Alexander H. Harding. Comic Book Metadata and Database Design. A Master's Paper for the M.S. in IS degree. April, 2014. 45 pages. Advisor: Brad Hemminger

Current systems of metadata storage for American superhero comic books in the digital realm are limited in scope and capabilities. This study reviews eight systems of comic book metadata storage to determine the importance of metadata included and propose a more optimal system design in a relational database format for the storage of digital comic book information. As part of this work, a proposal for standardization of the comic book metadata fields is generated. Further development will be publicly maintained and shared through a Github repository.

Headings:

Comic Books, Strips, Etc – History & Criticism

Database Design

Metadata

Information Storage & Retrieval Systems

Popular Literature – History & Criticism

COMIC BOOK METADATA AND DATABASE DESIGN

by
Alexander H Harding

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2014

Approved by

_____

Brad Hemminger

Table of Contents

Introduction 2
Statement of Purpose 9
Systems Review 10
- Markup Languages 11
- Personal File Managers 12
- Comics Databases 13
- Retailers 15
- Comparison 16

Methodology 17
Results 18
Metadata Descriptions and Proposed Standards 21
System Design 31
System Description 32
System Design: Entity Relationship Diagram 35
Discussion 36
- Innovations 36
- Weaknesses 39

Bibliography 41

Introduction

Since its inception, the American comic book industry's stable of heroes and universes has grown exponentially. Heroes of varying backgrounds and abilities came to co-exist within universes created by the collection of heroes under the banners of the publishers they belonged to. As the industry expanded, giving multiple heroes multiple titles published on a monthly basis year-round, an editorial crisis came to light. Given the expansive nature of the comic book storytelling medium, in which characters' stories span decades of chronology and hundreds and thousands of single issues, there comes a point in which there becomes too much to keep track of without retreading ideas and contradicting previous events. Characters' origin stories would change with alarming frequency, and significant events in the chronology of characters would be included or ignored as was most convenient to the author currently in charge of a character's story. This led to an inconsistent and often confusing experience for readers both new and dedicated, as they could never truly be sure of who, what or where they were in a particular story.

After fifty years of chaotic storytelling, DC Comics, one of the two major publishers of comic books, produced *Crisis on Infinite Earths*. *Crisis* took the established notion of the DC Universe as a "multi-verse" containing an infinite number of Earths, and, through a sequence of events, merged the majority of universes into a single, unified

continuity. To date, the DC Universe is divided into "pre-crisis" and "post-crisis" continuity, and the idea of a universal continuity has been in place since *Crisis on Infinite Earth*'s release in 1985. This is further complicated in numerous universe-wide alterations to continuity (often referred to as "crises"), and the 2011 cancellation of all series and subsequent rebooting for their "New 52" initiative.

Although a unified continuity is not a new or innovative concept, the execution of the concept in the hands of the comic book industry is novel due to the unprecedented size of the corpus. Typically, monthly comic books are published as series, which focus on a singular character or singular team of characters. DC Comics currently publishes anywhere between 40-60 series each month, and often limited series and promotional tie-ins are added to this monthly list as well. Each issue is typically between 24 and 36 pages, which makes for a tremendous amount of story occurring simultaneously on a monthly basis. Multiply that by the 75 year history of the company (or, for the sake of convenience, the 28 years since *Crisis on Infinite Earths*), and the amount of literary story is truly staggering.

Similar to the common television structure of episodes containing portions of story that are completed over the course of a season, comic book issues often contain small portions of story that are more cohesively read as a "story arc." Some arcs are told over a single series, however it is also common practice to employ what is called a "crossover" event, in which a story is told over multiple issues of multiple series. This is commonly done in order to facilitate interactions between characters and teams, as well as synergistically cross-promote series for higher sales figures (ie: if a reader reads one series, a crossover event would necessitate purchasing two or more series in a single

month, which may lead the reader to then continue purchasing those other series on a monthly basis). Regardless of series, the traditional procedure is to collect story arcs into larger books akin in size to graphic novels, often referred to as "trades," in reference to the physical publication of a "trade paperback." From a literary standpoint, a story is contained in these trades significantly more coherently than by a single issue. Additionally, these trades are published in perpetuity, whereas single issues are published only once during the month they are new. So for a reader to read a story without following on a monthly subscription, common wisdom is to purchase trades from any retailer (local comics stores, big box retailers like Barnes & Noble or Amazon).

Given the enormity of the universe, a major hurdle for readers new to comic books is the sheer amount of history present in the universe. Browsing any comic book forum or online community, the history of individual characters has grown to include tremendous lineages.

As comics publishers release an increasing number of issues and stories, comic book collection has become a major hobbyist market. The collectability of certain issues and creation of storylines and variant covers has led to a significant amount of metadata existing on a per-issue level, as certain comics with certain attributes place themselves differently in comics collection. Furthermore, as the comics industry transitions into the digital age, numerous attempts have been made to create metadata schemas for the cataloguing of digital comics, some going as far as to attempt to mark up the physical page into a machine-readable format. Various implementations for sorting digital collections exist, and a number of significant comics databases exist on the public internet dedicated to the collection and documentation of comics both in their physical

and digital manifestations. Some go as far as to include characters and story arcs in their relational model, allowing for research into character appearances across titles chronologically and encounters therein. However, many of these implementations are lacking in scope of completeness, and leave out crucial details either by design or as a result of the limited scope of their design prior to implementation.

The comics storytelling medium is extremely fast paced, and almost every story, be it a single issue or an arc covering a longer span of time, has a significant impact on the universe as a whole. Characters die and are resurrected, are crippled and changed, have minds erased or controlled, and teams are created and broken on an almost monthly basis. Given that malleable nature, a timeline of character appearances in chronological order of appearance does little good to a researcher. Knowing which characters appear in a certain issue is certainly important data, but knowing who each character is in relation to other characters and the aliases they take as heroes or villains at that point in history is equally valuable data.

Given the relational nature of this data, further research on current comics database implementation for analysis on strengths and weaknesses is an important topic to the comics industry and readership. Discovering a relational model that allows for recordable metadata corresponding to storytelling techniques and plot points could be key in the visualization of data regarding characters' arcs and chronologies in their decades of existence in the medium.

As the comics industry transitions from paper books to digital sales, new possibilities arise from the use of story-centric metadata. Predictive recommendations based on reading history, automatically curated collections based on character

appearances and even data-driven storytelling are all made possible by the inclusion of literary metadata in an ordinary database system. Similar to Netflix's system of categorical tagging (Madrigal, 2014), a comic book system of plot elements would enable predictive ratings and recommendations, and could even be used to inform content creation.

In recent years, the comic book has become increasingly popular for academic study. There is an abundance of literature concerning the creation of collections of comics and graphic novels for use in public and academic libraries, yet there is a noted dearth of any academic work focusing on comics either digitally or from the perspective of innate organization. Most academic literature focuses on the importance of graphic novels as a medium in a popular fiction collection, or the ways in which popular titles can be obtained to round out a well-curated collection. However, given the popular nature of comic books, the public internet is filled with opinions and research ranging from personal blogs to pieces authored by industry professionals and even software designed for the organization of private collections of digital comics. The phenomenon of popular opinion eclipsing academic research is fairly common in literature reviews of library-focused research papers; Catherine Clements quotes Will Savage's notion that histories of the medium are largely "popular" and not "academic," and in their own structure "contain no known documentary apparatus to support them (Clements, 26). Unfortunately for this study, the rising trend of research into comic book literature focuses on areas of research having very little overlap with the innate organizational characteristics of superhero comics themselves, as they focus on existing trade paperback comicbooks, which are then catalogued in library systems. Similarly, almost all attempts

at cataloguing comics in the digital realm focus not on the metadata inherent to comics as a storytelling medium, but instead are intent on cataloguing the physical properties of a comic even in digital form. Given the relational nature of the comic book storytelling, a wealth of connectivity calls for an organizational schema by which comics can be related to one another. Comic book universes "weave together a cohesive, metatextual tapestry" (Friedenthal, 2012), which, if properly tracked using modern digital informational organization schemas, could be invaluable to analyzing the medium. The idea of a "universal continuity," in which characters interact via crossover events and guest appearances, guarantees a consistency from one title to the next. "Significant changes or occurrences in any single book that takes place within the DC Universe are, by the strictures of this continuity, reflected in every other DC Universe title. Thus, a simple, working definition of this kind of continuity could be as follows: Continuity in a comic book superhero universe is the meta-narrative created out of the sum total of meetings, relationships, battles, births, deaths, and other twists of plot and characterizations that have taken place within that universe" (Friedenthal, 2012). By using these connections to create a schema of inherent connectivity, a database of comic book issues could be sorted and analyzed in entirely new fashions, discovering trends and timelines for charting character continuity while also offering the physical sort features regarding publishing expected of a comics database. A system involving character tracking and plot element storage could also be crucial in Netflix-like categorical ratings and suggestions, which in a digital retail environment could drive sales and properly cross-promote products. However, it seems as if research on the medium is firmly rooted in the past, and finding future-forward projects and research has become increasingly difficult. This may stem

from the industry's nature, in that physical comic book collection has come to be as important as the stories they contain. Frederick Wright's notion that "For many comic book readers and collectors, the idea that a comic book can exist other than in print may be difficult to accept" (Wright, 2008) is validated by the lack of research, either public or academic, concerning comics in any sense other than the organization of comics by purely physical characteristics. Similarly, while many systems exist as repositories of metadata for comic books, many are incomplete and lacking in key metadata fields. Digital retailers present very little metadata about the comic books they sell, which makes navigating their inventories often difficult and confusing for new readers. An aggregation of metadata fields across a number of different systems could prove to be invaluable in informing the design of a more thorough system.

Statement of Purpose

Given the problems discussed in the introduction, this study aims to complete four goals:

1. Review and critique current systems of digital comic book metadata storage
2. Analyze the metadata storage fields present in the systems to determine the frequency of their appearance.
3. Propose standards of definition for the metadata fields described.
4. Use goals 1-3 to inform the design of an optimal relational database model including new innovations based on the perceived weaknesses of the systems reviewed.

Subsequent discussion of the results of these goals will follow after the design of the proposed system and definitions of metadata standards, as well as discussion of new innovations in the system and weaknesses of the proposal and areas for further research.

Systems Review

Many databases and software packages exist both in open-source and commercial implementations aimed at cataloguing comics. Although all of these exist in the digital realm, a closer examination finds a significant bias towards the digital record keeping of physical comic books.

Digital comic book systems can be broken down into four rough categories: markup languages, personal file managers, databases and retailers. Markup languages aim to use XML markup to describe a physical comic book, turning the entirety of the content of a comic into a machine-readable format. Personal file managers attempt to provide a metadata system for managing personal file collections, typically comprised of DRM-free issues (or, unfortunately, pirated material). Databases attempt to act as thorough repositories of comic book information, and retailers provide a space for the sale of digital comics and limited comics exploration.

Both the markup languages and personal file managers base their systems on XML, either by creating an entire document representation in the markup languages or by creating small XML metadata files included in the comic book archive files used in the personal file managers. XML is an excellent language for the storage of metadata, as the language is designed to store information in a human- and machine-readable format for easy data processing. It is also easily validated to conform to a custom written template (XSD), which can allow for standardization across the formats.

Database systems use a relational database format, typically MySQL, to store the data in a much more versatile format. Most databases include a frontend interface using HTML/CSS and PHP to connect to the database and display or edit information. This is the eventual design goal, and will be the format of the product of this research.

Retailers' systems are not made publicly available, and therefore all research involving retailers' systems is derived entirely from the frontend interface. Retail systems likely use an underpinning of traditional relational databases, but are focused more on content delivery and purchasing than storing thorough information. However, more complete information could allow for easier discovery of new comics for a user to purchase.

Markup Languages

John A. Walsh's recent implementation, *Comic Book Markup Language* (CBML) is an attempt to mark up the physical comic into an XML format for machine readability. CBML concerns itself with completely digitizing the written page, highlighting "the boon-ness and bookishness of these documents, their material properties and bibliographic characteristics. Graphic narratives typically manifest as "books," stapled or otherwise bound leaves, perhaps thirty-six pages, with an interesting and complex structure, incorporating the graphic narrative –the sequential art and text or 'comics' content – alongside a rich assortment of paratexts: advertisements, fan mail and so on" (Walsh, 2012). CBML is an incredibly fascinating and valuable resource, yet the focus is again on the physical texts themselves. A similar project, *Advanced Comic Book Format* (ACBF) also focuses on marking up the full text of a comic and focusing on representing print comics in the digital medium more so than the idea of the comic's inherent sort properties. Both CBML and ACBF include some story-related metadata such as general

"characters" fields, but mostly again focus on the physical. Both offer similar schemas (with CBML being significantly more thorough) for organizing the text and images on a page, but other than general title and publisher information do not offer any form of story sorting.

Personal File Managers

The most popular, open-source filetype for DRM-free digital comics is that of the comic book archive. Available in .cb7, .cba, .cbr, .cbt, and .cbz (in which the last digit corresponds to a different method of compression), a comic book archive is essentially a compressed archive of digital images intended to be viewed in sequence (Wikipedia). Many implementations of this standard allow for an XML file to be added into this archive to track comics metadata, some with varying schemas. Comicbooklover is an application developed by BitCartel, available at http://www.bitcartel.com/comicbooklover/ . Comicbooklover attempts to create an iTunes-like interface for organizing digital comics files, and includes a variety of available sort fields. Unfortunately, because the metadata schema is not standardized, organization done within comicbooklover is only compatible with other comicbooklover products (such as their iOS app), and does not translate into other viewers and readers. Comicbooklover's metadata schema is fairly comprehensive, and has an interesting system of user input to make up for its weaknesses. Comicbooklover includes a system of user tagging, which allows for creative users to form their own search indexes. Although the system does not track characters or story arcs, one could add a "Batman" or "Dark Knight Returns" tag to issues involving the character or story, and use that as a sort field in later discovery. Another attempt at creating a standardized comics XML structure is

CoMet, by DenVog, available at http://www.denvog.com/comet . DenVog accurately summarizes the problem with non-standard formats in their specification documents: "Currently programs, web sites, databases, and applications all use their own naming schema for describing comic books. This makes it difficult, particularly for newcomers, to get comics into software programs or web sites they wish to use to track their collections. It becomes even more complicated when users wish to share or move information about their collections across platforms (i.e. PC to iPad), between software programs, or packaged as comic book archives (typically CBR or CBZ files)" (DenVog). CoMet does a great job of creating an ease-of-use XML schema with a fair amount of metadata for describing a comic. It however also does not catalogue any story data other than being able to include character names. Both comicbooklover and CoMet both have levels of completeness acceptable for the cataloguing of simple digital comics, yet both are focused on the single-issue level. As metadata schemas, the intent and practice of these implementations is to allow the user to add metadata on a per-issue basis, which severely limits any data generation other than basic sorting. This is inherent to metadata as a concept; for relational data, a traditional database is much better suited to the task.

Comics Databases

General comic book databases exist as well, some with exhaustive completeness in terms of stored metadata. However, these databases again focus on providing a digital platform for reference to print comics and their collection. *The Grand Comics Database* (GCD), available at http://www.comics.org is a major online database for comics collection. GCD makes their documentation freely available, and as such a physical schema whitepaper is available for examination. GCD currently contains 935,000 single issues. *The Comic*

Book Database, accessible at http://www.comicbookdb.com, is another prominent database, containing almost 300,000 single issues. Both of these databases go to great length to add copious amounts of metadata on multiple levels, and accomplish a much higher degree of completeness than the markup languages previously discussed. Both databases include all publishing information, ranging from author and ISBN to printing type and paper stock. Both also relationally model characters and story arcs separate from the issues themselves, giving them the flexibility to see what characters were in what issues and arcs, who characters are affiliated with by group, and even show a chronological list of issues appeared in by character. This is a significantly more comprehensive story view than the attempts of markup languages like CBML and ACBF, which are much more focused on cataloguing the page than the story. Databases like CBD and GCD are fantastic tools for reference on characters, but are very much focused on physical comic book collections. Both databases have "my collection" implementations, which function as a digital repository for the cataloguing of your own personal collection of comics. This collection is done both in retrospect and in perpetuity, as they offer "pull-list" like functionality, where users can note which series they subscribe to and have the database automatically add new issues to their personal collections, much as pull-lists work in a local comics shop. A definite issue on both databases is the lack of clear lines of query, as user search is restricted to basic keyword matching and full entity views. A user could search for "Batman" (although each database has multiple Batman listings for multiple Batmen), and a user could search for "Joker," but a user can't search for comics involving both Batman and the Joker. However, given the relational nature of these databases, these lines of query are

absolutely possible, if not immediately presented to the user. Both databases also attempt to categorize issues both by the series they belong to and the story arc they belong to. This can be very helpful in organizing a story arc that is not collected trade paperback form. However, because both databases are curated by their users, data varied across issues and story arcs, with vacillating tagging of characters and arcs to issues in the systems.

Retailers

In the United States, the largest digital comics retailer is Comixology. Both Marvel Comics and DC Comics, as well as many of the larger independent labels, sell their digital catalogues and new monthly issues through Comixology. Comixology has a web app, and standalone apps for iOS, Android, Windows 8 and the Kindle Fire. Comixology has a basic search function, and a similarly typical amount of metadata stored per issue. Comixology also stores story arcs in a similar fashion to many of the databases previously discussed, which is inherently more useful when reading actual comics as the application is able to tell you what's next in the series and what's next in the story arc upon completion of an issue. Otherwise, the metadata used in Comixology's user-facing collections adheres to similar standards of creator-based metadata.

Another major repository of digital comics in Marvel Comics' Marvel Unlimited. Marvel Unlimited is a subscription service that offers the entirety of Marvel's back catalogue for a monthly or yearly fee. Like comixology, Marvel Unlimited offers a web app or applications for the major tablet operating systems, and stores the same set of standard metadata per issue. Marvel Unlimited also supports story arc functionality when appropriate.

Comparison

In order to better inform the design of a new system, a comparison of captured metadata across multiple implementations is essential. Although most of the systems described do not encompass the fully relational model in their user-facing implementations, they all share a similar set of overlapping metadata fields that would serve as the backbone for any future comics database. Comparing both the systems described and creating a full list of metadata structures appearing across all systems, then analyzing this data both by system type and across the systems in total would provide an understanding key to informing future systems design. Additionally, detailed descriptions of the metadata fields themselves and what they aim to capture is important to fuller understanding of the purposes of each and the merits of their inclusion in a system. This is also an opportunity to propose new standards for the metadata fields present in the system design, which can prove beneficial for future designers to better understand the fields that they include in their designs.

Methodology

The eight systems reviewed and selected were examined to better catalogue the metadata included in each. If further information such as whitepapers or schemas were available, these were used for analysis; however, only two offered this information. Comic book markup language (Walsh, 2012) and the Grand Comics Database (Grand Comics Database, 2013) offered these deeper looks at their systems, but all other systems were analyzed by registering for or installing the implementations and examining the available metadata fields either by examining user-facing data or the tools available for editing or adding entries.

A spreadsheet was created, with columns for each system. As new metadata fields were discovered, rows were created for the discovered fields. A system of binary weighting was used across all eight implementations, where a column received either a "1" or a "0" in a metadata field row, indicating that it either did or did not include that metadata, respectively. After all eight systems were analyzed, each field of metadata was itself reviewed to determine the meaning of the metadata field. Similar fields or fields with similar purpose and intent were merged into one row, with exceptions, as described in the "findings" section of this paper.

The data was then filtered by system type, rather than by individual systems. This allowed for a greater understanding of what metadata fields are traditionally included where, and which are too purpose-built to be included in further systems design.

Results

After further analysis of the systems described in the literature review, twenty-eight types of metadata were found across the eight systems. Some of these data types were purpose-oriented, and would not warrant inclusion in a database system, such as full document markup and volume information.

Second, it became readily apparent that the two full-document markup systems, while interesting for discussion, contained little to no metadata. These systems were concerned with describing the content of a comic more so than the context within which a comic book fits, and in their purpose-built systems a significant portion of metadata was not included.

To that end, the decision was made to exclude the two from defining what metadata structures were considered significant enough to be used in informing system design. Now given a baseline total of six systems, any metadata appearing in three or more systems (50% of total or higher) was considered significant in informing system design. Similarly, it was important to note which metadata types appeared in each of the three types of systems surveyed. As the intended goal is informing the design of a database system as a repository of information, some information included in personal file managers may not be significant. Additionally, identifying metadata present in databases considered key for search yet not appearing in retailer systems can be used to

analyze flaws in retailer metadata presentation as more thorough information can be greatly beneficial to customers attempting to navigate the enormity of the collection available for purchase. By nature, a database environment attempts to provide the most specific level of granularity and the most thorough level of completeness, and this is reflected in the data. After separating the data to reflect which appears in which type of system, the new data was used to create figure 1, a Venn diagram depicting the systems in which each metadata field appears. Eleven fields appeared in all three types of systems, two only in databases, three only in file managers, and none were exclusive to retailers. One appeared in only file managers and retailers, four were exclusive to file managers and databases, and five were shared between only databases and retailers. The metadata fields that only appeared in a file manager system type were typically purpose-driven, and likely will not be important to future system design. The next section

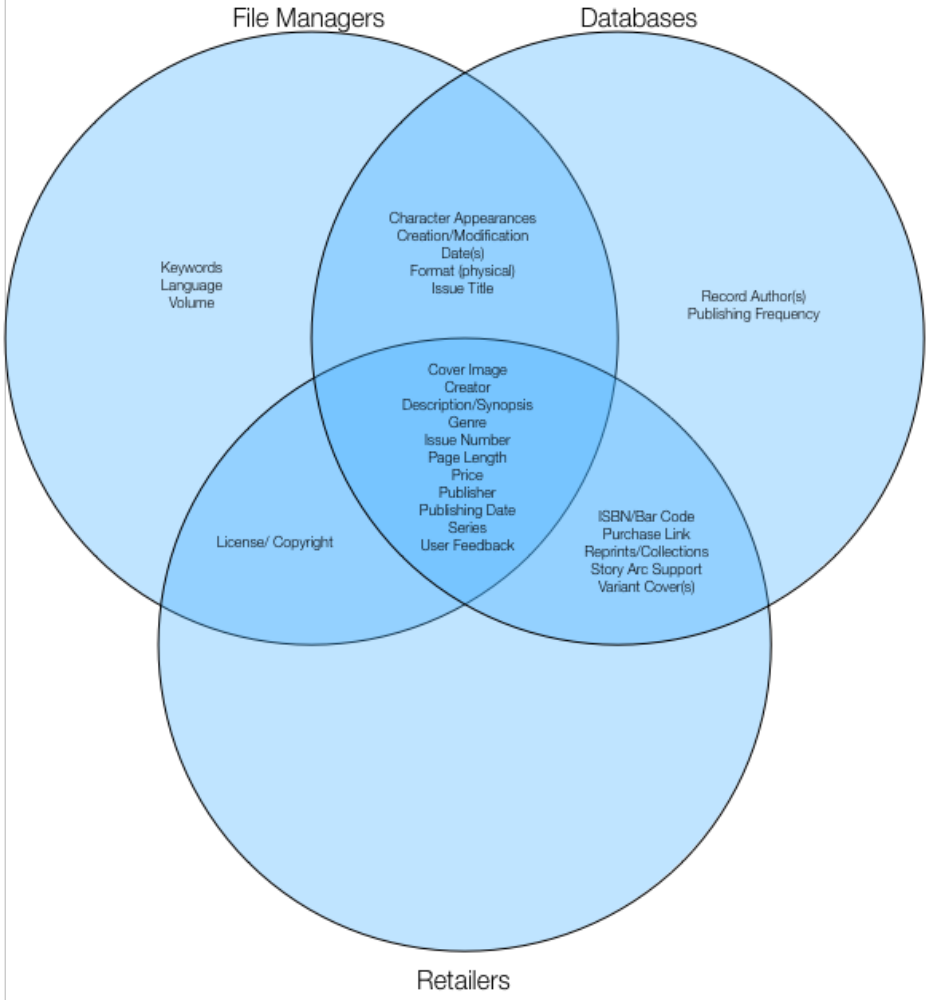


Figure 1: Metadata Inclusion by System Type

provides fuller explanation of each type of metadata, and includes discussions when necessary analyzing the necessity of inclusion of a particular piece of metadata.

	CoMet	Comicbooklover	CBML*	ACBF*	GCD	CBD	Comixology	Marvel Unlimited
Character Appearances	X			X	X	X		
Cover Image	X	X		X	X	X	X	X
Creator	X	X		X	X	X	X	X
Description/Synopsis	X	X		X	X	X	X	X
Record Author(s)				X	X	X		
Creation/Modification Date(s)		X		X	X	X		
Format (physical)	X				X	X		
Full Document Markup			X	X				
Genre	X	X		X	X		X	
ISBN/Bar Code				X	X			X
Issue Number	X	X			X	X	X	X
Issue Title	X	X		X	X	X		
Keywords		X		X				
Language	X	X		X				
License/ Copyright	X			X			X	
Page Length	X				X	X	X	X
Price	X				X	X	X	X
Publisher	X	X		X	X	X	X	X
Publisher City				X				
Publishing Date	X	X		X	X	X	X	X
Publishing Frequency					X			
Purchase Link						X	X	X
Reprints/Collections					X	X		
Series	X	X			X	X	X	X
Story Arc Support					X	X	X	X
User Feedback	X	X				X	X	X
Variant Cover(s)					X	X	X	X
Volume	X	X						

*Markup language, not included in analysis
BOLD indicates appearance in three or more systems

Metadata Descriptions and Proposed Standards

Character Appearances

Appears in 3 systems

Currently implemented differently across systems. Proposed standard would include the ability to store characters (by real name) as well as their aliases, and organizations that they belong to and appear in. Appearances of characters/aliases/organizations should be chronicled by issue, and can use the issue dates to track chronology.

Cover Image

Appears in 6 systems

An image file containing the cover of an issue. This can be stored in the system or separately and referred to via file path/URL. A thumbnail can be displayed on the main interface, which would act as a link to the full image.

Creator

Appears in 6 systems

Creators are any persons involved with the creation of the comic. Typically the names of these people are listed on the cover, indica and/or splash page. Creators can be writers, pencillers, inkers, colorists, editors and many other roles. The system should allow for any number of creators to fill any number of roles in an issue.

Description/Synopsis

Appears in 6 systems

A small blurb describing the issue or collected edition. Proposed system should use the publisher provided description provided in the monthly solicitation for the issue.

Record Author

Appears in 2 systems

An entry in the system denoting which user edited a tuple in the database, as well as the change made.

Creation/Modification Date

Appears in 3 systems

An extension of the record author. Records a timestamp along with the user information.

Format (physical)

Appears in 3 systems

This field records the physical format of the comic book issue or collected edition as it exists in the physical realm. This information is not particularly important for monthly issues, but determines whether a collected edition is a softcover or hardcover edition.

Full Document Markup

Appears in 2 systems

This field refers to the full markup of a document, as specified in the Advanced Comic Book Format and Comic Book Markup Language. This is not an essential field for a database system or even general metadata, as it is its own purpose-built type of metadata system for storing comic pages, and it will not be used in the proposed system.

Genre

Appears in 4 systems

A field recording the genre in which an issue is included. Often this can just be “superhero,” but many comics fall under the headings of other genres such as “horror,” “romance,” “supernatural” and many others. This field in the proposed system should be a single word or short phrase describing the genre.

ISBN/Bar Code

Appears in 2 systems

This field should capture the ISBN of collected editions, or the barcode numbers on the front of an issue. The barcode information on an issue can sometimes be indicative of other factors (variant, printing), but many issues do not include a bar code.

Issue Number

Appears in 6 systems

The issue number is one of the major pieces of metadata in identifying an issue. The issue number is typically just an integer, although sometimes issues have a special qualifying feature that may require a string (such as “director’s

cut"). The issue number, in combination with the series, is typically capable of forming a composite key, although a separate ID is still recommended.

Issue Title

Appears in 4 systems

Often an issue will have a title or subtitle naming the issue. This title is sometimes just the name of an arc and the number of the issue within the arc, and sometimes is just a title on its own. This field can also store any other text on the front of an issue if there is no title.

Keywords

Appears in 1 system

Two of the systems allowed for user tagging of keywords to assist search within the system. These two systems were ACBF and Comicbooklover, and is more intended for personal collections than for online databases. The keyword tagging acts as a surrogate to make up for the lack of other metadata fields, under the logic that the absence of any field can be recreated using keyword tagging. This will not be included in the proposed system design.

Language

Appears in 2 systems

A general field to denote the language that the issue is written in. Many of the systems are intended for American superhero comics, so they do not include this field. This will not be included in the proposed system, as the proposed system is intended for American superhero comics and it can be assumed that they are written in English.

License/Copyright

Appears in 2 systems

This is another aggregated field for a type of metadata that varies greatly among systems. Unfortunately, most systems did not make it clear what they intended this field to be used for, and the thoroughness of its completion throughout the systems was left wanting. This will not be included in the proposed system design.

Page Length

Appears in 5 systems

The number of pages in a particular issue, typically 24 or 36. Larger for collected editions. Stored as an integer.

Price

Appears in 5 systems

The cover price of an issue. Typically included on the bar code or top left corner of an issue. This is different from the sale price for issues on digital retailers such as comixology, but instead a reflection of the original price listed on the issue. Collected editions typically store a MSRP price on their barcode as well.

Publisher

Appears in 6 systems

The publisher responsible for the issue. Publishers should be stored as a separate entity in the system, so that browsing by publisher can be accomplished in the interface.

Publisher City

Appears in 1 system

Likely a relic of an older time, when publisher city was a common listing among comic books. This is no longer a common practice, and is not included in the majority of systems. This will not be included in the proposed system deisgn.

Publishing Date

Appears in 6 systems

This field typically refers to the cover date on a comic book issue, with some exceptions. Comic book cover dates often appear on the cover, yet sometimes appear on the indicia inside the book itself. Typically, cover dates are dated between one and three months ahead of the actual date published, and is only displayed in a month/year format. This comes from an old tradition, in which dating covers later than actual publishing can make consumers think an issue is current rather than older, and also functions as a pull date (the date on which the issue can be sent back to the publisher for a refund)(Wikipedia). For collected editions, this is just a publication date.

Publishing Frequency

Appears in 1 system

Although this field only appears in one system, it does represent a valid field of metadata. Traditionally all comics are published monthly, although some special comics have been published weekly and delays can break up the once

a month cycle. However, many new “digital-first” series are published bi-weekly or weekly, and are collected in a single physical issue monthly containing the three or four digital issues (White, 2012). This will not be stored in the proposed system, but warrants further study as to the best way to include it.

Purchase Link

Appears in 3 systems

This field is a given for digital retailers like comixology, but for other systems such as the public databases including a purchase link using an affiliate code for websites such as amazon offer a great way to monetize their system in a way other than advertisements. This will not be included in the proposed design, as it is an academic project and not for profit.

Reprints and Collections

Appears in 2 systems

Different from story arcs, this metadata field records where issues have been reprinted and/or collected in other works that would not otherwise be readily apparent from the other metadata. This includes foreign translation reprints and collection in “best of” collections and anniversary collections. This is not necessary in the proposed system, as the collections are themselves stored and issues are related to them. A thoroughly completed data set including issues and collected editions would allow this information to be derived through basic SQL, and is not a necessary field to be stored itself.

Series

Appears in 6 systems

As previously discussed in "issue number," this is a crucial component to identifying an issue. To solve the issue of multiple series having the same name (for example, most of DC has relaunched their series with the same series name but beginning with issue number 1 in 2011), series are differentiated by including the date range or start date in the title. The "New 52" Batman series is listed as "Batman (2011)" to differentiate it from the series before it. This will be stored in the system in this format, as a separate entity with its own attributes.

Story Arc Support

Appears in 4 systems

Support for including information about a story arc was limited and different across the systems it appeared in, but warrants mention as a single aspect. Some systems chose to relate issues to a story arc, and some related issues to the collected edition that encompasses the story arc (a weakness, as often many issues are included in a story arc that do not get collected in the paperback versions). However, story arc support is key to a database system for digital comics, and an improved aspect of this will be crucial to designing a thorough database.

User Feedback

Appears in 5 systems

Many systems included a form of user feedback, which varied from a rating using five stars or out of 10, or sometimes even just a comments section in which a user can make comments about an issue. Some also included a section for "reviews," in which users could write a fuller review of an issue or collected edition. Ratings systems appeared to be more frequently used than full text reviews. The proposed design will include a system of user ratings intended for a rating out of five, likely displayed as five stars in the interface.

Variant Cover(s)

Appears in 4 systems

An important aspect to physical comics collection is variant covers. Often an issue will be printed with a variant cover drawn by a different artist, or a pencils-only cover of the original (that is, without being inked and colored). These are usually manufactured with rarity in mind and function as a retailer incentive. If a retailer orders (typically) twenty issues of a given title, they'll receive one copy of a variant cover. If they order one hundred copies, they'll receive a pencils-only cover. In addition to traditional variants, often multiple printings will receive a slightly different color (typically the same cover as the first printing tinted a different hue) to differentiate. (Wikipedia). Many systems treat variant covers as a separate issue forked from the first print issue, and clone all metadata from the original to the variant. Variant covers are only important to physical issues, as digital copies traditionally include all

covers in the digital file. In the proposed system design, variants will be stored as separate issues related to the issue they are variants of.

Volume

Appears in 2 systems

The volume field is specific to collected editions, not issues. It refers to collected editions that are broken up into volumes, a common practice in collected editions. A traditional book contains between five and ten single issues (150-300 pages), and events that span more issues are broken up into multiple collected editions. Larger collected editions that encompass the entirety of events, including all issues and tie-in issues (issues not important to the main story but involved in the event) are referred to as omnibuses, and often eclipse the thousand page mark. This will not be stored in the system, as most volumetrically collected editions include the volume in the title.

System Design

Given the data acquired from the metadata survey, the twenty-two fields appearing in three or more systems (50% or greater from the six systems included for analysis) warranted inclusion in system design. With a SQL database end goal in mind, a written description was necessary to fully understand the desired capabilities of the system to be designed. The design is aimed to not only contain the metadata elements supported by the research, but also to add new enhancements to existing system designs for capturing new or different metadata structures. Some of these improvements come from the system design itself, and some come from more extensive use of SQL as a language for displaying related data.

The goal for this system is to create a MySQL database structure, and use PHP to create a front-facing and user-accessible interface for adding and editing data in the database. This system design is a foundation for that project, and further development of the project can be seen or contributed to at https://github.com/JaguarPhD/comicsdb .

The design described is an attempt to convey the metadata structures from the research captured in a relational format. As such, some metadata fields supported by the research may not appear or may appear in different ways, as some are intended to be displayed via SQL and derived information instead of duplicated storage.

System Description

Publishers have a name. Publishers publish series. A series must be published by a publisher, and a series can only be published by a single publisher.

A series has a title. A series contains issues, and the dates of these issues can be used to determine the start and end dates of a series. Issues must be in a series. Issues have an issue number, a cover price, a cover date and a cover image. An issue that is a variant cover of another issue will be related to that issue in a separate table. An issue can have many variants.

Creators have a first and last name. There are multiple roles that a creator can have (author, penciller, inker, colorist, letterer, etc.). An issue must have at least one creator, but can have any number of creators. All creators must have a role.
Story arcs contain issues, and have a title. A story arc contains many issues, and an issue can appear in more than one story arc. Story arcs contain plot elements. A plot element functions like a keyword, and can be used in many arcs (and a single arc can have many plot elements).

Story arcs and issues are both (separately) collected in collected editions. A collected edition is any printed collection (colloquially referred to as a "trade") in trade paperback, hardcover or digital form collecting a larger story. Collected editions have a title, price and cover image, as well as an ISBN.

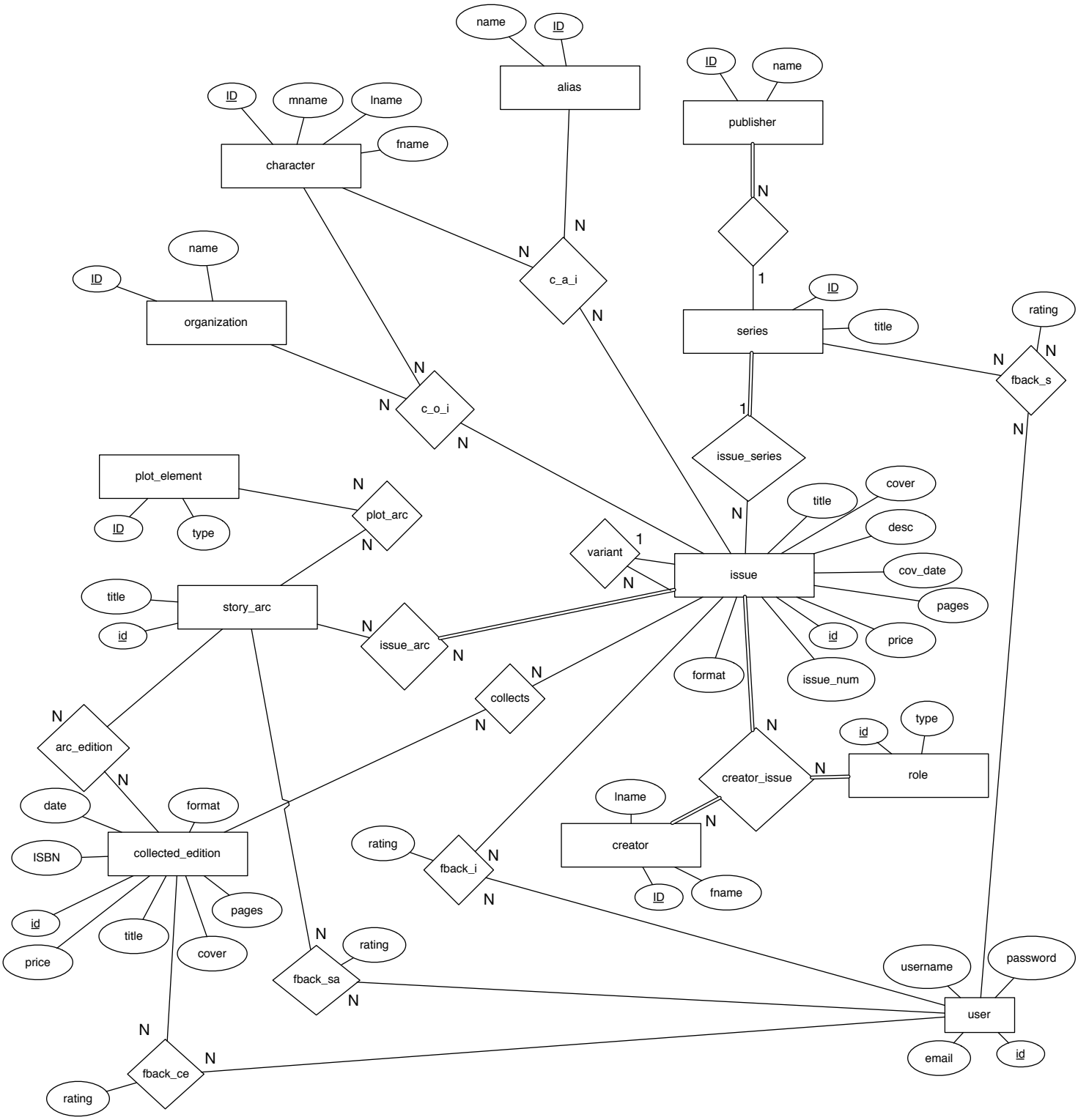
Characters appear in issues. Many characters appear in a single issue, and characters appear in many issues. A character has a first, middle and last name. If a character's real name is unknown, they gain an entry with null fields. A character often has an alias. An alias refers to the "nom de guerre" of a character, such as "Batman" for Bruce Wayne or "Superman" for Clark Kent. A character can have many aliases, and an alias can be used by many characters. When capturing a character appearance in the database, a ternary relationship between character, alias and issue is best. This way, using the cover dates of the issues, a character's relationship with their aliases can be tracked over time.

Organizations are also stored by the system. Often organizations and teams (such as the Justice League, Teen Titans and Avengers) appear in issues as well. Organization appearances in issues are stored as a ternary relationship between character, organization and issue. This ternary relationship functions similarly to that of the issue, character and alias, by tracking organizational membership across issues. This would not only allow for tracking of an organization's appearances over time, but also storing group membership over time.

Users in the system have a username, password and email address. Users are capable of giving a rating (out of ten) to issues, series, story arcs and collected editions. Each rating is saved as a relationship with the item they are rating. In this way, a user can only rate an entity once. For the sake of simplicity, creation and modification records are not saved in this system. In production, keeping record of who creates and modifies entries in the system is very beneficial, however in order to maintain readability in the system design these additional relationships (one

relationship for each entity in the database) have been omitted from the design visualization.

System Design: Entity Relationship Diagram



Discussion

Given the research, the proposed metadata fields as laid out in the relational design provides for an optimally thorough repository of comic book information. The system encompasses the entirety of the physical comic characteristics that current systems aim to capture, while embracing the digital and relational nature of the data to provide new and innovative data storage to enrich the content contained in the database. Similarly, the research indicated an extreme lack of standardization of definitions for metadata fields. Although the essential idea of certain types of metadata was shared across multiple systems, the definition of each field and the manner in which it was stored or implemented differed greatly across the different systems that contained them. The proposed standards for metadata definitions, along with the manner in which they are stored in the system design, will hopefully be influential in design of future systems and proliferation of completeness of information in digital retailers.

Innovations

Story Data: The system design includes plot elements as a table related to story arcs. Although this idea isn't explored in tremendous detail, adapting the metadata element of "keywords" to support a pre-defined list of plot elements can be beneficial for recommendations based on the system of user ratings. By examining the plot elements present in stories rated favorably in the user ratings system for a particular user, better

recommendations can be made in a retailer system to drive sales by accurately predicting a user's taste. This area of the system warrants further examination and further research in order to determine a proper list of keywords and the manner in which they relate, similar to the system implemented by Netflix (Madrigal, 2012).

Character Tracking: To this point, systems have included character tracking as an arbitrary binary relationship: a character appears in an issue or story. In the proposed system design, this is greatly expanded and improved. In the comic book world, characters change aliases relatively often, and often an alias is used by many characters. For example, Dick Grayson started as Robin before growing up and taking the name Nightwing. In the late 2000's, in the absence of Bruce Wayne (the original Batman), Grayson took the place of Batman. Additionally, Grayson is one of six characters to serve as Robin (also including Jason Todd, Tim Drake, Carrie Kelly and Stephanie Brown) , and one of four to serve as Batman (including Bruce Wayne, Jean-Paul Valley and Terry McGinnis). There are also multiple versions of each alias, as used in different types of comics. Traditionally, database systems store each version of Batman differently, titled to be specific. For example, Comicbookdb lists 51 separate characters starting with the term "Batman." This covers different versions of the character as well as different characters using the alias. In the proposed system, storing the different versions of an alias as needed as aliases separate from characters, and storing each value as a ternary relationship between character, alias and issue allows for much richer tracking of characters and aliases. Using this system, one can track a character's appearances over time using the cover dates of the issues they've appeared in, and even determine the

character's changing aliases over time, as well as how many times they've appeared as each alias. Similarly, the same can be done by alias instead of by character, and chart the alias's appearances over time and the breakdown of who appeared as that alias over time as well as how many times each character has appeared as that alias. This system is significantly more powerful than the existing system of binary relationship storage between character and issue, and allows for a large increase in clarity when determining the history of a character or alias. This system could also allow for a visualized timeline of a character, as the character's appearances and alias can be charted over time as they appear in issues by cover date. The system can also be used to track interactions between characters by determining which issues characters appear in together.

Additionally, this system replicates the same process of relations between characters and aliases for characters and organizations. Previous systems include binary relationships between characters and organizations and organizations and issues, respectively. A similar ternary relationship is in place in the proposed system design to relate characters, issues and organizations. This allows a user to not only know that an organization appears in an issue or story and that a character is a member of an organization, but also to use the cover dates to know *when* a character appears as part of an organization and when they don't.

Further thought is being put towards the end of combining these ternary relationships to a single quaternary relationship between character, alias, organization and issue. However, this presents problems in that more often than not characters appear without organization and creates a large amount of null fields stored, and reduces flexibility in terms of storing characters separate from their organizational appearances

for situations in which changes in organization or alias occur simultaneously for the benefit of reduced storage capacity.

Creator storage: The way in which storage of creators is handled also varied greatly from system to system. Some had a predefined list of roles that you could input any name to, some had predefined lists of creators (with their roles attached, stored as a table for each role) and each role/creator entry could be tagged to an issue. In the proposed system, a table of creators (first and last name) and a table of roles exist and are tagged in a ternary relationship with issues. This allows for more flexibility, as oftentimes creators fill different roles in different issues. Many writers are also artists, and sometimes even do lettering/inking/coloring themselves. The same is true of the opposite, often artists can write their own series. The proposed system allows for complete flexibility of any combination of creator and role.

Weaknesses

As previously discussed, the proposed system does not currently support recording of users along with creation and modification of entries in the database. However, this is a fairly simple concept and can be done with relative ease. All that is needed is a table for changes that is in a relationship with the users table and every respective table in the database. This table can record the user's ID, the table's ID, the field changed and a timestamp.

The "plot element" field also requires significant research on its own. Although conceptually the field has merit, the list of plot elements to be used in the system warrants further study in order to properly function in the system. If it is left to users to determine the content of this table, duplication and incorrect information will likely

become prevalent quickly. This table should be used internally and using predetermined and well researched ideas for optimal prediction of recommendation.

The proposed system also does not include purchase links or internal storage for digital comic book files. As this is an academic project, purchase and profit was not a goal of this system. However, common storage techniques or even URL links to an affiliate page on a reseller like Amazon could provide for easy monetization of the system, as users who find an issue or collected edition via the system can have a way to purchase the content from an official retailer while still providing income to the system.

Bibliography

Amatriain, X., & Basilico, J. (2012, April 6). [Web log message]. Retrieved from http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

Bell, M (2006). The Salvation of Comics: Digital Prophets and Iconoclasts.*Review of Communication, 6(1/2)*, 131-140. Doi 10.1080/15358590600763417

BitCartel. (2013, November 01). *Comicbooklover*. Retrieved from http://www.bitcartel.com/comicbooklover/

Canonical, L. (2012, April 01). *Abcf - file format specifications 1.0*. Retrieved from https://launchpad.net/acbf/trunk/1.0/ download/ACBF - File Format Specifications-1.0.zip

Canonical, L. (2012, April 01). *Abcf-xmlschema-1.0*. Retrieved from https://launchpad.net/acbf/trunk/1.0/ download/ACBF-XMLschema-1.0.zip

Christianson, J. E. (2011, November 29). [Web log message]. Retrieved from http://buquad.com/2011/11/29/the-comiquad-judging-a-book-by-its-cover/

Clements, C. (1999). *Building a comics collection: A plan for academic libraries*. (Master's thesis, University of North Carolina)Retrieved from https://cdr.lib.unc.edu/indexablecontent?id=uuid:6a2cdd05-de67-4e44-9cfa-c27ebf7f0f82&ds=DATA_FILE

ComicbookDB. (2013, November 01). *The comic book database*. Retrieved from http://comicbookdb.com/

Comic book archive. (n.d.) In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Comic_book_archive

Comixology. (2014, February 20). *Comixology: Batman (2011)* Retrieved from https://www.comixology.com/Batman-2011-1/digital-comic/13928

Cover Date. (n.d.). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Cover_date

Bibliography, Continued

DenVog LLC. (2013, November 01). *CoMet specification*. Retrieved from http://www.denvog.com/comet/comet-specification/

DenVog LLC. (2013, November 01). *CoMet about page*. Retrieved from http://www.denvog.com/comet/

Friedenthal, A. (2012). Monitoring the past: Dc comics' crisis on infinite earths and the narrativization of comic book history. *ImageTexT: Interdisciplinary Comics Studies*, *6*(2), Retrieved from http://www.english.ufl.edu/imagetext/archives/v6_2/friedenthal/

Grand Comics Database. (2013, October 09). *Grand comics database - new schema*. Retrieved from http://dev.comics.org/additional_wiki_files/new_schema.html

Grand Comics Database. (2013, November 01). *Grand comics database*. Retrieved from http://www.comics.org

Madrigal, A. C. (2014, January 02). How netflix reverse engineered hollywood. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

Marvel Comics. (2014, March 20). *Amazing spiderman #121 (1963)* Retrieved from: http://marvel.com/comics/issue/6507/amazing_spider-man_1963_121

McCloud, S. (2000). *Reinventing comics: How imagination and technology are revolutionizing an art form*. (1 ed.). William Morrow Paperbacks.

McCloud, S. (1994). *Understanding Comics: The Invisible Art*. (1 ed.). William Morrow Paperbacks.

Oliveros, J.C. (2007). VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html

Variant Cover. (n.d.). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Variant_cover

Walsh, J. (2012). Comic book markup language: An introduction and rationale. *Digital Humanities Quarterly*, *6*(1), Retrieved from http://www.digitalhumanities.org/dhq/vol/6/1/000117/000117.html

Bibliography, Continued

White, M. (2012, August 14). A look around the digital-first comics landscape. *Publishers Weekly*, Retrieved from http://www.publishersweekly.com/pw/by-topic/digital/content-and-e-books/article/53552-a-look-around-the-digital-first-comics-landscape.html

Wright, F. (2008). How can 575 comic books weigh under an ounce?: Comic book collecting in the digital age. *The Journal of Electronic Publishing*,*11*(3), doi: 10.3998/3336451.0011.304