would likely be more sustainable with only annotations being added as analysis work was completed. To facilitate this method, an RDF/SPARQL server is needed.

Fuseki2¹² is a linked data server with a SPARQL¹³ endpoint. It is run on Apache Tomcat, which is a java-esque application.¹⁴ Fuseki2 works on a Linux server. Fuseki2 has better documentation than other options, such as blazegraph,¹⁵ and is under active maintenance/development. Fuseki2 is recommended by scholars working with linguistic linked data.

In contrast to the RDF native approach, another approach is to run a translation layer over an SQL database so that the database can be accessed via a SPARQL query. D2RQ¹⁶ is software which allows read-only access of SQL relational databases as if they were SPARQL endpoints. A MySQL + D2RQ approach may have certain advantages if developers are not familiar with SPARQL or RDF based technology. It would allow *The Bridge* to have a SPARQL endpoint while maintaining a more traditional technology stack.

There are also down-sides to a MySQL + D2RQ approach. It is not a data-centric approach, meaning the data comes first, rather it is a technology-stack centric approach. The data is not portable as RDF/linked-data and when dealing with texts and character counting offsets. As is done in NIF and POWLA, one would need to use either a MEDIUMTEXT (4mb) or a LONGTEXT(4gb) field in MySQL.¹⁷ Exact performance specifications would need to be evaluated and tested. In an ideal development context, these performance specifications would be considered in development rather than dictated to developers. Using these long text field types in MySQL could drastically reduce web-application performance due to their possible size. The issue is how the application and the database tell caching systems to fill their memory.

User Data Server

User data could be modeled in RDF; however, this would be an unnecessary complexity. I recommend that user data be stored in a separate database such as MySQL or SQLite. One efficient way to do this is to use the Django¹⁸ framework to build the front end and user account

¹² <https://jena.apache.org/documentation/fuseki2>

¹³ <https://en.wikipedia.org/wiki/SPARQL>

¹⁴ <https://tomcat.apache.org>

¹⁵ <https://blazegraph.com>

¹⁶ <http://d2rq.org>

¹⁷ <https://dev.mysql.com/doc/refman/8.0/en/blob.html>

¹⁸ <https://www.djangoproject.com>

components. This architecture allows the user data security model, and issues like GDPR¹⁹ compliance (portability of the user's data, privacy policy, and terms of service) to be managed independently from other data. It is often the case that MySQL servers are independent from other servers for performance and security reasons.

Bibliographic Data Server

For bibliographic data and associated file management during the language data development workflow, a Django/MySQL based system is recommended. I am not currently aware of any such open source system. Every organization I have worked with has implemented add-hoc solutions. The development of a flexible, Dublin Core and WEMI-centric system as described elsewhere in this proposal may be well positioned for an Institute of Museum and Library Services (IMLS) grant.

Web Application Server

Since Haverford already has a web presence, this is not discussed in detail further. My recommendations are that the web-views and data processing logic are separated and maintained independently. One way to do this in a sustainable way is to use the Django framework which is implemented in Python. The reduction of bespoke code unique to *The Bridge* will increase the long term sustainability of the interactive service by reducing costs and required developer time. Adoption of a framework such as Django reduces bespoke code authorship and technical debt.

Language Data Organization

Linked Data Models²⁰ vary in the way they assert facts about language information. These variations are influenced by several factors. While it is simple to say that they vary by the goals of a project, a more objective assessment is useful. Bosque-Gil et al (2018) already pointed out that there are influencing factors based on which level or subfield of linguistics is in focus (their list is replicated in the section [RDF & JSON-LD](#)). However, even within these sub-fields, there are different approaches to studying language and the models presented represent these

¹⁹ <https://gdpr-info.eu>

²⁰ Bouda, Peter, and Michael Cysouw. 2012. "Treating Dictionaries as a Linked-Data Corpus." In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, edited by Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, 15–23. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-28249-2_2.

Fiorelli, Manuel, Armando Stellato, John P. McCrae, Philipp Cimiano, and Maria Teresa Pazienza. 2015. "LIME: The Metadata Module for OntoLex." In *The Semantic Web. Latest Advances and New Domains*, edited by Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann, 9088:321–36. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-18818-8_20.

Herold, Axel, Lothar Lemnitzer, and Alexander Geyken. 2012. "Integrating Lexical Resources Through an Aligned Lemma List." In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, edited by Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, 35–44. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-28249-2_4.

biases. For example, *OntoLex*²¹ (inclusive of its accompanying modules)²² is the most widely adopted Linked Data Model for lexicographic representation of language data. However, *OntoLex* prescribes for its users a linguistic grammatical theory as well as a linguistic lexicographical theory. These theoretical assumptions must be followed by users. Few problems with the *OntoLex* are encountered if one makes the assumptions that its designers do and use it with languages which have grammatical functions and categories like the languages the designers know well. Figure 1 presents the conceptual categories in OntoLex, while Figure 2 presents a larger scope when some of the modules are included. I find that looking at Entity Relationship Diagrams can rapidly provide a high-level understanding of the described entities and their descriptors.

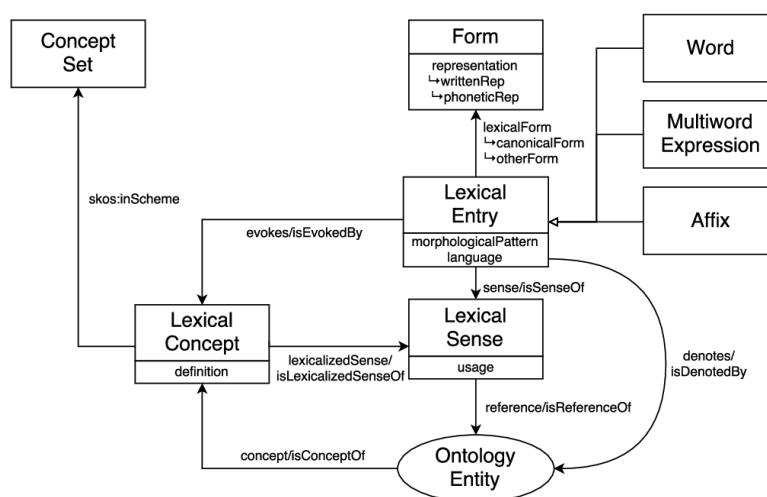


Figure 1: OntoLex ERD as presented in McCray et al 2017.²³

²¹ <https://www.w3.org/2019/09/lexicog>

²² These modules are OntoLex-Lemon: core; OntoLex-Lime; OntoLex-SynSem; OntoLex-Decomp; OntoLex-VarTrans; OntoLex-LiMe; OntoLex-Lexicog; OntoLex-Morph; and OntoLex-FrAC.

²³ McCrae, John P, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. 2017. "The OntoLex-Lemon Model: Development and Applications." In *Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference*, 587–97. Brno, Czech Republic: Lexical Computing CZ s.r.o. <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf>.

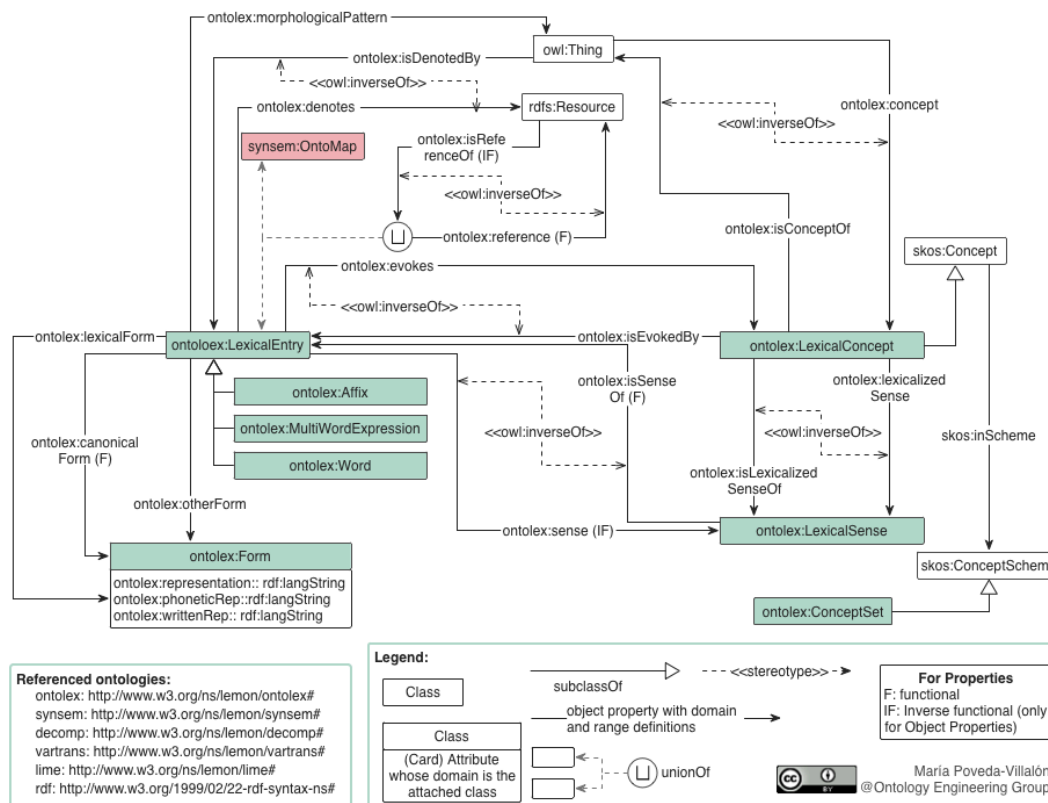


Figure 2: A perspective of Ontolex with some modules.²⁴

Text Based Approaches

NIF and *POWLA* contrast with *OntoLex* which prescribes categories and their meanings to database values--and is more useful in some lexicographical contexts than others. *NIF* and *POWLA* opt to model the whole text as a single string. A character encoding is defined, and the string is transformed to that encoding. Then each character (i.e., Unicode Character) is assigned a place based position. Within the same file, annotation-layers are defined with specific purposes. These layers define segments along the string and give meanings to those segments. In this way, we see that lexicographic analysis can exist as an annotation layer over texts. This is demonstrated when treating OCR'd dictionaries as "text" in Bouda and Cysouw (2012). An ERD of *NIF* applied in a project is shown in Figure 3, while Figure 4 illustrates various aspects of *POWLA*.

²⁴ <https://thepetiteontologist.wordpress.com/2018/03/25/brief-analysis-of-ontolex>

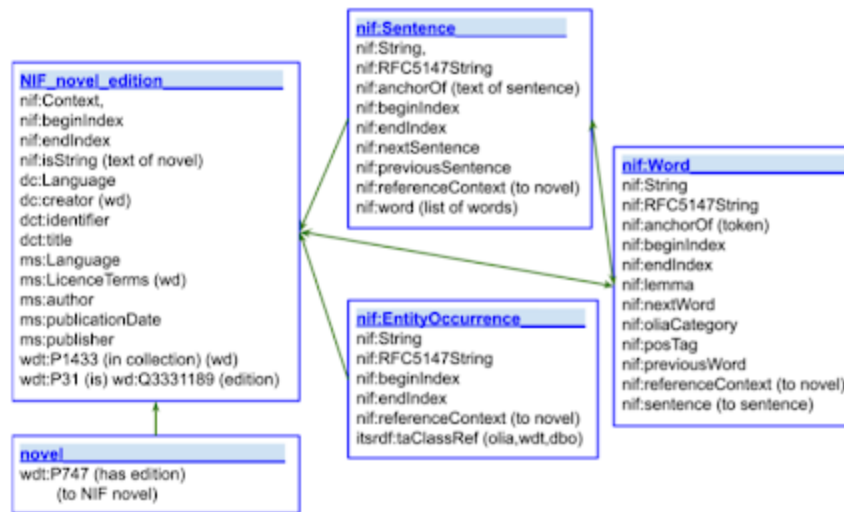


Figure 3: Diagram of the entities modeled and their metadata fields within the NIF model/ontology.

NIF allows the identification of subunits within the string just as POWLA does.

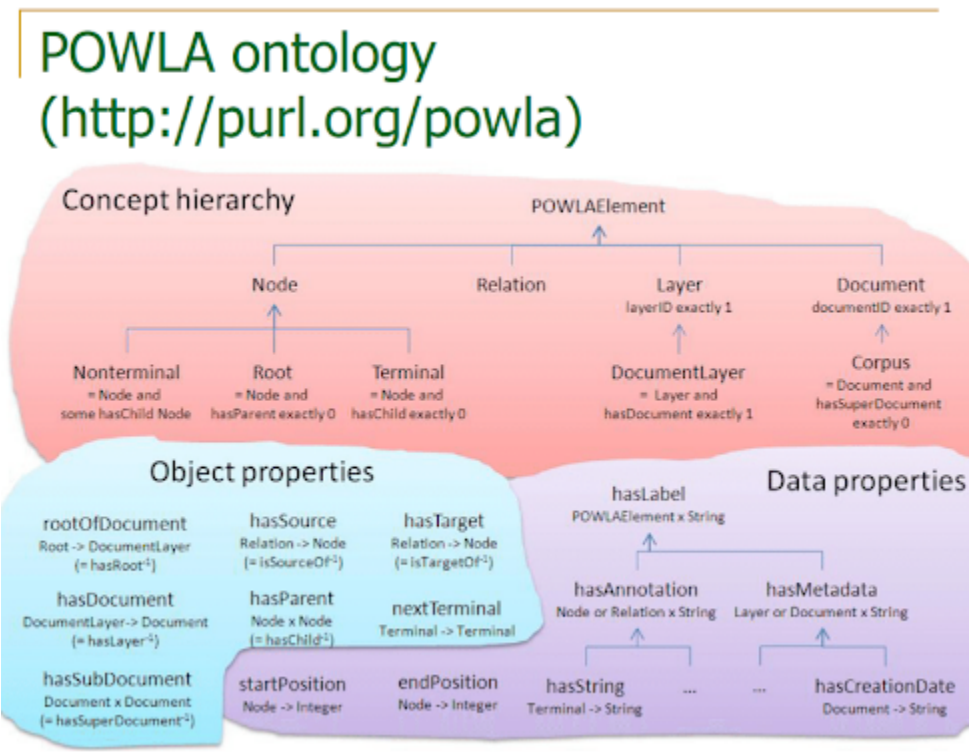


Figure 4: POWLA illustration.²⁵

²⁵ <http://nachhalt.sfb632.uni-potsdam.de/powla.html>

Current Approach, Limitations, and Opportunities

The current approach to data management for data facilitating *The Bridge* is to separate full-texts from analyzed tokens (analysis) by storing these two textual entities in separate documents. Analysis is stored as arrays of tokens (lemmas) per text. While *The Bridge* shows end-users the results of “compared texts”, it operationally functions by comparing arrays of reduced tokenized elements. This is similar to a “bag-of-word” or a “bag-of-lemmas” based comparison. This does have some speed enhancements.²⁶ This method reduces the required storage space for data accessed by the web-based application. An NIF or POWLA model would combine source text and analysis into the same storage document. This would increase storage space requirements,²⁷ but would reduce the quantifiable number of managed files.

Another major difference which makes current data less compatible with NIF or POWLA is that the current analysis data does not have an absolute marker to where it was found in the source string. Some data do have word order indicators, and, theoretically, could be reconstructed in those orders—inserting a space character between words. Until now, this has not been seen as a big challenge as current data processing and analysis has focused on word based tokenization, and then the transformation of tokens to lemmas. The additional layers of the full string and the words have seen less value for retention in the business model and resultitive praxis of work focused on *The Bridge*.

There are certain trade offs and compromises due to the technical implementation of *The Bridge* in its current state. Storage requirements and query speeds are two very important factors to consider. A third important factor to consider is that *The Bridge* is a digital service to a consumer-centric target audience. If *The Bridge* is a consumer centric service-as-product, then what is to be gained by providing a linked data interface? That is, *by providing its data as linked data to other consumers what problems does it solve?* The converse is also true, *by consuming other Latin Linked Data Project's data what features and interactions does it offer to the consumers of its services?* These two critical questions will help guide development strategy with regards to linked data and the future of *The Bridge*.

With language learners and their language teachers in mind, there are at least two reasons to consider full text linked data data-management approaches over current approaches to data management.

1. Linked Data gives a way for the sustainable development of data and analysis preservation through the management of a single-file resource per FRBR (Group 1) entity based object.

²⁶ An interesting computational experiment would be to create arrays attached to each unique lemma and put the IDs of each textual unit in the array. In this way the arrays would be smaller even if there are more arrays. The potential approach stands in contrast to the current method of creating arrays of lemmas per textual unit. By flipping the indexed unit look-up times might be reduced. Profiling and experimenting in this nature is something that an aspiring computer science professional could do.

²⁷ An experiment with the LiLa base API returning the text in the POWLA format saw an increase from 123kb to 15.5MB.

2. Using a Linked Data model for the whole text enables language learning tools to be developed beyond the current approaches and even Linked Data approaches depending on the generative/lexical model of linguistics.

With regards to point number one above, FRBR is explained in more detail in the section [Record Scope](#). With regards to point number two, the remainder of this section expounds on how a full text encoding can support language learning and teaching activities.

The current value proposition to end-users made by *The Bridge* is that by comparing the lemmas found in two texts that one can assess the percentage of new vocabulary relative to a student's current knowledge—assuming that the student is versed in the vocabulary of one of the two texts. However, there are several assumptions to consider here. The first is that lemmas are a viable approach for comparison in teaching environments. Second, there is a general agreement among scholars on how lemmas are to be constructed. Both of these issues impact the potential utility of *The Bridge* to end users, search-and-retrieval engineering, as well as data storage formats.

Lemmas²⁸ are one way that word instances and diverse word forms can be associated. The theoretical notion of lemmas suggests that these parts of words are stored in human memory and that “fuller” words are derived (“generated”) from the lemmas via mental processes. In contrast to generative models based on the idea of an internal lexicon,²⁹ other theories suggest that sound symbols exist within context called constructions.³⁰ Latinists frequently use the tools of Construction Grammar in their analysis of Latin.³¹ This is to say, strong support for informatic models presuming generative approaches to linguistics have limited value in the field of linguistics and in neighboring disciplines such as pedagogy. Pedagogy in second language

²⁸ Knowles, Gerry, and Zuraidah Mohd Don. 2004. “The Notion of a ‘Lemma’: Headwords, Roots and Lexical Sets.” *International Journal of Corpus Linguistics* 9 (1): 69–81. <https://doi.org/10.1075/ijcl.9.1.04kno>.

²⁹ Pustejovsky, James. 2002. *The Generative Lexicon*. Cambridge, Mass.: MIT Press.

³⁰ Croft, William. 2021. *Ten Lectures on Construction Grammar and Typology. Distinguished Lectures in Cognitive Linguistics 11*. Leiden; Boston: Brill. <https://doi.org/10.1163/9789004363533>.

Croft, William. 2023. “Philosophical Reflections on the Future of Construction Grammar (or, Confessions of a Radical Construction Grammarian).” <http://www.unm.edu/~wcroft/Papers/FutureCxG-rev1.pdf>.

Fried, Mirjam. 2014. “Construction Grammar.” In *Linguistics*. Oxford Bibliographies. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/obo/9780199772810-0061>.

Goldberg, Adele E. 2009. “Constructions Work.” *Cognitive Linguistics* 20 (1): 201–24. <https://doi.org/10.1515/COGL.2009.013>.

Haspelmath, Martin. 2023. “On What a Construction Is.” *Constructions* 15 (1): Article: 539. <https://doi.org/10.24338/CONS-539>.

Michaelis, Laura A. 2006. “Construction Grammar.” In *The Encyclopedia of Language and Linguistics*, edited by K Brown, 2nd ed., 3:73–84. Oxford: Elsevier.

³¹ Guardamagna, Caterina. 2018. “Type Frequency, Productivity and Schematicity in the Evolution of the Latin Secundum NP Construction.” In *Grammaticalization meets Construction Grammar*, edited by Evie Coussé, Peter Andersson, and Joel Olofsson, 169–201. *Constructional Approaches to Language* 21. Amsterdam: John Benjamins Publishing Company. <https://doi.org/10.1075/cal.21.c7>.

Lahey, Holly A. 2015. “The Grammaticalization of Latin Nē + Subjunctive Constructions.” *Journal of Latin Linguistics* 14 (1): 65–100. <https://doi.org/10.1515/joll-2015-0004>.

Pompei, Anna. 2016. “Construction Grammar and Latin: The Case of Habeo.” *Pallas. Revue d'études Antiques*, no. 102 (November): 99–108. <https://doi.org/10.4000/pallas.3601>.

acquisition is increasingly applying insights from usage based models and Construction Grammar.³² Tokenization and lemmatization processes do not help pedagogy specialists to identify *Multi-Word Expressions* or *constructions* which are the aspects of language that the literature claims language learners grasp first. The notations of Multi-Word Expressions and Constructions crossing word boundaries are still an unapproached issue in big Latin Linked Data projects such as LiLa.³³

Lemmatization as a process is of interest to scholars of a variety of languages. There are professional disagreements among Latinist about how these Latin lemmas should be constructed (e.g., between the LASLA and LiLa projects). For some information processing purposes, the lemma becomes less important than an identifier for the lemma. A great deal of informatic work has been conducted as part of CTS to create identifiers for these lemma.³⁴ Considering the differences between LiLa and LASLA lemmatization processes, it is important to point out the alignment work between the two datasets which has been carried out. Alignment work was carried out as part of the LiLa project.³⁵ However, based on interviews with Haverford-based Latin experts there is still significant work to do in this area.

In summary, first the use of linked data approaches which encode the whole text enable theory-biases (or theory informed) annotation of the text. This approach allows different theoretical views to be annotated within the same framework and support many more of the topics important to language teaching. Second, Latinists don't completely agree on lemma formation.

Latin Text Sources

There are quite a few sources for digital Latin texts. It is not clear how many of these collections of texts contain unique texts versus different formats or annotated versions of the same texts.

³² Endresen, Anna, Valentina Zhukova, Elena Bjørgve, Daria Demidova, Natalia Kalanova, Zoia Butenko, George Lonshakov, and David Lavén. 2022. "Construxercise! Implementation of a Construction-Based Approach to Language Pedagogy." *Russian Language Journal* 72 (1). <https://scholarsarchive.byu.edu/rj/vol72/iss1/4>.

Herbst, Thomas. 2016. "Foreign Language Learning Is Construction Learning – What Else? Moving towards Pedagogical Construction Grammar." In *Foreign Language Learning Is Construction Learning – What Else? Moving towards Pedagogical Construction Grammar*, edited by Sabine De Knop and Gaëtanelle Gilquin, 21–52. Applications of Cognitive Linguistics [ACL] 32. Berlin; Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110458268-003>.

Tomasello, Michael. 2009. "The Usage-Based Theory of Language Acquisition." In *The Cambridge Handbook of Child Language*, edited by Edith Laura Bavin, 69–88. Cambridge Handbooks in Linguistics. Cambridge: Cambridge university press.

³³ When I interviewed LiLa project staff I specifically asked about these issues.

³⁴ <https://opengreekandlatin.org/what-is-a-cts-urn> ; <https://sites.tufts.edu/perseusupdates/2021/01/05/what-is-a-cts-urn> ; <http://capitains.org/pages/guidelines>

³⁵ Fantoli, Margherita, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. "Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin." In *Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, 26–34. Marseille, France: ELRA. <http://www.lrec-conf.org/proceedings/lrec2022/workshops/LDL/pdf/2022.ldl2022-1.4.pdf>.

Mambrini, Francesco, and Marco Carlo Passarotti. 2023. "The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms" *Journal of Open Humanities Data* 9 (1): 28 pages. <https://doi.org/10.5334/johd.145>.

- 5 Latin Treebanks³⁶: Perseus³⁷, Late Latin Charter Treebank,³⁸ Index Thomisticus Treebank, UDante,³⁹ PROIEL^{40,41}
- LASLA^{42,43}
- LiLa⁴⁴
- Perseus Digital Library⁴⁵
- Database of Latin Dictionaries⁴⁶
- Corpus Thomisticum⁴⁷
- Latin Library⁴⁸
- Patrologia Latina⁴⁹
- Open Greek and Latin⁵⁰
- Classical Latin Texts (A Resource Prepared by The Packard Humanities Institute)⁵¹

-
- ³⁶ Gamba, Federica, and Daniel Zeman. 2023. "Universalising Latin Universal Dependencies: A Harmonisation of Latin Treebanks in UD." In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, edited by Loïc Grobol and Francis Tyers, 7–16. Washington, D.C.: Association for Computational Linguistics. <https://aclanthology.org/2023.udw-1.2>.
- ³⁷ Bamman, David, and Gregory Crane. 2011. "The Ancient Greek and Latin Dependency Treebanks." In *Language Technology for Cultural Heritage*, edited by Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, 79–98. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-20227-8_5.
- ³⁸ Cecchini, Flavio Massimiliano, Timo Korkiakangas, and Marco Passarotti. 2020. "A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, et al., 933–42. Marseille, France: European Language Resources Association. <https://aclanthology.org/2020.lrec-1.117>.
- ³⁹ Cecchini, Flavio Massimiliano, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2021. "UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works." In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-It 2020: Bologna, Italy, March 1-3, 2021*, edited by Johanna Monti, Fabio Tamburini, and Felice Dell'Orletta, 99–105. Collana Dell'Associazione Italiana Di Linguistica Computazionale. Torino: Accademia University Press. <https://doi.org/10.4000/books.aaccademia.8653>.
- ⁴⁰ Haug, Dag TT, and Marius Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34. <https://dev.syntacticus.org/proiel.html>
- ⁴¹ <https://dev.syntacticus.org/proiel.html>
- ⁴² De Jong, J R, and P C Masereeuw. 1986. "Using a Latin Computer Corpus for Linguistic Research." *Revue, Informatique et Statistique Dans Les Sciences Humaines* 22 (1–4): 7–22.
- Denooz, Joseph. 2007. "Opera Latina: Le Nouveau Site Internet Du LASLA." *Journal of Latin Linguistics* 9 (3): 21–34. <https://doi.org/10.1515/joll.2007.9.3.21>.
- Verkerk, Philippe, Yves Ouvrard, Margherita Fantoli, and Dominique Longrée. 2020. "L.A.S.L.A. and Collatinus: A Convergence in Lexica." *Studi e Saggi Linguistici* 58 (1): 95–120. <https://doi.org/10.4454/ssl.v58i1.275>.
- ⁴³ https://www.lasla.uliege.be/cms/c_8508894/fr/lasla
- ⁴⁴ <https://lila-erc.eu/#page-top>
- ⁴⁵ <http://www.perseus.tufts.edu/hopper>
- ⁴⁶ <https://about.brepolis.net/database-of-latin-dictionaries>
- ⁴⁷ <https://www.corpusthomisticum.org> ; <https://mdr-maa.org/resource/corpus-thomisticum>
- ⁴⁸ <https://www.thelatinlibrary.com>
- ⁴⁹ <https://www.lib.uchicago.edu/efts/PLD> ; <https://patristica.net/latina>
- ⁵⁰ <https://opengreekandlatin.org>
- ⁵¹ <https://latin.packhum.org>

- Alpheios Project⁵²
- St. Thomas Aquinas' Works in English [and Latin]⁵³

Software

There are a lot of attempts at creating Latin NLP tools. In general, there is not a good comparison matrix for when one might want to use one of these tools over another one. The utility of the tools varies by task, method implemented to accomplish the task, and code language.

Latin Text Processing Software Tools

- Classical Language Toolkit⁵⁴
- Deucalion Latin Lemmatizer⁵⁵
- The Bridge Lemmatizer: “With Bridge/Lemmatizer you can create a lemmatization spreadsheet for any Latin or Greek text. To begin, either upload a .txt file or just paste the text. Lemmatizer will match all the words that have only one possible lemma, and will return a csv file for you to complete. All you have to do is disambiguate the remaining words and you have a fully lemmatized text. You can also add custom definitions and/or principal parts!”⁵⁶
- CST's Lemmatiser⁵⁷
- LEMLAT⁵⁸
- Collatinus⁵⁹ is a “lemmatiser and a morphological analyser for Latin texts: if a conjugated or declined form of a word is entered, it is capable of finding the correct root word to search for in the dictionary and then displaying its translation into another language, its different meanings, and any other information usually found in dictionaries” (Website description).
- LatinBert⁶⁰

⁵² <https://alpheios.net>

⁵³ <https://web.archive.org/web/20191109052048/https://dhspriority.org/thomas>

⁵⁴ <http://cltk.org>

⁵⁵ Clérice, Thibault. 2017. “Dire La Sexualité En Latin Classique et Tardif : ‘Une Étude Lexicographique Par Apprentissage Profond.’” PhD, École Doctorale 3LA, Laboratoire HISOMA, Université Lyon 3.

Clérice, Thibault. 2020. “Deucalion Latin Lemmatizer.” Zenodo. <https://doi.org/10.5281/zenodo.4043059>.

Clérice, Thibault. 2022. “Latin Deucalion, a Model for the Lemmatization and Morphosyntactic Tagging of Classical and Late Latin.” Python. <https://doi.org/10.5281/zenodo.3773327>.

“Deucalion, a Lemmatization Service - École Nationale Des Chartes.” n.d. Accessed December 11, 2023. <https://dh.chartes.psl.eu/deucalion/latin>.

Manjavacas, Enrique, Ákos Kádár, and Mike Kestemont. 2019. “Improving Lemmatization of Non-Standard Languages with Joint Learning.” In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 1493–1503. Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1153>.

⁵⁶ <https://bridge.haverford.edu/lemmatizer>

⁵⁷ <https://cst.dk/online/lemmatiser/uk>

⁵⁸ <http://nevenjovanovic.github.io/lemmatize-neo-latin-lemlat>

⁵⁹ <https://outils.biblissima.fr/en/collatinus> ; <https://github.com/biblissima/collatinus>

⁶⁰ Bamman, David. (2020) 2023. “Latin-Bert.” Shell. <https://github.com/dbamman/latin-bert>.

- LatinCy⁶¹
- Lamon, The Latin POS Tagger & Lemmatizer⁶²
- PyWORDS: Latin Word Look up tool⁶³

Relevant Python Modules

For scripting with python in the Latin linked data space the following Python modules are useful.

- *Latin Databases*⁶⁴ is a library for accessing a variety of Latin Lexical Databases.
- TheLatinLibrary⁶⁵ is a Python module for accessing TLL texts.
- *RDFLib*⁶⁶ is a library for manipulating and using RDF data.
- *SPARQL Wrapper*⁶⁷ is a Python library for making SPARQL queries more API like.

Bamman, David, and Patrick J. Burns. 2020. "Latin BERT: A Contextual Language Model for Classical Philology." *arXiv*. <https://doi.org/10.48550/arXiv.2009.10053>.

Lendvai, Piroska, and Claudia Wick. 2022. "Finetuning Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae." In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, edited by Michael Zock, Emmanuele Chersoni, Yu-Yin Hsu, and Enrico Santus, 37–41. Taipei, Taiwan: Association for Computational Linguistics. <https://aclanthology.org/2022.cogalex-1.5>.

⁶¹ <https://spacy.io/universe/project/latincy>

⁶² <https://github.com/bab2min/lamonpy>

⁶³ <https://github.com/sjgallagher2/PyWORDS>

⁶⁴ <https://pypi.org/project/latin-databases>

⁶⁵ <https://pypi.org/project/thelatinlibrary>

⁶⁶ <https://rdflib.readthedocs.io/en/stable>

⁶⁷ <https://sparqlwrapper.readthedocs.io/en/latest/main.html>

Legal Considerations

Across the field of Latin scholarship, data sets are distributed under different licenses.⁶⁸ Because the context of Latin projects are most frequently *texts-of-antiquity*, copyright is likely not able to apply. This should call into question the legal legitimacy of the Creative Commons licenses⁶⁹ and underlying copyright claims on these texts. My investigation suggests that copyright is not applicable to the texts. A remaining question exists if marked-up or annotated texts are subject to copyright. Further, if the annotations are simply identifiers within an XML schema, and the XML schema is openly licensed independently, then the annotations may not meet a creativity threshold for sustaining a copyright claim.

A second argument is also possible. First, mark-up encoded texts are often referred to as data. Data in-and-of itself is not copyright-able as data represents facts not a creative work. Facts are exempted under copyright law. Therefore, if viewing marked-up texts as “data”, there may not be any legitimate copyright claim, thus the licenses with their basis in copyright are nullified.

For these reasons, I have real doubts about the legal legitimacy of creative commons licenses applied to linked data. The over-licensing of data within the field of linguistics and associated fields of the classics and philology ought to be a grave concern to practitioners as it has the potential to reduce the kinds of collaborations, services, and engagement that scholars desire. The use of the CC0⁷⁰ or a Public Domain Dedication⁷¹ can make copyright-free statuses for

⁶⁸ For example, LASLA is released under the creative commons BY-CC-NC-SA license. <http://ancientworldonline.blogspot.com/2023/10/the-lasla-latin-corpus-has-been.html>. Non-commercial does not mean anything-but-for-profit businesses. It means that it can not be used in a formal organizational context. Organizations conduct activities and these activities are commerce. Institutions of higher education are in the business of selling learning experiences. Integrating LASLA into these environments would violate the terms of use because it is part of the commercial activities of the organization (university). Additionally, the resource can only be integrated in contexts in which other resources are NC-SA due to the SA (share-alike) clause of the license. This means that a CC-BY resource mixed with the NC-SA resource would need to be re-licensed with the NC-SA clauses. This becomes a more restrictive “over licensing” of the originally more open content. Additionally, the LASLA and LiLa teams in a joint statement published by *Ancient World Online* call the LASLA resources “Open Access” (also a Poster: Fantoli, Margherita, and Marco Passarotti. 2022. “Linked Open Data as a Path towards Open Science.” In Proceedings 2022. PubPub. <https://doi.org/10.21428/1192f2f8.1d70c293>.). This is flagrant disregard for the scholarly community accepted definition of Open Access which specifically states that there is no barrier to the resource except the internet itself. In some cases, the LASLA resources are restricted by the necessity to log-in. Further, the Budapest Open Access Initiative excludes resources which are CC-BY-NC-SA from being open access by virtue of their license. Open Access must be: Copyright Free, CC0, or CC-BY. <https://www.budapestopenaccessinitiative.org/read>. In contrast to LASLA’s license requirements to only be used in Non-Commercial contexts, the LiLa Lemma-Bank’s current license does not have a Non-Commercial clause (it is CC-BY-SA), meaning that it is more permissive, but also not Open Access (BOAI) compliant. LiLa’s materials would become only usable in a Non-Commercial context if they were mixed with LASLA content. (LiLa Lemma-Bank prior to July 2023 had an SA clause in their license). There is no copyright claimant mentioned in the LiLa Lemma-Bank license so it might not actually be licensed legitimately.

⁶⁹ A Creative Commons license is only as valid as the copyright claim which supports the license. Creative Commons licenses require a valid copyright claim in order to grant the privileges to users under a copyright-based legal framework.

⁷⁰ <https://creativecommons.org/publicdomain/zero/1.0/deed.en>

⁷¹ <https://opendatacommons.org/licenses/pddl/1-0>

published data sets clear for downstream re-uses. These legal tools should be explored with regard to their possible application to Haverford projects.

Licensed Data is data that *The Bridge* and associated project have imported or elected to use under license from other projects. The license particulars of these data need to be tracked so that *The Bridge* does not violate the terms of the license. For example, some Latin projects allow data re-distribution and others do not. My recommendation for tracking these license terms is to keep track of the [license information in the bibliographic record](#).

Future Publishing Opportunities

There are five direct areas of application from my work done via LEADING Fellowship which deserve professional public engagement. These are outlined below:

1. The Bridge's genre list can shed light on the applicability and utility of the Open Language Archive Community's genre list. A comparison of text genres based on developing interactive experiences around a collection of Latin texts would be of interest to the 2024 LangDoc conference.
2. Latin is a complex set of speech varieties. There is no succinct way to indicate these via BCP-47 codes. There should be. An informed proposal should be made to the ISO 639-3 registrar. Following an unsuccessful proposal, a subsequent application should be made to the IANA language subtag registrar.
3. To the best of my knowledge, very little has been published about technical aspects of workflows related to the digitization pipelines in digital humanities projects. A paper on using Dublin Core in a Digital Humanities pipeline process would exposit the utility of the WEMI model and how to use Dublin Core to manage various aspects of a project. This could easily be published via the DCMI conference and Digital Humanities Quarterly.
4. Across the scholarship of the classics and linguistics more broadly, how texts are controlled is a big issue. The major Latin Linked Data projects all have different licenses. The legal illegitimacy of the licensed content could be investigated if a collaboration with a lawyer were pursued. A review paper of these licenses, their legitimacy, and the situational impact on scholarship with a better solution could be published in a classics journal and a law journal. E.g., a possible title/topic could be: *Latin Texts and Legal Constraints in the Open Access Linguistic Linked Data Context*.
5. The NEH and other funders are increasingly asking for insights into the sustainability of projects. However, what does sustainability mean for digital humanities projects? Data might be preserved in a repository, software may be made open access, but what about the interactive functionality and the social impact resulting from those interactions? The digital humanities is often dependent on software to implement research and social impact goals, but there is very little guidance on how to select software to reduce technical debt over time. Open source software selection is not enough to reduce the investment cost for digital humanities projects. The best software is open source and commercially used and contributed to. A publication on sustainability management in digital humanities projects would greatly contribute to the field and beyond. Many grant

winners do not have years of experience in software project management or lifecycle management. Concrete recommendations would provide applicants to the NEH and other funders with actionable information to bring into their praxis.

Import Workflows

The goal of this section is to document current and recommended terminology for the categories of information imported-to or generated-through the Haverford processes related to making Latin texts compatible with *The Bridge*. There are two general approaches, a top-down approach which (a) evaluates categories required by the User Interface of *The Bridge* and (b) “squeezes” data into these categories for the benefit of users. Or a second bottom-up approach which investigates the categories in the existing data and then provides definitions for those. The approach taken here is a bottom-up approach.

There are seven current processes. Each is outlined in its own section below. The seven processes include:

1. [File scraped from Perseids](#)
2. [Simple List](#)
3. [Textbooks](#)
4. Simple locally generated lemmatized text
5. [Text from LASLA](#)
6. Text from a Concordance
7. Text from PROIEL Treebank

Perseids

In this section, I discuss files scraped from Perseids (XML converted to spreadsheet and aligned with Bridge Titles)... except that I only see XML files not spreadsheets in the particular folder I found on Box.com. Maybe there are other files in a spreadsheet format.

- **Description:** Perseids⁷² is a classics humanities project which is built upon a collection of texts (Perseus Digital Library),⁷³ and adds to it tools,⁷⁴ and resources presenting lexical analysis⁷⁵ for a variety of languages—including Latin. The Latin texts from Perseids in the possession of Haverford have yet to be integrated into *The Bridge*. These include resources such as those from the Harrington Tree Bank (urn:cts:latinLit:phi0975.phi001.perseus-lat1.tb.xml).⁷⁶

⁷² <https://www.perseids.org>

⁷³

<https://www.perseus.tufts.edu/hopper/collection?collection=Perseus%3Acorpus%3Aperseus%2CLatin%20Text>

⁷⁴ <https://www.perseids.org/libraries-tools>

⁷⁵ There are several different Treebanks.

⁷⁶ <https://app.box.com/file/1211702227254>

- **Column Headings:** *The Bridge* generally imports data from CSV or XLS files. The Perseids files are currently in a custom XML format. They contain an XML header with a significant amount of bibliographic and provenance metadata including information about the digital source and contributors to the analysis. The XML files identify the:
 - The Source Document (the encodingDesc element: <encodingDesc> </encodingDesc>)
 - The Sentence element has the following attributes:
 - ID (**id**="1". The numeric ID value is incremented sequentially across the whole document and is not subject to restart for the subdoc sequence.)
 - The Current Document ID (**document_id**="urn:cts:latinLit:phi0474.phi013.perseus-lat1").
 - An internal document structure (**subdoc**="1.1").
 - Word (**id**="1" within the word element, which is nested within the sentence element. The ID element presumably is the order of the word within the sentence.)
 - The string is understood to be the "word" in an edited orthographic form (**form**="abutere").
 - A lemma with a conjugation type numeral (**lemma**="quo1").
 - A Part-of-Speech Tag (**postag**="v2sfid---" All strings are the same length. Each position of the string signifies one of several oppositional values.)
 - Relation (**relation**="AuxY") An exact linguistic analysis of each term within the Treebank is outside of the scope of the current project.
 - Head (**head**="4") This is likely the word ID of the term which is the head for the word upon which this attribute value appears. This is a type of grammatical relationship specific to a linguistic analysis.
- **Discussion:** In the Perseids context, lemma strings should be replaced with identifiers from within the LASLA or CTS systems with preference given to the CTS system. The lemma field seems to be indicating several different things. A lemma and (presumably) a conjugation via numeral. These two indexical components should be independent. It is hard to know what the various values in the relation field map to in other theories of grammar. These terms should be re-conceived in another more widely used theory of grammar (e.g., construction grammar). However, the various layers and analysis expressed via the treebank can be expressed within a POWLA conformant file via a series of analysis layers in conjunction with new layers representing a more accessible linguistic analysis.
- **Suggested Terms:** Use *Head Word* instead of *Lemma*.
- **Additional Terms:** N/A
- **Source Documentation:** Values possible in the *POSTag* attribute are documented in the internally accessible documentation.⁷⁷ Values of the *Relation* attribute are documented in the internally accessible documentation.⁷⁸ Values for the attribute *Head* are not clearly defined but are likely to be related to the grammatical framework called

⁷⁷ <https://app.box.com/file/1211692585058>

⁷⁸ <https://app.box.com/file/1211686591252> & <https://app.box.com/file/1211690408109>.

Dependency Grammar⁷⁹ which is used to create the Treebank's grammatical relations between lexical elements. A logical suggestion then is that the value of *Head* contains the Word ID of the element's head word linking the two words together. The specific application of Dependency Grammar is explained in a handbook for the treebank available via the internet archive.⁸⁰ The use of Dependency Grammar as an analytical framework does lead to some "atypical" grammatical categories within the POSTag and Relation tags. The Treebank's authors suggest that Latin has free variation word order. This may appear to be true, however, functionalist and usage based approaches to grammar analysis suggest that constructions will place bounds on the absolute freedom and provide structure. Two examples of greater attention being paid to grammatical categories in opposition to "traditional" grammatical categories (and syntactic structures) are constructions around reported speech,⁸¹ and the often divergent treatment of "particles" in African languages. For example, STAMP morphemes as described by Anderson⁸² "*STAMP morph (Anderson 2012; 2015). This is mnemonic for what these elements largely are, portmanteau morphs that encode the referent properties of semantic arguments that typically play the syntactic role of 'S[ubject]'*—that is, the person, number and gender properties of such an actant—in combination with categories of *T[ense], A[spect], M[ood] and P[olarity]. Such elements have also been previously called the tense-person complex (Creissels 2005), and pronominal predicative markers or pronominal auxiliaries (Vydrine 2011; Ėrman 2002) in the Africanist literature.*"⁸³

⁷⁹ De Marneffe, Marie-Catherine, and Joakim Nivre. 2019. "Dependency Grammar." *Annual Review of Linguistics* 5 (1): 197–218. <https://doi.org/10.1146/annurev-linguistics-011718-011842>.

⁸⁰

<https://web.archive.org/web/20170421060451/http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf>

⁸¹ Spronck, Stef, and Tatiana Nikitina. 2019. "Reported Speech Forms a Dedicated Syntactic Domain." *Linguistic Typology* 23 (1): 119–59. <https://doi.org/10.1515/lingty-2019-0005>.

⁸² Anderson, Gregory. 2017. "STAMP Morphs in the Macro-Sudan Belt." In *Diversity in African Languages: Selected Papers from the 46th Annual Conference on African Linguistics*, edited by Doris L. Payne, Sara Pacchiarotti, and Mokaya Bosire, 513–39. Contemporary African Linguistics 1. Berlin: Language Science Press. <http://langsci-press.org/catalog/book/121>

⁸³ Anderson, Gregory D. S. 2012. S/TAM/P morphs in the history of Benue-Congo and Niger-Congo conjugation. (Paper presented at Niger-Congress, Paris, September 2012.) Anderson, Gregory D. S. 2015. STAMP morphs in Central Sudanic languages. In Angelika Mietzner & Anne Storch (eds.), *Nilo-Saharan: Models and descriptions*, 151–167. Köln: Rüdiger Köppe Verlag.

Ėrman, Anna V. 2002. Sub"ektnye mestoimenija v dan-blovo i modal'no-aspektnotemporal'nye znachenija [Subject pronouns in Dan-Blowo and (their) modalaspectual-temporal meanings]. In Valentin F. Vydrin & Aleksandr Ju Zheltov (eds.), *Juzhnye mande: Lingvistika afrikanskikh ritmakh. Materialy peterburgskoj ekspeditsii v Kot d'Ivuar (k 50-letiju Konstantina Pozdnjakov)*, 154–82. Sankt-Petersburg: Evropejskij Dom.

Vydrine, Valentin F. 2011. Ergative/absolutive and active/stative alignment in West Africa. *Studies in Language* 35(2). 409–443.

Creissels, Denis. 2005. A typology of subject and object markers in African languages. In F. K. Erhard Voeltz (ed.), *Studies in African linguistic typology (Typological Studies in Language 64)*, 43–70. Amsterdam: John Benjamins.

Simple List

These resources were created by hand.

- **Description:** Several “Simple Lists” are imported into The Bridge. Presumably these lists are generated via consultation with a published resource. Consultation of the “Latin/texts/lists” folder⁸⁴ shows between 10 and 15 lists generated from various resources. The contents of these lists vary as do the categories of data present within the lists. For example, Mahoney 200 contains 5 columns while 50MOST only contains two columns. Most files in the “List Folder” are of the format XLSX. Some are of the format CSV. Some of the XLSX files contained several “tabs” in their workbooks. It is strongly recommended that XLSX files be avoided and CSV files be embraced. The two reasons for this include: 1) greater file portability by using open, non-proprietary file formats,⁸⁵ and 2) a move away from XLSX allows for the consistent and overt management of data through the workflow. It removes the temptation to stack multiple steps of a workflow on the same data into a single XLSX workbook. In addition to those files already mentioned, a folder “Diederich” exists. In contrast to the Simple Lists elsewhere in the folder, “Diederich” seems to be a project with multiple processing needs and objectives. This is a significant departure from the assumed processing workflow for the other resources discussed in this section. It is highly recommended that folders within the data structure be organized by workflow and the various stages within the workflow. Workflow suggestions are discussed in the section on [Import Workflows](#).
- **Example Documents:** Bridge_Latin_List_mahoney_200_words_prep.xlsx⁸⁶ and Bridge-Vocab-Latin-List-50MOST.xlsx⁸⁷
- **Column Headings:** The 50MOST list uses two header terms “TITLE” and “50 Most Important Latin Verbs”, while the Mahoney200 list employs the header terms: “TITLE”, “TEXT”, “LOCATION”, “SECTION”, “RUNNING COUNT”. These work in different ways, for example the Term “TITLE” seems to correspond with the concept of “lemma” or “Head Word”. While it is not clear what the numeric value of the content of the field for “50 Most Important Latin Verbs” is supposed to represent, an indexed sort order is apparent in Mahoney200 using the “LOCATION” field. However, it is not clear what “RUNNING COUNT” is trying to index or represent.
- **Suggested Terms:**
 - For “TITLE” a clearer column header consistent with the project would be “Lemma” or the even more preferred term “Head Word”.
 - For “TEXT” a clearer column header would be “Orthographic Form”.
- **Discussion:** When compared with data from other data sources, *Simple List* data appears to be impoverished. That is, some lists don’t even contain the full orthographic form of the Latin words. The spreadsheets use the field “TITLE” which I appears to be equivalent to “lemma” and a ‘reduced’ form.

⁸⁴ <https://app.box.com/folder/207180598647>

⁸⁵ Bird, Steven, and Gary F. Simons. 2003. “Seven Dimensions of Portability for Language Documentation and Description.” *Language* 79 (3): 557–82. <https://doi.org/10.1353/lan.2003.0149>.

⁸⁶ <https://app.box.com/file/1211601699183>

⁸⁷ <https://app.box.com/file/1211665217829>

In lexicography, Head Words are used to organize a variety of related content. Because *Column Headings* vary by compiled data source (XLSX file) it is hard to know if the same semantics has been applied to each vocabulary list. For example, why is “RUNNING COUNT” needed at all? Or what is the anchor that “SECTION” points to? Will TITLE always be a lemma?

A third major issue is the lack of a defined lexicographic model in which to integrate list terms. The following diagram presents a lexicographical approach as presented for use in multilingual dictionaries.⁸⁸

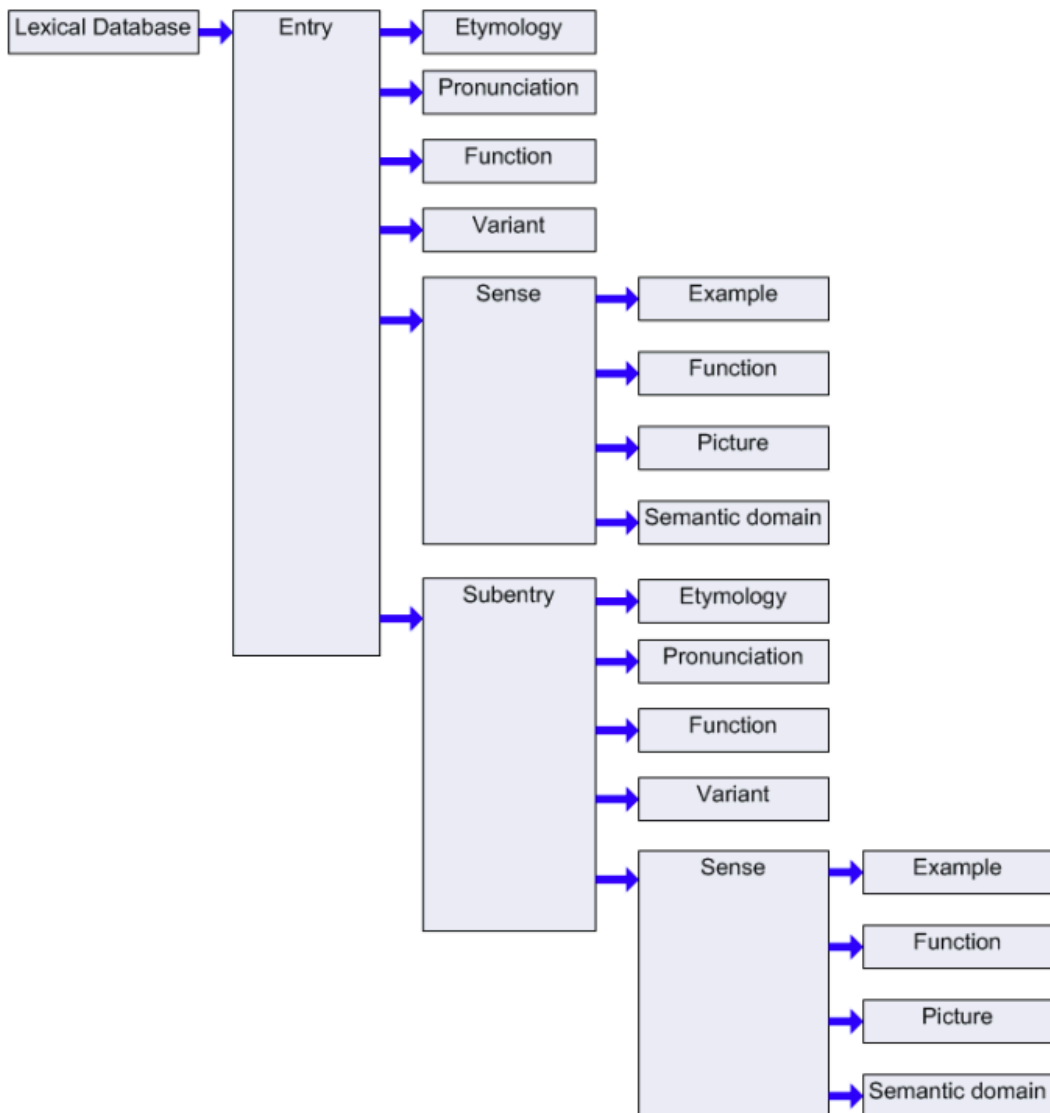


Figure 5: Multi-lingual Lexicographic Model used in SFM/ToolBox and FLEx

⁸⁸ Zook, Ken. 2013. “Technical Notes on SFM Database Import.” SIL International. https://downloads.languagetechnology.org/fieldworks/Documentation/Technical_Notes_on_SFM_Database_Import.pdf.

Adopting a clear lexicographic model would allow lists based data to be matched up to specific fields within the model. When all extracted texts are reduced to “lemmas” and also don’t include orthographic forms, it is hard to know if nominalized verbs are indexed with nouns or verbs as only the lemma is recorded.⁸⁹

Based on interviews with Haverford staff, a preliminary assessment about Latin Lexicography practices used within the project indicates a root-based organizational structure is desired. Such an approach is discussed by Moe (2016).^{90,91} However, the root is known within *The Bridge* project as a “lemma”. Organization by root has many practical advantages in languages like Latin; however, it does require a model which explicitly indicates variants which might be morphologically complex or easily (systematically) derived.⁹²

- **Additional Terms:** N/A
- **Source Documentation:** There was no self-evident source documentation for any of the “list files”. File names were descriptive but are a bad proxy for proper provenance information. The *why* and *how* of the creation of these lists would improve the provenance details and understanding of the creation context for future use for these files in the collection.

Textbooks

Textbook with Text-Specific Principal Parts & Definitions (generated locally via Lemmatizer app + Human completion)

- **Description:** Textbooks have a variety of texts and lists and usually center around learning units/chapters. These subdivisions can have as their base one or more texts.

⁸⁹ While I use illustrations and examples from SIL International’s Open Access and well documented MDF/SFM and FLEx based lexicography tools, it may be that the MDF based options do not represent a complete solution for Haverford’s work. The Haverford requirements and the potential for effective use must be investigated more closely.

⁹⁰ Moe, Ron. 2016. “Types of Dictionaries.” In Introduction to Lexicography for FieldWorks Language Explorer, §3.5. Dallas, Texas: SIL International.

<https://downloads.languagetechnology.org/fieldworks/Documentation/Intro%20to%20Lexicography/Introduction%20to%20Lexicography.htm#sTypesofDict>.

⁹¹ See also the discussion for the */lx* or *lexeme* field within the SFM framework for making dictionaries: Coward, David F. and Charles E. Grimes. 2000. *Making Dictionaries: A Guide to Lexicography and the Multi-Dictionary Formatter*, (page: 13). Waxhaw, North Carolina: SIL International.

⁹² A number of years ago a company in Switzerland, Canoo.net (now defunct), produced a website which would generate and break-down any German word based on its morphology, word class, and other grammatical features inherent to the word. (Example of the kind of breakdown: <https://web.archive.org/web/20061104215804/http://www.canoo.net/services/WordformationRules/Derivation/To-N/N-To-N/Suffig.html#Anchor-ohne-49575>). This was an extremely useful tool for parsing German words as a student and as a commercial text translator. I have yet to see something similar for Latin, but it seems to be something which would be possible. Dr. Stephan Bopp of the University of Zürich is the most knowledgeable of Canoo.net’s technologies (<https://blog.leo.org/wer-ist-dr-bopp>). An interview or correspondence with Dr. Bopp might provide some fodder for considering a similar tool for Latin. No such tool or interface has been seen in the LASLA, LiLa, Alpheios, or other projects.

Approximately 35 individual data files (usually XLSX) with various tabs and header columns exist along with approximately an additional 13 folders with additional projects.

- **Example URL:** Folder⁹³, LLPSI-VOCAB.xlsx,⁹⁴ LLPSI Vocab.xlsx,⁹⁵ Wheelock_Latin_Exercitationes.xlsx⁹⁶
- **Column Headings:** From LLPSI Vocab.xlsx⁹⁷: “CHECK”, “TITLE”, “CHAPTER (add number of chapter; e.g. for Chapter 1, add a 1 to this column)”, “DISPLAY LEMMA IVY L”, “DISPLAY LEMMA BRIDGE”, “PROBLEM”. The file LLPSI-VOCAB.xlsx⁹⁸ is similar except that it has additional tabs/worksheets of data—most of these without column headers. In contrast to the previous two mentioned files lock_Latin_Exercitationes.xlsx uses different header titles: “CHECK”, “BRIDGE”, “TEXT”, “LOCATION”, “SECTION”, “RUNNING COUNT”, “DISPLAY LEMMA”, “SHORT DEF”, “LONG DEF”, “PROBLEM”.

Still other textbooks employ even more headings. For example, in a Textbook titled LNM⁹⁹ within file LNM1_Readings_Lemmatization-completed.xlsx¹⁰⁰ several additional fields are used including: “DISPLAY LEMMA MACRONLESS”, “SIMPLE”, “SHORT DEF”, “LONG DEF”, “LASLA Combined”, “Decl”, “Conj”, “Reg Adj/Adv”, “Proper”, “Part Of Speech”, “LOGEION LINK”, “PROBLEM”. Some of these fields (e.g., Decl, Conj, Reg Adj/Adv, Proper, Part Of Speech) are likely to refer to paradigms and grammatical forms of the lexical item as used. These are likely to be similar to Perseids’s POSTag and the kinds of information contained therein. How these two kinds of information should be reconciled requires someone with Latin Language expertise.

- **Suggested Terms:**
 - Instead of “CHECK” consider “Status”. Devise a simple ontology to use as a controlled value in the new status field possible terms in the ontology include: “Verified”, “Congruent”, “Needs Review”, “Different”, etc. The values of the PSO ontology may be useful here.¹⁰¹
 - Instead of “CHAPTER” break apart the larger resource into individual files per Chapter. Use the FRBR/WEMI model to inform file boundaries.
 - Instead of “PROBLEM” consider a more generic “Note” or “Status note” field.
 - The use of ‘Display Lemma’ in two different column headings is confusing. The column headings have unique titles and the values are different “aberrō -āre” vs. “aberrō -errāre” but the similarity in the column heading doesn’t clearly identify the purposes.
 - Instead of “SHORT DEF” consider “Gloss”. The content requirements for this field are not clearly defined.

⁹³ <https://app.box.com/folder/207181807386>

⁹⁴ <https://app.box.com/file/1211669439203>

⁹⁵ <https://app.box.com/file/1211678587447>

⁹⁶ <https://app.box.com/file/1211674779387>

⁹⁷ <https://app.box.com/file/1211678587447>

⁹⁸ <https://app.box.com/file/1211669439203>

⁹⁹ <https://app.box.com/folder/207177610640>

¹⁰⁰ <https://app.box.com/file/1211677975617>

¹⁰¹ <https://sparontologies.github.io/psocurrent/psocurrent.html>

- Instead of “LONG DEF”, consider the role of the field and perhaps just use the term “Definition”. The content requirements for this field are not clearly defined.
- “Parts of Speech” values should be identifiers to a dereferenceable controlled vocabulary. Within Linguistic Linked Data the GOLD ontology has had a significant presence.¹⁰² Regardless of ontology used, POS definitions must be grounded in theory and clearly defined, as there are diverse definitions for common POS terms.
- **Discussion:** In contrast with other resources some of these resources have diacritics. The difference between diacriticed and non-diacriticed Latin is an orthographical difference. Lexicographic tools like linked data and FLEx use BCP-47 variant tags to indicate difference in orthography usage on a per field basis. This allows a single field to have multiple string values each in a different orthography. The distinction between “Display Lemma” and “Lemma” is not clear. “BRIDGE” as a column header is not clear because it is not clear with what it contrasts.
- **Additional Terms:** BCP-47 variant tags should be employed to identify the orthography of strings.
- **Source Documentation:** Source documentation was not available on a resource specific basis.

Simple locally generated lemmatized text

Simple locally generated lemmatized text (generated locally via Lemmatizer app + Human completion)

- **Description:** These texts were lemmatized with a Haverford developed lemmatizer and completed.
- **Example URL:** Bridge_Latin_Text_Vulgate_Genesis_37_43_prep.xlsx¹⁰³
- **Column Headings:** “TITLE”, “DISPLAY LEMMA”, “DISPLAY LEMMA MACRONLESS SIMPLE”, “SHORT DEF”, “LONG DEF”, “LASLA Combined”, “Decl”, “Conj”, “Reg”, “Adj/Adv”, “Proper STOPWORD”, “Part of Speech”, “LOGEION LINK”, “FORCELLINI LINK”.
- **Suggested Terms:** Consider using ID instead of LINK.
- **Discussion:** LOGEION LINK and FORCELLINI LINK would be more efficiently saved as IDs rather than full URLs. Additionally these IDs might be better managed in another table which is matched to a Bridge ID rather than on a direct lemma record.
- **Additional Terms:** N/A
- **Source Documentation:** No resource specific documentation was observed.

Text from LASLA

Texts from LASLA originate in the LASLA project and are converted to The Bridge format.

¹⁰² <https://web.archive.org/web/20200201190950/http://linguistics-ontology.org/version>

¹⁰³ <https://app.box.com/file/1211691868965>

- **Description:** Text from LASLA (converted from fixed-width .BPN into CSV)
- **Example URL:** Plautus_Captiui_PICapt.xlsx¹⁰⁴
- **Column Headings:** “TITLE”, “TEXT”, “LOCATION”, “SECTION”, “RUNNINGCOUNT”, “PRINCIPAL_PARTS”, “SHORT_DEFINITION”, “LONG_DEFINITION”, “LOCALDEF”, “PROBLEM”, “CASE”, “GRAMMATICAL_CATEGORY”, “GRAMMATICAL_CATEGORY_SUB”, “SUBORDINATION_CODE”, “_merge”.
- **Suggested Terms:** Alignment of grammatical concepts and terms needs to be addressed between various tree banks, LASLA, LiLa, and linguistic theory. This work goes beyond looking at just the lemmas.
- **Discussion:** Some terms and titles have already been addressed in previous sections. New terms which need a Latin specialist to evaluate include: “GRAMMATICAL_CATEGORY”, and “GRAMMATICAL_CATEGORY_SUB”, “SUBORDINATION_CODE”. These codes and their content should be compared with the meanings inferred by Perseids. This comparative work requires additional knowledge in Latin, Linguistics, specifically Linguistics as understood or taught from a Classics perspective.
- **Additional Terms:** N/A
- **Source Documentation:** CodesSubordinationBPN.doc¹⁰⁵ StructureBPNFiles.DOC¹⁰⁶ CodesBPN.DOC¹⁰⁷

Text from a Concordance

- **Description:** These types of files have undergone a process where they were OCR'd, then human-proofed, then scraped by Python scripts into a spreadsheet, which is then aligned to Bridge lemmas.
- **Example URL:** Lucretius¹⁰⁸ and Bede¹⁰⁹
- **Column Headings:** Lucretius files such as “Bridge_Latin_Text_Lucretius_DeRerumNatura_LucrDRN_all_7_2020.xlsx”¹¹⁰ contain headers which have been discussed already: “TITLE”, “LOCATION”, “SECTION”, “RUNNINGCOUNT”, “TEXT”. Other files in the Bede via instance Lucretius set contain different column headings. Other concordances and their associated files such as “Bede_Book1_Prep.xlsx”¹¹¹ contains the following: “CHECK”, “TITLE”, “TEXT”, “LOCATION”, “SECTION”, “RUNNING COUNT”, “LOCALDEF”. The terms here also seem to be congruent with terms already presented and discussed. However, other files in the Bede folder contain other Column Headings. This suggests that files are products of various workflows or representative of various stages in a single workflow. The file hosting platform shows the date the files were uploaded to the server, not the file

¹⁰⁴ <https://app.box.com/file/1211691065306>

¹⁰⁵ <https://app.box.com/file/1211692360904>

¹⁰⁶ <https://app.box.com/file/1211691838385>

¹⁰⁷ <https://app.box.com/file/1211690549084>

¹⁰⁸ <https://app.box.com/folder/207175904003>

¹⁰⁹ <https://app.box.com/folder/207177595653>

¹¹⁰ <https://app.box.com/file/1211596665889>

¹¹¹ <https://app.box.com/file/1211609134764>

creation or modification date. One new column heading found in the file “Bede_combined-7-1-2017.xlsx”¹¹² contains a column titled DCC.

- **Suggested Terms:** Articulate a specific semantic domain strategy. Preferably one which can handle several semantic domain models.
- **Discussion:** DCC seems to be related to the semantic domain of the word. Semantic domain¹¹³ and Frames¹¹⁴ are closely related to WordNet classifications.¹¹⁵ The relevant academic discussion around semantic domains in lexicography center on emic and etic views of semantic domains relative to the language analyzed. That is, should semantic domains be attached to cross-linguistic and global ontologies or should they be derived from the concepts expressed within the language and bound by the language-culture context. It is not clear to which view, emic or etic, DCC lends its analysis. When The Bridge includes semantic domain information, it should have a field for each ontology it is mapped to. Ideally IDs to locations in semantic domain ontologies should be used.
- **Additional Terms:** DCC could be generalized to “semantic domain” but if DCC is a specific ontology then generalization may not be recommended.
- **Source Documentation:** No resource specific documentation was found.

Text from PROIEL Treebank

- **Description:** PROIEL texts are annotated texts mostly from the Perseus Digital Library with other information added by the PROIEL project. No text documents were found in the Box.com repository. However, texts were found in the PROIEL repository via Github.

¹¹² <https://app.box.com/file/1211623109195>

¹¹³ Moe, Ronald. 2003a. “Compiling Dictionaries Using Semantic Domains.” *Lexikos* 13: 215–23. <https://doi.org/10.5788/13-0-731>.

Moe, Ronald. 2003b. “Dictionary Development Program.” *Word & Deed* 2 (1): 55–65.

Moe, Ronald. 2016a. “Collecting Words Using Semantic Domains.” In *Introduction to Lexicography for FieldWorks Language Explorer*, §4.1.1. Dallas, Texas: SIL International. <https://downloads.languagetechnology.org/fieldworks/Documentation/Intro%20to%20Lexicography/Introduction%20to%20Lexicography.htm#sCollectSemD>.

Moe, Ronald. 2016b. “Using Semantic Domains to Define Words.” In *Introduction to Lexicography for FieldWorks Language Explorer*, §4.3.1. Dallas, Texas: SIL International. <https://downloads.languagetechnology.org/fieldworks/Documentation/Intro%20to%20Lexicography/Introduction%20to%20Lexicography.htm#sDefSemDom>.

¹¹⁴ Fillmore, Charles J. 1982. “Frame Semantics.” In *Linguistics in the Morning Calm: Selected Papers from SICOL-1981*, edited by The Linguistic Society of Korea, 111–37. Seoul, Korea: Hamshin Publishing Company. https://brenocon.com/Fillmore%201982_2up.pdf.

Fillmore, Charles J. 1985. “Frames and the Semantics of Understanding.” *Quaderni Di Semantica* 6 (2): 222–54. <https://www1.icsi.berkeley.edu/pubs/ai/framesand85.pdf>.

Ziem, Alexander. 2014. *Frames of Understanding in Text and Discourse*. Human Cognitive Processing 48. Amsterdam: John Benjamins Publishing Company. <https://benjamins.com/catalog/hcp.48>.

¹¹⁵ Dacanay, Daniel, Atticus Harrigan, Arok Wolvengrey, and Antti Arppe. 2021. “The More Detail, the Better? – Investigating the Effects of Semantic Ontology Specificity on Vector Semantic Classification with a Plains Cree / Nêhiyawêwin Dictionary.” In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, edited by Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, 143–52. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.americasnlp-1.15>.

- **Example URL:** “Jerome's Vulgate”,^{116,117} which is sourced in 2014 from Perseus_text_1999.02.0060.xml
- **Column Headings:** (from screenshot) “LEMMA_PROIEL”, “TITLE”, “LOCATION”, “SECTION”, “RUNNINGCOUNT”, “TEXT”, “PP”, “Short Def”, “LEMMA_PRO...”, “POS”, “token_id”, “morphology”, “head-id”, “relation”, “information-...”, “presentation-...”.
- **Suggested Terms:** N/A
- **Discussion:** Because many texts are based on the Perseus Digital Library text, some of the information structure will be the same. See discussion and notes for Perseus texts and their content structure. Some annotations and the methods of annotation are different.
- **Additional Terms:** N/A
- **Source Documentation:** XML Format documentation,¹¹⁸ and Lemma, Part of Speech, and Morphology.¹¹⁹

¹¹⁶ <https://raw.githubusercontent.com/proiel/proiel-treebank/master/latin-nt.xml>

¹¹⁷ <https://raw.githubusercontent.com/proiel/proiel-treebank/master/latin-nt.conll>

¹¹⁸ <https://dev.syntacticus.org/development-guide/#the-proiel-xml-format>

¹¹⁹ <https://dev.syntacticus.org/development-guide/#lemma-part-of-speech-and-morphology>

Data Spreadsheet Header Summary Table

Simple List	Textbooks	Simple locally generated lemmatized text	Text from a concordance	Text from LASLA	Text from PROIEL Treebank	Perseids	Suggested Terminology
Haverford Created Resources				LASLA Source	Same Treebank Source		
Title	Title	Title	Title	Title	Title	Lemma	<i>Head Word</i>
Location	Location	Location	Location	Location	Location		
	Section	Section	Section	Section	Section		
	Running	Running	Running	Running	Running		
	Text	Text	Text	Text	Text	Word-form	<i>Orthographic Form</i>
	Local_de			Local_def			
	Local_Lemma						
	Original PP						
				Display_Lemma			
				Short_Def	Short Def		<i>Gloss</i>
				Long_def			<i>Definition</i>
				Problem			<i>Note</i>
				Case			
				GRAMMATICAL_CATEGORY			
				GRAMMATICAL_CATEGORY_SUB			
				SUBORDINATION_CODE			
				_merge			
					PP		
					Lemma_pro		
					POS	postag	
					Token_id		
					morphology_...		
					head-id	head	
					relation	relation	
					information-...		
					presentation-...		
						Sentence-ID	
						Sentence-document_id	
						Sentence-subdoc	
						Word-ID	

Data File Organization Suggestion

The organization of files shows evidence of several systems of organization. For example, some works were directly nested under the folder “Latin” while others were nested under “Textbooks”, “Texts”, “Lists”, “Inscriptions”, etc. At the same time “Lists” contains a folder with a resource which appears to be more than a “list”.¹²⁰ A consistent approach to data file management will improve team efficiency and performance across sub-projects. The following suggestion is workflow-based.

For each workflow, map out the stages necessary and document those stages in the root folder. Each lower folder within the hierarchy should contain a README.txt file with a defined structure describing the quantity and titles of the resources within the folder as well as critical provenance or workflow stages or procedures applied to the resources. Within the root folder, establish a new folder for each resource which will be placed through the workflow. Sometimes each resource may have components which need grouping. I use the term “Work” (borrowing from the WEMI ontology) to reference these items.

- Workflow based folder
- > Resource based folder
- >> Work based folder

To exemplify these we might use the following structure for textbooks. A README.txt would appear in each folder with relevantly scoped provenance information. This README.txt could be structured with specific topics or fields to fill in or it could be more free-form text. With a more structured document parts of it may be added programmatically.


- Workflow for Textbooks
- > Wheelock Textbook
- >> Wheelock Text1

Within the *Wheelock Text1* folder several files may be needed. For example, the original file, the OCR'd file, a corrected OCR file, and an extraction of the content. The folder for *Wheelock Text1* would then contain five files:

- README.txt for Process
- Text1-ori
- Text1-ocr
- Text1-cor (rected)
- Text1-ext (racted)

¹²⁰ I emphasize that the processing of language resources is by workflow-type as dictated by the processing steps necessary to structure the workflow. It is important to keep in mind that this means that processing is not necessarily based upon the genre (nature) of the resource. For Example, resources within the “Simple List” workflow should not be there because they are “lists” in structure/genre but because they need the same processing steps as the other items within the workflow. It may be true that the workflow for simple lists is specialized or accommodated to a specific genre of resources.


Bibliographic Record Management

The goal of this section is to review and make suggestions regarding the classification of terms in the bibliography spreadsheet.  Bridge About Text Dataset

Reviewed materials

The following terms are used in the headings of columns and therefore appear as lower level sections below. In contrast to the order of appearance in the spreadsheet, I have taken liberty to regroup the terms into what I perceive as mini-groups of references: *Text, Data Format, Provenance, Source Text, Special Notes about Format, etc., Local Definitions?, Genre, Language, Meter, Sample Formats, Type, Era, External Link, Other Notes, Problem?, Divisions, Date Created, Last Modified, Date Published, Status-OLD, Status-NEW, Kind, Author, Includes (Range of Text Present in Database), ShortName, Syntax Data?*

Terminological Sources

These terms were investigated with regard to their purpose and function within the Haverford workflow context (see Figure 2). These terms were taken from the “Bride Texts” tab of the spreadsheet: “ Bridge About Text Dataset ”.¹²¹ An attempt was made to use Qualified Dublin Core term titles as a superset. Supplemental terminology was sourced from the application profile for describing language resources used by the Open Languages Archives Community (OLAC).¹²² This approach allows the data to easily become linked data if the project were to choose such a direction. It also allows *The Bridge* Project to also become a data contributor to the OLAC aggregator for language resources. OCLC (WorldCat) and CLARIN.EU (The Virtual Language Observatory)¹²³ both harvest OAI-PMH records from OLAC. This means that as a Data Provider to OLAC *The Bridge* and its content will have greater exposure in language subject-specific search engines.

Record Scope

As a general note, it was observed that not all resources were linked to bibliographic references in the same way. *Record Scope* should be regularized. For example, some bibliographic references may reference a collection of independent works in a similar way to an edited volume. This stands in contrast to items which are independent works—analogueous to a single chapter or a single-author (single story) book. This list of bibliographic entities needs to be regularized. I suggest it is regularized around the WEMI concepts presented in the OpenWEMI ontology.¹²⁴ WEMI stands for Work-Expression-Manifestation-Item, and is a core concept in

¹²¹ This spreadsheet is sometimes referred to as the “Bibliographic Record Spreadsheet”.

¹²² <http://olac ldc.upenn.edu/documents.html>

¹²³ <https://vlo.clarin.eu>

¹²⁴ Coyle, Karen. 2022. “Works, Expressions, Manifestations, Items: An Ontology.” *The Code4Lib Journal* 53 (May). <https://journal.code4lib.org/articles/16491>.

library and information science (IFLA 1998).¹²⁵ The International Federation of Library Associations (IFLA 1998) presented the WEMI entities as part of the Functional Requirements for Bibliographic Records (FRBR). Other Latin libraries have centered their indexing (bibliographic) practices around FRBR.¹²⁶ More broadly there are project management benefits for organizing resources with FRBR concepts.¹²⁷ That is, consumers can benefit via *The Bridge* as well as Haverford project staff.

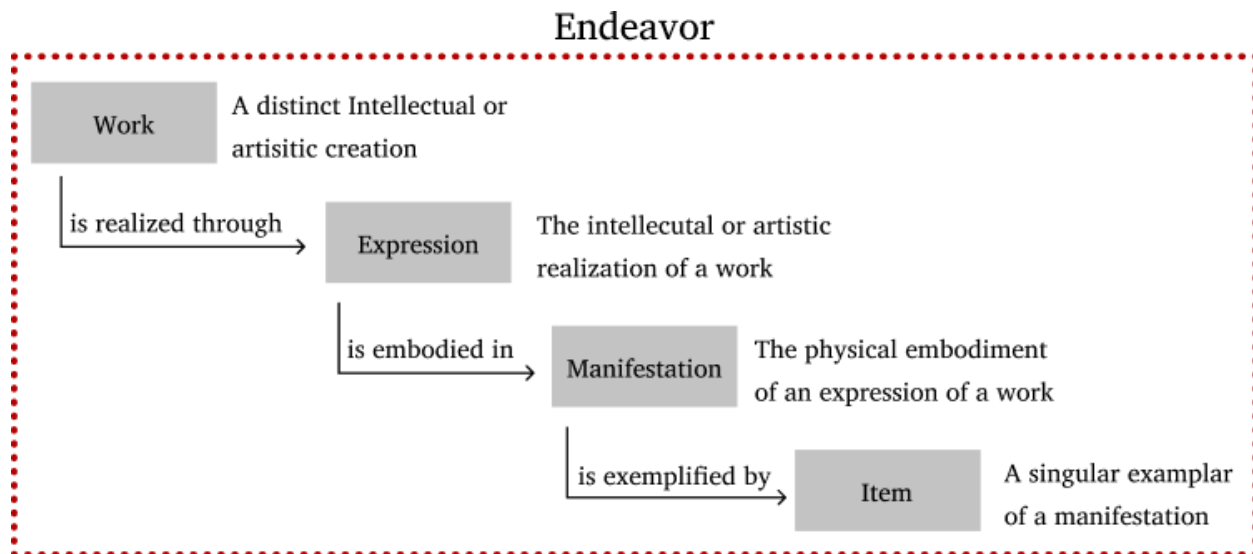


Figure 6: Classes and relationships in OpenWEMI.

¹²⁵ IFLA Study Group on the Functional Requirements for Bibliographic Records and Plassard, Marie-France. 1998. "Functional Requirements for Bibliographic Records: Final Report." 19. 2nd ed. [UBCIM Publications, New Series] IFLA Series on Bibliographic Control. Munich, Germany: K.G. Saur. <http://www.ifla.org/VII/s13/frbr>.

¹²⁶ Babeu, Alison. 2019. "The Perseus Catalog: Of FRBR, Finding Aids, Linked Data, and Open Greek and Latin." In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 53–72. Age of Access? Grundfragen Der Informationsgesellschaft 10. De Gruyter. <https://doi.org/10.1515/9783110599572-005>.

Huskey, Samuel J. 2019. "The Digital Latin Library: Cataloging and Publishing Critical Editions of Latin Texts." In *The Digital Latin Library: Cataloging and Publishing Critical Editions of Latin Texts*, 19–34. De Gruyter Saur. <https://doi.org/10.1515/9783110599572-003>.

Bamman, David, and David Smith. 2012. "Extracting Two Thousand Years of Latin from a Million Book Library." *Journal on Computing and Cultural Heritage* 5 (1): 2:1-2:13. <https://doi.org/10.1145/2160165.2160167>.

Crane, Gregory, Bridget Almas, Alison Babeu, Lisa Cerrato, Anna Krohn, Frederik Baumgart, Monica Berti, Greta Franzini, and Simona Stoyanova. 2014. "Cataloging for a Billion Word Library of Greek and Latin." In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 83–88. DATeCH '14. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2595188.2595190>.

¹²⁷ Signoles, Aurélie, Corinne Bitoun, and Asuncion Valderrama. 2012. "Implementing FRBR to Improve Retrieval of In-House Information in a Medium-Sized International Institute." *Cataloging & Classification Quarterly* 50 (5–7): 402–21. <https://doi.org/10.1080/01639374.2012.681603>.

The classes of Work-Expression-Manifestation-Item cascade inside of each other. For example, a *Work* may have several *Expressions*, and an *Expression* may have several *Manifestations*, etc. The point here is that by aligning with WEMI *The Bridge* has a regularized way of consistently referencing textual units. It is often the case that textual units may be reprinted or appear in several publications. By organizing bibliographic entities around WEMI, it allows bibliographers to side-step issues related to physical or digital bindings. For example, a particular Latin textual unit might have a very early origin as a specific leather manuscript. There is a theoretical *Work* as well as a *Physical Item*. This *Item* infers an *Expression* and a *Manifestation*. If we take a photograph of that same manuscript, we have not changed the *Work* or the *Expression* but we have created a new *Manifestation*—the digital photograph. Likewise, if we then type the text into a computer we have the same *Work* and *Expression* but a third *Manifestation*. Let's assume though that this old textual work was re-printed in 1741s with editorial changes to the text. This would still be the same *Work* but a new *Expression* and the physical printed book would be a new *Manifestation-Item* of that new *Expression*.

The most important point for *The Bridge's* bibliography is to index by works and to “surface” *Works* which appear in aggregates (bindings or couplings with other works, e.g., edited volumes, compilations, etc.). This is standard practice in citation and referencing strategies such as APA or Chicago which generally require the citation and referencing at the *Manifestation* level of the WEMI model. However, identifying specific *Manifestations* may not necessarily be standard practice in the classics. Identifying manifestation also facilitates easier resource management through the data production workflow. This workflow is generalized in Figure 7.

By way of an example, the current record for *Brevissima (Gibbs)* has several *Eras* associated with its current record: “Imperial,Late Antique,Medieval,Neo-Latin”. This seems to indicate that *Brevissima (Gibbs)* might be a collection of works created by different authors across a large time range (i.e., an edited volume). Under the WEMI model the therein contained *Works* would each get their own record (allowing for a single application of *Era* to each new record). Then these resultant records would then be related to a single record for the collection (edited volume). But the individual records would serve *The Bridge's* target audience more efficiently by pointing them directly to the literary portion which is most relevant to their needs.

Created	Republished	Identified for Inclusion	Digitized (Scanned)	Converted to Text (OCR)	Edited / Checked	Released to Platform	Released to Open Access
"Original" Manuscript	Slight Editing assumed	An Expression-Manifestation combo is identified for inclusion in The Bridge	A scan is produced or acquired			Full Text	Full Text
						Expression 4 Manifestation 5	Expression 4 Manifestation 5
	Might be part of another compilation					Summary Text	Summary Text
Expression 1 Manifestation 1	Expression 2 Manifestation 2		Expression 2 Manifestation 3	Expression 3 Manifestation 4	Expression 4 Manifestation 5	Expression 5 Manifestation 6	Expression 5 Manifestation 6

Figure 7: The “normal” eight stages in the Haverford Text pipeline process

As the process is laid out in Figure 7, at each stage the question needs to be asked: “is this a new *Work*, *Expression*, or *Manifestation*?”. Developing and implementing an ID scheme which becomes the file name can help with this in systems which do not have overt tracking WEMI type management. In Figure 7, *Expressions* and *Manifestations* are counted based on their stage in the process while assuming that they are referencing the same *Work*.

Figure 8, illustrates some accepted guidance in the domain of libraries on how to determine if two items are associated with the same *Work*. To this guidance I have overlain the Dublin Core relationships most appropriate to the *Work*’s record. In library practice three broad categories exist: Equivalent, Derivative, and Descriptive. This is a way to provide artifacts with the most appropriate records and relationships to similar artifacts. For example, two records which are an exact reprint or Facsimile of each other would be associated with a *hasFormat* record. For example, this may be useful in pedagogical contexts where one resource is behind a paywall and another is not. In technical studies, it is important that equivalence be understood down to the Unicode character level. By way of another example, removing diacritics from a text would create derivative work and the original manifestation record would have a *hasVersion* relationship with the record for the new (derivative) text without diacritics.

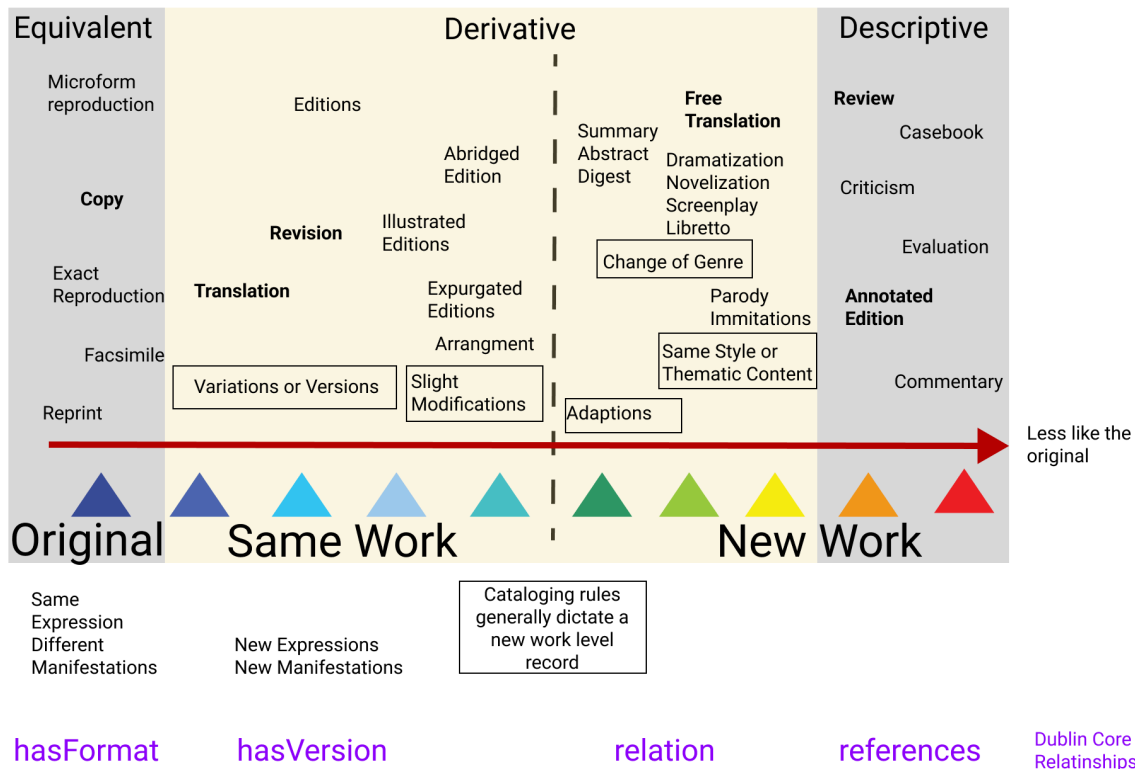


Figure 8: The Same-New divide aligned with Dublin Core Relationships

It is strongly recommended that bibliographic records be created on the basis of *Manifestations*. With conceptual records created for *Expressions* and *Works*. *Expression* and *Work* records are critical for end-user discovery and navigation tasks.

New Metadata Elements

Identifier

Current *Texts* do not have an apparent stable system-internal identifier. It is important (foundational) to give each bibliographic entity a unique identifier which can be both system-internal and system-external. I suggest an ID pattern be defined. An ID pattern may be as simple as an increasing numerical pattern. I suggest an OAI compliant scheme, which is a subset of the URI/URL compliant scheme.¹²⁸ The bibliographic entity's bibliographic information (and maybe some statistics) would usefully be accessible via a stable URL such as *thebridge.lat/bibliographic-unit/id/123*. This is an API based approach and is also the foundation for making The Bridge data "Linked Data".

¹²⁸ <http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm>

See note under [OpenWEMI](#) for further consideration around identifiers.

Old term: N/A

New term: Identifier

Linked Data value: dct:identifier

URI for value: <http://purl.org/dc/terms/identifier>

Comment: It is reasonable to assume that some records in the bibliography of *The Bridge* represent external published resources. These might be identified with ISBN,¹²⁹ ARK,¹³⁰ or DOI persistent Identifiers. Bibliographic records for The Bridge can have multiple identifier elements. For example, a system internal one and a system external one (the ISBN). ISBNs are usually issued on the basis of a (WEMI) *Manifestation*. Within the identifier element the following syntax can be used and is recommended when including ISBNs. *dct:identifier=URN:ISBN:978-0-86516-289-1*.

Using prefixes in the URN syntax¹³¹ allows for the identifier field to hold multiple identifiers. For example, urn:isbn:123-456-798 and urn:doi:10.123/456.

Type:openWEMI

The current bibliographic spreadsheet does not have an indication for type in the sense described here. Here we are declaring the type of record rather than the type of the resource. Here we are describing the type of bibliographic entry according to the WEMI model. The recommendation is to limit the values for this field to the controlled vocabulary: *Work*, *Expression*, *Manifestation*, *Item*.

Note: if an identifier scheme such as W1E2M45 (or w2-e4-m31, w1.e3.m2, etc.) were to be implemented, then this element might not be necessary. The Work, Expression, Manifestation information would be encoded in the identifier. Further, when sorting entries by identifier (or sorting files, assuming files were named with their identifier) then those which are part of the same *Work* or *Expression* would group together.

Note 2: LiLa uses a WEMI-like ontology called The FRBR-aligned Bibliographic Ontology (FaBiO).¹³² However, observant information architect specialists fluent in FRBR and its variations have detected that FaBiO is not exactly the same as the library model it was based upon (FRBR). The nuanced differences are in the entity definitions. This is why OpenWEMI is recommended instead of FaBiO.

Old term: N/A

¹²⁹ <https://www.iana.org/assignments/urn-formal/isbn>

¹³⁰ <https://arks.org> ; https://en.wikipedia.org/wiki/Archival_Resource_Key

¹³¹ <https://datatracker.ietf.org/doc/html/rfc8141>

¹³² <http://www.sparontologies.net/ontologies/frbr>

New term: Type:openWEMI

Linked Data value: dct:type

URI for value: <http://purl.org/dc/terms/type>

Comment: N/A.

Type:DCMIType

The current bibliographic spreadsheet does not have an indication for type in the sense described here.¹³³ This applies to the type of resource being described. Here we are declaring the interactivity and materiality of the resource. For this we use the DCMIType vocabulary. It has twelve terms, but it is likely that *The Bridge* will only use two of them: *Text* and *Collection*.

The complete set of DCMIType vocabulary terms includes: *Collection*, *Dataset*, *Event*, *Image*, *InteractiveResource*, *MovingImage*, *PhysicalObject*, *Service*, *Software*, *Sound*, *StillImage*, *Text*.

Old term: N/A

New term: Type:DCMIType

Linked Data value: dct:type.dcmitype

URI for value: <http://purl.org/dc/terms/DCMIType>

Comment: Note that each value in the DCMIType vocabulary also has its own URI. Only one term is used at a time.

Rights Management

Currently there is no indication in the bibliographic record for the license under which Haverford utilizes a text. There is also no indication of who claims rights such as copyright or what their rights actually cover. Only legal entities can claim copyright.

Rightsholder

Old term: N/A

New term: Rightsholder

Linked Data value: dct:rightsHolder

URI for value: <http://purl.org/dc/terms/rightsHolder>

Comment: This is the name or identifier of the Right's Holder.

¹³³ The spreadsheet does have a column labeled "type" and this is discussed later.

Rights

Old term: N/A

New term: Rights

Linked Data value: dct:rights

URI for value: <http://purl.org/dc/terms/rights>

Comment: To what do they claim the rights, the legal jurisdiction in which the claim is made, the legal basis of the claim, and what rights are claimed.

License

Old term: N/A

New term: License

Linked Data value: dct:license

URI for value: <http://purl.org/dc/terms/license>

Comment: Usually a link to the license document if it is a standard license.

Titles

Text

The current way of identifying bibliographic entities is via their “title” which uses the label *text*. This should change to be more consistent with industry practice. As a secondary point, bibliography entities should be referenced via their ID not their *Title*. All records receive an ID but not all works are given a title by their creators. Some titles are assigned later by publishers or archivists. It might be the case that *The Bridge* project needs to assign titles in some cases. For information architecture, reference by ID is preferred.

Old term: Text

New term: Title

Linked Data value: dct:title

URI for value: <http://purl.org/dc/terms/title>

Comment: If The Bridge must assign Titles to resources, consult DACS¹³⁴ for best practices.

ShortName

Application-specific (bridge internal) short names can be useful. This should change to be more consistent with industry standards. ShortName may be abbreviations. If short names are exclusive to The Bridge, are unique, and cannot change then they are behaving more like identifiers and should be reconsidered in the context of identifiers.

Old term: ShortName

New term: Alternative

Linked Data value: dct:alternative

URI for value: <http://purl.org/dc/terms/alternative>

Comment: N/A.

Contributors

Author

The current spreadsheet has a column “author”. This should change to be more consistent with industry standards. In this case, the information needs to change to two values: a name and a role.

When considering records representing the creator of the work, the role of author is applicable. When considering records representing derivative works, such as digitized copies or summaries, a different role may be more applicable. Distinguishing roles allows for explicit indication of the kinds of contributions to a resource. See the [Role Appendix](#) for relevant MARC and OLAC roles. Further investigation is needed for the most appropriate contributor roles based on the stage of the resource within the workflow. MARC contributors can include roles like publishers and data annotations. A consistent format for names should be chosen.¹³⁵ Considerations should include corporate names, personal names, names with dates, and person IDs. Generally, personal names are ‘last name <comma> first name’. IDs may be from a variety of sources.¹³⁶

¹³⁴ Society of American Archivists. 2013. Describing Archives: A Content Standard. 2nd ed. Chicago, Illinois: Society of American Archivists. http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf.

¹³⁵ <https://original.rdatoolkit.org/document.php?id=lcpschp9.pdf> ;

https://www.loc.gov/catworkshop/courses/nametitleauth/pdf/Name-Title_Instr_Manual.pdf

¹³⁶ <https://orcid.org> ; <https://authorities.loc.gov> ; <https://viaf.org> ; <https://www.wikidata.org>

Old term: Author

New term: Contributor

Linked Data value: dct:contributor

URI for value: <http://purl.org/dc/terms/contributor>

Comment: N/A.

Relationships

Source Text 1a

The current use of the “Source Text” column appears to be ambiguous. For example: *Aesop, Fables 1-53* has its source indicated as *Aesopi Fabulae by Aemilius Chambry*. There are several structural issues here. First, with a WEMI-based approach, each of the Fables would constitute its own *Work*, thus receiving its own bibliographic entry. Then each of those would be indicated as “isPartOf” and related to a 54th bibliographic entry for the whole **Work** *Aesopi Fabulae by Aemilius Chambry*. E.g., *Aesop Fable 1 is part of Aesopi Fabulae by Aemilius Chambry*. In contrast to the current practice of using full text strings in this column, the bibliographic entity ID should be used. E.g., *TB_01 is part of TB_02*. As an implementation note, the URI *thebridge.lat/bibliographic-unit/id/TB_01* could be used or the internal ID could be used. The ID is shorter and may not be search engine friendly, but the URI would be “more linked data like”. The search engine friendly component could be circumvented by assigning each entity two URIs. The first is an SEO friendly URL, and the second is an identifier-based permanent URL which is set to resolve to the SEO friendly URL.

Old term: Source Text

New term: Is Part Of

Linked Data value: dct:isPartOf

URI for value: <http://purl.org/dc/terms/isPartOf>

Comment: N/A.

Source Text 1b

A second option also exists for indication of the whole-part relationship. The above method recommends marking the part for the whole it belongs to. The alternative strategy is to mark the whole for the part it belongs to. Such an indication would use the “hasPart” relationship.

Old term: Source Text

New term: Has Part

Linked Data value: dct:hasPart

URI for value: <http://purl.org/dc/terms/hasPart>

Comment: N/A.

Source Text 2

In the first paragraph I mentioned two senses of “Source Text”. The first relationship focuses on source in the sense of taking a part out of a whole which is a “collection”. However, in a second sense source is the beginning state of a resource which undergoes some sort of change. For example, a text without capitalization, paragraphing, or punctuation may become edited to include those new editorial changes. There are now two (WEMI) Expressions of this resource which are part of the same work, and one of those expressions is the source for the other one. These two resources exist within a derivative relationship.

Old term: Source Text

New term: Source

Linked Data value: dct:source

URI for value: <http://purl.org/dc/terms/source>

Comment: This field should occur on the record for the derived resource and should contain the ID of the original resource.

Language

Language

The current bibliography uses a full text term for the name of the language. This would ideally be replaced with a controlled vocabulary which is compliant with BCP-47.¹³⁷ At this time RFC5646 should be used.¹³⁸

There are two fields in the spreadsheet which work together to “refine” or “classify” the language; these two are Language and Era. My current understanding is that the values in Era

¹³⁷ <https://www.rfc-editor.org/info/bcp47>

¹³⁸ <https://www.rfc-editor.org/info/rfc5646>

would fluctuate with the values in Language. That is, Era is a sub-categorization scheme to the broader field *Language*. This is inline with classifications of Latin such as those which are presented in *Fowler's History of Roman Literature*, Wilhelm Sigismund Teuffel's first edition (1870) of *History of Roman Literature*, and *Freund's Lexicon of the Latin Language*. Further work with the IANA group and the ISO 639 Committee should be conducted to specify clearer Latin Languages.

Old term: Language

New term: Language

Linked Data value: dct:language

URI for value: <http://purl.org/dc/terms/language>

Comment: N/A.

Era

The field Era seems to sub-divide the Latin language along a time depth dynamic. The taxonomically unique values in the column include: *Archaic*, *Classical*, *Imperial*, *Late Antique*, *Medieval*, *Modern*, *Neo-Latin*, and *Republican*. I think this field is redundant and can be removed even though it serves an obvious purpose. There are four possible ways to do this.

1. The first way to remove the field would be to apply to the IANA list for a valid variant code to Latin for each of the values in the used taxonomy.¹³⁹ The result of this process would result in valid BCP-47 codes and these distinctions can be made directly via the language field.
2. The second way to resolve this would be to apply to the ISO 639-3 registrar for separate language tags, treating these various stages of the Latin “language” as separate languages. This is the common approach in contexts such as Old English, Old High German, Old Chinese, etc.
3. The third way is to treat these text values as a controlled vocabulary and use them as a date. Essentially, they are referencing a date range.
4. The fourth way would be to use specific dates (even if they are inexact, e.g., ‘circa 75AD’) and put these in a date field. Then create Era “buckets” (ranges) in the web application’s user interface and describe these “buckets” to pull from all texts with a date range of ‘x’ through ‘y’. In this way, what is part of the record is a date, and what is part of the user interface is what to do with that date.

¹³⁹ In fact this has already been started. See the archives at: <https://www.ietf.org/mailman/listinfo/ietf-iana>

If an informatic structure such as that presented in number three were to be selected, then the most appropriate metadata title for this information would be to use *dct:created*. A fuller discussion on the use of dates with relationship to records and workflows is presented in the section on dates.

Old term: Era

New term: Date Created

Linked Data value: *dct:created*

URI for value: <http://purl.org/dc/terms/created>

Comment: N/A.

Dates

Dublin Core has nine separate fields for different types of dates. The most appropriate to use in any given scenario is highly dependent on the requirements and assumptions set out in the application profile. In this section, I present a best-guess model of workflows at Haverford. I then apply the WEMI model to this workflow model and use WEMI-based terms for the different records at different stages of the workflow. Finally, I show which date fields I perceive as being the best semantically for any given record within the modeled context. Dublin Core date elements implicitly invoke an event type. For example, a modification event, or an acceptance event. If a narrowly scoped understanding for the need for an event is contextualized to the stage of the workflow, then date elements can be “reused”. Assuming that a resource receives a new identifier (e.g., W1E2M2) for each stage of the workflow it goes through then there is no conflict for date application to records. For example, a record for a *Work* might have a date 75AD because that is the date of the earliest known manuscript. That manuscript record would have a *Manifestation* record with a created date. A digital image of that manuscript might have a *Manifestation* record with a created date of 2022. Finally, a digital text based on the image (e.g., OCR) might have another *Manifestation* record with a creation date of 2023. Each of the manifestations may be locked in a 1-to-1 relationship with an *Expression* which is linked to the *Work* record.

The nine date types possible per record are as follows:

1. [date](#)
2. [dateAccepted](#)
3. [dateCopyrighted](#)
4. [dateSubmitted](#)
5. [issued](#)
6. [modified](#)

7. [valid](#)
8. [created](#)
9. [available](#)

Created	Republished	Identified for Inclusion	Digitized (Scanned)	Converted to Text (OCR)	Edited / Checked	Released to Platform	Released to Open Access
"Original" Manuscript	Slight Editing assumed Might be part of another compilation	An Expression-Manifestation combo is identified for inclusion in The Bridge	A scan is produced or acquired			Full Text Expression 4 Manifestation 5 ----- Summary Text	Full Text Expression 4 Manifestation 5 ----- Summary Text
Expression 1 Manifestation 1	Expression 2 Manifestation 2		Expression 2 Manifestation 3	Expression 3 Manifestation 4	Expression 4 Manifestation 5	Expression 5 Manifestation 6	Expression 5 Manifestation 6

Figure 7 (repeated): The “normal” eight stages in the Haverford Text pipeline process

The three date values most relevant to several stages of the Haverford workflow include *Date* on *Work* level records and *Item* level records representing manuscripts. *Created* (and *dateCopyrighted* if applicable) on *Manifestation* records for facsimiles of manuscripts, Neomodern Latin works and editorial editions. Meanwhile, *issued*, *modified*, *created*, and *available* dates can apply to *Manifestation* records in various stages of the product lifecycle management pipeline (such as editing, or release). The exact semantics of each date value as they apply to each pipeline needs to be documented. Each *Manifestation* record can reference this documentation via the *dct:conformsTo* value.

Existing Dates

Within the bibliographic metadata spreadsheet there are three date columns. These include: *Date Created*, *Last Modified*, *Date Published*. These need to be contextualized within the scope of the record. For example, the record for a manuscript, for an OCR'd resource, or a released resource. Last Modified should only be rarely used in contexts where the resource has not moved on to another stage in the workflow but has been adjusted in some way. This is useful when adjustments are not intended to be carried downstream. At each modification of a resource, a new record should be made in the provenance field. In this sense, the provenance field and the date modified field are connected. For records representing new additions to collections, it is the provenance and modification of the collection record which should be considered rather than the item. Of course, if policy dictates, the inclusion of an item in a collection can also trigger an update to the item's provenance field.

Date Created

Old term: Date Created

New term: Created

Linked Data value: dct:created

URI for value: <http://purl.org/dc/terms/created>

Comment: N/A.

Last Modified

Old term: Last Modified

New term: modified

Linked Data value: dct:modified

URI for value: <http://purl.org/dc/terms/modified>

Comment: N/A.

Date Published

Old term: Date Published

New term: issued

Linked Data value: dct:issued

URI for value: <http://purl.org/dc/terms/issued>

Comment: Contrast Date Issued with Date Accessible as the Date Accessible is a technical publication date.

Description

There are several columns in the spreadsheet “[+ Bridge About Text Dataset](#)” which logically belong to the category *description*. These include: *Syntax Data?*, *Data Format*, *Special Notes about Format*, *Meter*, *Local Definitions*, *Sample Formats*, and *Other Notes*. It is my suggestion that these concepts be divided into two “classes” which are each placed in their own column/field. First, there are some columns which describe the technical format or substance of

the artifact.¹⁴⁰ (This is closer to an is-ness description.) Second, there are things which describe the content, style, or composition of the artifact—often from a linguistic or literary view. Those which describe the technical format or substance of the artifact should appear in *dct:description* as a **Format Description** while those which describe the content style or composition should appear in *dct:abstract* as a **Form Description**. However, any information which can be conveyed in the *medium*, *format*, *extent*, or *conformsTo* fields should not appear in a description or abstract field at all.

Form Description

Meter

Meter is the rhythmic and syllabic structure of a poetic unit. For poetry, meter and poetic structure indexing is like indexing a song for a key signature. Such indexing is conducted in MARC records representing musical pieces.

Old term: Meter

New term: Description

Linked Data value: *dct:description*

URI for value: <http://purl.org/dc/terms/description>

Comment: While Meter in the bibliographic metadata spreadsheet is a sortable field, here the meter information is part of a broader text field which may contain several different types of things. The Haverford Application Profile could specify how to craft a multi-topic description statement or several description elements could be added per record with one of them containing the first word “Meter”. Meter could be indicated via a controlled vocabulary and the *dct:conformsTo* element.

Format Description

Syntax Data?

In the current bibliography metadata spreadsheet, there is a column titled *Syntax Data?* This is filled with a value of “yes” or “no”. The reference point here is does the data file contain a specific type of data. Several of the source formats have syntax data. For example, treebank data from PROIEL or Perseids has this type of data as some data from LiLa or LASLA. These files may be sourced via an XML or tabular format such as CSV. A metadata field such as [dct:format](#) can be used to indicate the XML or CSV file-format but it takes an additional bit of information to indicate the classification of data within these structures. This could be indicated via prose in an abstract description, or it could be indicated in a *dct:conformsTo* field. For

¹⁴⁰ Those things describing the processing of the text should be moved to *provenance*.

example, PROIEL XML is versioned and a field and value of *dct:conformsTo=PROIEL XML 2.1* would indicate that specific fields such as syntax or part of speech would be available.

Old term: Syntax Data?

New term: Abstract

Linked Data value: *dct:abstract*

URI for value: <http://purl.org/dc/terms/abstract>

Comment: The abstract field can be repeated. If the abstract field were used to specify a single item then this might be how that looks: *dct:abstract=SyntaxYes*. However, this seems to be an excellent use case for a custom taxonomy and the use of the *dct:conformsTo* property. E.g., *dct:conformsTo=SyntaxYes*. The benefit of using *dct:conformsTo* over *dct:abstract* is that it reserves *dct:abstract* for prose based structures.

Data Format

The data format column indicates the subdivision structure of the resource. I suggest this column is unnecessary as a separate column. It is good to have this information in a prose description. However, it would be even better if resources were divided according to WEMI-based units. Then the remaining formats would be more consistent. (Separate files instead of 1.1.1, the remaining files would be more consistent with a 1.1 pattern). Formats labels would ideally be managed via a controlled vocabulary. In a POWLA file, structured content units are independently managed as pointers to a range of Unicode characters. This can permit multiple systems of reference to exist in the same file. For example, Lines in poetry, page numbers of a specific published version, or even section and stanza numbering. Data format terms currently sometimes contain an extent number, e.g., POEM (507). Indications of extent should be separated out to an [extent column](#).

Special Notes about Format

Special notes should be added to a description or a provenance field depending on their nature. Notes about format would be best written in prose within the description. This specific field can be removed.

Local Definitions

The local definitions field within the bibliography spreadsheet either has a *yes* or *null* value. I suspect that the field indicates whether there is a present field in a locally held spreadsheet with Haverford generated word glosses.

Word glossing could easily be an overt stage in a workflow. So a whole Expression-Manifestation artifact stage (as shown in Figure 7) could be dedicated to this type of

record. Within a Dublin Core framework, the use of *dct:conformsTo* along with a value such as *localGlossYes* would fit this use scenario.

Sample Formats

Examples illustrating format would fit well within the prose context of a description.

Other Notes & Problem?

These two columns were essentially empty so maybe they are not needed at all. It is likely that their contents should be moved to a [description](#) or [provenance](#) statement(s).

Type Description

Type

The *Type* column in the Bibliographic spreadsheet has seven values: “List”, “Poetry”, “Prose”, “Prose, Poetry”, “Prose,Textbook”, “Text”, “Textbook”. I suggest that this column is redundant between the *Kind* and *Genre* columns. See further discussion on the columns *Genre* and *Kind*.

Genre

The *Genre* column in the Bibliographic spreadsheet has fifty-one values (as comma separated strings). Assuming that some bibliographic records do not represent a single WEMI *Work* some of these values may not represent a single work rather may apply “in Collection” to an aggregate of works. When thus considered, only thirty-seven unique values exist. This also presumes that commas are not indicating a subtype (e.g., “Epic,Didactic”, as being a subtype of “Didactic”). The thirty-seven unique values are: “Biography”, “Cento”, “Christian”, “Comedy”, “Comprehensible”, “Dialogue”, “Didactic”, “Drama”, “Elegy”, “Epic”, “Epideictic”, “Epigram”, “Epigraphy”, “Epistle”, “Epithalamium”, “Epyllion”, “Ethnography”, “Fable”, “History”, “Hymn”, “Inscription”, “Invective”, “List”, “Lyric”, “Medical”, “Mixed”, “Natural History”, “Novel”, “Oratory”, “Panegyric”, “Pastoral”, “Philosophy”, “Satire”, “Textbook”, “Tragedy”, “Various”.

Some terms among the thirty-seven seem to have redundant, overlapping, and/or non-contrastive meanings. For example, “Mixed” versus “Various” seem to have overlapping and redundant meanings. “Lyric” and “Hymn” could also have very similar meanings. “Medical” seems to be a subject term rather than a genre term. Terms such as “Christian” are not clear in their implied opposition(s). For example, *Christian* religious texts can also (among others) have the genres: *Pastoral*, *Dialogue*, *Epistle*, and *Oratory*. However, along the religious dimension it is challenging to know if those texts marked “Christian” are *religious texts* or are they more specifically *non-Muslim texts*? Translations of Arabic texts (in various subject areas) as well as Islamic religious texts also exist in Latin, yet there is no genre term present at Haverford for

these.¹⁴¹ From a linguistic perspective (both translation studies and discourse studies) a gap exists in the genre indication related to translated texts. That is, no genre term exists for Latin texts which exist as the result of being translated into Latin from another language. It is well known that translated texts can carry over linguistic structures or vocabulary from the original language to the target language (Latin).

There are many controlled vocabularies for Genre terms (e.g., MARC Genre terms,¹⁴² and Library of Congress Genre terms¹⁴³). Genre terms often combine the ideas of a compositional form and a publication format. In contrast to looking at the composition and materiality components, the Open Language Archives Community has approached the issue of genre through the lens of discourse studies. The OLAC *Discourse Type Vocabulary*¹⁴⁴ offers the following ten terms whose definitions should be compared with the meanings implied by Haverford developed terms: “dialogue”, “drama”, “formulaic”, “ludic”, “oratory”, “narrative”, “procedural”, “report”, “singing”, “unintelligible_speech”. The categorization of texts through linguistic discourse similarity may serve educators and learners better. It has been shown that different types of speech acts use different kinds of grammar structures.¹⁴⁵ However, it may also

¹⁴¹Akasoy, Anna. 2011. “Arabic Texts: Latin Translations of Philosophy.” In *Encyclopedia of Medieval Philosophy*, edited by Henrik Lagerlund, 92–97. Dordrecht: Springer Netherlands.
https://doi.org/10.1007/978-1-4020-9729-4_47.

Ferrero Hernández, Cándida, and John Tolan. 2021. *The Latin Qur'an, 1143–1500: Translation, Transition, Interpretation*. The European Qur'an 1. Berlin; Boston: De Gruyter.
<https://doi.org/10.1515/9783110702712>.

Martínez Gázquez, José. 2014. “Translations of the Qur'an and Other Islamic Texts before Dante (Twelfth and Thirteenth Centuries).” In *Dante and Islam*, edited by Jan M. Ziolkowski. Fordham University Press. <https://doi.org/10.5422/fordham/9780823263868.003.0004>.

¹⁴² <https://www.loc.gov/standards/valuelist/marcgt.html>

¹⁴³ <https://www.loc.gov/aba/publications/FreeL/CGFT/freelcgft.html>

¹⁴⁴ Johnson, Heidi, and Helen Aristar Dry. 2012. OLAC Discourse Type Vocabulary. OLAC Recommendations. Dallas, Texas, USA: The Open Language Archives Community.
<http://www.language-archives.org/REC/discourse.html>.

¹⁴⁵ Biber, Douglas, Jesse Egbert, Daniel Keller, and Stacey Wizner. 2021. “Towards a Taxonomy of Conversational Discourse Types: An Empirical Corpus-Based Analysis.” *Journal of Pragmatics* 171 (January): 20–35. <https://doi.org/10.1016/j.pragma.2020.09.018>.

Bühlig, Kristin. 2005. “‘Speech Action Patterns’ and ‘Discourse Types’” *Folia Linguistica* 39 (1–2): 143–71.
<https://doi.org/10.1515/flin.2005.39.1-2.143>.

Caselli, Tommaso, Rachele Sprugnoli, and Giovanni Moretti. 2022. “Identifying Communicative Functions in Discourse with Content Types.” *Language Resources and Evaluation* 56 (2): 417–50.
<https://doi.org/10.1007/s10579-021-09554-4>.

Dijk, Teun A. van. 1982. “Episodes as Units of Discourse Analysis.” In *Analyzing Discourse: Text and Talk*, edited by Deborah Tannen, 177–95. Georgetown University Round Table on Languages and Linguistics 1981. Washington, DC: Georgetown University Press.

Dorgeloh, Heidrun, and Anja Wanner, eds. 2010. *Syntactic Variation and Genre*. Topics in English Linguistics 70. Berlin ; New York: De Gruyter Mouton.

Esser, Jürgen. 2014. “Taxonomies of Discourse Types.” In *Pragmatics of Discourse*, 443–62. De Gruyter Mouton. <https://doi.org/10.1515/9783110214406-017>.

Fludernik, Monika. 2000. “Genres, Text Types, or Discourse Modes? Narrative Modalities and Generic Categorization.” *Style* 34 (2): 274–92. <https://www.jstor.org/stable/10.5325/style.34.2.274>.

Virtanen, Tuija. 2010. “Variation across Texts and Discourses: Theoretical and Methodological Perspectives on Text Type and Genre.” In *Syntactic Variation and Genre*, edited by Heidrun Dorgeloh

be the case that the OLAC *Discourse Type Vocabulary* should be improved as at one time they did have poetry included as its own term, but that has since been removed. [This is a possible area of future work.](#)

With regards to translations as a genre, it is possible to indicate that a text is a translation via the [language tag](#) by using the BCP-47 -t subtag.¹⁴⁶ For example, [la-t-ar]. This use of BCP-47 to indicate translation is under-documented in language archives and text collections. Future scholarly work could include describing the use of the -t subtag in conjunction with the identification and introduction of Latin language variant tags—especially Ecclesiastical Latin.

Old term: Genre

New term: Discourse Type

Linked Data value: olac:discourse-type

URI for value: <http://www.language-archives.org/REC/discourse.html>

Comment: Within an OAI-PMH expression of the metadata Discourse Type is an xsi type refinement of the Dublin Core type element:

<dc:type xsi:type="olac:discourse-type" olac:code="primary_text"/>. Some Haverford Grene terms may be subtypes of discourse terms. This needs investigation.

Kind

The *Kind* column in the Bibliographic spreadsheet has four values: *Null*, *Textbook*, *Text*, and *List*. I suggest that we look at *Lists* as a type of lexicographical resource – perhaps a simple list of words and their glosses. *Texts* then we could conceptualize as being documents evidencing the language in the first order. That is, they are primary evidence of language usage - a *primary text*. Finally I suggest that *Textbooks* are a type of analysis and *Language Description* which presents the language to students/learners. If recasting these distinctions are acceptable then the Open Language Archive Community's *Linguistic Data Type Vocabulary*¹⁴⁷ fits the need perfectly.

Old term: Kind

New term: Linguistic Data Type

Linked Data value: olac:linguistic-type

URI for value: <http://www.language-archives.org/REC/type.html>

and Anja Wanner, 53–84. Topics in English Linguistics [TiEL] 70. DE GRUYTER MOUTON.
<https://doi.org/10.1515/9783110226485.1.53>.

¹⁴⁶ <https://www.ietf.org/rfc/rfc6497.txt>

¹⁴⁷ Aristar Dry, Helen, and Heidi Johnson. 2006. OLAC Linguistic Data Type Vocabulary. OLAC Recommendations. Dallas, Texas, USA: The Open Language Archives Community.
<http://www.language-archives.org/REC/type.html>.

Comment 1: The old controlled vocabulary values were: *Textbook*, *Text*, and *List*. The new values are *lexicon*, *primary_text*, and *language_description*.

Comment 2: Within an OAI-PMH expression of the metadata Linguistic Data Type is an xsi type refinement of the Dublin Core type element:

```
<dc:type xsi:type="olac:linguistic-type" olac:code="primary_text"/>.
```

Divisions

Divisions as a category is a bit ambiguous in what it refers to. If it refers to separate Works from a “Collection” perspective it might be good to only indicate these divisions via the *hasPart* relationship. However, another option exists via the *dct:tableOfContents* element. There are two ways to make use of this option. First is to put an ordered list within the field. Second is to add a Table-of-Contents entry for each “division” or component. There is a strong preference for the single-element ordered list method (Example 1), but if a link or ID is to be provided then the second method is the preferred method (Example 2). The Table of Contents method can be used with both “Collections of WEMI Manifestations” (e.g., an edited volume of papers) and single WEMI Manifestations (a single scholarly paper). However, a consistent approach should be used by The Bridge. Not every bibliographic record may need a Table of Contents field.

Example 1:

```
<tableOfContents>Part 1, Part 2, Part 3</tableOfContents>
```

Example 2:

```
<tableOfContents>Part 1</tableOfContents>
<tableOfContents>Part 2</tableOfContents>
<tableOfContents>Part 3</tableOfContents>
```

Table of Contents

Old term: Divisions

New term: Table of Contents

Linked Data value: *dct:tableOfContents*

URI for value: <http://purl.org/dc/terms/tableOfContents>

Comment: N/A.

Provenance

Like the description or abstract statements, the provenance field may benefit from a more formal composition or structure to include: *when*, *who*, *how*, *what-change*, *source*, *acting tool*, *result*, and *comment* for each change event (See Example 1). With a WEMI aligned bibliography which includes overt relations, some provenance information will be programmatically retrievable because it will be embedded in the relationships of the records to each other. Derivative expressions should contain the provenance information present in the records describing their source material. That is, *Provenance* information should travel across records from sources to derivatives whereas most other information will be newly generated for each record. Some provenance information already exists in the bibliography spreadsheet (Example 2). An additional column in the spreadsheet “Problem?” seems to be a good candidate for wrapping into a prose based provenance field.

Example 1:

Event 1: 13 July 2022. Professor Plum¹⁴⁸ applied the Text Encoding Workflow (v1.2) stage 2 to Perseus Digital Library file 1234abc.xml to create file 1234abc.txt. The results need to be manually checked.

Event 2: 17 July 2022. Professor Plum manually checked and corrected file 1234abc.txt. Most corrections related to an improper treatment during conversion of diacritics from the source source ISO/IEC 8859-1 encoding to the finished Unicode UTF-8 encoding.

Example 2: *manually coded data (LASLA); additional vocabulary by B. Mulligan, summer 2019 (Haverford)*

Provenance

Old term: Provenance, Problem?

New term: Provenance

Linked Data value: dct:provenance

URI for value: <http://purl.org/dc/terms/provenance>

Comment: N/A.

¹⁴⁸ Professor Plum has also recently been accused of being in the library with the computer...

External Link

There is a column in the bibliography spreadsheet with the designation “external link”. It is not clear what the possible range of usage is for external links. It seems that most links point to a possible acquisition point for digital copies of the resource. What is not clear is if this is the same copy as the source for the textual content Haverford has used. Two approaches show promise: 1) if the link is to the source’s location then the URL can be used as an identifier for WEMI Expression 1: Manifestation 1 (the workflow source document). 2) if the URL merely points to another expression of the same work, then a new expression record should be created and the URL added as a URL Identifier for that new record. Use of the hasFormat or hasVersion relationships (as appropriate) with the value of the URL is less preferred than a link to another record in the bibliography. Multiple expression manifestation records can exist in the bibliography without requiring them to be a part of the workflow source chain.

Old term: External Link

New term: Identifier.URL

Linked Data value: dct:identifier

URI for value: <http://purl.org/dc/terms/identifier>

Comment: identifier.URL is a qualified version of the general dct:identifier property.

Extent

Includes (Range of Text Present in Database)

Extent is a measure of size. The field “Includes (Range of Text Present in Database)” is a multifunction field which has some Extent data in it. Extent could be measured in Word Count, Lemma Count, Character Range, Line Range, Section Range, or WEMI Entity Range. If records in the bibliographic spreadsheet are based on WEMI entities, then I suspect that the field “Includes (Range of Text Present in Database)” might not be needed at all. The *Distribution Status* field would cover the functional need for assessing which content is included in or

excluded from *The Bridge*. Meanwhile an Extent field could choose a more meaningful unit such as **total word count** or **unique lemma count**. As the field currently exists there are 313 values with 104 unique values. Even across the “unique values” some of the terms are overlapping in meaning. For example, “all” and “entire text”, or “Volume 1” and “Book 1”. For those records which represent resources with multiple parts (such as Book 1, etc.) if each book section has its own record as well as a collective record pointing to the various parts then an appraisal can be made for which parts of the collective resource is included in *The Bridge*.

Old term: Includes (Range of Text Present in Database)

New term: Extent

Linked Data value: dct:extent

URI for value: <http://purl.org/dc/terms/extent>

Comment:

Distribution Status

A variety of uses for `dct:conformsTo` have already been suggested (Meter, Syntax Data, Compliance to specific data structure formats, etc). Two additional columns in the bibliographic spreadsheet include Status-OLD and Status-NEW these relate to the release of resources on The Bridge platform. It is suggested that the *dct:conformsTo* property be used along with a controlled vocabulary. Perhaps five values are sufficient: “undistributed”, “unconfirmed”, “confirmed”, “error”, and “withdrawn”.

Status-OLD, Status-NEW

Old term: Status-OLD, Status-NEW

New term: Distribution Status

Linked Data value: `dct:conformsTo`

URI for value: <http://purl.org/dc/terms/conformsTo>

Comment: The values in a `conformsTo` must be unique to all other possible values in the same field. For this reason it is recommended that URIs be used as much as possible. These URIs could/should be to either an adopted ontology (e.g., PSO¹⁴⁹) or a Haverford published ontology as described in the section [Ontology Server](#).

¹⁴⁹ <https://sparontologies.github.io/psocurrent/psocurrent.html>

Role Appendix

There are two resources for roles which are relevant to this work. The Open Language Archive Community Roles¹⁵⁰ and the MARC Relator Roles Vocabulary.¹⁵¹ In most cases these two vocabularies are compatible, but in several cases care must be taken. These incompatible cases are explained in a thorough blogpost on the issue.¹⁵²

OLAC Roles

Here relevant roles from the OLAC Roles Vocabulary are replicated.

annotator

Name: Annotator

Definition: The participant produced an annotation of this or a related resource.

author

Name: Author

Definition: The participant contributed original writings to the resource.

compiler

Name: Compiler

Definition: The participant is responsible for collecting the sub-parts of the resource together.

Comments: This refers to someone who creates a single resource with multiple parts, such as a book of short stories, or a person who produces a corpus of resources, which may be archived separately.

Examples: A compiler of a book of short stories or a CD with several songs on it; a collector of a corpus of recordings in some language or on a given topic; a person who assembles a suite of software tools.

¹⁵⁰ <http://www.language-archives.org/REC/role.html>

¹⁵¹ <https://www.loc.gov/marc/relators/relaterm.html>

¹⁵² <https://hughandbecky.us/Hugh-CV/post/olac-roles>

data_inputter

Name: Data Inputter

Definition: The participant was responsible for entering, re-typing, and/or structuring the data contained in the resource.

MARC Relator Roles

Here relevant roles from the *MARC Relator Roles* are replicated.

Analyst [anl]

A person or organization that reviews, examines and interprets data or information in a specific area.

Annotator [ann]

A person who makes manuscript annotations on an item

Associated name [asn]

A person or organization associated with or found in an item or collection, which cannot be determined to be that of a Former owner [fmo] or other designated relationship indicative of provenance.

Attributed name [att]

An author, artist, etc., relating him/her to a resource for which there is or once was substantial authority for designating that person as author, creator, etc. of the work.

Compiler [com]

A person, family, or organization responsible for creating a new work (e.g., a bibliography, a directory) through the act of compilation, e.g., selecting, arranging, aggregating, and editing data, information, etc

Data contributor [dct]

A person or organization that submits data for inclusion in a database or other collection of data.

Data manager [dtm]

A person or organization responsible for managing databases or other data sources.

Editor of compilation [edc]

A person, family, or organization contributing to a collective or aggregate work by selecting and putting together works, or parts of works, by one or more creators. For compilations of data, information, etc., that result in new works, see compiler.

Collection registrar [cor]

A curator who lists or inventories the items in an aggregate work such as a collection of items or works.

Collector [col]

A curator who brings together items from various sources that are then arranged, described, and cataloged as a collection. A collector is neither the creator of the material nor a person to whom manuscripts in the collection may have been addressed.

Compiler [com]

A person, family, or organization responsible for creating a new work (e.g., a bibliography, a directory) through the act of compilation, e.g., selecting, arranging, aggregating, and editing data, information, etc

Data contributor [dct]

A person or organization that submits data for inclusion in a database or other collection of data.

Data manager [dtm]

A person or organization responsible for managing databases or other data sources.

Depositor [dpt]

A current owner of an item who deposited the item into the custody of another person, family, or organization, while still retaining ownership.

Host institution [his]

An organization hosting the event, exhibit, conference, etc., which gave rise to a resource, but having little or no responsibility for the content of the resource.

Markup editor [mrk]

A person or organization performing the coding of SGML, HTML, or XML markup of metadata, text, etc.

Publication place [pup]

The place where a resource is published.

Publisher [pbl]

A person or organization responsible for publishing, releasing, or issuing a resource.

Photographer [pht]

A person, family, or organization responsible for creating a photographic work.

Organizer [orm]

A person, family, or organization organizing the exhibit, event, conference, etc., which gave rise to a resource.

Repository [rps]

An organization that hosts data or material culture objects and provides services to promote long term, consistent and shared use of those data or objects.

Research team head [rth]

A person who directed or managed a research project.

Research team member [rtm]

A person who participated in a research project but whose role did not involve direction or management of it.

Researcher [res]

A person or organization responsible for performing research.

Reviewer [rev]

A person or organization responsible for the review of a book, motion picture, performance, etc.

Sponsor [spn]

A person, family, or organization sponsoring some aspect of a resource, e.g., funding research, sponsoring an event.

Transcriber [trc]

A person, family, or organization contributing to a resource by changing it from one system of notation to another. For a work transcribed for a different instrument or performing group, see Arranger [arr]. For makers of pen-facsimiles, use Facsimilist [fac].

Translator [trl]

A person or organization who renders a text from one language into another, or from an older form of a language into the modern form.

University place [uvp]

A place where a university that is associated with a resource is located, for example, a university where an academic dissertation or thesis was presented.