Alison Babeu

The Perseus Catalog: of FRBR, Finding Aids, Linked Data, and Open Greek and Latin

Abstract: Plans for the Perseus Catalog were first developed in 2005 and it has been the product of continuous data creation since that time. Various efforts to bring the catalog online resulted in the current Blacklight instance, first released in 2013. Currently, both the XML data behind the Perseus Catalog and the digital infrastructure used to support it are undergoing a significant revision, with a focus on finally making the bibliographic data available as Linked Open Data (LOD). In addition, work is underway to develop a digital infrastructure that is not just open source but that is more easily extensible and better supports navigating the complex relationships found in that data. This article describes the history of the Perseus Catalog, its use of open metadata standards for bibliographic data, and the different open source technologies used in building and putting it online. It also documents the challenges inherent in the creation of open bibliographic data and ends with a discussion of the move towards LOD and other planned future directions.

1 Introduction

The Perseus Catalog[1] at its beta release in 2013 declared the broad purpose of providing systematic catalog access to at least one open access edition of every Greek and Latin author from antiquity to around 600 CE. This ambitious announcement was vastly different in scope from its initial modest goals when

[1] http://catalog.perseus.org (last access 2019.01.31).

Note: The Perseus Catalog, in all its iterations, owes its beginnings to David Mimno and its growth to Bridget Almas, Anna Krohn, and Greg Crane. Special thanks to Cliff Wulfman for pushing my thinking on all things metadata, to Monica Berti for finally making me write this, to Sam Huskey and Paul Dilley for providing inspiration for a broader catalog world, and to Lisa Cerrato for everything.

Alison Babeu, Perseus Project, Tufts University

https://doi.org/10.1515/9783110599572-005

the creation of metadata for collections outside of the Perseus Digital Library[2] (PDL) first began in 2006. Over its thirteen year history, the Perseus Catalog has grown from a classical text finding aid to an expanding component of the infrastructures of both its parent project the PDL and related projects such as Open Greek and Latin (OGL).

2 Overview of key standards for the Perseus Catalog

The central standard underpinning the Perseus Catalog is the FRBR (Functional Requirements for Bibliographic Records) entity-relationship model, which was designed as a conceptual framework to assist in the creation of bibliographic records independent of any one set of cataloging rules (IFLA 1998). Of particular importance to the Perseus Catalog are the FRBR model Group 1 entities (works, expressions, manifestations, and items), which were proposed as one potential means of organizing bibliographic data. While a work is defined as a “distinct intellectual or artistic creation,” an expression is the “intellectual or artistic realization of a work,” a manifestation physically embodies the expression of a work, and an item is a “single exemplar of a manifestation.” To illustrate, Homer’s *Iliad* is a work; a critical edition by Thomas Allen is an expression; a 1931 Oxford publication of that edition is a manifestation; and an individual library copy of that publication is an item.

The other key standard behind the catalog metadata and architecture is the Canonical Text Services Protocol (CTS)[3] and the related CITE (Collections, Indexes, Texts and Extensions) Architecture, both developed by the Homer Multitext project.[4] While CTS defines a network service to identify and retrieve text fragments using permanent canonical references expressed by CTS-URNs, the CITE Architecture supports discovery and retrieval of texts or collection of objects.[5] CTS has been influenced by the FRBR model and defines several key concepts utilized by the Perseus Catalog for its data architecture. To begin with, the CTS hierarchy has created *textgroups* above the work level. Textgroups support more strategic grouping of texts because they are used not just for literary

[2] http://www.perseus.tufts.edu (last access 2019.01.31).

[3] http://cite-architecture.org (last access 2019.01.31).

[4] http://www.homermultitext.org (last access 2019.01.31).

[5] For further discussion of CTS and recent implementations see Tiepmar and Heyer (2017) and their contribution in this volume.

authors but also for corpus collections, and they also require unique identifiers. While *works* are defined as in the FRBR model, CTS has defined *editions/translations* instead of *expressions*, a practice the catalog has followed to indicate a particular published version of a work.

CTS-URNs are used in the catalog to uniquely identify editions and translations and form the basis both for version identifiers and for canonical edition URIs. They utilize work identifiers from three classical canons: the Thesaurus Linguae Graecae (TLG), the Packard Humanities Institute (PHI), and the Stoa Consortium list of Latin authors.[6] For example, consider the URN: *urn:cts:greekLit:tlg0012.tlg001.perseus-grc1*,[7] "tlg0012" is the *textgroup* identifier for Homer, author 0012 in the *TLG Canon*; "tlg001" is the *work* identifier for the *Iliad* assigned by the TLG; and "perseus-grc1" is the *version* identifier for the 1920 Oxford *edition* by Thomas Allen available in the PDL.

The Perseus Catalog also currently contains two kinds of metadata: bibliographic records for editions/translations of works and authority records for its authors/textgroups. In order to increase the interoperability and extensibility of the catalog data, two standards from the Library of Congress (LC) were chosen: the MODS (Metadata Objection Description Standard)[8] XML schema was used for bibliographic metadata and MADS (Metadata Authority Description Standard)[9] was used for all authority records.

In addition, the Perseus Catalog also includes what has often been referred to internally as *linkable data*, rather than fully Linked Open Data (LOD).[10] While there was not sufficient time to implement full LOD prior to the May 2013 beta release, resources published within the catalog do use Perseus data URIs under the http://data.perseus.org URI prefix. This prefix is followed by one or more path components indicating the resource type, a unique resource identifier, and an optional path component identifying a specific output format (Almas et al. 2014). The general catalog pattern is http://data.perseus.org/catalog/<textgroup urn>[/format], with URIs for catalog records distinguished from PDL text records

[6] TLG (http://stephanus.tlg.uci.edu); PHI (http://latin.packhum.org/about); STOA (https://github.com/paregorios/latin-authors/blob/master/fodder/StoaLatinTextInventory.csv) (last access 2019.01.31).

[7] See http://catalog.perseus.org/catalog/urn:cts:greekLit:tlg0012.tlg001.perseus-grc1 (last access 2019.01.31).

[8] http://www.loc.gov/standards/mods/ (last access 2019.01.31).

[9] http://www.loc.gov/standards/mads/ (last access 2019.01.31).

[10] For more on linked data, see https://www.w3.org/DesignIssues/LinkedData.html (last access 2019.01.31).

by the catalog path element.[11] There are published URIs for textgroups, works, and edition/translation level records, with full CTS-URNs used for texts in catalog record URIs. Additionally, users can also link to an ATOM feed for the catalog metadata for any textgroup, work or edition/translation by appending the format path to the URI.

3 Related work

Three research areas in particular have influenced the recent evolution of the Perseus Catalog, namely: the development of semantic bibliographic metadata/ ontologies and LOD models for other catalogs; the use of CTS-URNS and other semantic identifiers in similar digital classics projects; and the development of classical text knowledge bases and online work catalogs that include similar data.

First, as the Perseus Catalog transformation work is currently using the FRBRoo ontology[12] to rethink its metadata, relevant research includes how bibliographic ontologies[13] might be used for mass conversion of legacy bibliographic records into LOD (Chen 2017), and how the use of bibliographic ontologies can move metadata workflows towards the creation of LOD (Guerrini and Possemato 2016, Clarke 2014). Other influential work (Fuller et al. 2015, Jett et al. 2016) has been conducted by the HathiTrust Digital Library affiliated Research Center (HTRC)[14] that investigated how bibliographic ontologies could be used to remodel traditional bibliographic data in their large-scale digital library so that it better supported scholars in citing and accurately referencing specific editions in the collection.

A second area of related research involves how other digital classics projects have made use of CTS-URNs or other semantic identifier systems to implement and support stable identification of digital objects within their collections. The Coptic Scriptorium[15] faced related challenges in its efforts to

[11] Thus the textgroup URI for Homer's *Iliad* would be: http://data.perseus.org/catalog/urn:cts:greekLit:tlg0012.tlg001 (last access 2019.01.31).

[12] http://www.cidoc-crm.org/frbroo/home-0 (last access 2019.01.31). See Le Boeuf (2012) for an overview of the ontology and its potential for bibliographic data conversion to the Semantic Web.

[13] For a comprehensive overview and comparison of four major data models (FRBR, FRBRoo, BIBFRAME, Europeana Data Model) see Zapounidou et al. (2016).

[14] https://www.hathitrust.org/htrc (last access 2019.01.31).

[15] http://copticscriptorium.org (last access 2019.01.31).

uniquely identify the expressions of texts and other types of linguistic objects in its collection as well as in its need to expand its category of "digital expressions" to include various visualizations and annotations on objects such as manuscripts (Almas and Schroeder 2016). Similar data modeling and identifier issues have also been encountered by Syriaca.org,[16] and Michelson (2016) and Gibson et al. (2017) have discussed both this project's digital infrastructure (TEI-XML, LOD, GitHub) and its extensive work in assigning stable URIs to all the entities found in their digital reference works.

The third and most important area of related work involves two new digital classics canons/catalogs with which the PDL team is actively collaborating: the Iowa Canon of Ancient Authors and Works and the Digital Latin Library (DLL) Catalog.[17] The Iowa Canon, in development since 2015, will offer extensive metadata for Greek and Latin texts, such as genre, time and place of composition, as well as links to other canonical references.[18] It includes additional metadata on both lost and fragmentary authors and works.[19] In the summer of 2018, the DLL released a beta interface to their collection of classical author and textual metadata. The DLL Catalog[20] focuses on helping users find openly available Latin texts online from the classical era up to neo-Latin texts. Its metadata collection (including authority records for authors and works) has made use of data from both the Perseus Catalog and the Virtual International Authority File (VIAF)[21] and includes item records both to digitized books and to digital texts in numerous collections.

[16] http://syriaca.org (last access 2019.01.31).

[17] https://catalog.digitallatin.org (last access 2019.01.31).

[18] Earlier relevant work in integrating data from Greek and Latin canons is that of the Classical Works Knowledge Base (http://cwkb.org/home), which is also an important component of the HuCit ontology, a domain-specific ontology and knowledge base of metadata involving ancient authors and work titles (Romanello and Pasin 2017) (last access 2019.01.31).

[19] Fragmentary authors are those authors whose texts have only survived through the quotation and transmission of other authors and texts (Berti et al. 2015). And for more on the the Perseus Catalog and the Iowa Canon's complementary work, see (Babeu and Dilley, forthcoming).

[20] Before releasing the catalog, the DLL team conducted two information behavior studies (Abbas et al. 2015; 2016) that helped inform its design.

[21] http://viaf.org (last access 2019.01.31).

4 History of the Perseus Catalog and its development

4.1 Perseus Catalog 1.0 (2005)

The first inspiration for what became the Perseus Catalog grew out of a Perseus software developer taking a cataloging class (Mimno et al. 2005) that introduced him to the FRBR conceptual model. Mimno decided to investigate how FRBR could be used to organize the PDL classics collection since it was small in size, highly structured, and already roughly cataloged.

This initial catalog design utilized pre-existing unique identifiers available for a large majority of Perseus texts. Called abstract bibliographic objects or ABOs, these identifiers were central at the time to the PDL document management system. ABOs were designed to represent distinct "units of intellectual content in the digital library" or, in other words, works.[22] Along with ABOs, MODS were used for bibliographic records for expressions (the editions used for PDL texts) and manifestations (the TEI-XML versions) and MADS for authority records for works and authors. Since all of the PDL texts were digital and there were no physical items, the first Perseus Catalog only implemented the first three levels of the FRBR hierarchy. The experimental system also made use of the open source XML database eXist.[23]

Two key observations from this hierarchical catalog design are particularly relevant. First, this experiment illustrated the challenge of representing the part-whole relationship among different works, manifestations and expressions. Within the PDL classics collection, many manifestations of short works were part of larger volumes, such as poetic anthologies or collected Greek orations. The solution that was implemented involved automatically creating a single manifestation level record for a multi-work volume and then linking it to multiple expression-level works. While this plan worked in 2005 for the relatively small PDL collection, it presented serious scalability issues as the catalog data collection grew exponentially.

Secondly, the creation of the eXist system involved several searching and indexing problems. Searching a hierarchical catalog can require very complicated queries as it may need to draw on information from multiple levels. The solution that was employed was to maintain two parallel versions of the catalog. While each version contained the same records, the first set was

[22] For more on ABOs see Smith et al. (2001).

[23] http://exist-db.org/exist/apps/homepage/index.html (last access 2019.01.31).

a collection of individual records (one for each work, expression and manifestation) which served as the editable source code; the second set contained composite records and served as the compiled version, with one XML document for each work containing all its expressions and the manifestations of those expressions. This compiled version was then utilized as a “flat” catalog optimized for searching in eXist and required over 50 XSLT stylesheets to control the display in response to queries. These composite versions also made use of the custom tags <work> <expression> and <manifestation> in order to maintain the FRBR hierarchical structure, a practice that did not continue in the next stage of metadata creation.

4.2 Perseus Catalog 2.0 (2006–2012)

4.2.1 Mass book digitization, new partnerships, and new goals

The experimental system described above was only briefly online and never intended to scale beyond the PDL classics collection. Subsequent developments expanded its scope. Firstly, two massive book digitization projects, starting with Google Books[24] and soon afterwards followed by the Open Content Alliance (OCA) of the Internet Archive[25] began providing access to thousands of Greek and Latin editions in the public domain. Secondly, a grant from the Andrew W. Mellon Foundation for the Cybereditions project led the PDL team to reconsider what type and level of data to include within the Perseus Catalog. The experimental catalog of 2005 only included records and links to PDL editions, but the additional funding supported greatly expanded metadata creation. A decision was made therefore to create an extensible and growing catalog, inspired by FRBR, that would bridge the gap between the deep but narrow coverage of disciplinary bibliographies such as the TLG and the much broader but shallower metadata found within library catalogs regarding classical editions.

From 2006 to 2009, the PDL actively participated in the OCA and created a bibliography of editions to be digitized. The ultimate goal was to provide granular intellectual access to individual works by classical authors at the online page level in these editions. In creating this initial bibliography we focused on editions that were fully in the public domain because we wanted to develop an open collection of primary sources that could be utilized without any

[24] http://books.google.com (last access 2019.01.31).

[25] https://archive.org (last access 2019.01.31).

restrictions. Since the PDL did not expect at the time to be able to create full TEI-XML digital editions of these many authors and works, it was ultimately decided that the catalog should provide analytical level detail not only to the OCA editions but also to a comprehensive canon of Latin and Greek authors. This decision led to the creation of an extensive open access bibliography[26] of Greek and Latin authors and works with a list of standard editions that could be used to guide future digitization. The list was created by combining the standard lists of authors, works and reference editions from a number of prominent classical Greek and Latin lexicons and is still continuously updated as new authors and works are added to the catalog.

4.2.2 The Perseus Catalog metadata and authority records

Between 2006 and 2013, large amounts of metadata[27] were created for numerous digital editions found within Google Books, the OCA, and eventually the HathiTrust. Six basic types of editions were identified with slight variations as to how they were cataloged.[28] The typical cataloging practice was to create single MODS manifestation level records for each volume (rather than for an entire edition), and for those volumes that contained more than one author/work entry, <relatedItem type="constituent"> component records for the individual works were created *within* those MODS records. The constituent records included relevant work identifiers, page numbers and online page level links to digital manifestations.

Separate duplicate expression level MODS records were also created that were linked to these top-level manifestations through the use of <relatedItem type="host">. While this provided a way to both quickly gather up individual expression records for an author in one folder as they were cataloged and to add them to the spreadsheets used for collection management, it also meant that a significant amount of redundant data was created at the same time. The only type of edition with a slightly different practice were multi-volume editions for single works (e.g. a multi-volume edition of Livy's *Ab Urbe Condita*). MODS records were created for each volume with unique descriptive metadata

[26] https://tinyurl.com/y86ttntv (last access 2019.01.31).

[27] For a full description of the MODS/MADS records including XML examples see Babeu (2008; 2012).

[28] See the catalog wiki: *The Different Types of Editions and the Addition of Analytical Cataloging Information* https://git.io/fp7CY (last access 2019.01.31).

such as volume number, extent of the work, and publication dates, but there was no collocation other then being saved in the same folder.

Whether MODS and MADS records were created from scratch using a template or downloaded from different sources, certain types of information were typically added or enhanced. For MODS records this included standard identifiers/headings from library systems for author names and work titles; unique work identifiers from standard canons; structured metadata for all author/work entries; links to online bibliographic records, digital manifestations and page level work links. For MADS records this included lists of variant names with language encoded; standard identifiers (e.g. VIAF number); lists of work identifiers for linking to MODS records; and links to online reference sources.

4.2.3 First experiments with open source system

In the fall of 2011, with a growing mass of metadata and no user interface, PDL staff began active discussions regarding the Perseus Catalog metadata and what type of interface it would require. One key challenge was that the metadata was very granular with thousands of deeply hierarchical XML records to be indexed. It was eventually decided that supporting a native XML database would require more time and resources than were available. In addition, while an open source and adaptable system was preferred, most of the open source library systems that were examined did not provide support for MODS records. Despite not having MODS support, the eXtensible Catalog (XC)[29] system was ultimately chosen as the first test interface.

After an initial test data conversion was conducted in fall 2011,[30] a first XC prototype catalog interface was made available for internal testing. This prototype utilized the Fedora Repository[31] (to store the catalog records) and made use of the XC Drupal and Metadata Services toolkits. The Metadata Services toolkit supported the XC interface and allowed it to present "FRBRIzed, faceted navigation across a range of library resources", and it was this FRBRIzed support with which we most wanted to experiment. Due to the lack of MODS support, however, all metadata had to be reverse transformed into MARCXML for

[29] http://www.extensiblecatalog.org (last access 2019.01.31).

[30] For more on the 2011–2012 work, see http://sites.tufts.edu/perseusupdates/beta-features/catalog-of-ancient-greek-and-latin-primary-sources/frbr-catalog-sips/ (last access 2019.01.31).

[31] https://duraspace.org/fedora/ (last access 2019.01.31).

import into the XC environment.[32] Extensive internal testing of this interface revealed a number of issues, largely due to the reverse transformation, which caused significant data loss and strange duplication issues. The PDL team therefore concluded another implementation solution would need to be found.

4.3 Perseus Catalog Beta (2013–2017)

4.3.1 New metadata practices and workflows: moving to Blacklight and GitHub

In 2012, it was decided that the XC instance could not fully exploit the catalog's XML data and a digital library analyst was hired to assist in the catalog development process. Consequently, active work to get the catalog data online began in earnest. This work would involve a transition from previously closed workflows to a new open and collaborative environment, largely through the use of GitHub.

For a number of years, metadata had been managed on a restricted CVS server and Eclipse software was used for adding data and committing changes. The move of catalog metadata to GitHub was part of a larger transition from closed to open environments that the PDL had undertaken. All catalog metadata was now downloadable and all new data also became publicly viewable upon committing,[33] in addition, the source code was also made available soon after the live release.[34] The adoption of GitHub best practices thus offered a new level of transparency. Extensive documentation was also created for both the code[35] and for catalog usage.[36]

Along with the move to GitHub, it was decided to use project Blacklight[37] as an interface to the catalog's data. Blacklight is an "open source, Ruby on Rails Engine that provides a basic discovery interface for searching an Apache Solr[38] index,"[39] all of which could be customized used Rails. Out of the box, Blacklight

[32] For full technical details, see http://sites.tufts.edu/perseusupdates/beta-features/catalog-of-ancient-greek-and-latin-primary-sources/frbr-catalog-sips/ (last access 2019.01.31).

[33] Available at https://github.com/PerseusDL/catalog_data and https://github.com/PerseusDL/catalog_pending (last access 2019.01.31).

[34] https://github.com/PerseusDL/perseus_catalog (last access 2019.01.31).

[35] https://github.com/PerseusDL/perseus_catalog/blob/master/doc/PerseusCatalogDocumentation.docx (last access 2019.01.31).

[36] Blog with FAQ, usage guide, and other data at http://sites.tufts.edu/perseuscatalog/ (last access 2019.01.31).

[37] http://projectblacklight.org (last access 2019.01.31).

[38] http://lucene.apache.org/solr/ (last access 2019.01.31).

[39] https://github.com/projectblacklight/blacklight/wiki (last access 2019.01.31).

provided a standard search box, faceted searching, and stable document urls, all features which made it an excellent candidate for an interface. Over the spring of 2013, the PDL team converted the XML data into ATOM feeds for reviewing, tracked problems, and developed customized Ruby subclasses. The catalog first went live in May 2013 and included MADS records[40] for authors/textgroups with lists of works, and MODS edition/translation records that were grouped under top level work records.[41]

One major change to metadata practices after the release was that every MODS record now contained an automatically assigned and unique CTS-URN to serve as a version identifier. This new practice was unrelated to Blacklight and had to do instead with the PDL's prior decision to follow the CITE and CTS standards. In addition, where once there had been one MODS record created for each individual work/expression even if the that record also included a translation, the system automatically split these expressions into two MODS edition/translation records, each with their own URN, as required by the CTS model. While this had the positive effect of finding and splitting translations apart from editions in the browsing environment and data tables, it also had the negative effective of creating additional metadata.

At the time of the beta release, the catalog system also automatically created CTS-URNs for all the individual expressions in the data, and generated expression level records for all the author/work constituent records in the large composite MODS editions that did not have them. This system would also continue to create CTS-URNs for MODS records it ingested from catalog_pending on GitHub, the location for all newly created records. To enable collaborators to make contributions to this repository, record templates and a form to reserve a CTS-URN and/or create a base level MODS record were added. In addition, now that all records and versions had published CTS-URNs, an additional data correction pass was involved using the CITE Collection tables[42] if records were deleted or if a published version was incorrect. When the first catalog data set was generated, four relevant CITE_Collection tables were created for all the data in the repository (authors, textgroups, works, versions) as was required by the CTS/CITE standard and certain types of data changes had to be registered here manually.

[40] See the authority record for Cicero: http://catalog.perseus.org/catalog/urn:cite:perseus:author.364 (last access 2019.01.31).

[41] Such as Cicero's *De Amicitia*: http://catalog.perseus.org/catalog/urn:cts:latinLit:phi0474.phi052 (last access 2019.01.31).

[42] For a full explanation of the CITE Collection tables and the Perseus catalog see https://git.io/fp7W5 (last access 2019.01.31).

The only other major cataloging change involved how single work multi-volume editions were cataloged. Originally each volume had its own MODS record with a work identifier and it thus had a CTS-URN generated for it, so a seven volume edition of Livy in the beta catalog ended up with seven URNs instead of one. Hundreds of invalid CTS-URNs were thus created in the beta catalog so, from 2013 onwards, the new practice was still to create MODS records for each volume (using the ID attribute to indicate volume number) but then to save all the records (each with the same CTS-URN) in a single modsCollection file.[43]

4.3.2 OGL and new collections for metadata

In 2013, another major development would change the goals of the Perseus Catalog once again when the PDL's editor in chief, Gregory Crane, became a Humboldt professor and established the Digital Humanities Chair at the University of Leipzig (DH Leipzig) in Germany. One of the major projects begun at DH Leipzig was OGL,[44] which sought to produce at least one open source digital edition – ideally, multiple editions – of every Greek and Latin text from antiquity through approximately 600 CE.[45] In addition, DH Leipzig also worked with the Saxon State and University Library Dresden (SLUB) to digitize several hundred Greek and Latin volumes.[46] Many of these new collections grew exponentially before even basic metadata creation[47] or cataloging, other than a basic TEI header and the creation of a CTS-URN, could be accomplished. While there was some brief experimentation in the automatic creation of metadata through the use of a CSV sheet,[48] only one collection, a highly-structured Arabic language corpus, was ever imported into the catalog using this method.

[43] This process also involved extensive data cleanup as a large number of records had to be manually collated and CTS-URNs redirected.

[44] https://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/ (last access 2019.01.31). The Humboldt Chair ended in 2018, but the OGL continues forward as part of an international collaborative partnership: http://opengreekandlatin.org (last access 2019.01.31).

[45] A full list of available collections can be found here: https://github.com/OpenGreekAndLatin (last access 2019.01.31).

[46] http://digital.slub-dresden.de/en/digital-collections/127/ (last access 2019.01.31).

[47] For further discussion of OGL metadata and the Perseus Catalog see Crane et al. (2014).

[48] https://git.io/fp7lT (last access 2019.01.31).

4.4 Current work in remodeling the data (2017–present)

Changes in staffing in 2016 coupled with the lack of dedicated funding to maintain and update the Perseus Catalog have led to the current status: a significant backlog of metadata that has not been ingested into the final data repository; corrections to metadata within the final data repository that have not been pushed to the database underlying the Blacklight instance; and numerous technical issues with the way that interface represents the catalog metadata documented and unresolved. Therefore in the fall of 2017 the PDL contracted with the Agile Humanities Agency (Agile) to thoroughly review and enhance the current catalog metadata formats and to investigate whether the Blacklight instance should be updated or if a new interface should be developed instead.

4.4.1 Blacklight interface and updating issues

After its 2013 release, three updates were made to the Blacklight instance, each with their own technical challenges and unresolved metadata issues. The time between updates led to large amounts of new and revised data being stored in catalog_pending making it difficult to keep track of the different types of metadata changes and to test whether errors had been fixed. Nonetheless, the use of Blacklight as an interface to the Perseus Catalog had been reasonably successful, and has served as the beta – and, indeed, only – interface to the data for over 5 years. As the senior Perseus software developer noted in 2016, however, the custom programming approach that adapted Blacklight to support pre-existing data creation workflows led to long-term sustainability issues and a hard to maintain idiosyncratic codebase.

This codebase had in fact made updating the catalog nearly impossible for as Agile noted in their review, previous data ingestion had required catalog developers to twice build the tool's index by hand and internal tables often had to be manually managed. Blacklight handles the indexing of MARC and other fielded bibliographic records quite well and uses the Rails framework to allow Ruby developers to write sub-classes to support other formats as had been done for MODS in the beta release. The underlying database is SQL, however, and modeling the catalog's metadata in ActiveRecord (Ruby's object front-end to SQL) had proven difficult and time consuming. Since any modification of the ActiveRecord format required a Rails developer to write new code to migrate the database, Agile staff concluded that while Blacklight could possibly be updated, this would require both a programmer with Ruby expertise and more stable and clearly defined metadata.

4.4.2 Agile assessment of current metadata

As identified by Agile's analysis, one major issue with the Perseus Catalog bibliographic records is that MODS records served as both records of bibliographic *manifestations* and as records of the abstract *works* contained within them. Further complicating matters was not just how expressions had been defined as versions/translations but also the large number of bibliographic items that could be versions (epigrams, plays, whole books, etc). Because distinctions between abstract works and their editions and translations were not well established, they had found it difficult to automatically extract different properties and relationships. In addition, Agile noted that using MODS records to encode non-bibliographic text aggregations (e.g. editions containing dozens or hundreds of works) and creating individual MODS records for expressions had also led to a number of serious problems: large amounts of data duplication, inconsistency in the records as the MODS standard evolved, increasingly complex MODS records, and the inability to specifically address many items within the catalog.

Due to all of this semantic complexity, Agile recommended utilizing the FRBRoo ontology to represent the underlying relational structures and FRBR level information found within the records. In FRBRoo, editions and translations are individual works that are members of a larger complex work, and MODS records could be recast as encodings of manifestations that carry expressions of one or more editions or translations of one work or many works. Thus the work of the Perseus Catalog began to move from more routine metadata creation into the needed – if somewhat nebulous – world of conceptual and ontological modeling of bibliographic data.

4.4.3 Agile recommendations for new metadata practices

After the suggestion was made and accepted to use FRBRoo, Perseus catalog staff also began implementing a number of Agile recommendations in terms of converting the metadata records. MODS records were still going to be used to encode traditional bibliographic information, and a plan was created to work from the existing records to generate statements about "Manifestation Product Types that carry expressions." One challenge this approach introduced was that a way was needed to address all of the MODS records as unique manifestations with identifiers that could be referenced. The current plan is to use OCLC identifiers where available with the possibility of using CITE-URNs for all top level manifestation records also being explored.

The version and expression level data found within the MODS records also needed to be better encoded. The first step was to remove all work identifiers and CTS URNs from the top-level manifestation records and the second step was to use the <relatedItem> tag to separately encode works and expressions. Thus for an edition of Herodian's *Ab Excessu Divi Marci Libri Octo*, instead of having <identifier type="cts-urn">urn:cts:greekLit:tlg0015.tlg001.opp-grc1</identifier> in the top level record, this identifier has now been relocated to a separately encoded constituent statement using "otherType="work" and "otherType"="expression".[49] This new format has also made it both quicker and easier to encode multiple language expressions (or even both Perseus and OPP[50] expressions) within the same manifestation.

The way single work *multi-volume* editions are cataloged has also been greatly changed again. Instead of creating large modsCollection files with one MODS record for each volume, Agile proposed creating one MODS record instead for the whole edition and to expand the use of the <relatedItem type="constituent"> element again. In this case <otherType= "structure"> was used to encode the physical structure of a work found *within* each volume with only unique manifestation level details given. This allowed the top level manifestation record to then represent the entire edition and the constituent records to encode unique volume level information (e.g. publication years, the section of a work it contains, online links, different editors, etc.). Encoding all of this information in a single MODS records makes it much easier to quickly determine what content of a work is in a given volume.[51]

The final type of change to MODS records impacted records for multi-work manifestations (either single or multi-volume). Previously, the catalog update system would take multi-work manifestation MODS records and automatically create edition/translation level records but would then eliminate the record of the entire manifestation. It was decided for the moment to stop this separate record creation process and the top level manifestation records that had been split apart in the beta and subsequent data creations were recompiled automatically. These newly recreated manifestations included full lists of encoded constituent works, albeit with only top level information (page numbers and page level links to online manifestations were not included). Re-inverting the data once again enabled us to quickly count how many works were within a volume

[49] To see the full MODS record: https://git.io/fp7WE (last access 2019.01.31).

[50] OPP stands for the Open Philology Project at Leipzig, a version identifier chosen to represent non-PDL editions.

[51] For a sample two volume edition of Tacitus *Annales*, see https://git.io/fp7WV (last access 2019.01.31).

and to more easily answer the question of how many editions have actually been cataloged. One unresolved and important challenge introduced by this approach, however, is that of where and how to store the expression level data left behind in the separate records. This data is not currently found within the newly revised catalog data files but will be "added back in" once an appropriate format and structure is decided upon.

Another unresolved metadata challenge, in terms of adding new editions to catalog_data, was the inability to relaunch the system that automatically created CTS-URNs for MODS records. At the end of the Agile revision project, the data within catalog_pending was not ingested but only converted to the newer formats, with a number of errors due to the varying types of works found within this repository. Manual revision of these records, including correcting errors and creating CTS-URNs for new work/expressions is ongoing. On the other hand, all of the new MADS authority records within catalog_pending were successfully ingested. In addition, as a further enhancement, all of the author name files were renamed to their CITE URNs as a first step towards LOD compliance.

A number of other changes were also suggested and implemented by Agile in terms of MADS authority records. Agile suggested that the most comprehensive data listing of authors, works and expressions maintained by the PDL was not the catalog itself but was instead the open access bibliography first created in 2005. A MADS RDF database was thus created from this spreadsheet, with MADS authority records created not just for all of the works on the list but also for the many authors not yet in the Perseus Catalog (as there had been no editions cataloged for them). These MADS records contain CTS-URNs which can then be used to potentially link MODS expression constituents to expanded work level data. MADS work authority records were created again because the lack of them not only limited automatic reasoning about works but also meant there was no metadata space for work description (variant name titles, uncertain dates, contested authorship attribution), and no way to pull in data from other sources about a work. Interestingly, this practice of creating work authority records was implemented in the Perseus experimental catalog but the sheer volume of data creation made it impossible to continue manually.

4.4.4 LOD at last? Commitment to openness and future directions

Over the course of almost a year's work, it was determined that the amount of metadata revision needed and the inability to update/modify the Blacklight instance required a rethinking of what could be accomplished. While the metadata is still being actively converted and edited, work on a new interface has been put off for the time being until there are further funds for infrastructure, deployment and testing. At the same time, work is also still ongoing to represent the metadata found in the both the catalog and its related bibliographic spreadsheets and adapt it in such a way that captures the complex relationships between works, expressions, and manifestations. It has been decided that RDF due to its relational nature and its logical foundation, would make it the ideal format to which the catalog data could be transformed.[52]

An initial RDF knowledge base of statements about authors, works, expressions and manifestations has been developed and an additional knowledge base of statements relating expressions to manifestations has been generated from the converted MODS records. This RDF data can be loaded into any triple store and queried using SPARQL.[53] It is hoped that this knowledge base upon completion and release can be efficiently linked to tools or bibliographies that will allow librarians and scholars to update and correct it easily. In addition, the creation of such a knowledge base will allow for machine-readable applications to make use of the data. By encoding bibliographic knowledge as RDF, we seek to integrate our work with the semantic web and the larger global work of scholars and librarians who have already captured bibliographic information in RDF.[54]

An extensive amount of Linked Open Data about ancient authors and works has been generated within the last few years, and, ideally, partnerships with the Iowa Canon, DLL, and OGL will continue. At the same time, the Perseus Catalog RDF does provide something unique: expression and manifestation level metadata that links works to their published editions and translations. It may turn out that the need for a separate interface to the Perseus Catalog becomes redundant as its most useful part is its bibliographic data about actual works with links to their online expressions and manifestations. If that data can be packaged up and better searched through other projects' APIs and interfaces, then work will likely exclusively focus on the development of more metadata as LOD for sharing with other digital classics projects. Much of the effort of the next year will be to try to both design and implement a system that will enable Perseus catalog metadata creators to curate authority metadata about ancient authors and works and, similarly, to collect and

[52] We are closely following the work of the MODS to RDF mapping group. See https://tinyurl.com/yaud3gmt (last access 2019.01.31).

[53] Access to this knowledge base is currently only available through an experimental web application.

[54] See for example the work of OCLC: https://www.oclc.org/research/themes/data-science/linkeddata.html (last access 2019.01.31).

curate references to both online editions and links to specific portions of them. There is a need for a new metadata management system that allows not only for efficient creation of metadata but supports collaborative workflows between the different projects.

5 Conclusion

So after thirteen years, the goal of the Perseus Catalog has evolved once again: having shifted from 1) a FRBR-based interface to the PDL classics collection, 2) to an online finding aid both for PDL texts and for all Greek and Latin works produced up until 600 CE, 3) to a metadata source for OGL and its component projects, and now 4) to the aim of producing a comprehensive, extensible and machine readable knowledge base about Greek and Latin texts.

Whatever future path the development of the Perseus Catalog takes in terms of infrastructure and data creation, the leaders of this effort remain committed to openness. This is not simply limited to the distribution of data and any code, but more importantly extends to a desire to collaborate with the growing number of digital classics projects exploring the same issues.

Bibliography

Abbas, J.; Baker, S.R.; Huskey, S.J.; Weaver, C. (2015): "Digital Latin Library: Information Work Practices of Classics Scholars, Graduate Students, And Teachers". In: Proceedings of the American Society for Information Science and Technology. Wiley Online Library. 52, 1–4.

Abbas, J.; Baker, S.R.; Huskey, S.J.; Weaver, C. (2016): "How I Learned to Love Classical Studies: Information Representation Design of The Digital Latin Library". In: Proceedings of the 79th ASIS&T Annual Meeting. Access Innovations Inc. Volume 53, 1–10.

Almas, B.; Schroeder, C. (2016): "Applying the Canonical Text Services Model to the Coptic SCRIPTORIUM". Data Science Journal 15. http://doi.org/10.5334/dsj-2016-013.

Almas, B.; Babeu, A.; Krohn, A. (2014): "LOD in the Perseus Digital Library". ISAW Papers 7: Current Practice in Linked Open Data for the Ancient World. New York, NY: Institute for the Study of the Ancient World. http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/almas-babeu-krohn/ (last access 2019.01.31).

Babeu, A. (2008): "Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience". Perseus Digital Library. http://www.perseus.tufts.edu/publications/PerseusFRBRExperiment.pdf (last access 2019.01.31).

Babeu, A. (2012): "A Continuing Plan for the "FRBR-Inspired" Catalog 2.1? (Fall 2012)". Perseus Digital Library. http://sites.tufts.edu/perseusupdates/files/2012/11/FRBRPlanFall2012.pdf (last access 2019.01.31).

Babeu, A.; Dilley, P. (forthcoming): “Linked Open Data for Greek and Latin Authors and Works.” In: Linked Open Data for the Ancient World: Standards, Practices, Prospects, ISAW Papers.

Berti, M.; Almas, B.; Dubin, D.; Franzini, G.; Stoyanova, S.; Crane, G. (2015): “The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors”. Journal of the Text Encoding Initiative 8 https://jtei.revues.org/1218 (last access 2019.01.31).

Chen, Y.N. (2017): “A Review of Practices for Transforming Library Legacy Records into Linked Open Data”. In: E. Garoufallou; S. Virkus; R. Siatri; D. Koutsomiha (eds): Metadata and Semantic Research. MTSR 2017. Cham: Springer, 123–133.

Clarke, R.I. (2014): “Breaking Records: The History of Bibliographic Records and their Influence in Conceptualizing Bibliographic Data”. Cataloging & Classification Quarterly 53:3–4, 286–302.

Crane, G.; Almas, B.; Babeu, A.; Cerrato, L.; Krohn, A.; Baumgart, F.; Berti, M.; Franzini, G.; Stoyanova, S. (2014): “Cataloging for a Billion Word Library of Greek and Latin”. In: DATeCH ‘14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. New York, NY: ACM, 83–88.

Fuller, T.N.; Page, K.R.; Willcox, P.; Jett, J.; Maden, C.; Cole, T.; Fallaw, C.; Senseney, M.; Downie, J.-S. (2015): “Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications”. In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. New York, NY: ACM, 169–172.

Gibson, N.P.; Michelson, D.A.; Schwartz, D.L. (2017): “From Manuscript Catalogues to A Handbook of Syriac Literature: Modeling An Infrastructure For Syriaca.Org”. Journal of Data Mining & Digital Humanities. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages (May 30, 2017). http://arXiv:1603.01207 [cs.DL].

Guerrini, M.; Possemato, T. (2016): “From Record Management to Data Management: RDA and New Application Models BIBFRAME, RIMMF, and OliSuite/WeCat”. Cataloging & Classification Quarterly 54:3, 179–199.

IFLA. (1998): Functional Requirements for Bibliographic Records. Final Report. Volume 19 of UBCIM Publications-New Series. München: K.G. Saur. https://www.ifla.org/publications/functional-requirements-for-bibliographic-records (last access 2019.01.31).

Jett, J.; Fuller, T.N.; Cole, T.W.; Page, K.R.; Downie, J.S. (2016): “Enhancing Scholarly Use of Digital Libraries: A Comparative Survey and Review of Bibliographic Metadata Ontologies”. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL ‘16. New York, NY: ACM, 35–44.

Le Boeuf, P. (2012): “A Strange Model Named FRBRoo”. Cataloging & Classification Quarterly 50:5–7, 422–438.

Michelson, D.A. (2016). “Syriaca.org as a Test Case for Digitally Re-Sorting the Ancient World”. In: C. Clivaz; P. Dilley; D. Hamidović (eds.): Ancient Worlds in Digital Culture. Leiden and Boston: Brill, 59–85. http://dx.doi.org/10.1163/9789004325234_005.

Mimno, D.; Crane, G; Jones, A. (2005): “Hierarchical Catalog Records Implementing a FRBR Catalog”. D-Lib Magazine 11:10. http://www.dlib.org/dlib/october05/crane/10crane.html (last access 2019.01.31).

Romanello, M.; Pasin, M. (2017): “Using Linked Open Data to Bootstrap a Knowledge Base of Classical Texts”. In: Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017). CEUR, 3–14.

Smith, D.A.; Mahoney, A.; Rydberg-Cox, J. (2001): "Management of XML Documents in an Integrated Digital Library". Extreme Markup Language 2000. https://people.cs.umass.edu/~dasmith/hopper.pdf (last access 2019.01.31).

Tiepmar, J.; Heyer, G. (2017): "An Overview of Canonical Text Services". Linguistics and Literature Studies 5, 132–148.

Zapounidou, S.; Sfakakis, M.; Papatheodorou, C. (2016): "Representing and Integrating Bibliographic Information into the Semantic Web: A Comparison of Four Conceptual Models". Journal of Information Science 43:4, 525–553.