Semiotic Principles for Metadata Auditing and Evaluation

Erik Radio

Introduction

The effectiveness of an information system is dependent on the quality of the metadata it indexes. While there are other critical components that contribute to effectiveness, how queries are modeled and resources indexed matters little if the metadata is of insufficient quality. Advances in information retrieval technology have to an extent lessened the risk of unsatisfactory retrieval due to inaccurate or incomplete metadata, for example through automated query expansion[1] and an increased focus on the functional tasks related to a query.[2] However, this is not to say that metadata quality can be ignored given the prospects for more sophisticated retrieval mechanisms. By contrast the increase of non-textual resources that rely on descriptive metadata for discoverability necessitate thoroughly descriptive records. In an effort to improve metadata in an information system, reviewing the types of data and ways that they are being used for particular elements in records can be a way of isolating particular issues involved with the metadata creation process.[3] Hopefully this would lead to an improvement in the overall consistency of records.

In the context of libraries, the large number of bibliographic records makes any sort of record-by-record analysis and correction or enhancement an unsustainable use of time and resources. To best address the issue of assessing consistency of data in records,

metadata auditing, by which a representative sample of the corpus is selected and reviewed, has been widely used to guide data assessment and fuel solutions for large scale remediation.[4] As the amount of metadata continues to increase, this will likely continue to be a popular solution to the problem of addressing data consistency.

Defining metadata quality becomes an important consideration and one necessary before an audit should be undertaken, particularly as an evaluation without criteria may lead to misinterpretation. Yet even though evaluation can only happen when working with concrete values, quality frameworks are necessarily abstract. This rift suggests that a different model for understanding metadata by the functions it serves may be another way of informing the act of evaluation. The field of semiotics provides such a framework for analyzing the role of signs and sign-functions as they apply to metadata records.

For semiotics to provide a useful lens for viewing metadata, it is important to determine what on the semiological level constitute metadata's particular functions. In other words, what does the choice of semantic units used to populate a schema imply for an understanding of 'consistency' or 'accuracy' in records. Problems that arise from taking metadata out of its original context are well known and become more acute when performing a general audit over a large corpus.[5]

If a collection of metadata records can be understood as a discrete collection of signs, or sign body, then the issue of what actually constitutes the body must be considered. The ways in which the catalog, or corpus, can be divided to fit a particular administrative need, like auditing, allows for the creation of new formulations with their own structural implications for interpretation. Understanding the ways in which a corpus

can be shaped as a sign-producing object is necessary to ensure that transformations are both accurate and beneficial.

As a structural inquiry and analysis into the semiotic processes at work in metadata and records, this paper will first provide a brief overview of the semiotic landscape from its modern origins through the poststructural criticisms articulated in the latter half of the 20th century. From this perspective, the catalog will be reviewed as a particular kind of sign-producing object on several semiotic levels, each with their own particular sign-functions. Finally, common auditing practices will be examined to identify ways in which auditing is impacted by the catalog's sign-functions and semiotic boundaries, as well as what implications this holds for evaluation and remediation.

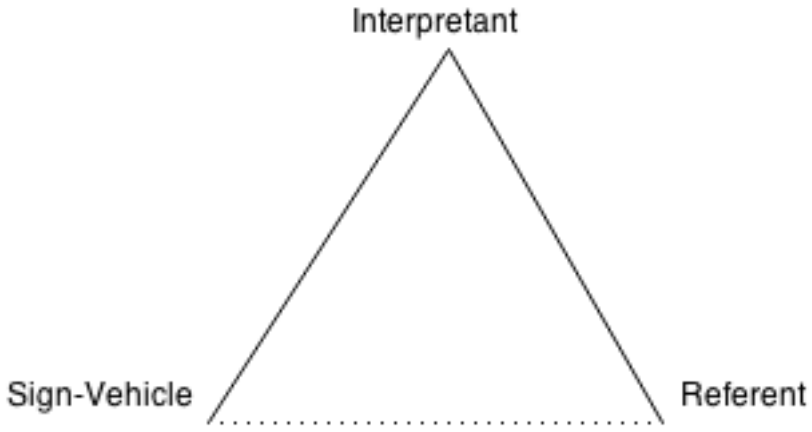
Sign Models

Semiotics is the study of everything that can be interpreted as a sign.[6] While the philosophical study of signs has ancient origins, it was its formalization by Charles Sanders Peirce in the 19th century, as well as a relatively synchronous development by Ferdinand de Saussure that created the environment in which the study of semiotics and sign systems was robustly developed.[7] A primary difference between these two thinkers relates to their models of signs, but both models can be seen as interpretations of a signified-signifier relation.

Peirce's model is triadic consisting of an object, a representamen, and an interpretant.[8] As illustrated in Figure 1, the representamen (sign-vehicle) is analogous to

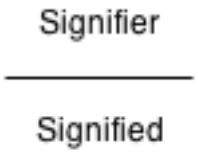
the signifier and is the form in which the sign is transmitted. The object (referent) is the signified, or the real-world entity from which the signifier is derived. The interpretant is the effect the signifier produces on the receiving entity.

Figure 1 Triadic Sign Model



By contrast, Saussure's model is dyadic, reflecting a more direct relationship between signified and signifier.[9] In Figure 2 the absence of the object present in Peirce's model is an important characteristic indicative of this model's origins in idealism as for Saussure it is language that determines the order of the world and that it is "the viewpoint that creates the object."[10] In other words, meaning arises in language through relations of association and opposition. There is no singular sign-function that an object can always be said to be producing as all is dependent on the interpreting entity.

Figure 2 Dyadic Sign Model



These two models form the basis of what can be understood as structural semiotics. The primary difference between both theories is that Peirce is concerned primarily with the production of meaning while Saussure is interested in the structure of the system in which meaning takes place.[11] Though ontologically different, both rely on a perceived system of structures that underlay their models and from which all sign systems can be understood. However, there have been many criticisms leveraged against purely structural semiotics. A criticism of poststructuralist semiotics was a lack of concern for how signs changed over time and as the result of social processes.[12] As Sturrock notes,

> '[Structuralism] is concerned to study particular systems…under artificial and ahistorical conditions, neglecting the systems or structures out of which they have emerged in the hope of explaining their present functioning.'[13]

The idea of reducing a complex system of signs to a few structures is a flawed strategy as one can never be objectively removed from the system under analysis.[14] It is the role of context in a sign system that will have important considerations for viewing a catalog as a sign body.

Catalog as Sign Body

Language represents our most sophisticated sign system. It follows then that metadata records, like any text, are comprised of signs. As such, the use of a particular term to describe a given resource is a type of sign-function. Triadically modeled, the term (representamen) conveys some meaning to the user (interpretant) about the object to which it refers**.** While descriptive metadata in records are primarily used for discovery, viewing the record as a signifying entity conveys additional information about its resource based on the manner in which it is described. This idea is essential to understanding a record or catalog as a signifying body.

Schema/element and Langue/Parole

The dichotomy between language and speech is a central concept to semiotics, specifically the Saussurean strain, and one for which there is a unique parallel in the cataloging realm. Barthes describes language (*langue*) as language minus speech, or a system of rules and values defining a structure within which all speech happens.[15] It is what allows for the construction and combination of signs to make meaningful statements. An individual cannot change it; 'it is a collective contract which one must accept in its entirety if one wishes to communicate.' [16]

Speech (*parole*) is how one uses language to code what one wishes to communicate. It is the way in which individuals choose to express themselves through *langue* and the structures it entails. While an individual cannot change language, it is through the evolution of speech over time and endless iterations of sign combinations that *langue* is gradually changed.[17]

An appropriate and serendipitous example of the *langue/parole* dichotomy is the schema/value relationship familiar to cataloging and metadata. The MARC schema can be considered a type of *langue*. As such, the schema outlines what can be said about an information resource and how it can be said in the document. For example, the uniform title of a resource can be recorded but only in 240$a.[18] Additional restrictions on the formatting of the data values (ex. AACR2, RDA) are another part of what may be considered the language as it is an additional framing of how a resource can be described.

Speech, then, is represented by the freedom a cataloger has to use certain elements and document particular values. A schema like Dublin Core, which has very few restrictions on what values should look like, could be described as having a much greater degree of freedom in this regard albeit at the cost of consistency. On another level speech also consists of the values a cataloger chooses to use to describe a resource, and impacts how thoroughly that description is achieved. Over time the dwindling use of particular elements, even their misuse, may affect the language, or schema's emphasis on capturing particular features of a resource.

Syntagmatic Levels in Metadata

While metadata falls rather transparently into the *langue/parole* structure, understanding metadata as a unit or units for analysis requires an additional semiotic framing. The syntagm is a combination of signs, or interacting signifiers, that form some kind of cohesive unit.[19] For example, as a set of words organized sequentially, a sentence is a type of syntagm in that it is a group of signifiers that collectively conveys some kind of information. In the context of a resource record we would say that the following Dublin Core (DC) statement is also a syntagm:

<dc.title>L'archéologie du savoir</dc.title>

While it does not follow the syntax of a spoken sentence, it is a collection of interacting signifiers conveying information, specifically the title of a book. It is a sentence within the context of the system.

Where syntagms differ from signs is in their ability to absorb other syntagms to create a larger syntagmatic unit.[20] With this comes the ability for a larger syntagm to express a new sign-function. To extend on the previous example we might say the following record has consumed several syntagms via additional elements and values to create a new syntagm, a record instead of a statement:

<dc.title>L'archéologie du savoir</dc.title>
<dc.creator>Foucault, Michel</dc.creator>
<dc.date>1969</dc.date>
<dc.language>French</dc.language>

<dc.identifier>2-07-026999-X</dc.identifier>

<dc.publisher>Éditions Gallimard</dc.publisher>

This example could be expanded to the level of an entire catalog as multiple records comprising a single whole. Syntagmatic relations emphasize the dependent relationships between the part and the whole. One cannot have a record with statements, and without a collection of records there is no catalog, and so the largest syntagmatic unit is reliant on the most atomic unit of which it is composed.

A sign-function cannot be reduced to a single denotation even if that is its objective; in most cases a signifier transmits several types of information. Accordingly a record transmits information about a resource and as such is a signifier for a signified, just as a catalog conveys information about a collection of resources. Deconstructing sign-functions on various syntagmatic levels allows one to determine the exact nature of a catalog's sign-function.

A sign exists when there is a meeting between expression and content for the purpose of enabling a coded correlation.[21] Expression refers to the actual perceived form of the signifier, for example words on the page or auditory phenomenon. Content is the paradigmatic structure of the signified. Both can also be described as functives and their meeting can give rise to another sign-function.[22] All of this is to show that signs are based on transitory correlations within a coded framework, and are not fixed identities with fixed meanings. Signs dissolve '…into a highly complex network of changing relationships.'[23]

The catalog as a large corpus of records has semiotic functions on several levels. As a signified a catalog does not refer to a tangible object but rather to its mental representation. A catalog has a type of unity in that it binds together information about resources common to a collection. Analyzing the accuracy of that denotation is a primary consideration behind metadata auditing.

Catalog Unity

Questioning the catalog's unity causes it to lose its self-evidence as a unified object, as it is really a node within a network. Here we may draw a parallel to Foucault's observation of the book as a series of networked references to other forms of information.[24] Extending this to the catalog which is itself a kind of book and more clearly reflective of Foucault's description as it is a book of signifiers on the syntagmatic level of the record, we can also observe that its unity as a collection of records is purely self-referential. Self-contained systems allow for patterns and regularities to be observed but only because they have been isolated from a greater context.

Setting aside the previous quality of unity which concerned the plane of content on the collection level, for purpose of metadata auditing we must ask what unity there is between the records, or the signifiers, on the plane of expression. We can further dissect this unity by casting records into forms and substances of expression.[25] Form of expression is the paradigmatic and syntactic rules governing signs. For records it is a linguistic system. The substance of expression for records is digital text. More specifically we might say that the form of expression is that of an *element:value* binary.

All records can be said to have the same form of expression. That metadata (when serialized in XML) follows the structure of

<dc.title>L'archéologie du savoir</dc.title>

instead of

The title is ‘L'archéologie du savoir’

is indicative of its machine-readable purpose and is a characteristic that greatly impacts auditing methods. Yet if one abstracts this structure it can be observed that within this form of expression there are really two subforms of expression that are meeting to express the syntagm. At its most skeletal the first form consists of a binary relationship, or a yes/no situation expressing whether a data field does or does not have a value. This is not without serious implications for the auditing framework as the presence or absence of a data value can drive a transformation process based on metrics of completeness.

The other subform is the language that complements the yes/no binary. The combination of element and value that creates a statement also creates, when joined to others, a record. If a record consisted of statements independent of a binary structure as in the sentence above, then a transformation could be to ensure that a particular syntax for statements was followed. But joined to a schema one must also consider how accurately the data value reflects what is designated by the element. This presents another more complicated arena in the realm of metadata enhancement. The intersection

of a linguistic system and a logical system is where one finds metadata statements/syntagms. This meeting of a human and machine language creates a third hybrid language that is the object of auditing efforts.

Metadata fields are a kind of syntagm that combine to create records and eventually a catalog or corpus. That the unity of a catalog can be called into question indicates that it is necessary to examine the goals of auditing, its methods, and how the artificial unity of a catalog or corpus should be the starting point for auditing efforts. As we will see, it is where the unities break down that one may find different sign bodies demarcated by various semiotic boundaries that warrant different auditing perspectives and efforts.

Quality Definitions

Defining metadata quality is an important component of digital library stewardship since the growth of resources requiring effective metadata for discovery **is** an obvious need. The purpose of metadata auditing is to gain a clearer picture of the state of metadata in a system so as to guide transformations that enhance quality with the end goal of facilitating retrieval.

Frameworks for assessing metadata quality vary considerably in scope and the type of metrics they seek to capture. Stvilia et al. identify three dimensions affecting information quality.[26] Intrinsic quality is measured in relation to a reference standard (e.g. spelling, validation, currency), and is largely independent of context. Relational

quality measures the relation between an object and its usage or how accurately it reflects its surroundings (i.e. collection title in a federated collection). Finally, reputational quality measures the position of a resource in a cultural or activity structure. The implementation of their framework uses bench line representations meeting minimum requirements for their three categories to produce an aggregated quality ranking. They note that information 'may have different kinds and levels of quality and value in different contexts of use'.[27] This important consideration emphasizes not only the different roles of information systems but also the transitory nature of information as signifiers.

Bruce and Hillman outline several broad metrics for assessing metadata quality noting that some may be more important for particular communities or collections than others.[28] Two are directly related to auditing. Accuracy, that records are factual, may not be directly verifiable due to the high labor involved in reviewing all records in a large corpus. As such sampling techniques or statistical profiles are used to assess accuracy. The second, timeliness, refers to the fact that metadata loses quality over time if it loses synchronicity with its external context. The shift from static to dynamic metadata modeling is a necessary factor contributing to its long-term effectiveness. This element of temporality is a critical component of auditing and will be discussed below.

The vague but widely used 'fitness for a task' definition of metadata quality, is refined by Ochoa and Duval as the ability for one to find, identify, select, and obtain resources in a digital repository.[29] Yet while they note that measuring quality should be schema-agnostic when possible, the question of whether this is ever possible as the choice of schema and its underlying ontology directly affect not only the metadata value

used, but also how it is framed, is critical. 'Title' in Dublin Core and VRA Core are not semantically equal, so examining their values together independent of schema may lead to a false interpretation of quality.

Price and Shanks developed their quality framework from a Peircean semiotic framework that identifies three levels of quality that align with the representamen (sign), referent (relation to object), and interpretant (use of the sign): syntactic, semantic, and pragmatic.[30] They maintain that a datum serves as a sign in an information system as it refers to some external phenomenon, and that its use requires some sort of interpretation that results in action, the process of semiosis. Since their framework is concerned with database systems they can confine themselves to a structuralist framework. But as metadata is shared and aggregated from various providers the corpus or catalog boundary becomes increasingly blurry and auditing techniques require a different set of considerations.

Auditing and Evaluation Techniques

Defining metadata quality and identifying a framework for assessing it is a necessary precursor to any evaluation process. Without an idea of what to look for it is not possible to effectively find or improve existing metadata. While those mentioned above are not the only frameworks available, they do share a common thread in emphasizing the importance of context when evaluating metadata and defining use.

Auditing and evaluation are sometimes considered equivalent but they are actually two different processes. Auditing is the process by which one selects a sample

that will be evaluated. Evaluation refers to the process of assessing how metadata measures against quality metrics. What follows is an analysis of several common evaluation and auditing practices and initial points as to how semiotics could further inform this process.

Random sampling is a widely used method in metadata evaluation since the number of records in a corpus is usually large enough to make an evaluation of each one untenable. Hillman notes that when MARC was the dominant schema, random sampling evaluation was frequently used to examine the quality of shared records since the cataloging environment was more tightly controlled.[31] While one can argue that the goals of cataloging are the same across institutions, the particular nature of those goals differs depending on the type of institution. As Robertson notes, a metadata record for a book will look different in a library, museum, and archive and that of these three, one is not objectively better than the others.[32] Did the record providers customize records for the type of institution with whom they would be shared? As Hillman has observed this problem has only become more acute in the metadata world as evaluation techniques are increasingly aligned with the metadata functions needed for a particular application.[33]

Random sampling was also used by Stvilia et al. in applying their framework for evaluating metadata.[34] Their IMLS project harvested simple DC records via OAI-PMH from 16 different providers including libraries, museums, and historical societies. All of the records were considered incomplete as they did not use the full set of 15 simple DC elements. Among a litany of flaws, 94% contained redundant information across fields.[35] Though relational/contextual considerations are a part of their information quality framework, their sample does not account for this in their evaluation. Compounding the

problem is that these records, having been harvested through OAI-PMH, had likely been transformed into DC and possibly were not originally cataloged in that schema. The high redundancy of values may be a result of this transformation since, for example, a schema as complex as VRA does not translate well to DC terms. Ignoring the original context and schema and proceeding with a random sampling imposes a type of false unity on the corpus. Similar audits also involved random sampling though issues of context remain unclear in these as well.[36] [37]

As the volume of digital objects continues to grow at a rapid rate, random sampling will likely continue to be seen as a viable auditing method. But with this method it has been seen that manual analysis still plays a critical role in determining the level of quality. To ensure though that quality goals can be achieved upon ingest it is necessary for some sort of automating mechanism. As Ochoa and Duval note, manual analysis is meaningful but not scalable.[38] Hillman mentions the NSDL's use of a more batch-oriented method of sorting records via graphical software that allows for visual pattern recognition.[39] Contrasted with manual analysis, statistical analyses of this type (determining usage use of element, length of value, data type of value) are scalable but generally not as meaningful in shaping an evaluation. It is necessary to see both the forest and the trees simultaneously. The current discussion is not concerned with determining the better method suffice it to say that a combination of the two is probably the best course. Instead we have tried to identify particular areas where the idea of collection as a unity provides a misleading understanding that can negatively impact an analysis' conclusions. Addressing the sign-function and nature of these collections is one way to ensure a more systematic and meaningful evaluation.

Semiotic Auditing

By analyzing various quality metrics, auditing tactics, and evaluation methods, it has been shown that these vital concerns and processes can be hindered by various assumptions as to the nature of the collection, the corpus, or catalog. That these three units are composed of interrelating sign-functions on both the microscopic and macroscopic syntagm planes indicates that a more nuanced interpretation of a corpus is required for an evaluation to be maximally successful. Unfortunately a systematic method that could be used for all auditing scenarios is not possible; its utility would ultimately be fleeting as it would have to account for all specifics in the wide range of evolving collection types that exist. Instead, principles based on the nature of sign-functions will be the most useful tools.

Data Absence

A recurring aspect of the evaluations mentioned above is the use of statistical analysis to determine how frequently elements were used. As Ochoa and Duval observe, given the enormous quantity of records that exist this kind of analysis is likely to be used as it is scalable in a way that manual analysis is not.[40] Owing to its smaller sample size, the University of Houston conducted an audit that contained a completeness metric that was defined as possessing some data from a core list of required elements.[41] Similarly, Stvilia

et al's audit relied heavily on the presence of elements as a statistically significant component related to quality.[42] But here the meaning of completeness must be called into question as one that can be misapplied and negatively direct an enhancement effort.

Returning to forms of expression, data expressed in XML can be understood as having two subforms, one of which is built on a yes/no binary. (Metadata serialized in other machine-readable formats is also subject to this characteristic.) Seemingly neutral in objective, the absence of an element is a significant consideration in evaluation processes, either as a metric for completeness or to identify critical fields for usage. What does the absence of a value signify?

The California Digital Library's metric for metadata evaluation posits among other things that metadata analysis should inform which fields are present, and what percentage of the total number of records have each field.[43] Clearly here the absence of a field is interpreted as contributing to a lack of completeness. Reducing it to the binary, an absent field maps to a 'no' with all of the negative connotations that may signify. But it is unfortunately not as simple as just that, for the syntagm

<dc.date>No date</dc.date>

is not semantically equivalent to

<dc.date/>.

While both can be interpreted to mean the same thing, the sign-function of these syntagms are opposites as binary subforms, which is precisely how it will be interpreted when dealt with in the context of a large corpus. This subform is inescapable in metadata, but the sign-functions it produces are, as with all signs, variable and depending on context. Yes/no does not necessarily correlate to complete/incomplete. This being so, the absence of a field or fields does not necessarily indicate an absence of quality. In most cases an absent value is to be preferred to an incorrect value.

To use an example, *dc.title* is in many systems a required element as it is essential for retrieval. Yet if there were missing titles in records describing computer code, this absence may just reflect a mismatch as a title is a more bibliographically oriented field. What then of the single record for computer code in a catalog missing a title? Is the record incomplete or misrepresentative? Many born-digital resources do not have clear titles (e.g. tweet, email). In these cases producing

<dc.title>No title</dc.title>

could actually negatively impact overall quality as it would just be noise.

If something as seemingly uncontroversial as a the presence of a title can be shown to not be universally applicable, other less commonly used fields in a corpus may likewise not necessarily be in need of enhancement. As a guiding semiotic principle, the connotative properties of data absence are not indicative of data quality or completeness. Statistical analysis of field usage is useful and necessary but it can draw patterns that drive false assumptions. As such it is deeply reliant on the boundaries of the corpus.

Unity and Boundary

It has been noted that when discussing a catalog one is not referring to a physical object but to a mental representation. The connotation of 'unity' that the representation provides has already been brought into question. A record can be understood as a unit, or a collection of statements that collectively describe a resource. Its unity is not difficult to defend but its boundaries can be vague. The innocuous *relatedItem* field in MODS is where this unity could break down.[44] Though it would be unusual in practice, one could chain along the number of related items indefinitely. The catalog's unity is harder to defend as it is based on possession. While it cannot be denied that possession is a significant quality, the syntagms on every level of granularity are not homogeneous. If the catalog consists of smaller unities like records, then it follows that it is as disparate as it is unified. Identifying the boundaries that make up the internal structure of the catalog or collection is necessary prior to auditing.

Defining a boundary is the process of determining meaningful unities. Whether an evaluation of a subcollection is to be by random sampling, manual analysis, or another method, it will only be as meaningful as the coherence of the subcollection, or rather, the accuracy of its boundary. In the current context a semiotic boundary might be considered a constructed aggregation indicating a level of coherence and substantiality. It is important to recognize though that boundaries can overlap as a given resource may meet the designated criteria of coherency for several aggregations, while others perhaps none.

The particular type of unity these subcollections represent can overshadow other parts of the collection. Since different collections may and will likely require different evaluation methods, recognizing that these shadows can hinder enhancement techniques asks the question of the necessity for different instances of a record, a matter addressed below.

The boundary of a catalog is rather weak as its defining quality is its own existence. Recognizing the superficiality of the boundary is in line with the poststructuralist view that an objective structure is an artificial construction. However, the act of metadata evaluation is one that requires the presence of a collection; it is the reason an evaluation happens at all and it is only by identifying patterns in the structure that metadata can be fixed. Auditors must then recognize the artificiality of the collection while using it as its basis for a functional need. Perceived patterns, while possibly meaningful, are still just perceptions. Recarving a collection with a new boundary can call a pattern into question.

Since metadata falls very neatly, even rigidly, into the *langue/parole* structure, identifying boundaries can begin from this characteristic in various ways. Perhaps the most obvious demarcation that can be made is with the particular *langue*, or schema, that has organized the data in records. Different granularities between schemas and underlying data models make this a critical boundary. Thoroughness has different implications when a DC record is placed beside a MARC record; by what metric can one be deemed to be more complete than the other (noting again that use of an element is not indicative of quality)? Additionally, while the semantic field of *mods:typeOfResource* and *dc:type* overlap, they are not exactly equivalent as indicated by the nature of the vocabularies provided by each.[45] [46]Similarly, since MODS was created with a mapping

from MARC in mind, it would seem that auditing a collection that used both schemas would be an uncontroversial proposition, but upon closer inspection it is not quite so simple. Names, for example, are more granularly enunciated in MARC, or at least differently organized. While they can be used to capture the same type of data, the ways in which the data are stored and related are not congruent. The numerous issues that arise from crosswalking between schemas are very much at play here. Since problems of inconsistency have already been identified, it follows that evaluating a collection with mixed schemas would also be susceptible to misdirection via the same channels. As a boundary, auditing must require homogeneity by schema of its original cataloging.

Granularity of description in a catalog is also of necessary concern. Collection level vs. item level records, even when captured using the same schema do not capture the same types of data given that the interpretive lens has changed and the signifying function of the records is different as well. For example, *dc:creator* when applied to a letter would likely lead to a value of the author of the document. But for a collection of letters by various people (under a different semiotic carving) this field becomes less immediately applicable, and it wouldn't be unreasonable to leave it blank. How, then, does its absence discovered through a statistical analysis drive an enhancement effort? If flagged for attention, one might try to fill a creator field for every author in the corpus, which could certainly increase the noise in the collection. Similarly, the nature of a subject term for a collection versus an item would be of a different granularity, and while one isn't more accurate than the other, when terms are viewed *en masse* and without this distinction, a collection level subject term may appear more vague, e.g. 'Civil War' instead of 'Battle of Gettysburg'. These are of course hypothetical situations but ones that

highlight very likely scenarios if a semiotic boundary is not drawn to realize granularity distinctions.

Resource type is another place where a demarcation should be in order. Especially as one moves beyond the bounds of descriptive content, metadata needs begin to diversify rather dramatically as, for example, an article and a website have very different structural characteristics. The types of values that might get found in something like *mods:extent* would make various patterns less decipherable and obfuscate any kind of possible programmatic enhancement. Similarly, a title for a book is very different in function than the name of a computer model. Even auditing serials and monographs together could lead to misconstrued analyses of the data quality based on the different documentary needs of the resource type and their functional purposes. Educational resource metadata also tries to convey information about its intended audience, a goal that is not explicitly shared by many other types of resources.[47] Compare metadata for a textbook captured using Dublin Core fields versus Learning Object Metadata fields and the functional goals of the latter become clear (though it should be noted that the Dublin Core community has now drafted its own educational resource schema).[48] Through separating by type, it is clearer to see the types of functions a particular resource type serves and which should form an important part of any evaluation. Much more about the documentary needs of different resources can be said, but for now it is enough to outline it as a boundary.

Guidelines put in place for the cataloging of a part of a larger collection should also be considered as a boundary. An obvious example is the split between AACR2 and RDA; a random sampling drawing from records that could have used either is poised to

create a false impression of the accuracy and consistency of particular values. Likewise, the semantic vagaries of many DC elements has allowed for a wide spectrum of interpretation, making documentation of how certain elements should be used in a particular context all the more crucial for a responsible evaluation.[49] Knowing that a record has a high level of quality given its original context and guidelines is a very different matter than saying that it currently has a low information quality level. Original contexts may not even be collections, but a digital exhibition with its own set of functional needs that informed the creation of the records. Knowing those guidelines can also take out some of the work in identifying patterns and allow for a greater accuracy in transformation scenarios.

Structural and Semantic Drift

Metadata is a system of signs with language as its primary form of expression (though it is the merging of two forms of expression that creates this language). The Saussurean emphasis on linguistic signs and their semantics bring us back to the *langue/parole* structure. *Langue*, as the system that governs what can be said must be accepted for *parole* to exist, is a necessary precondition that cannot be changed by an individual or an atomic instance of speech. It is only by the continued use of speech that *langue* gradually changes. To draw the parallel with cataloging, MARC was the *langue* for the latter half of the 20th century, but over time and for a variety of reasons it has increasingly been found wanting and necessitated the emergence of new schemas.[50] Part of this change is no doubt technologically based as MARC was born in a substantially different age in the

history of computing, but it is also much broader than that. Information objects are beginning to look less like traditional books, and the rigidity of the MARC standard, which is so closely tied to the bibliographic format, has required new and broader ways of documenting information. This is what might be deemed structural drift; the *langue/parole* relationship stands, but the former is changing as a result of the use of the latter.

The underlying factor here and which also plays into many of the boundaries mentioned above is the aspect of time. The changing nature of information as both languages and contexts develop makes a static view of records a flawed perspective. Semantic drift, or the changing meaning of the linguistic form of expression, is a dynamic process. What is currently an appropriate data value may no longer be so in a century's time. Ascertaining when it happens is no doubt a difficult task but this reinforces it is an important temporal boundary. How, then, is the best way to demarcate a corpus given this particular phenomenon?

Homogeneity is the ideal in auditing. As such subcollections must be as synchronic as is possible and reasonable. A varied but temporally limited corpus is preferable to one stretched over a long period of time to ensure it is an accurate cross-section of the state of the corpus. Finding that fine line is difficult, but there are some built in indicators. As Barthes explains:

> "Some systems establish their own synchrony of their own accord…but for others one must choose a short period of time, even if one has to complete one's research by taking soundings in the diachrony. These initial choices

> are purely operative and inevitable in part arbitrary: it is impossible to guess the speed at which systems will alter, since the essential aim of semiological research…may be precisely to discover the systems' own particular time, the history of forms."[51]

Auditing is of course not exactly the same as semiological research, even if we are applying its considerations to it. However the notion that collections have their own particular time is provocative. While already having noted that schemas change over time based on functional needs, it is an interesting idea to consider how metadata for a collection of images created and then independently cataloged again after a number of years would look. More specifically what values had stayed the same and which had evolved, which fields were more widely used or emphasized, and perhaps most interestingly what does this signify?

Conclusion

Metadata is comprised of signs with various signifying functions. This paper has argued that these functions are present in the smallest syntagmatic unit and that when syntagms are combined, as with records, entirely new sign-functions are created. The catalog or corpus is such an aggregation with its own particular sign-function that is different in nature than the smallest unit of which it is comprised. Auditing metadata in a system is much more complex of a task than a statistical analysis of used fields or compliance with a content standard. Data absence, semiotic boundaries of various kinds, and semantic

and structural drift, to name just the three discussed, are all critical factors that impact the interpretation of a metadata statement. The act of interpretation is the very process of semiosis.

A prominent motif that has surfaced in relation to metadata quality, auditing, and subsequent methods for evaluation is that of context. The broad definition of quality as 'fitness for task' can only be useful when measured against the application that will use the metadata. But here one might designate two shades of that particular definition: a metric by which metadata is accepted by a system, and by what the metadata conveys. The first is of course much easier to determine, but it is the latter that we have dealt with most closely as it is what drives audits. The unity of a collection is a dotted line more than a firm boundary, but identifying metadata that is of high quality in a given context does not mean it will have that same level of quality in another. Context changes meaning. For valid reasons of efficiency the idea of the master record continues to persist, but in reality is it only a master record for its context? Its sign-function has the potential for changing when it is ingested into a different system. Should there be iterations of metadata for different contexts and should a master record be replaced by the object as an idea? Of course, this necessitates the acceptance of a name or title as an unchanging statement to refer to which is already the basic structure that the linked data environment is founded on with identifiers and subsequent triples as statements. And perhaps the idea of control as coordination, which usually applied to the user's ability to more effectively navigate a system, can be used in the cataloging context in which the function/context the record must serve/exist in can be more finely tailored from all possible statements that can be

made about a resource.[52] The record will persist but its form may be increasingly mutable.

A final mention of an additional boundary that may help auditing efforts pertains to application profiles. As an example of the type of documentation that increasingly complex world of digital curation demands, application profiles provide a specific key to unlocking not only the mechanisms that drove the creation of certain metadata values, but also the scope of the given collection which can be used to identify boundaries. Ideally in the realm of shareable metadata this documentation would come as a part of the metadata package, but even then the ability to determine how well one profile matches its new collection's context is a manual process. For scalable reasons, machine readable application profiles would allow to see more explicitly what type of remediation work would have to be done and to which values they would apply.

That the rapidly growing body of digital resources requires quality metadata appropriate to its various contexts is clear. Since manual analysis and remediation of records on an individual basis is in many cases an unfeasible endeavor, auditing in all its forms may remain the most scalable solution for evaluating metadata quality. While sampling a catalog or other large corpus would seem like a straightforward endeavor, the perceived unity of a catalog, or any collection, can be deconstructed along several semiotic lines. It is necessary to consider what the collection as a sign body produces as a sign-function, especially given the changing nature of language, which is a point this discussion has only briefly explored.

The evolving world of digital collections and the continued aggregation of metadata from various providers brings the question of context into sharper focus, and

the nature of the sign-function gains even greater significance in these shifting contexts. Without tying a sign-function to a given context, the issue of metadata and information quality is one that will continue to persist.

The field of semiotics is one that is rich in ideas for examining critically the areas of metadata and knowledge organization. This paper has only examined the particular topic of auditing, but a logical extension of this topic would pertain to the subject of semantic overlap between various metadata schemas as a way of measuring metadata loss that occurs when transforming between schemas. Additional, but by no means exhaustive, topics of future work would concern a semiotic analysis and deconstruction of the RDF, specifically triple syntax and what the nature of its sign production entails for linked data integration. Finally, the issue of temporality is one that would also be illuminated by a semiotic analysis of various ontologies that incorporate time into their fabric and how this relates to an endurant versus perdurant approach to knowledge organization.

Notes

1. Michael Symonds, Peter Bruza, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Ian Turner, "Automatic Query Expansion: A Structural Linguistic Perspective," *Journal of the Association for Information Science and Technology* 65, 8, (2014): 1578, http://dx.doi.org/10.1002/asi.23065.

2. Dirk Lewandowski, "Evaluating the Retrieval Effectiveness of Web Search Engines Using a Representative Query Sample," *Journal of the Association for*

Information Science and Technology 66, 9, (2014), 1764, http://dx.doi.org/10.1002/asi.23304.

3. Merkourios Margaritopoulos, Thomas Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris, "Quantifying and Measuring Metadata Completeness," *Journal of the Association for Information Science and Technology* 63, 4, (2011): 725, http://dx.doi.org/10.1002/asi.21706.

4. Diane I. Hillman, "Metadata Quality: From Evaluation To Augmentation," *Cataloging & Classification Quarterly* 46, no. 1 (2009): 71-72, http://dx.doi.org/10.1080/01639370802183008.

5. Sarah L. Shreeves, Jenn Riley, and Liz Milewicz, "Moving Towards Shareable Metadata," *First Monday* (2006), http://dx.doi.org/10.5210/fm.v11i8.1386.

6. Eco, Umberto Eco, *A Theory of Semiotics* (Bloomington: Indiana University Press, 1976), 16.

7. Daniel Chandler, *Semiotics: The Basics* (New York: Routledge, 2007), 13.

8. Ibid., 30.

9. Ibid.,14.

10. John Sturrock, *Structuralism* (London: Paladin, 1986), 86.

11. Alon Friedman and Richard P. Smiraglia. "Nodes and arcs: concept map, semiotics, and knowledge organization," *Journal of Documentation* 69.1 (2013): 34, http://dx.doi.org/10.1108/00220411311295315.

12. Chandler, *Semiotics,* 218.

13. John Sturrock, *Structuralism and Since: From Levi-Strauss to Derrida* (Oxford: Oxford University Press, 1979), 9.

14. Chandler, *Semiotics,* 218.

15. Roland Barthes, *Elements Of Semiology*, trans. Annette Lavers and Colin Smith (New York: Hill and Wang, 1968),14.

16. Ibid.

17. Ibid., 16.

18. "MARC 21 Format for Bibliographic Data," Library of Congress (September 22, 2015), http://www.loc.gov/marc/bibliographic/.

19. Barthes, *Elements of Semiology,* 58.

20. Ibid.

21. Eco, *A Theory of Semiotics,* 48-49.

22. Ibid.

23. Ibid.

24. Michel Foucault, *The Archeology of Knowledge* (London: Routledge, 1989), 23.

25. Barthes, *Elements of Semiology,* 40.

26. Besiki Stvilia, Les Gasser, Michael Twidale, Sarah L. Shreeves, and Timothy W. Cole. "Metadata quality for federated collections," In *Proceedings of ICIQ04—9th International Conference on Information Quality. Cambridge, MA, 2004*: 113–114,

http://hdl.handle.net/2142/721.

27. Ibid.

28. Thomas R. Bruce and Diane I. Hillmann, "The continuum of metadata quality: defining, expressing, exploiting." In *Metadata in Practice*, Edited by Diane I. Hillmann and Elaine L. Westbrooks. (Chicago: American Library Association, 2004): 5-7.

29. Xavier Ochoa and Erik Duval, "Automatic Evaluation of Metadata Quality in Digital Repositories," *International Journal on Digital Libraries* 10, no. 2-3 (2009): 68, http://dx.doi.org/10.1007/s00799-009-0054-4.

30. Rosanne Price and Graeme Shanks, "A Semiotic Information Quality Framework: Development and Comparative Analysis," *Journal of Information Technology* 20, no. 2 (2005): 91, http://dx.doi.org/10.1057/palgrave.jit.2000038.

31. Hillman, "Metadata Quality: From Evaluation To Augmentation," 68.

32. R. John Robertson, "Metadata Quality: Implications for Library and Information Science Professionals," *Library Review* 54, no. 5 (2005): 295–300, http://dx.doi.org/10.1108/00242530510600543.

33. Hillman, "Metadata Quality: From Evaluation To Augmentation," 69.

34. Stivila et al.,"Metadata quality for federated collections,"122.

35. Ibid.

36. R. Nicole Westbrook, Dan Johnson, Karen Carter, and Angela Lockwood, "Metadata Clean Sweep: A Digital Library Audit Project," *D-Lib Magazine* 18, no. 5/6 (2012), http://dx.doi.org/10.1045/may2012-westbrook.

37. Julie Weagley, Ellen Gelches, and Jung-Ran Park, "Interoperability And Metadata Quality in Digital Video Repositories: A Study of Dublin Core," *Journal Of Library Metadata* 10, no. 1 (December 2010): 38, http://dx.doi.org/10.1080/19386380903546984.

38. Ochoa and Duval, "Automatic Evaluation of Metadata Quality in Digital Repositories," 68.

39. Hillman, "Metadata Quality: From Evaluation To Augmentation," 72.

40. Ochoa and Duval, "Automatic Evaluation of Metadata Quality in Digital Repositories," 88.

41. Westbrook, "Metadata Clean Sweep: A Digital Library Audit Project."

42. Stvila et al., "Metadata quality for federated collections," 122.

43. Hillman, "Metadata Quality: From Evaluation To Augmentation," 72.

44. "MODS Elements and Attributes," Library of Congress (September 4, 2014), http://www.loc.gov/standards/mods/userguide/generalapp.html

45."DCMI Metadata Terms," Dublin Core Metadata Initiative (June 14, 2012), http://dublincore.org/documents/dcmi-type-vocabulary/.

46. "Top-level Element: <typeOfResource>," Library of Congress (February 18, 2011), http://www.loc.gov/standards/mods/userguide/typeofresource.html.

47. "IMS Meta-data Best Practice Guide for IEEE 1484.12.1-2002 Standard for Learning Object Metadata," IMS Global Learning Consortium (August 31, 2006), http://www.imsglobal.org/metadata/mdv1p3/imsmd_bestv1p3.html.

48. “LRMI Version 1.1,” Dublin Core Metadata Initiative (October 23, 2014), http://dublincore.org/dcx/lrmi-terms/1.1/.

49. Ann Windnagel, “The Usage of Simple Dublin Core Metadata in Digital Math and Science Repositories,” *Journal of Library Metadata* 14, no. 2 (May 2014): 79, http://dx.doi.org/10.1080/19386389.2014.909677.

50. Roy Tenant, “MARC Must Die,” *Library Journal* (October 2002): http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/#_.

51. Barthes, *Elements of Semiology,* 98.

[52] Manolis Peponakis, “Libraries’ Metadata as Data in the Era of the Semantic Web: Modeling a Repository of Master Theses and PhD Dissertations for the Web of Data,” *Journal of Library Metadata* 13, no. 4 (November 2013): 344, http://dx.doi.org/10.1080/19386389.2013.846618.