D. Grant Campbell, University of Western Ontario, Canada, gcampbel@uwo.ca
Alex Mayhew, University of Western Ontario, Canada, amayhew@uwo.ca

A Phylogenetic Approach to Bibliographic Families and Relationships

Abstract

This presentation applies the principles of phylogenetic classification to the phenomenon of bibliographic relationships in library catalogues. We argue that while the FRBR paradigm supports hierarchical bibliographic relationships between works and their various expressions and manifestations, we need a different paradigm to support associative bibliographic relationships of the kind detected in previous research. Numerous studies have shown the existence and importance of bibliographic relationships that lie outside that hierarchical FRBR model: particularly the importance of bibliographic families. We would like to suggest phylogenetics as a potential means of gaining access to those more elusive and ephemeral relationships. Phylogenetic analysis does not follow the Platonic conception of an abstract work that gives rise to specific instantiations; rather, it tracks relationships of kinship as they evolve over time. We use two examples to suggest ways in which phylogenetic trees could be represented in future library catalogues. The novels of Jane Austen are used to indicate how phylogenetic trees can represent, with greater accuracy, the line of Jane Austen adaptations, ranging from contemporary efforts to complete her unfinished work, through to the more recent efforts to graft horror memes onto the original text. Stanley Kubrick's *2001: A Space Odyssey* provides an example of charting relationships both backwards and forwards in time, across different media and genres. We suggest three possible means of applying phylogenetics in the future: enhancement of the relationship designators in RDA, crowdsourcing user tags, and extracting relationship trees through big data analysis.

Introduction

This paper represents an initial exploration into applying the principles of phylogenetic classification to the phenomenon of bibliographic relationships in library catalogues. The FRBR paradigm supports hierarchical bibliographic relationships between works and their various expressions and manifestations. The FRBR-based *Resource Description and Access (RDA)* offers relationship designators; however, we need a different paradigm, both to support the effective use of these designators, and to explore alternative means of representing associative bibliographic relationships of the kind detected in previous empirical studies. We offer three possible means of applying phylogenetics in the future: enhancement of the relationship designators in RDA, crowdsourcing user tags, and extracting relationship trees through big data analysis.

Background and Objectives

Descriptive cataloguing, both as a profession and as a field of study, has devoted continuing and urgent attention to the problem of representing bibliographic relationships in library catalogues. If the catalogue is to be anything more than a mere inventory of materials, it needs to provide users with the means to navigate a large information space by encoding meaningful connections between resources that would normally be physically dispersed, both on shelves and in alphabetical listings (Svenonius 2000, 20). Traditionally, subject cataloguing has addressed this need through the syndetic references in subject headings, connecting resources with related content in relationships of equivalence, hierarchy and association (Chan 1995, 119). Descriptive cataloguing, by contrast, has focused not on subject content but on bibliographic identities and relationships. Since Panizzi, cataloguing rules have aimed, to greater or lesser degrees, "to bring together under an author's name all his works and under the title of a work all the editions and translations" (Lubetzky 1961, 233). The IFLA Report on the Functional Requirements of Bibliographic Records used an extensive entity-relationship analysis of catalogues as databases (Delsey 2016) to produce a four-tiered paradigm of one-to-many relationships: work, expression, manifestation, and item (IFLA 1998). By founding the new cataloguing standard upon this paradigm, the designers and steering committee of *Resource Description and Access (RDA)* hope to facilitate the creation of new kinds of bibliographic records grounded on hierarchical relationships, in which the abstract work is described once, and then related to various expressions, manifestations and items. Such relationships would make the bibliographic records more amenable to representation using the linked data techniques which form the basis of the emerging BIBFRAME initiative (Library of Congress 2016).

While this effort promises to improve the catalogue's ability to represent fairly traditional bibliographic relationships, numerous studies have shown the existence and importance of bibliographic relationships that lie outside that hierarchical FRBR model. Considerable research undertaken in the 1990s and 2000s showed the importance of bibliographic families: sets of bibliographic works that are related by their derivation from a common progenitor (Smiraglia & Leazer 1999, 494).

These relationships present challenges to cataloguing design for three reasons. First, FRBR's hierarchies, while containing many of these relationships (Smiraglia 2007, 75), do not cover them all. Derivative relationships can be simultaneous or successive, and can range from translations and performances to adaptations, extractions and amplifications (Smiraglia & Leazer 1999, 495), intra- and extratextual references (Green 2001), content relationships, and whole-part or part-to-part relationships (Tillet 2001). RDA acknowledges the importance of these derivative relationships by providing an ambitious range of relationship designators. While records for serial publications have traditionally used linking fields due to the complexity of serial publishing, with RDA have we now have a systematic vocabulary of designators for recording relationships in the access points of bibliographic records generally.

Second, these relationships are not always represented in the bibliographic evidence used to create descriptions. Works that form a series do not always advertise themselves as such, and relationships may be important only to particular cultural, temporal or geographic contexts. As such, these ephemeral relationships defy the efforts of even the most dedicated and conscientious cataloguer to enter and maintain. It is relatively simple to record publications as a series if the series is explicitly named, as with Oxford World Classics. It is more difficult to record the relationship between the separately-published volumes in Mary Stewart's Merlin trilogy, for example, since *The Crystal Cave*, *The Hollow Hills* and *The Last Enchantment* were all published independently. Disney's *The Lion King* and Tom Stoppard's *Rosencrantz and Guildenstern are Dead* both share significant plot similarities with Shakespeare's *Hamlet*. But the nature of the debt differs between the two instances, and not all user communities would value both equally. And, of course, with *Hamlet*, the inter-textual relationships could extend indefinitely.

Finally, these relationships, for all their ephemerality, are often important to users, who are struggling to connect to other resources based on specific, bounded connections which may not be evident in other contexts. As Harold Bloom acknowledges in *The Anxiety of Influence*, much of our understanding of our cultural record involves detecting the echoes of one work that appear in succeeding works, often as a vexed and uneasy relationship in which one author battles with the oppressive heritage of a predecessor and subjects that predecessor to creative misreadings (Bloom 1997, 5).

If, as Bloom suggests, authors "misread" their predecessors, then sequences of creative misreading could be meaningful to users. We would like to suggest phylogenetics as a potential means of gaining access to those more elusive and ephemeral relationships. Phylogenetics, as a branch of biological systematics, involves the reconstruction of genealogical history, generally in the form of diagrammatic trees which display the evolution of organisms over time from a common ancestor (Velasco 2013, 990). Phylogenetics, owing to its roots in biological taxonomies and classification, has attracted some interest in Knowledge Organization circles, primarily in relation to subject classification. James Duff Brown attempted to build a subject classification based on a serial approach of subjects evolving from their more primitive levels, and although his approach failed to gain traction, phylogenetic principles have been explored as a possible means of isolating main classes in approaches to classification based on the theory of integrative levels (Gnoli 2006, 139). We would like to suggest extending this interest into the area of bibliographic description, specifically to provide a means of representing those more elusive, and possibly more important bibliographic relationships that FRBR does not capture.

Methodology

Using phylogenetics outside the domain of biology and paleontology is unusual, but not unprecedented. O'Brien and Lyman (2003) use phylogenetics in archaeology, "reconstructing networks of evolution relationships among cultural phenomena" (4). Using cladistic visualizations based on perceived similarities and perceived lines of influence provides a means of visualizing bibliographic relationships. In order to explore the initial possibilities of this method, we selected two works: *Pride and Prejudice* by Jane Austen, and Stanley Kubrick's film, *2001: A Space Odyssey.* In each case, we assembled a list of possibly relevant relationships to other bibliographic entities, and visualized them using the standard clade visualization, as can be seen in the figures to follow. In so doing, we used the following definitions:

- *Phylogenetics***:** the practice of studying "evolutionary interrelationships," in an attempt to map diversifications and changes over time (Gale 3075);
- *Cladistics*: a particular method of phylogenetics which isolates those entities which descended from a common ancestor (Gale 3075).

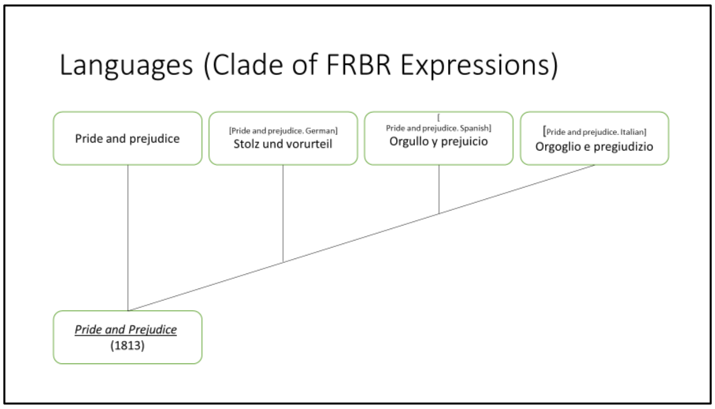
According to these definitions, a "clade" is a tree structure which stems from a single root. But phylogenetic relationships may involve multiple clades.

We began by mapping clades based on obvious relationships, many of which can be represented within the FRBR Framework or by subject headings:

- Genre (in which "Fiction" gives rise to "Regency Fiction," "Gothic Fiction," "Victorian Fiction," and "Modern Fiction."

- Similar authors (in which “Regency Fiction” gives rise to “Austen, Jane, 1775-1817,” “Scott, Walter, 1771-1832,” “Ferrier, Susan, 1782-1854,” and “Edgeworth, Maria, 1768-1849”);
- Works (in which “Austen, Jane, 1775-1817” gives rise to *Sense and Sensibility* (1811), *Pride and Prejudice* (1813), *Mansfield Park* (1814), and *Emma* (1815));
- Expressions (in which the work, *Pride and Prejudice* (1813) gives rise to the English, German, Spanish and Italian translations) (See Figure 1);
- Superworks (in which the work, *Pride and Prejudice* (1813), gives rise to adaptations on film, television, radio and theatre.

Figure 1: Languages Clade


Languages (Clade of FRBR Expressions)
Pride and prejudice
[Pride and prejudice. German]
Stolz und vorurteil
[
Pride and prejudice. Spanish]
Orgullo y prejuicio
[Pride and prejudice. Italian]
Orgoglio e pregiudizio
Pride and Prejudice
(1813)

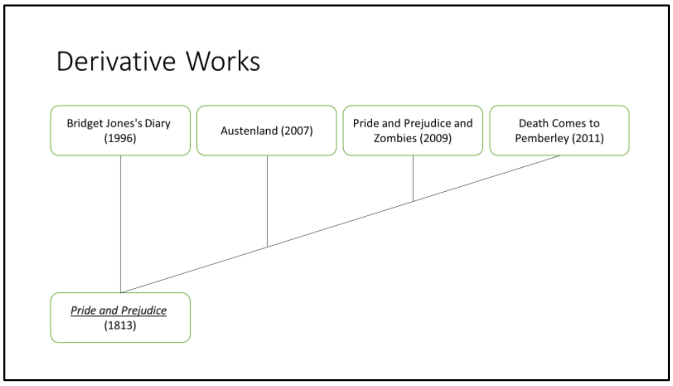
In none of these cladistic visualizations did we attempt to be comprehensive or authoritative; we merely attempted to see how the visualizations would look, and if they could conceivably be useful. In these conventional relationships, we detected some potential benefits, mainly deriving from the fact that the FRBR structure of RDA would make it fairly simple to extract these entities and represent them in this fashion. However, we noted certain problems that would prevent them from being enthusiastically adopted by either cataloguers or interface designers.

To begin with, definitions of genre and of author similarity are matters of domain expertise, and often hotly contested. Indeed, many of these clades are based on scholarly and educational warrant, and in a general catalogue, such warrant is likely to be highly diverse for different areas and subjects, and often subject to debate and scholarly revision. Asking cataloguers to acquire such expertise is hardly reasonable.

More important, the cladistic visualization is vulnerable to a misleading heuristic. Because a clade by definition presents the descendants of a single source, the visualization implies that the single source is the root of the tree: “When a tree is drawn in the rectangular format, some node must be drawn at the extreme left. The problem is that our eye interprets that leftmost node as the root, when in fact there is no root” (Hall 2011,81). The visualization, when founded on relationships defined by educational and scholarly warrant, ends up recreating the very authority of “the work” that this project sought to question and problematize.

When we move to more informal relationships, however, based on influence and derivation, the visualization becomes more powerful. The influence of *Pride and Prejudice* on modern culture can be vividly represented through a clade that offers links to *Bridget Jones's Diary*, *Austenland*, *Pride and Prejudice and Zombies*, and *Death Comes to Pemberley* (see Figure 2).

Figure 2: Clade Based on Derivative Works



Derivative Works

Bridget Jones's Diary (1996)

Austenland (2007)

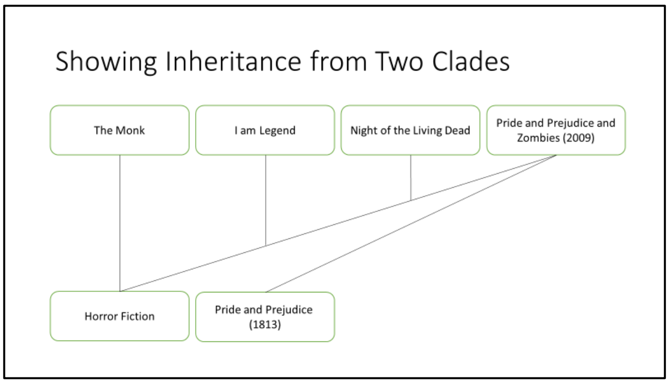
Pride and Prejudice and Zombies (2009)

Death Comes to Pemberley (2011)

Pride and Prejudice (1813)

We can also see the emergence of a single entity from more than one clade, as *Pride and Prejudice and Zombies* can be seen to descend, not just from Austen, but from the horror genre stretching as far back as *The Monk* and across different media to embrace *I am Legend* and *Night of the Living Dead* (see Figure 3).

Figure 3: Multiple Inheritances



Showing Inheritance from Two Clades

The Monk

I am Legend

Night of the Living Dead

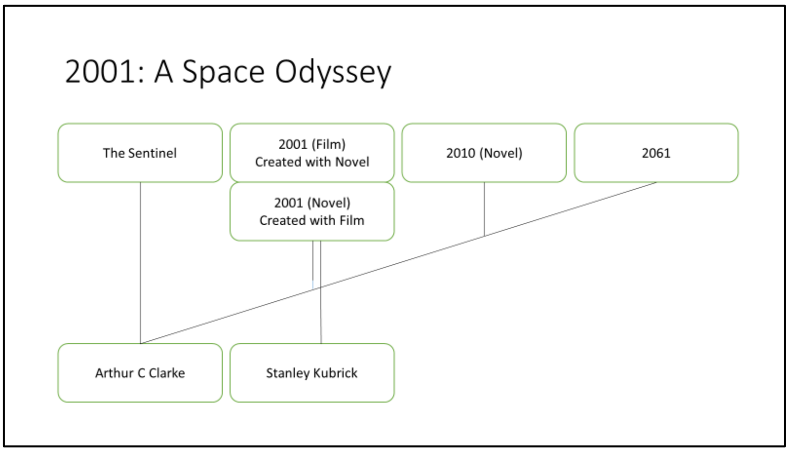
Pride and Prejudice and Zombies (2009)

Horror Fiction

Pride and Prejudice (1813)

With *2001: A Space Odyssey*, the web of derivations and inheritances becomes even denser. The Kubrick movie was inspired by Arthur C. Clarke's short story, "The Sentinel." Clarke's novel appeared simultaneously with the initial release of the movie; the sequel to the novel, *2010: Odyssey Two* gave rise to a movie of the same name (see Figure 4).

Figure 4: The Case of 2001: A Space Odyssey



The clade visualizations for derivations and influences across media and time suggest two entirely different qualities from those exhibited by the more traditional clades. First, the relationships appear far more overtly satisficing than the relationships of genre, author, work and expression. The clade makes no pretense at being a definitive, thorough or authoritative representation of relationships produced by subject specialists; rather, they have the provisional air of relationships that are meaningful to users within specific use contexts, built in the process of use, reference and exploration. Second, the root node(s) in these cases are blatantly arbitrary. There is no suggestion that "The Sentinel" occupies any startlingly significant place in the bibliographic universe; it is simply a place where a user would presumably begin.

Results and discussion

Phylogenetic analysis has three primary advantages over FRBR as a means of representing these relationships. First, such analysis does not follow the Platonic conception of an abstract work that gives rise to specific instantiations; rather, it tracks relationships of kinship as they evolve over time. This enables us to explore kinship relationships across works, rather than establishing the abstract work as the ultimate source. While it is undoubtedly useful to link all expressions and manifestations of Shakespeare's plays to their abstract identities as works, we also need a method of linking these plays to their source material: European plays in various languages, Holinshed's *Chronicles*, and various source poems and stories. We may also want to trace the origins of these works in new works that adapt and transform them: sometimes obviously, as in

West Side Story as a retelling of *Romeo and Juliet*, and sometimes covertly, as in the overtones of *Hamlet* in *The Lion King*.

Second, phylogenetics as a field of practice sustains multiple levels of precision and definition. While phylogenetics is widely used in biology and bioinformatics as a precision instrument for analyzing and providing evidence of ancestry (Binet, et al. 2016), there is a telling and highly advantageous ambiguity at its heart: "exactly what a phylogeny represents is a matter of debate and is arguably a context-sensitive matter" (Velasco 2013, 991). Far from being a weakness, this ambiguity enables us to construct multiple trees based on diverse needs, and on diverse conceptualizations of the "origin."

Finally, not all contexts need be permanent. Trees of kinship and descent can be provisional. Recent political events, for example, have triggered a resurgence in the lineage of twentieth-century dystopian fiction: a lineage of great interest at present, but whose long-term significance is uncertain. Important current events, significant anniversaries, and longitudinal analysis of pressing policy issues can all create a need for the extraction of temporary phylogenetic trees as conceptual and navigational aids.

These examples lead to three suggested means of applying phylogenetics to bibliographic description. First, these relationships could be encoded at the time of description. At present, the relationship designators in RDA provide a useful list of terms for relationships between resources. These terms are chiefly designed to clarify relationships within the FRBR paradigm: indicating, for instance, that a translation is an expression of a particular work. However, they can be used to extend beyond the FRBR paradigm. While this option presents daunting challenges to already overworked cataloguing services, it may be viable in the case of evolutionary relationships that the library considers of permanent value.

Second, external resources could be harvested to indicate relationships of emerging interest, at both broad and local levels. In the case of the more conventional relationships that require educational and scholarly warrant, academic resources such as *The Victorian Web* (http://www.victorianweb.org/), which document precise and authoritative relationships, could conceivably be converted to linked data and imported into the catalogue as a means of orienting and arranging the library's existing resources into meaningful patterns of influence and descent. For less formal relationships, user tagging could conceivably be used to detect connections and to produce cladograms that could orient the library's resources around a provisional, culture-specific or emerging pattern of descent.

Finally, big data analysis could be used to detect relationships of correlation between multiple resources: correlations that could be analyzed for possible further visualizations in catalogue displays. Search engine queries, catalogue search queries, publication lists, database queries, social media activity, and even such seemingly unhelpful user tags as "to read" and "to do" might conceivably contribute to emergent cladistic patterns that not only assist in catalogue navigation, but serve collection development and reference activities as well.

In recent decades, we have witnessed two phenomena of great significance to libraries: the fracturing of information resources into increasingly diverse media, and a growing tendency away from ownership of physical resources to access of digital resources (Kelly 2016, 109). The proliferation of multiple media is creating a greater demand for methods of deriving relationships of lineage across these various media: plays, books, movie adaptations, television adaptations, games, and even apps and operating systems. As these media migrate inexorably from physical collections to virtual ones, libraries have the opportunity of using knowledge organization principles and practices as a service which provides meaningful pathways through the complexities of digital access. We suggest that phylogenetic analysis and visualization provide a flexible, agile and fruitful means of mobilizing knowledge organization in the service of this opportunity.

References

Binet, Manuel, Olivier Gascuel, Celine Scornavacca, Emmanuel J.P. Douzery, and Fabio Pardi. 2016. “Fast and Accurate Branch Lengths Estimation for Phylogenomic Trees.” *BMC Bioinformatics* 17, no. 23. http://doi.org/10.1186/s12859-015-0821-8

Bloom, Harold. 1997. *The Anxiety of Influence: A Theory of Poetry*. 2nd ed. Oxford: Oxford University Press.

Chan, Lois Mai. 1995. *Library of Congress Subject Headings: Principles and Applications*. 3rd ed. Englewood: Libraries Unlimited.

Delsey, Tom. 2016. “The Making of RDA.” *Italian Journal of Library, Archives, and Information Science* 7, no. 2: 25-47.

Gale Research Group. 2004. "Phylogeny." *The Gale Encyclopedia of Science*, edited by K. Lee Lerner and Brenda Wilmoth Lerner, 3rd ed., vol. 5. Detroit: Gale: 3075-3076.

Gnoli, Claudio. 2006. “Phylogenetic Classification.” *Knowledge Organization* 33, no. 3, 138-152.

Green, Rebecca. 2001. “Relationships in the Organization of Knowledge: An Overview.” *Relationships in the Organization of Knowledge*. Ed. R. Green & C. Bean. Boston: Kluwer, 1-3.

Hall, Barry G. 2011. *Phylogenetic Trees Made Easy.* 4th ed. Sunderland: Sinauer.

IFLA Study Group on the Functional Requirements for Bibliographic Records. 1998. *Functional Requirements for Bibliographic Records*. München: Saur.

Kelly, Kevin. 2016. *The Inevitable: Understanding the 12 Technological Forces that will Shape our Future.* New York: Viking.

Library of Congress. 2016. *Overview of the BIBFRAME Model*. Retrieved from https://www.loc.gov/bibframe/docs/bibframe2-model.html

Lubetzky, Seymour. 1961. “The Function of the Main Entry in the Alphabetical Catalogue—One Approach.” Reprinted in *Seymour Lubetzky: Writings on the Classical Art of Cataloging*. Ed. E. Svenonius, D. McGarry. Englewood: Libraries Unlimited: 231-237.

Smiraglia, Richard. 2007. “Bibliographic Families and Superworks.” *Understanding FRBR: What it is and how it will affect our retrieval tools*. Ed. Arlene G. Taylor. Westport: Libraries Unlimited, 73-86.

Smiraglia, Richard and Gregory Leazer. 1999. “Derivative Bibliographic Relationships: The Work Relationship in a Global Bibliographic Database.” *Journal of the American Society for Information Science*, 50.6, 493-504.

Svenonius, Elaine. 2000. *The Intellectual Foundation of Information Organization*. Cambridge: MIT Press.

Tillet, Barbara. 2001. "Bibliographic Relationships." *Relationships in the organization of knowledge*. Ed. R. Green & C. Bean. Boston: Kluwer, 3-19.

Velasco, Joel D. 2013. "Philosophy and Phylogenetics." *Philosophy compass*, 8.10: 990-998.