2017

Extending the FRBR model: A proposal for a Group 4

Alex Mayhew
Western University, amayhew@uwo.ca

Extending the FRBR model: A proposal for a Group 4

Alex Mayhew
Supervisor: Grant Campbell
Course: LIS9891
September 6

Abstract:

The three Entity-Relationship Groups (E-R Groups) have formed the conceptual framework of cataloguing in FRBR since being established by IFLA in the late 1990s. These three groups define the entities and relationships of interest to FRBR cataloguers. Group 1 describes the Work, Expression, Manifestation, and Item, the parts that make up the whole of a catalogue record. Group 2 describes the responsibility relationships, ensuring proper attribution and ownership to people and corporations. Group 3 describes the subject relationships which include the entities of groups 1 and 2, as well as Concept, Object, Event, and Place, and function like subject headings.

However, the growing demands put upon catalogues by their users present cataloguers with an opportunity to expand the functionality of their catalogues. This paper will argue for the creation of a "Group 4" consisting of flexible entity-entity relationships, modeled after biological phylogenetics, anchored temporally, and generated by users.

Introduction:

Functional Requirements for Bibliographic Records (FRBR) was at the time of its creation the latest effort to create a list of basic requirements for national bibliographic records. "The formal charge for the IFLA study involving international bibliography standards was to delineate the functions that are performed by the bibliographic record with respect to various media, applications, and user needs." (Madison, 2000) The explosion of the online environment has radically changed expectations regarding media, applications, and user needs. This challenge presents the cataloguing community with an opportunity to expand the possible functions of catalogues.

This paper will describe one possible approach to realizing this aim, broken into four parts. First, the current 3 Groups of the FRBR Entity-Relationship Model will be described. This includes their purpose and structure. The next section will discuss the limitations of the current groups included in the FRBR model. Particular attention will be directed at the sorts of functions a variety of users may find valuable. Following from that will be a description of the Phylogenetic Model, as well as an explanation of how it maps onto the bibliographic universe and how it can meet the demands of the previous section. Finally, this paper will outline three stages of implementation of a fully realized Group 4 based on a Phylogenetic Model.

Part 1: FRBR Entity-Relationship Model

FRBR is based on the principle of Entity–relationship modelling, setting out three distinct groups of entities and their relationships of concern. These groups serve as conceptual

frameworks for guiding the creation of catalogue records. It is the extendability of this framework to other relationships that is being questioned in this paper.

Group 1: Entities and "Primary" Relationships: Work, Expression, Manifestation, Item

"The entities in the first group (as depicted in Figure 1.1) represent the different aspects of user interests in the products of intellectual or artistic endeavour. The entities defined as work (a distinct intellectual or artistic creation) and expression (the intellectual or artistic realization of a work) reflect intellectual or artistic content. The entities defined as manifestation (the physical embodiment of an expression of a work) and item (a single exemplar of a manifestation), on the other hand, reflect physical form." (IFLA)

Describing entities and relationships of Group 1, William Denton explains that "FRBR uses its 4 level hierarchy to move from an abstract work to an item one can hold in one's hand, but other people have other arrangements, such as work, version, adaptation." (Denton, 49) Alternatively, Group 1 can be thought of as a collection of Whole-Part relationships, the totality forming the final bibliographic record. Regardless, Group 1 breaks the formerly unified display of the bibliographic record into 4 entities in a series of one-to-many relationships.

Group 2: Entities and "Responsibility" Relationships: Entities: Person, Corporate Body

"The entities in the second group (outlined in bold in Figure 1.2) represent those responsible for the intellectual or artistic content, the physical production and dissemination, or the custodianship of the entities in the first group. The entities in the second group include person (an individual) and corporate body (an organization or group of individuals and/or organizations)." (IFLA)

This second group is effectively concerned with capturing ownership or attribution of the various Group 1 entities. All texts in the catalogue have authors and owners.

Group 3: Entities and "Subject" Relationships: Entities: Concept, Object, Event, Place

"The entities in the third group (outlined in bold in Figure 1.3) represent an additional set of entities that serve as the subjects of works. The group includes concept (an abstract notion or idea), object (a material thing), event (an action or occurrence), and place (a location)." (IFLA)

Group 3 is based on subject "aboutness", similar to thesauri, subject headings, and other classification schemes. In practice cataloguers may find themselves encouraged to use the minimum number of subject headings to capture the content of a text.

Part 2: Limitations and Expectations

An important point of note is that while the FRBR entities can have countless examples, FRBR concerns itself with only eight defined relationships. Group 1 has the 'is realized through', 'is embodied in', and 'is exemplified by' relationships. Group 2 has the 'is created by', 'is realized by', 'is produced by', and 'is owned by' relationships. Group 3 has the 'has as a subject' relationship. See appendix A for more details.

This has made the creation of bibliographic records more efficient, but has also, paradoxically, entrenched the status of the single, isolated bibliographic record, and has diverted cataloguers from an important fact mentioned in LIS research: that relationships between the resources represented by these records are important, and underserved. Recognition of the value of these relationships can be seen the proposal to capture in records 'bibliographic families': "groups of works that share common intellectual content." (Smiraglia, 73) Indeed, Smiraglia's concept of the progenitor work foreshadows extended biological metaphor used in this paper.

The card catalogue and main entry in particular seem to be the origin of modern catalogues' use of a limited number of relationship types. The historical technological situation simply posed too great an impediment to the creation of more inclusive or dynamic systems. This idea of a limited set of relationships was even embodied in the Paris principles with its authors, works, and editions. FRBR is likewise bound to its own list of entities and relationships.

While this often resulted in a functional catalogue, users may want more from a catalogue then locating a text through only predefined access points, such as authorship or subject. The connections between texts, such as sequels, new editions, and parodies all fall outside of FRBR's eight relationships. Some implementations of FRBR have recognized this as evidenced by the Appendices provided in RDA to connect one record to another.

These limitations suggest that a supplementary model is needed. The biological taxonomy schema of phylogenetics may in fact provide cataloguers with an enhanced means of encoding bibliographic relationships.

Part 3: Phylogenetic Model for Group 4

Before phylogenetics can be shown to be a useful model for creating a Group 4 of Entity-Entity relationships, it must first be explained. In biology phylogenetics is the method of organizing species by lines of descent into groups called clades. There are several terms that need further clarification.

A species is a group of organisms that is able to interbreed with each other, but are generally unwilling or unable to interbreed with others. Species can be considered roughly analogous with the Entities of Group 4. While for the most part these will be texts, those entities could be expanded at a later point in the development of the model.

Clades are groups of species that descend from a common ancestor. In a catalogue this might be all the books by a single author, or all the books part of a series. Any grouping, no matter how large or small, can be a clade, as long as the entities in the group are bound by common descent.

The next concept of importance is the line of descent. In biology each organism is descended from an unbroken chain of ancestors that leads all the way back to the earliest life on the planet. Note that each species exists at a particular time and for a particular duration, it is impossible to be descended from a species that does not exist yet. "An ancestor must be totally primitive with respect to their descendants." (O'Brien and Lyman, 83) This means that any we must always be wary and ensure that we are identifying actual ancestors, not just earlier branches from another more distant common ancestor.

This reliance on temporality is the primary organizing principle of phylogenetics. The line of descent would be analogous to all the myriad relationships between entities that users and cataloguers may wish to record. While the full list of biological relationships is beyond the scope of this paper, for the sake of illustration two will be described along with their bibliographic counterparts: Parent-child descent and interspecies hybridization.

The concept of parent-child descent is familiar enough; two members of the same species commingle their genetic information and the resulting child inherits characteristics from both parents. In the biology example it should be noted that these characteristics need not be obviously expressed. Despite being rather common in nature, at least amongst the animals, an

exact parallel is somewhat rare in creation of texts, however an example can be found in the manuscript traditions. A scribe can work from two different manuscripts at once, producing a new copy with characteristics of both.

Hybridization occurs when two formerly distinct species merge. The result can either be a new species distinct from either of the ancestor species, or the absorption of one species into another, simply increasing its genetic diversity. Despite being fairly rare in nature, hybridization occurs fairly frequently in human created texts. An example of hybridization might be the works of Isaac Asimov. Originally Asimov kept his two fictional universes, Empire and Robots, entirely separate from each other. Eventually in his later Foundation novels he tied them all together and from then on most of his works were in that one shared universe.

Like FRBR's entity-relationship groups, phylogenetic trees have a standard visualization called a cladogram. "A cladogram is one kind of phylogenetic tree, a common ancestry tree." (Wiley, 6) Each phylogenetic tree typically only displays one relationship type at a time.

Cladograms are temporally oriented. For example, a cladogram could be set up such that the left of the image is the past and approaches the present to the right. The branch points of the tree represent the development of new traits, allowing one species to be distinguished from another. The individual species are located at the ends of the tree, in following related example this is near the top of the image. An important piece of information communicated by a cladogram is that the more recent the point of divergence, the closer to the top of the graph the branch point will be. It is important to remember that the branching denotes temporal sequence, not necessarily similarity, though the two are often highly correlated in biological species.

The true utility of cladograms and the phylogenetic model becomes apparent when bibliographic examples are used. Figure 2.1 below is an example of a cladogram depicting a selection of derivative works spawning from Jane Austen's *Pride and Prejudice*.
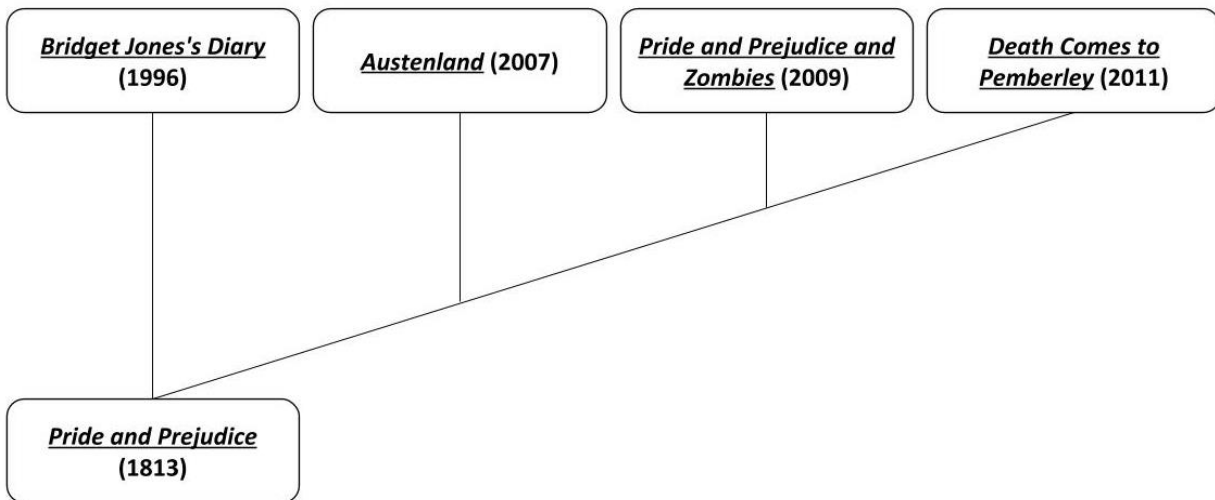

Figure 2.1 (Campbell and Mayhew, 2017)

In this example several texts are displayed as being descended from *Pride and Prejudice*, most likely through the 'adaptation' relationship. Once again, it is important to remember that the branching denotes temporal sequence, not necessarily similarity. *Death Comes to Pemberley* was created after *Bridget Jones's Diary*, but other than both being adaptations of *Pride and Prejudice* they need not have anything else in common.

Since each work has a date of creation, and the original has not gone extinct, each derivation can be modeled as branching off from the original at that time. This model has a simple elegance that is intuitive. However, this apparent simplicity belies its capacity to contain a great variety of relationships.

Of course, the temporality of the tree need not be limited simply to creation dates. Publishing dates, in story chronology, and even reading order by individuals can all potentially be useful to other catalogue users. Additionally, there will be cases when the dates are unknown, only approximately known, or when two texts are developed simultaneously, such as the movie and book version of Arthur C. Clarke's *2001:A Space Odyssey.* Each of these situations poses difficulties, but have have analogies in biology to draw upon.

Users, including cataloguers, publishers, and the public, would, at various stages, be able to add relationships they found useful between entities in the catalogue. By allowing users to create their own types of relationships and then apply them to entities in the catalogue the potential record of connections between texts will greatly expand. As mentioned above, clades can exist at any scale, and for any relationship type. A text that has no ancestors and no descendents for a particular relationship could exist in a clade on its own, though it would likely be difficult to identify such a text in the real world.

Another strength of this model is that the added connections are likely to be of value to more than just the person who added them; people share interests after all. Additionally this means that not all the added connections need to be permanent. 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.

Finally, another advantage of this model is that it has the potential for new types of conflict detection. Tied once again to the temporal nature of phylogenetics, certain logical impossibilities become easy to detect. For example, consider three texts, A, B and C where the catalogue records that A is descended from B, B from C, and C from A. Because this is impossible any system based on phylogenetics would be able to alert the cataloguers that something is inaccurate.

Part 4: Implementation

The devil, of course, is in the details. This paper lays out a three-phase plan for the general implementation of a phylogenetic catalogue. The first stage involves the creation of Group 4 for the use of cataloguers. The second stage opens up Group 4 to the public. Finally the third stage uses the connections created in stages one and two as the basis of a Big Data project to generate new connections.

Part 4a: Cataloguers

This first step of implementation rests on the cataloguers. By focusing efforts on using the existing relationship designator list in FRBR basic phylogenetic trees could be generated. The primary advantage of this is that there are no major changes required and the system remains under the purview of the cataloguers.

The main disadvantage of this plan is that cataloguers are already very busy with the standard workload. Extra work would make it unlikely that cataloguers would have time to go back and designate relationships of older works as well. The second issue is that the FRBR supplied relationship designators are not very expansive. In *From Complex Reality to Formal Description: Bibliographic Relationships and Problems of Operationalization in RDA* Henrik

Wallheim argues similarly, stating that "by means of the recorded relationships, the user should be able not only to find resources which are related to a given resource, but also to understand the connections that exist between different resources." (Wallheim, 487)

Wallheim believes that an ongoing cause of this is a focus on machine readable relationships at the expense of recording relationships that users may find meaningful. This means that the types of trees that could be generated would not be as interesting as the would otherwise be. However, adding in additional relationship links based on the phylogenetic model would just compound the original problem. To be as useful as possible cataloguers would likely need to massively update every record by hand.

Part 4b: Crowdsourcing

Once the cataloguers provide a proof of concept the Group 4 can be useful and descriptive, it would be possible to create a section on bibliographic records that would be open to public input. At that point in time these records would be interesting to people with specialized knowledge, such as professionals, enthusiasts and fandoms. The Victorian Web (http://www.victorianweb.org/), for example, would be able to make their efforts more broadly available. By tagging records with the relationships they know best they will all be contributing towards a global linked data project.

Some features like this already exist, such as the ability to add categories to Wikipedia pages or tagging on Facebook. Bringing these features into cataloging has the chance of reducing the burden put upon the cataloguers. There are ample enthusiasts of books, movies and television media that would be willing to add their knowledge to catalogues such as GoodReads or IMDB. There are also many professionals that would gladly contribute interesting connections they discovered for the benefit of themselves and their colleges.

These relationship tags would naturally take the form of Subject-Predicate-Object triples, the same as used in the Web Ontology Language (OWL), the basis of much of the semantic web. This would allow the computer to automatically generate relationship maps, even keeping up with the latest additions and updates.

Part 4c: Big Data

Much farther down the road the database of relationships generated by the crowd sourced project could acts as a training set for a Big Data project. By comparing entered relationships new relationships could be suggesting. Going farther, the Google Books corpus contains vast amounts world's published texts and would be an obvious candidate for this sort of automated relationship generation. Failing that there are several repositories of public domain works, such as Project Gutenberg, that may be willing to benefit from this.

At this point in time the technical requirements for this portion of the project are not known. The longer the crowdsourced linked data is generated the better the big data results are likely to be, but also the more resources they are likely to need in order to be generated. Only time will tell on this point.

Conclusion:

In conclusion, proliferation of multiple media is creating a greater demand for methods of deriving relationships of lineage across these various media. 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. A phylogenetic bibliographic model would provide a flexible, agile and fruitful method for accurately capturing the reality of the creation and use of texts.

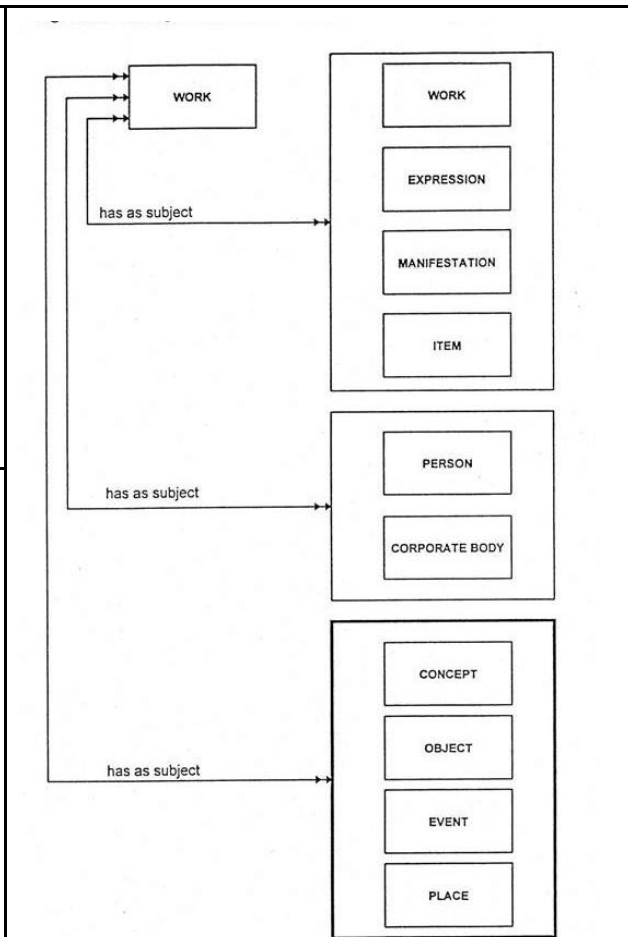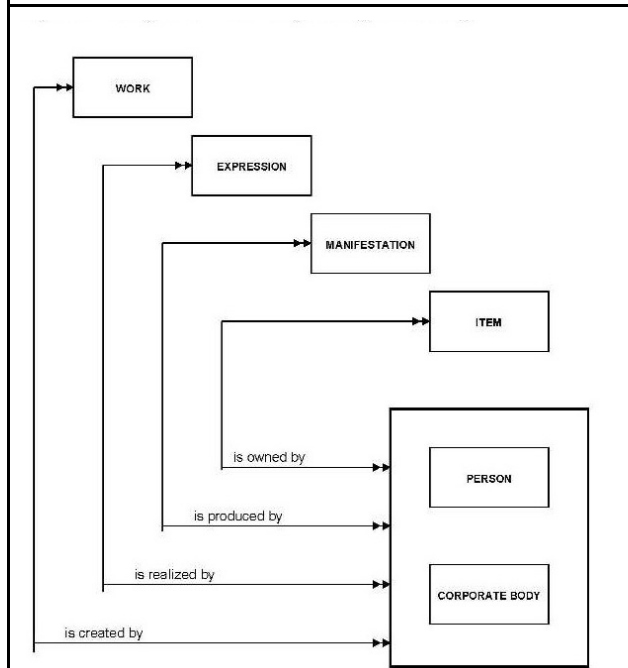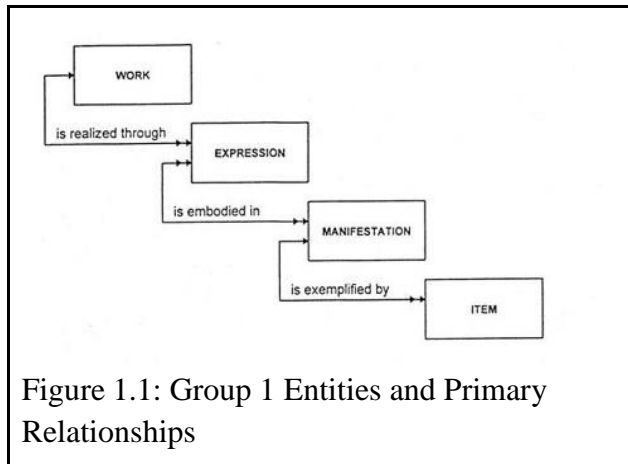Appendix A: *Group 1, 2, and 3 Entities and Relationships.*

Figure 1.1: Group 1 Entities and Primary Relationships

Figure 1.2: Group 2 Entities and "Responsibility" Relationships

Figure 1.3: Group 3 Entities and "Subject" Relationships

Figure 1.1, 1.2, and 1.3 (IFLA, 1998)

References (APA)

Campbell, D. G. and Mayhew, A. (2017). Phylogenetics as a replacement model for FRBR: Applications and implementation. *ISKO conference*. Champaign Illinois.

Denton, W. (2007). FRBR and the history of cataloguing. In *Understanding FRBR: What it is and how it will affect our retrieval tools.* Ed. Arlene Taylor. 35-58

IFLA. (1998). Functional requirements for bibliographic records. *International Federation of Library Associations and Institutions*. www.ifla.org

Le Boeuf, P. (2005). FRBR: Hype or cure-all? Introduction. *Cataloging & Classification Quarterly*, *39*(3-4), 1-13

Letunic, I., & Bork, P. (2011). Interactive tree of life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, *39*, W475–W478. http://doi.org/10.1093/nar/gkr201

Madison, O. (2000). The IFLA Functional Requirements for Bibliographic Records: International Standards for Universal Bibliographic Control. *Library Resources and Technical Services*. 44(3), 153–159

O'Brien, M. & Lyman R. (2003). *Cladistics and archaeology*. The University of Utah Press.

Smiraglia, R. (2007). Bibliographic families and superworks. In *Understanding FRBR: What it is and how it will affect our retrieval tools.* Ed. Arlene Taylor. 73-86

Wallheim, H. (2016). From complex reality to formal description: Bibliographic relationships and problems of operationalization in RDA. *Cataloging & Classification Quarterly*. 1-21

Wiley, E. O. et al. (1991). *The compleat cladist: A primer of phylogenetic procedures*. University of Kansas Museum of Natural History. www.amnh.org