Michael Henry Grant
Digital Preservation
December 2014

Modeling for a Database of Recordable VHS Stocks

Nearly a decade into what might be termed the "post-VHS era", it has become challenging and costly to maintain playback equipment for the format, and more and more collections are falling victim to the ravages of age and poor storage conditions. Produced with innumerable formulations by all-but-innumerable companies over the course of nearly four decades, VHS cassettes today exist in every conceivable state of preservation and deterioration. A tape may be shedding oxide, suffering from sticky shed syndrome, or growing mold, or it may be as clean and stable as if they were new. Some of these conditions are difficult to identify through simple visual inspection, and playing back an unstable tape may compromise both the content and the playback deck. Many unstable tapes can be stabilized for digital transfer, but this is costly work. You would probably spring for the treatment to preserve a newly discovered VHS recording of Abraham Lincoln delivering his second inaugural address, but you might take a pass on my circa 1984 off-air tape of *Garfield Goes Camping*. Yet many tapes are inadequately labeled, or not labeled at all; how could you *tell* a Lincoln inaugural from a Garfield special if nobody put a label on either one? You would need to play the tapes. You would need to put the tape and desk at risk to determine whether it was worth the expense to save the tape.

Of course, there are some known indicators of a tape's probable stability: If you know that a tape has been stored in a warm and/or damp climate, it is likely to be unstable. You may know that, past storage conditions being more or less equal, brand name tapes such as TDK, Fuji, Sony, and Maxell are typically safer bets than store and unfamiliar brands. You *may* be able to form certain hunches and suspicions based on a tape's approximate age. Perhaps you know that Ampex tapes are just bad news. But there are far greater levels of nuance to which batches and years of which brands are or are not likely to destroy themselves and your equipment. Yet there is no ready guide to look up this information, and with tapes that are both unlabeled and un*boxed* it can be hard even to know enough to intelligently query the experts. Yet this needn't quite be so: If somehow you could gather samples of every variant of tape produced under every brand name over the years, you might be able to brand and date a cassette based on

any number of characteristics, from the batch code stamped on the spine to the grippy patterns molded into the plastic to the color of the reels inside. If that data could be gathered about a very large sample of tapes, and correlated with preservation metadata from orgnizations that have had the tapes digitized, it is possible that you could learn a lot about a tape just from its unlabeled cassette.

Hopeful that this kind of data crunching can be a success, I am embarking on the project of building a database of cassettes and their sleeves, collecting as much information as practicable on each item. Additionally, I am gathering images of as many varieties of cassette and sleeve as I can to add to the database, and to build an Audubon-style color-organized guide to recordable VHS stocks. These actual images will likely be stored outside of the database, but each entry in the database will reference the set of images in the guide that represents it (this will not necessarily be an image of the item itself, but of a representative item that matches all of its significant visual characteristics.) In a later phase of information gathering, when I begin to receive preservation metadata on cassettes, that information will be added to the database. For purposes of this preliminary project, however, I have created a data model to describe the external physical characteristics of the tapes and sleeves, and created rules to govern the data fields necessary to that end. In this paper I will explain the structure of the data model, the nature of its entities, the attributes that describe them, and the tables needed to represent that information flexibly.

THE ENTITIES
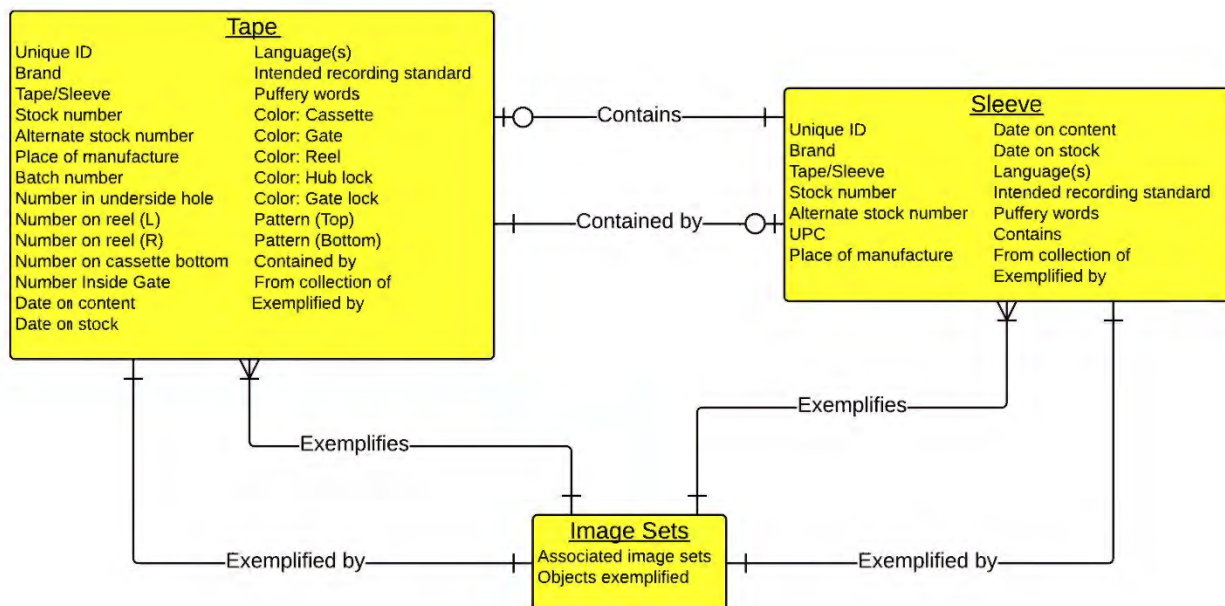
The database is designed to contain and describe three fundamental entities, each related to the others, and to which the all of the metadata that I gather will attach. These are:

- **Sleeves**: These are the boxes that contain (or once contained) blank recordable VHS cassettes. They may be true sleeves (paperboard or flexible plastic containers with one open side or end) or "library cases" (the snapping clamshell cases common in professional video environments) – any package provided by the manufacturer to house recordable stock is termed, for purposes of this database, a "sleeve".


Suspiciously professional-looking library cases seen in the Tanners' living room
in the Season 3 & 4 opening of *ALF*. Using this database, Willie Tanner might learn that,
twenty-five years later, it would be inadvisable to play that Ampex 189 in his VCR!

Sleeves and the tapes that they house are described separately because tapes frequently end up stored in the "wrong" boxes, making it desirable to describe the attributes of each independent from the other. Yet even a wrong box may contain valuable information – a tape and its sleeve are likely to come from the same approximate era, for instance – so containing sleeves and their contained tapes are connected in the database by a field for the unique identifier of the related object. They *contain* cassettes, and are *exemplified by* image sets.
- **Tapes**: These are the cassettes themselves, described in much the same way as (though with certain different attributes than) their sleeves. They are *contained by* sleeves, and *exemplified by* image sets.
- **Image Sets**: These are collections of images of examples of cassette models and sleeve designs. The goal is to collect examples of as many different models and designs as possible, but for each one to have only one set which is linked to each individual cassette that it exemplifies. Together they act as a sort of visual controlled vocabulary of tape stocks. Each item represented is depicted by a series of several images – typically six for every cassette and five

for every sleeve, assigned a five-digit identifier per item, with each attribute assigned a letter representing what surface it depicts, as laid out in the following chart. The numbers for a set of sleeve images are meant to match the numbers of a set of tape images representing the tape that was *originally packaged in that sleeve by the manufacturer*. Additional letters will be assigned later for less-frequently imaged (yet possibly very informative) surfaces, such as glued-down interior flaps and the back sides of record tabs. Missing image sets (for a sleeve of a model of tape whose model was first found sleeveless, for instance,) may be filled in later, and image sets may be replaced if a cleaner, more complete example of an object is found. Image sets *exemplify* sleeves and tapes.
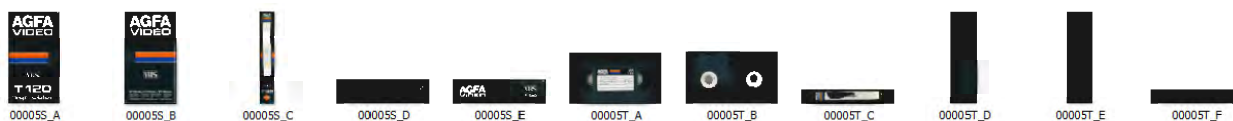
Sample image sets of a sleeve and the tape that it contains.

Tape:	Sleeve:
A: top	A: front
B: bottom	B: back
C: spine	C: spine
D: left edge	D: top
E: right edge	E: bottom
F: gate	

As with any physical object, the cassettes and sleeves display innumerable attributes, and the aim of this database is to describe the objects to a very fine level of detail. To this end, there are many descriptive fields for those objects. Since many fields – such as alternate stock numbers and "puffery words" – need to be repeatable, and since many controlled vocabularies will be needed to maintain consistent, flexible data, a number of join and look-up tables have been developed for the database. These are shown in the data model here (a larger copy of which is also attached separately), and are explained and have their rules spelled out, along with their non-tabled fellows, below.
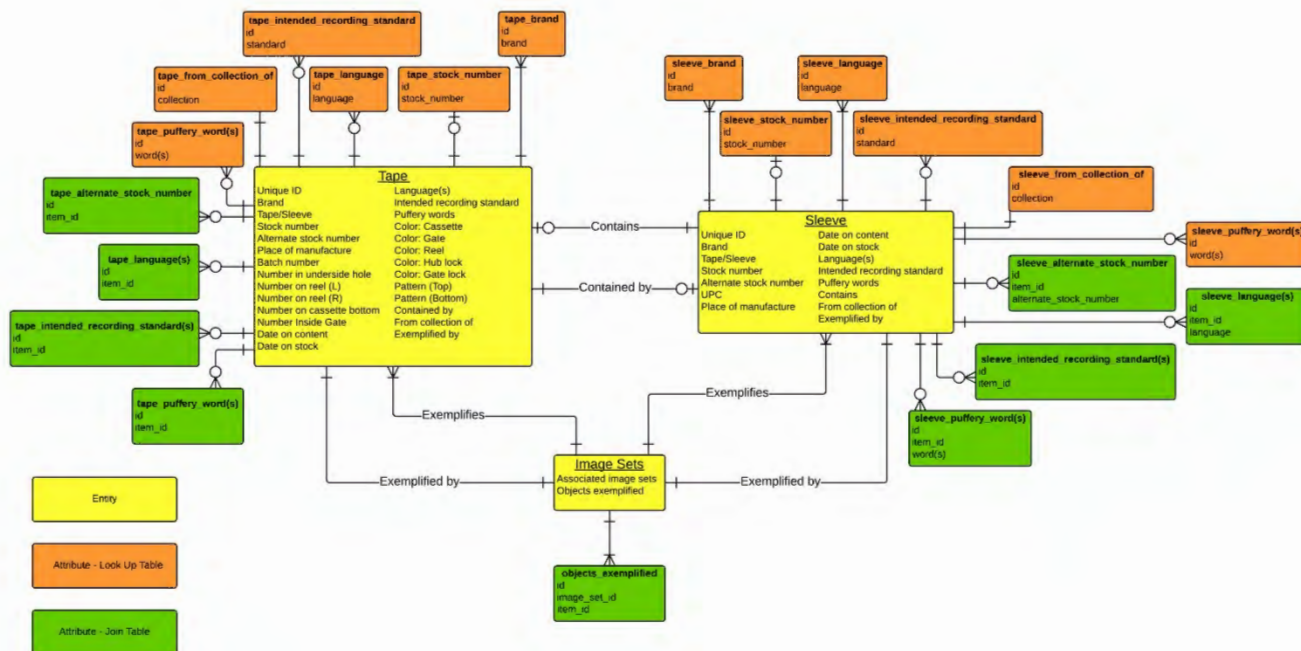
The Data Model

THE ATTRIBUTES

Sleeve Attributes:

- **Unique ID:** Unique identifying number for each sleeve, assigned sequentially at the time of entry, beginning with VHS_S_00001 and progressing upward. This field is mandatory and non-repeatable, as this will be the value by which the item is identified in the database. The digits of this string should match the digits of the ID of the tape contained in the sleeve; thus, if there is a tape with no sleeve, number the tape and skip that number in the sequence of the sleeves. For instance, if cassette VHS_T_24601 has no sleeve, there will be no sleeve VHS_S_24601.
- **Brand:** This necessarily refers to the brand name printed on the package or cassette, rather than to the actual manufacturer. Thus, "Scotch" and "3M" are different brands, as are "BASF" and "EMTEC". Frustratingly, but for the sake of consistency, "Fuji" and "Fujifilm" are also different "brands". Consistency in this matter may later become particularly critical if many different off-brand tape stocks (store brands, regional brands such as Trisonic, etc.) can eventually be linked to fewer larger manufacturers. This field is mandatory (although "Unknown" is an acceptable value), and while it generally will not be repeated, it may be when needed (the EMTEC and BASF brands, for instance, appeared together on tapes for a time).

- **Stock number:** This is the standard E- or T- number designating the length of the actual tape inside the cassette. This is an optional field, since there will be many instances (more with the cassettes than the sleeves) in which this number will not be able to be determined with certainty, but it should always be filled in when possible. To ensure that these numbers are reported in a consistent form (a two hour NTSC cassette should be reported as T-120, not T120, t120, or T 120), this field is governed by a controlled vocabulary which will grow only to accomodate unusual length numbers that emerge (T-3, perhaps, or T-170, etc.), but not for brand or grade related numbers, such as T-120HSN.
- **Alternate stock number:** This field is for all variations on the T- and E- formula, brand-related and otherwise. Some boxes may have multiple versions of this number, so the field is repeatable; others may not have any variations at all, so it is optional. Examples include T-120HSN, T120, and E-240HS B.
- **UPC:** This is simply the UPC code on the sleeve. It is optional, and repeatable for unusual instances in which two or more codes may appear on a sleeve. Bar codes printed on some very old sleeves may not be formal UPCs, but may still be entered in this field. Numbers from other bar codes, such as library stickers, should not be entered here.
- **Place of manufacture:** This is a text field to enter a sleeve's place of manufacture exactly as printed (examples: "MADE IN JAPAN/FABRIQUE AU JAPON", "Product of Korea"). This is an optional field, and not repeatable. The comparative benefits of a controlled vocabulary of countries are being weighed; if the field is changed to a controlled vocabulary, it would become repeatable for instances in which different components are manufactured in different countries.
- **Number(s) on concealed flap:** This is a text field for characters found on parts of a box that are ordinarily hidden by glued-down flaps. These may be highly informative, but as accessing them is destructive to the boxes, they are entirely optional in the database.
- **Date on content:** This is an optional, non-repeatable date field reflecting the earliest date indicated for recording of content. It refers to the **recording**, not the **production** of content, so that a copy of *The Maltese Falcon* recorded from TCM in 1995 should be dated 1995, not 1941.
- **Date on stock:** This field is for any date evident for production or acquisition of stock. It is optional and will usually be absent, but could derive from a copyright date on the box or labels, an invoice or receipt for purchase of the stock, or conceivably from a large number of colocated tapes of one type clearly recorded around the same time by a body that regularly bought blank tapes for regular recording.

Considering the breadth of those information sources, and their relative rarity, the options of either deleting this field, and adding a related field for the source of this date, are both being weighed.

- **Language(s):** This is a mandatory, repeatable text field reflecting the languages represented on the sleeve. It is governed by a controlled vocabulary that will grow whenever a new language is found.
- **Intended recording standard:** This is a repeatable, optional field for the recording standard for which the manufacturer intended the stock to be used. This may be indicated by statements on the sleeve and/or labels, or by T- (NTSC) and E- (PAL/SECAM) numbers.
- **Puffery words:** This field is for the often technical-sounding brand words used by companies to market tape (examples: Epitaxial, Beridox, Avilyn), and for the claims of high quality made on virtually all tape packages ("High Standard", "High Color", etc.) The field is optional and repeatable, and is governed by an ever-expanding but carefully curated controlled vocabulary.
- **Contains:** This field holds the Unique ID of the cassette contained by the described sleeve. Data entered in this field must be in the standard form of a tape Unique ID (example: VHS_T_10086), or in the case of an empty sleeve, a yet-to-be-determined null value. The field is mandatory, and repeatable in rare (potentially unheard-of) instances where a sleeve holds more than one videotape.
- **From Collection of:** This field is for the name of the source of the sleeve. Examples so far include "NYU Cinema Studies Techno Trash", "NYPL LPA Dance Division", "Kit Goldstein Grant Theatre Recordings", and "David Neary Home Videos". It is governed by a controlled vocabulary which will be added to for each successive collection from which data is gathered.
- **Exemplified by:** This field is to be filled by the five-digit, one-letter string of the series of images that illustrates the type of the sleeve described. Values must match these identifiers.

Tape Attributes:

- **Unique ID:** Mandatory, non-repeatable unique identifying number. Conforms to the same rules as the Unique ID for sleeves.
- **Brand:** Mandatory, repeatable field. Conforms to the same rules as the Brand field for sleeves. "Unknown" is an acceptable answer, and may be a frequent one in collections where tapes are missing boxes, or not convincingly associated with the boxes they came in.
- **Stock number:** Mandatory, non-repeatable field. Conforms to same rules as Stock number field for sleeves. Many tapes may have "unknown" for their stock length, and many will have non-standard forms of their T-

and E- numbers. "Unknown" is an acceptable value, but "T120SONY" is not. For this reason, this field is carefully governed by a controlled vocabulary. Options are being weighed for repeatability for some tapes, such as certain BASF models, which show plain evidence (such as E- numbers imprinted on the reels) of being PAL/SECAM tapes repackaged and relabeled for the NTSC market.

- **Alternate stock number:** Optional, repeatable field for all variations on the T- and E- formula, brand-related and otherwise. These are quite common on in the area near the batch code on a cassette.
- **Place of manufacture:** Optional, non-repeatable. Conforms to same rules as for sleeves.
- **Batch number:** This is a mandatory, repeatable text field for the batch code exactly as it appears on cassette, either in the upper left hand corner of the spine, or in the space where the label on the spine would go. "None" is an acceptable value.
- **Number in Underside Hole:** This is a mandatory, non-repeatable field for the number, typically molded in the plastic, that sometimes appears in the approximately ¼" hole on the underside of the cassette near the gate. If numbers are presented on more than one line, delimit the lines with " / ". "None" is an acceptable value.
- **Number on Reel (L)** and **Number on Reel (R):** These are mandatory, non-repeatable text fields for the numbers frequently molded into the plastic of the reels of a cassette. "None" is an acceptable value. For numbers separated by "spokes", leave three spaces between strings. For numbers on separate concentric rings, start from the center and delimit rings with " / ".
- **Number on Cassette Bottom:** This is a mandatory, repeatable text field for numbers molded into the plastic in various places on the bottom surface of the cassette. "None" is an acceptable value.
- **Number Inside Gate:** This is a mandatory, repeatable text field for a number molded into the plastic on the inside of the gate of the cassette. "None" is an acceptable value.
- **Date on content:** Conforms to same rules as the Date on content field for sleeves.
- **Date on stock:** Conforms to same rules as the Date on stock field for sleeves.
- **Language(s):** This is a mandatory, repeatable text field reflecting the languages represented on the tape. It is governed by controlled vocabulary that will grow whenever a new language is found. "None" is an acceptable value.
- **Intended recording standard:** Conforms to same rules as the Intended recording standard field for sleeves.

- **Puffery words:** Conforms to same rules as the Puffery words field for sleeves.
- **Color: Cassette, Color: Gate, Color: Reel, Color: Hub lock**, and **Color: Gate lock:** Mandatory text fields to describe the colors of the various components of the cassette. "Color: Cassette" refers to the main body of the cassette, and is hypothetically repeatable, since there are two pieces to the body. "Color: Reel" and "Color: Hub lock" may also be repeated in the rare event that the two pieces of a kind in one cassette are not the same color. All of these fields are governed by a controlled vocabulary that will expand whenever a new color is "discovered". Colors so far include Black, Beige, White, Orange, and Mint green, but options are being weighed with regard to the level of specificity with which a color should be described. There are many shades and intensities of white, ranging from solid and bright to slightly translucent and ivoryish; the "orange" items pointed to in the examples are the common TDK hub locks that are in fact closer to cadmium red. It is frustrating not to use these more precise terms, yet hard to imagine a large number of users successfully applying them with the desired accuracy.
- **Pattern (Top)** and **Pattern (Bottom):** Mandatory, repeatable text fields giving brief descriptions of the patterns molded in the plastic on the top and bottom surfaces of the cassette. This is meant to provide basic descriptive information and direct users to visual guide to find information on the cassette in hand, or to find the identifier of the image set to add to the cassette's database entry. As such, some descriptions will describe many varieties of tape. This field will be fed by an ever-expanding controlled vocabulary including entries such as "Diagonal matrix of raised squares", and "Horizontal matrix of raised rectangles on end opposite gate".
- **Contained by:** Conforms to same rules as the Contains field for sleeves.
- **From Collection of:** Conforms to same rules as the From Collection of field for sleeves.
- **Exemplified by:** Conforms to same rules as the Exemplified by field for sleeves.

Image Set Attributes:

- **Object depicted:** This is mandatory, non-repeatable field that records the Unique ID of the actual object depicted in the images.
- **Objects exemplified:** This is a table that records the Unique IDs of all of the physical objects exemplified by an image set, and correlates them to that image set's identifier.

FUTURE STEPS

In the coming months I will be developing this model further, adding tables for preservation metadata, and building the model into an actual database. It would be desirable in the long term to have the database online and receiving contributions from other users, but until it reaches that point of interest, it will be kept locally as a FileMaker database. Look for a working model in the spring of 2015, and further progress in the following year.