The Phylogeny of a Dataset

Andrea K. Thomer* **Nicholas M. Weber***
Center for Informatics Research in Science and Scholarship,
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
501 E. Daniel St. Champaign, IL 61820
{thomer2, nmweber} @illinois.edu
*both authors contributed equally to this work

ABSTRACT
The field of evolutionary biology offers many approaches to study the changes that occur between and within generations of species; these methods have recently been adopted by cultural anthropologists, linguists and archaeologists to study the evolution of physical artifacts. In this paper, we further extend these approaches by using phylogenetic methods to model and visualize the evolution of a long-standing, widely used digital dataset in climate science.

Our case study shows that clustering algorithms developed specifically for phylogenetic studies in evolutionary biology can be successfully adapted to the study of digital objects, and their known offspring. Although we note a number of limitations with our initial effort, we argue that a quantitative approach to studying how digital objects evolve, are reused, and spawn new digital objects represents an important direction for the future of Information Science.

Keywords
Phylogenetics, significant properties, digital artifacts, fitness for use, ICOADS, climate informatics

INTRODUCTION
How do digital artifacts evolve? Do they follow a linear progression, becoming more complex and durable over time, or do they splinter, becoming more diffuse, heterogeneous and divergent? Do they mimic biotic processes or do they have unique, abiotic "life" trajectories?

For that matter, is an evolutionary account of a digital object even possible without being purely metaphorical?

Think for instance of any piece of software that is marked by a release number:

77th ASIS&T Annual Meeting, October 31- November 5, 2014, Seattle, WA, USA.

Windows 97, 2000, XP
iOS 4, 5, 6, 7
Ubuntu 4.10, 5.04, 5.10, 6.06

Each of these names signifies a change in the state of a code base as it moves from one generation of an artifact to the next. Changes between release 1.0 and 1.3 may be subtle and indistinct from brand new versions released as 3.0 or 4.0.

These numbers mark important and distinct points in the life of a digital artifact; they help us understand how to troubleshoot particular problems we encounter as end-users, and to communicate with one another about which version of an artifact we are referring to exactly.

The question then is not whether digital artifacts evolve – their markings indicate that they do in a way that is more than metaphorical. Instead the questions to be asked are, how can differences between generations of a digital object (1.0 to 1.3; or 3.0 to 4.0) be studied more rigorously? Can approaches used to study the evolution of texts, narratives, or other material artifacts help us understand how software and other digital objects evolve from one version to the next? Can these methods model the way that digital objects are reused and reworked into new "species"? If so, what properties of a digital object must be preserved or expressed to facilitate this modeling? In a digital object, what properties lead to evolutionary fitness?

This paper proceeds as follows: first, we define some basic concepts in evolutionary biology, and note important limitations in adapting these ideas for the study of evolution in digital objects. We also review a number of previous studies from linguistics, archaeology and anthropology that have taken a similar quantitative approach to evolution. We then present our case study: a phylogenetic analysis the significant properties of 99 datasets derived from, or related, to the International Comprehensive Ocean and Atmosphere Dataset (ICOADS), a long standing, widely used dataset in climate science. Finally, we describe the methods used to create a tree that visualizes how ICOADS and related artifacts have evolved over a thirty-year period. We conclude with a discussion of the relevance of this work and its potential application to different digital objects.

EVOLUTIONARY CONCEPTS IN A DIGITAL REALM
Just as living organisms do not arise through spontaneous generation, digital objects and artifacts (e.g. software,

datasets, digital images, etc.)[1] are often created through modification of existing structures rather than through entirely novel innovations. Finding methods to better understand which particular traits of a digital object are reused, and how those carry forward from one generation to the next is the main concern of this paper. Though a one-to-one mapping from the biotic world of living creatures to the abiotic context of software and digital objects would be ill advised (and impossible), the analytical and conceptual framework of evolutionary biology can help explain how different "species" of digital objects change over time. Below, we define some of the biological concepts pertinent to this study and explain the limitation to extending this terminology to a digital realm.

Evolution: Most simply, evolution can be defined as "descent with modification" (Darwin, 1859): each generation of organisms is derived-but-slightly-altered from the one before it; and all organisms are descendent from one common ancestor. In Darwin's *Origin of the Species*, he distinguishes between two types of evolution, both of which have application to our study:

Anagenesis, in which a single lineage evolves over time, and

Phylogenesis, in which a single lineage splits into two or more new species following a ***speciation event*** (Figure 1).

Phylogenetics is the study the evolutionary relationships between organisms; there is a broad range of philosophical and statistical approaches used to determine these relationships, which are typically visualized as a genealogical-style ***tree***[2]. A group of organisms that a researcher seeks to place on a phylogenetic tree is called a ***taxon*** (plural, ***taxa***). Groups of taxa that are connected by a common node (representing a common ancestor) are called a ***clade*** **(**Figure 1**).**

All phylogenetic analyses first start with the identification of ***homologous characters*** within the taxa under consideration. A ***character*** is any "part or attribute of an organism that may be described, figured, measured, weighed, counted, scored, or otherwise communicated by one biologist to other biologists" (Wiley, 1981). Two characters are homologous if "they represent corresponding parts of organisms built according to the same body plan" (Wagner, 1989). By comparing corresponding body parts of a group of organisms (for instance, the number of digits on a foot), we can create hypotheses of the evolutionary history of that group (horses have one toe, but cows have two; therefore horses lost a toe or cows gained one at some point in their evolutionary history; furthermore horses are more closely related to other one-toed animals, and cows are more closely related to other two-toed animals). More recently, the A's, G's, T's and C's of genetic sequences are used as characters; changes in these nucleotide patterns have been found to correspond with evolution over time.

In a phylogenetic approach like the one adopted here, homologous characters are identified and coded for presence, absence, or other graded states, creating a ***character matrix*** (Figure 2). Different statistical models can then be used to calculate the probability of changes in traits represented in the matrix; for instance, some models favor loss of characters over gains (thereby favoring the hypothesis that horses lost a digit), and vice versa (thereby favoring the hypothesis that cows gained a digit). The resulting phylogenetic tree is a visualization of the genealogy of the organisms being studied; it represents a hypothesized account of the relationships between taxa. This account is derived from the researcher's assessment of the organisms' traits, as well as the statistical models applied to the dataset.

The Inference of Relatedness

In biology, phylogenetic analysis is necessary because evolution produces no explicit documentation: the historical relationships between organisms must be inferred from the genetic code, the fossil record, or modern observational data of living animals' body structure (morphology).

Similarly, digital artifacts often lack the type of documentation needed to clearly understand evolutionary

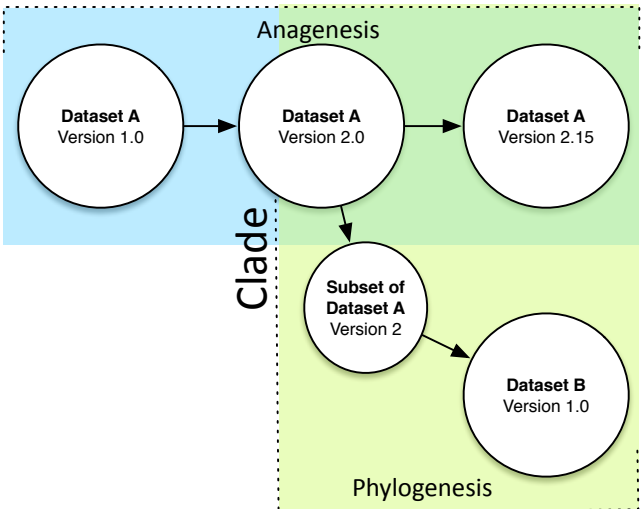

Figure 1. Anagenesis can be viewed as occurring through the sequential creation of different versions of a digital object (Dataset A 1.0 becomes Dataset A 2.15.), while Phylogenesis might occur when a subset of Dataset A is created, reanalyzed and combined with other data (Dataset C) for a new study. The subsequent datasets and their most common ancestor form a clade.

[1] Throughout we use "digital object" as a more generic term for artifacts that are absent a cultural context. .

[2] It's beyond the scope of this article to review each different approach to phylogenetic differentiation, but we justify our choice of an algorithmic approach in our Methods.

progression. When documentation about the past states of a digital artifact does exist, it often describes only a single object, and not the nuanced, networked relationships between different versions, iterations and derivations of an object. For instance, metadata about a cultural artifact may provide descriptive information (e.g. title, data created, author, material, etc.), structural information (e.g. physical or logical structure of the object) of administrative information (e.g. technical, intellectual property rights, provenance, etc.), but a single field such as "relation" in the Dublin Core schema does not support reference to the history of an object's derivations and changes. In short, most standard metadata schemas offer no way to express how a relation between objects has unfolded over time, and to what extent one artifact is genealogically related to another through those relationships.

Linked open data has the potential to express networked relationships between digital objects, but these relationships are structured in the form of a graph that is limited in its ability to express how relationships have changed over time. Similarly, provenance ontologies, such as the W3C PROV-O, are capable of expressing past relationships between objects by creating traceable accounts of actions and events, but these formalisms can only represent history in a single forward direction (Lebo, et al. 2013). Further, provenance isn't meant to track related offspring or express an overall genealogy in the same way that a phylogenetic approach can explicitly model heredity.

Provenance methods are also limited in that each requires intentional deployment; post hoc application of PROV-O, for instance, is difficult for a single project, and likely impossible for a repository's worth of digital objects. A phylogenetic approach, on the other hand, can be applied long after objects are created, in order to recover the "heritable continuity" of a network of related objects – thereby inferring the history of innovation even when explicit documentation is missing or nonexistent (O'Brien, Darwent & Lyman, 2002).

Descent With Modification

To return to biological study of evolution, inference is possible because all organisms are derived from a common ancestor: through the comparison of phylogenetically informative characters biologists can then extrapolate the order of descent.

Many digital objects are obviously not derived from one common ancestor, but cultural objects, digital or otherwise, often evolve through derivation, replication and modification in traceable ways. For example, the evolution of early email protocols followed file transfer standards over ARPANet (e.g. CPYnet), and only later split into separate, email specific user-agent protocols (e.g. POP (Post-office Protocol) and IMAP (Internet Message Access Protocol)). The clients developed to retrieve and send email, and the larger enterprise of email servicing and hosting can all be traced forwards and backwards by understanding the ways in which networks developed from these two early origins (Partridge, 2008).

This account of the history of email protocols was constructed from archival documentation, but this evolution could have similarly been compiled through comparison of different characteristics of the protocols, file transfer standards and other technologies – in other words, by examining essential material properties of each technology to infer how each "descendent" was modified from its "ancestor". As in biological organisms, we could qualitatively observe those properties, code them into a character matrix, and infer a phylogeny. In digital objects, though, the observation of homologous material properties is complicated by their lack of either an identifiable "body plan," skeleton or extractable DNA sequence. Therefore, extracting the "genetic" information of digital objects requires that we identify an analogous source of characters that are common across many digital objects. The concept of "significant properties" from digital preservation may provide us with just such a set.

Significant Properties as Informative Characters

Significant properties are defined as:

"...the characteristics of an information object that must be maintained to ensure that object's continued access, use, and meaning over time as it is moved to new technologies" (Knight & Pennock, 2009, p.163).

Traditionally, significant properties of digital objects have been seen as key to digital preservation; they are the necessary characteristics that will ensure maintenance of an object's unchanging "essence" (e.g. Lynch 1999) through different migrations. In prior work, we have focused on identifying significant properties that support the intelligible transfer of data content from one format to another (Thomer & Weber, 2012).

Here, we are interested in identifying significant properties that *survive through reuse in new forms and projects.* These properties may be additionally "significant" for their ability to help us reconstruct a set of digital objects' evolution, by acting as analogs to the informative characters used by evolutionary biologists in phylogenetic studies. Properties such as the content, context, rendering, structure and behavior (from Grace & Knight's typology, 2008) of a digital object can be coded into a character matrix, and used to build a phylogenetic tree. For many digital objects, these significant properties are expressed through an item's metadata, and in metadata record aggregations. Further, the work of identifying homologous properties is done through the process of normalizing metadata into a controlled vocabulary. Thus, character matrices and phylogenetic trees should be computationally derivable from metadata record aggregations (provided the metadata is sufficiently well curated). After creating a phylogenetic tree, identifying significant properties that *do* survive through reuse can increase our understanding of what, exactly, makes a digital

object fit for use beyond its initial creation (Palmer, Weber and Cragin, 2011).

PREVIOUS WORKS

Quantitative phylogeny of artifacts

This approach is not without precedence: application of quantitative phylogenetic methods to linguistics and textual criticism is almost as old as phylogenetic methods themselves; in fact, Platnick & Cameron argue that similar methods were accepted as standard practice in both fields before biologists came to embrace them (1977). There has been a recent resurgence of interest in phylogenetic approaches to non-biological problems partly due to computational advances in bioinformatics, which not only allow for faster and easier computation, but also support the use of "molecular clocks" to root known speciation times (sometimes called divergence points) in ways that were previously difficult or impossible (Mace and Holden, 2005; Mace, Holden, and Shennan 2005). To further emphasize the shift between biotic and abiotic studies of evolution, Howe and Windram coin the phrase "phylomemetics" in lieu of phylogenetics, "given the use of the word 'meme' to refer to a non-genetic principle that behaves in a genetic way" (2011). Though the differences between memetic and genetic evolution may have bearing on the models and algorithms used to study these processes, in this work, we use methods and software developed explicitly for phylo*genetic* work, and refer to our study as such.

Previous work in linguistics and textual criticism also borrows heavily from biogeography in coupling an analysis of linguistic divergence – how dialects differ from one generation to the next – with analysis of human migration routes (Rexová, Frynta, & Zrzavý, 2003). Similarly, phylogenies of historical texts have been constructed for literary works such as Chaucer's *Canterbury Tales* (Barbook, 1998) and *Little Red Riding Hood* (Tehrani, 2013). These approaches typically focus on finding divergence points to estimate when texts were altered, replicated or significantly changed by different authors or cultural groups.

Most immediately applicable to this study are phylogenetic applications by archaeologists and anthropologists who conceptualize artifacts as, "complex systems comprising any number of parts that act in concert to produce a functional unit," in which the "changes that occur over generations… are highly constrained, meaning that new structures and functions almost always arise through modification of existing structures and functions as opposed to arising de novo" (O'Brien, Lymen & Darwent, 2002).

This "system" view of artifacts is particularly applicable to digital objects, which may also be viewed as complex systems comprising any number of interactions between layers of information content and representation (Wickett et al., 2012). Bit sequences, encoded information content and information systems work together to produce a functional unit, and the changes that occur over generations of use are constrained by the practices and sociotechnical contexts of the groups using them.

Qualitative phylogeny & the biography of artifacts

Just as quantitative phylogenetics has a long history of application to the study of material and textual artifacts, so does the qualitative study of evolution as cultural diffusion. Anthropologists, economists and sociologists have each noted the importance of tracking the social "markings" of mundane objects that personalize, and make a given object individual to a period of time (Appadurai, 1986). In this vein, Igor Kopytoff proposed that tracking the movement of an artifact between different contexts of use required a *biographical approach* that could see "…a culturally constructed entity, endowed with culturally specific meanings, and classified and reclassified into culturally constituted categories" (1986, p. 68).

More germane to this study, Williams and Pollock describe a technique called the *biography of artifacts*, which takes a popular software platform as a unit of analysis (e.g. Microsoft Sharepoint), and attempts to trace the way it was modified, changed, and socially shaped by studying the different contexts in which it was used. The ambition of the biography of artifacts approach is to show the evolution of similar technical artifacts in different social contexts, including their adaptability (or evolutionary fitness) across diverse software ecosystems (Williams and Pollock, 2009).

Dosi and Nelson similarly relate evolutionary concepts from biology to behavioral economics and organizational theory (2003). In doing so, they relate technological change within private firms to environmental pressures in an ecology, effectively equating these externalities as selection mechanisms for evolutionary processes. Dosi and Nelson attempt to study links between organizational economics and evolutionary biology through qualitative observations of the practices, policies and technological adaptations of a firm.

A quantitative phylogenetic approach can add another dimension to each of these types of analysis. Though it cannot answer the same types of questions about how context or culture has shaped technical artifacts as used in different social settings, it can more rigidly answer when and to what extent an artifact has changed between cultures, and visualize those changes over time.

OUR TAXA OF INTEREST: ICOADS AND RELATED DATASETS

The different "cultures" under examination in this study are groups of researchers using and altering subsequent generations of a core set of digital objects; and the digital objects under examination are different versions of the International Comprehensive Ocean and Atmosphere dataset (ICOADS). ICOADS is a cooperative project between the National Ocean and Atmosphere Administration (NOAA), the National Science Foundation (NSF) and the National Center for Atmospheric Research (NCAR), which aims to provide historical marine climate

Entry_ID	Start Date	Stop Date	geo resolution	temporal resolution	Parameters
coads_1degree	1960-01-01	1997-12-31	1x1	monthly	Visibility; SurfaceWinds; SeaSurfaceTemp
EARTH_LAND_NGDC_PALEOCLIM1500AD	0000-01-01	1990-12-31	1x1		SeaSurfaceTemp
LDEO_IGOSS_MONITORING	1981-11-01		3x3	weekly	SurfaceWinds
SIPEX_II	2012-09-14	2012-11-16	5x5		MicroAlgae
REYNOLDS_MONTHLY_SST	1981-12-01		5x5	daily	SeaSurfaceTemp
SIPEX_II_Algae_Growth_Rates	2012-09-14	2012-11-16		monthly	MicroAlgae

Continous characters are "binned" into graded categories:

	StartDate	StopDate	GeographicResolution	TemporalResolution	Visibility	SurfaceWinds	MicroAlgae	SeaSurfaceTemp
coads_1degree	6	5	1	3	1	1	0	1
EARTH_LAND_NGDC	?	4	1	?	0	0	0	1
LDEO_IGOSS_MONITORING	7	?	3	2	0	1	0	0
SIPEX_II	9	7	5	?	0	0	1	0
REYNOLDS_MONTHLY_SST	7	?	5	1	0	0	0	1
SIPEX_II_Algae_Growth_Rates	9	7	?	3	0	0	1	0

Discrete states are split into individual characters for which presence or absence can be noted

In this matrix, "Visibility" is considered uninformative because only one "organism" shows a state change

Figure 2. The migration of metadata records (top) to a character matrix (bottom). Uninformative characters like the one shown ("Visibility") cannot show extensive relatedness between groups of organisms.

data to the earth science community. ICOADS is the largest source of historical in-situ weather observations, and has been used widely since its first release in 1987.

ICOADS data consist of marine surface measurements and observations (e.g. sea-surface temperature, sea-level pressure, wave swell, wind direction, etc.) that have been digitized from historical ship logs, or taken from floating buoys. As a result of the broad time periods that the dataset covers (approximately 450 years, 1662-2014) the quality and reliability of the data varies considerably.

Much like a piece of software, ICOADS is an evolving dataset with intermittent releases. Version 1.0 – called simply COADS – was publically released in 1987, and contained almost 100 million historical observations starting in 1854 and continuing to 1979.

In 2002, the project adopted a new name, the International-COADS, to reflect a growing multi-national collaboration. The latest version ICOADS (2.5) was released in 2011 and contains over 500 million historical weather observations from 1662-2014.

Each update to the dataset incorporates new data points, and also improves data processing techniques, error estimations and quality controls on these historical records. Additionally, different portions of ICOADS have been subset, reused, and integrated into new, and sometimes wholly different climatological datasets. It should be therefore possible to visualize the continuous evolutionary history of different versions of ICOADS, and their resulting offspring.

Tracing the history of ICOADS

Tracing the impact of ICOADS has proven difficult through traditional Information Science techniques like citation analysis. Each new versioned release of ICOADS results in the publication of a peer-reviewed journal article, however, the analysis of citations made to those publications offers few reliable indicators of the ways that ICOADS has been used as part of a new research project, or spawned related, derivative products (Weber et al., 2014).

ICOADS provenance records (typically in the form of metadata about ships) offer a detailed and important look at how the dataset was assembled at each stage of a new release, but this metadata fails to fully account for how data have changed between releases, and more importantly, hasn't been able to account for ways in which ICOADS has been used as a part of other derivative climate data products (Kent et al., 2007).

Thus, understanding the ways in which ICOADS evolved into new versions, and gave rise to "offspring" datasets over

a thirty-year period is the focus of the case study presented below.

METHODS

To conduct a phylogenetic study of a digital artifact we make three assumptions drawn from the discussion above:

1. The significant properties of an artifact, such as content, context, rendering, structure and behavior (from Grace & Knight's typology, 2008) are phylogenetically informative, and can be coded as characters.

2. Significant properties are homologous from one artifact to another; that is, we can compare two objects' encoding formats much in the same way that we can compare the number of toes in different animals.

3. In digital objects, a character matrix can be assembled by looking at a codebase, collection, or aggregation of standardized metadata records and noting the presence or absence of different homologous properties.

In operationalizing these assumptions we hypothesize that statistical models and software developed specifically for tracing the evolution of biological organisms will also be effective in studying the differences between versions and offspring of ICOADS as a digital artifact. We therefore expect to see datasets to form clades (clusters that stem from a common ancestor) based on their shared significant properties, and to show newer datasets as being "derived" from older datasets.

Data collection and processing

Through keyword searches, we retrieved XML-formatted metadata records from 99 different versions and subsets of COADS and ICOADS from NASA's Global Change Master Directory (GCMD; http://gcmd.gsfc.nasa.gov/), which catalogs 26,000 datasets produced by earth and space science research. These records are aggregated from federal agencies such as NOAA, NCAR, NASA and standardized using a GCMD-specific controlled vocabulary to describe the geographic coverage, temporal range and parameters contained in each dataset.

Of these 99 records, only 23 represented different versions or subsets of the ICOADS project, meaning that the remaining 76 records represented derivatives or offspring of ICOADS.

We then wrote a script to parse the XML and extract potentially informative fields (the significant properties being used as phylogenetic characters) from the metadata records. The fields harvested included: Entry Title, Entry ID, Summary, Geographic Coverage, Start Date, End Date, Geographic Resolution, Temporal Resolution, Scientific Keywords (often dataset parameters), Geographic Keywords, Sources (platform of data collection), and Instruments. Once collected, each field was converted into binary codes for "presence" or absence" of individual keywords (Figure 2). In some cases we coded additional "presence" or "absence" of characters based on the free text summaries of the records (for instance, in some cases, resolution was stated in the free text "Summary" field but not the "Geographic Resolution" field). Data were formatted as a NEXUS file for use with existing phylogenetic software (Maddison, Swofford & Maddison, 1997).[3]

Data analysis

After importing our NEXUS file into PAUP* (Phylogenetic Analysis Using Parsimony *and other methods) version 4.0a134 (Swofford, 2014), we first assessed what characters were and were not phylogenetically informative (PAUP* automatically calculates informativeness as when character state changes are shared by two or more taxa, and uninformativeness when all characters share the same state. See Figure 2 for further description).

We then created three proof-of-concept phylogenetic trees using each of models available in PAUP*:

1. A neighbor-joining distance-based tree (Saitou & Nai, 1987). These are often used as a heuristic method of assessing a dataset's quality, but are not considered the most accurate model of biotic evolution;

2. A parsimony tree, which aims to find the "least steps" tree that minimizes the number of changes (e.g. the amount of evolution) required to explain the observed characters (Farris 1970); and

3. A maximum likelihood (ML) tree (Felsenstein 1981), utilizing a statistical model specifically designed for use with morphological, or presence/absence data (Lewis, 2001). The ML algorithm essentially asks, "Given a group of taxa and a stated evolutionary model, what is the likelihood of observing a group of taxa?" The "best" tree is the one with the highest likelihood score (Page & Holmes, 1998).

Tree thinking

Phylogenetic trees must be read in a specific way to be revealing: all "branches" or forks in the tree represent speciation events (or, in this study, would mark a point at which a dataset was subset, versioned or reused by another project). In the tree presented here, branches further to the top left (closer to the tree's "root") represent older speciation events, and branches further from the root are more recent.

After producing the three trees, each was compared to a handmade timeline depicting the known history of ICOADS (e.g. was COADS version 1 shown as "older" than version two? Were COADS datasets shown to be older than ICOADS?). All three trees recovered this history with similar levels of success.

---

[3] The metadata records, coded matrix, Nexus formatted files and resulting trees can be obtained at: https://github.com/akthom/phylomemetics

In the biological community, the maximum likelihood model of evolution has been shown to be far more accurate than either the parsimony or neighbor-joining models. Therefore, in-depth analysis is focused on the ML tree. We discuss the applicability of this approach below, as well as our recognition that digital objects may evolve in a fundamentally different manner.

RESULTS

Informativeness of Significant Properties

In general, characters describing a dataset's *Resolution* and *Source & Sampling Instrumentatio*n were most informative; PAUP* also found characters describing *Parameters* to be approximately 45% informative. Characters describing geographic coverage were very rarely informative. No one category of significant property was more or less informative than others (though "structural" properties had a slight advantage) (Table 1). Overall, 42% of the characters harvested from GCMD were informative.

Trees and Clades

Several of the clades (clusters of ICOADS derivatives with a common "ancestor") produced by the ML tree (Figure 3) are intuitive, such as a cluster of COADS and its closest derivatives (Figure 3, Clade 1); the recovery of these clades shows that the ML algorithm was successful at recovering known groups of closely related ICOADS versions. This further demonstrates that the significant properties from metadata records can indeed function as phylogenetically informative characters. Additionally, the ML approach grouped ICOADS versions and subsets that were not explicitly "linked" through their metadata, implying that the algorithm was able to find similarities between versions even without an explicitly stated relationship.

The five main clades (numbered and marked in black bars in Figure 3) each have unique potential explanations for their clustering, or relationship to the earliest versions of COADS, which is a member of *Clade 1*.

Character type	# informative	# uninformative	Significant property type(s)
Resolution (timestep & gridding)	3	0	Structure
Parameters	140	169	Content, Behavior
Source & Instrument	26	16	Context, Structure
Geographic Coverage	8	59	Context

Table 1. Informative vs. uninformative characters

Clade 2 is composed of ICOADS input datasets that were originally collected from international archives. Input datasets include historical ship-logs, or entire archived catalogs provided by meteorological offices from the UK, Netherlands and Germany. Though ICOADS is a chronological descendent of these datasets collectively – each of these datasets contributed to making up the larger aggregate – here, they appear as derivatives of COADS 1. This demonstrates one of the primary drawbacks of using tree-based visualization (discussed further in the next section).

Similar to the software releases described at the beginning of this paper, the COADS input datasets are marked by a numbering system which mimics a genealogical relationship; ds540.0 is the current release of ICOADS, and most datasets that have contributed to its make up are in the ds530-ds539.9 range. The algorithm also successfully grouped two ICOADS input datasets, "NSIDC_0057" and "Indian_Ocean_Dipole" that are input datasets to later versions of ICOADS, but *not* explicitly identified as such by metadata harvested from GCMD.

Clade 3 includes a number of datasets that contain sea surface flux calculations, which capture the dynamic exchange of energy between the ocean surface and atmosphere (Friehe et al., 1991). Surface flux calculations are one of the most important applications of ICOADS data, and are used widely in the study of climate forcing, and seasonal weather events such as *El Nino* Southern Oscillation phenomenon.

Clade 4 is a cohesive cluster of COADS data products that were converted by NOAA from ASCII to the *NetCDF* (Network Common Data Form) format in preparing a highly influential climate reanalysis project. These datasets are similar in time scale, and geographic coverage, and should contain identifiable character types – such as keywords, and platforms. Though these datasets follow Clade 1 chronologically, they appear in this tree as being several additional "generations" removed from Clade 1 than we would have expected.

Clade 5 contains derivative COADS data products, most notably high-resolution sea-surface temperature (SST) datasets used extensively for sea surface flux calculations. Many of the datasets in this clade combine quality-controlled COADS data with satellite data to create high-resolution data products. For instance, a group of Reynolds SST datasets (named after their author) is heavily used in statistical hurricane models that calculate tropical cyclone intensity.

STRENGTHS AND LIMITATIONS OF A PHYLOGENETIC APPROACH

All three tree-building algorithms consistently grouped the earliest versions of COADS together (Clade 1 on Figure 3). As shown in Clade 2, the ML tree successfully clustered ICOADS input datasets, including those not specifically identified as such by the GCMD. This implies that a

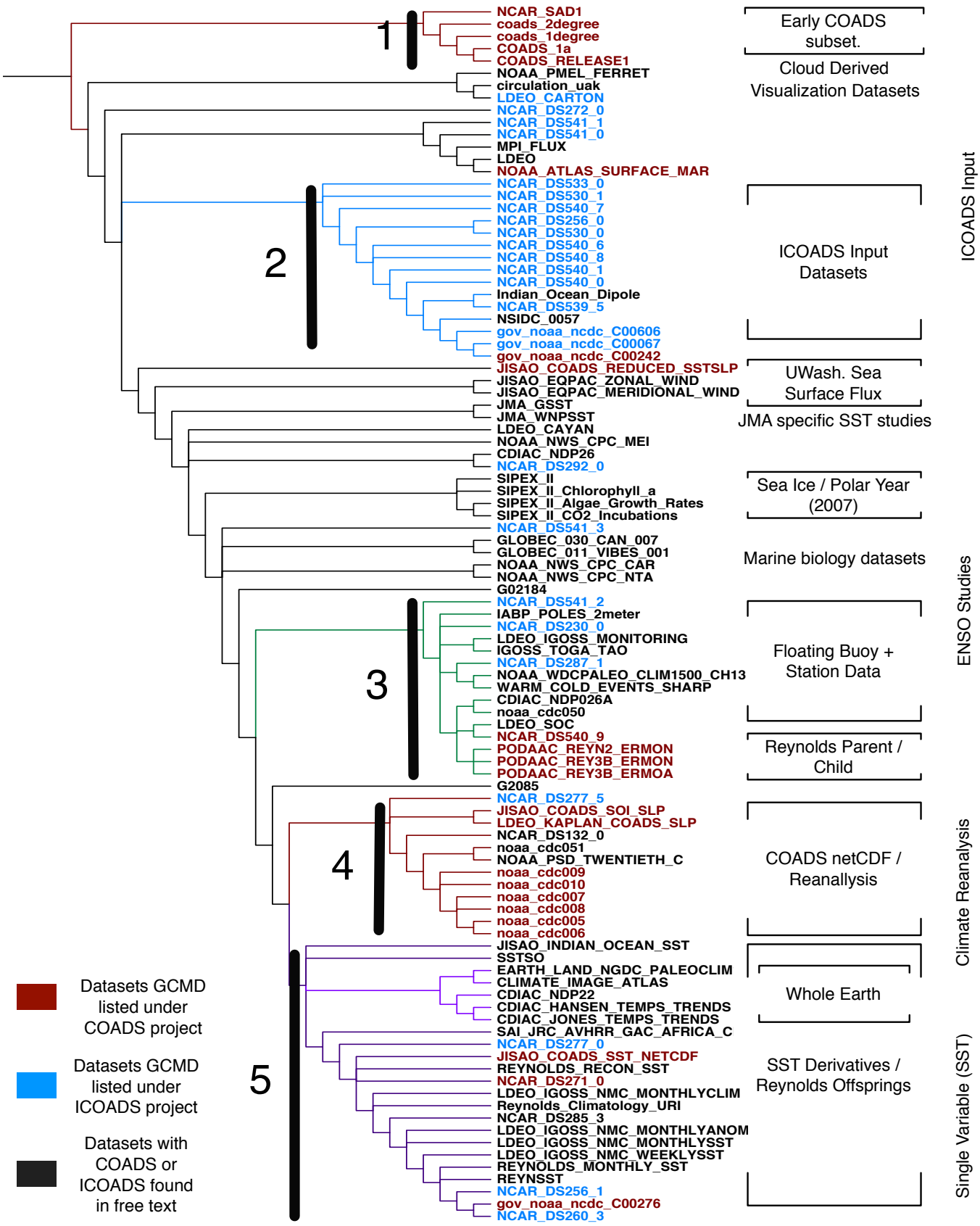

Figure 3. Phylogeny of COADS, ICOADS and derivatives created with a maximum likelihood algorithm in PAUP*

phylogenetic method has potential to not just retrieve and group records by similarity, but also can infer chronological relationships not explicitly documented in metadata records. We believe these results demonstrate the potential for extending the use of a phylogenetic method to the study the evolution of digital objects, but we first note some important limitations to this work.

Limitations in our findings

This phylogenetic approach retrieved just that: instances of phylogenesis, or of "speciation events" occurring in this collection of digital objects. However, it was less effective at showing anagenesis – versioning without speciation – which we recognize is a core method of "replication" in digital objects. A purely tree-based phylogenetic approach is also incapable of showing the exchange of traits between different lineages of digital objects, or cases in which several organisms merge into one; thus a reticulating network may be needed in lieu of a bifurcating tree.

Quality and Significance of Informative Characters

The large percentage of uninformative characters from the GCMD metadata implies that despite use of a controlled vocabulary, these records are too heterogeneous for creating a reliable character matrix. Many of the metadata records included keywords that were unique to those records and those records alone; in biology, these would be described as *diagnostic characters*: traits that could make a fine-grained, species-level classification, but could not help in the construction of a deeper phylogeny. These characters are likely integral to these objects' individual identity, but simply don't provide enough information to place them within a phylogeny.

In future attempts at deriving character matrices from metadata records, heterogeneity could be overcome by creating aggregate character groups, or binning detailed fields into broader categories. For example, geographic coverage characters such as "Canada" and "The United States" could be binned into a single category of "North America"; instrumentation characters such and "buoys" and "coastal stations" could be binned into "fixed data collection points."

This seeming disconnect between the significance of a property and its phylogenetic informativeness has implications for data curation best practices. Though it's tempting to privilege metadata describing the unique qualities of a digital object, we must also record relational metadata that situate an object within a group of related-but-different objects, thereby preserving that object's context. Significant properties of digital objects *within a collection or aggregation* are therefore not only those that set a digital object apart as unique, but also those that place an object within contextualizing groups.

On survival of the fittest for use

ICOADS is an excellent phylogenetic case study because it is a clear example of an evolutionarily "fit" digital object: not only has it adapted to new environments, but it has also given rise to numerous diverse offspring. However, the reasons for that fitness remain uncertain. It may very well be that COADS wasn't initially so much fit for use, but rather, was the only resource marine climatologists *could* use for comprehensive studies of the ocean-atmosphere interface. However, even without competition from other datasets, the subsequent versions and offspring of COADS – and the great pains taken by a large community of climate researchers to create them – represent clear, selective forces guiding the evolution of these objects. Thus, there may indeed be properties that make a dataset fit for use that have been nurtured into being, rather than selected through competition. This, again, points to the need for developing a memetic model of evolution that is unique to constructed cultural artifacts, rather than the closely related, but distinctly different models borrowed from biology.

One the most obvious properties shared by ICOADS and their derivatives is open availability: because these datasets are hosted in public repositories, anyone can freely access and use this data. A future hypothesis to test is whether open information can drive closed information to extinction: can we demonstrate that open datasets like ICOADS are widely reused, adapted and versioned that they actually cause similar, but closed datasets to atrophy in use?

FUTURE WORK

This case study relied on statistical models of biotic evolution – particularly, the maximum likelihood model which is currently largely considered the gold standard in evolutionary biology. Though these models are sufficient for a proof of concept, digital objects surely evolve and replicate in ways that are different from biological organisms, and even physical objects. This work supports O'Brien et al.'s conclusion that there is a need to review and adapt the underlying statistical models of replication and memetic procreation to better suit the mechanisms at play in cultural transmission (2000).

One particularly promising avenue for future work may be the use of reticulating networks to show not just the evolution but also the remixing of digital objects. Reticulation (exchange of traits back and forth between lineages) is much more common in cultural transmission than it is biology (Ibid.). However, the software currently designed for these studies is not as efficient as the software designed for bifurcating phylogenetic trees. More development is needed, as well as a more in-depth study of the statistical models that best mirror evolution and replication of digital objects.

Finally, future work should explore algorithms and visualization techniques that more clearly model and represent anagenesis and phylogenesis in the same graph. Future work will also need to consider ways of accounting for and identifying "spontaneous generation" in a set of digital objects: the creation of new artifacts not derived from existing materials.

ACKNOWLEDGMENTS

Thanks to Dr. Julie Allen for help with PAUP* and advice on tree interpretation, and thanks to Dr. Peter Fox for helpful conversations about provenance ontologies. Many thanks to our four anonymous reviewers for their careful readings and excellent feedback.

REFERENCES

Appadurai, A. (1986). Introduction: commodities and the politics of value. The Social Life of Things: Commodities in Cultural Perspective. A. Appadurai. Cambridge, Cambridge University Press: 3-63.

Barbrook, A. C., Howe, C. J., Blake, N., & Robinson, P. (1998). The phylogeny of the Canterbury Tales. Nature, 394, 839.

Darwin, C. (1859). On the Origin of species by means of natural selection: Or the preservation of favored races in the struggle for life. London.

Dosi, G., & Nelson, R. R. (2010). Technical change and industrial dynamics as evolutionary processes. Handbook of the Economics of Innovation, 1, 51-127.

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *American Naturalist*, 645-668.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution, 17(6), 368-376.

Friehe, C. A., Shaw, W. J., Rogers, D. P., Davidson, K. L., Large, W. G., Stage, S. A., ... & Li, F. (1991). Air-sea fluxes and surface layer turbulence around a sea surface temperature front. Journal of Geophysical Research: Oceans (1978–2012), 96(C5), 8593-8609.

Grace, S. & Knight, G. (2008) What are significant properties and why should I care? Presentation delivered at Digital Curation 101, October, 7 2008. Edinburgh, Scotland

Howe, C. J., & Windram, H. F. (2011). Phylomemetics—Evolutionary Analysis beyond the Gene. PLoS biology, 9(5), e1001069.]

Knight, G., & Pennock, M. (2009). Data Without Meaning: Establishing the Significant Properties of Digital Research. International Journal of Digital Curation, 4, 1.

Kent, E. C., Woodruff, S. D., & Berry, D. I. (2007). Metadata from WMO Publication No. 47 and an assessment of voluntary observing ship observation heights in ICOADS. Journal of Atmospheric & Oceanic Technology, 24(2).

Kopytoff, I. (1986). The cultural biography of things: commoditization as process. The social life of things: Commodities in cultural perspective, 64, 94.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., ... & Zhao, J. (2013). Prov-o: The prov ontology. W3C Recommendation, 30th April.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, *50*(6), 913-925.

Lynch, C. (1999). Canonicalization: A fundamental tool to facilitate preservation and management of digital information. D-Lib Magazine, 5(9).

Mace, R., & Holden, C. J. (2005). A phylogenetic approach to cultural evolution. Trends in Ecology & Evolution, 20(3), 116-121.

Mace, R., Holden, C. J., & Shennan, S. (Eds.). (2005). The evolution of cultural diversity: a phylogenetic approach. Left Coast Press.

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). NEXUS: an extensible file format for systematic information. Systematic Biology, 46(4), 590-621.

National Information Standards Organization (U.S.). (2004). Understanding metadata. Bethesda, MD: NISO Press.

O'Brien, M. J., Lyman, R. L., Saab, Y., Saab, E., Darwent, J., & Glover, D. S. (2002). Two issues in archaeological phylogenetics: taxon construction and outgroup selection. Journal of Theoretical Biology, 215(2), 133–50. doi:10.1006/jtbi.2002.2548

O'Brien, M. J., Darwent, J., & Lyman, R. L. (2001). Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the southeastern United States. Journal of Archaeological Science, 28(10), 1115-1136.

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. Proceedings of the American Society for Information Science and Technology, 48(1), 1-10.

Page, R. D. M., & Holmes, E. C. Molecular Evolution: A Phylogenetic Approach, 1998.

Partridge, C. (2008). The technical development of internet email. IEEE Annals of the History of Computing, 1(2), 3-29.

Platnick, N. I., & Cameron, H. D. (1977). Cladistic methods in textual, linguistic, and phylogenetic analysis. Systematic Biology, 26(4), 380-385.

Pollock, N., and R. Williams. 2009. Software & Organizations: The Biography of the Enterprise-wide System or How SAP Conquered the World. Vol. 5. London: Routledge.

Rexová, K., Frynta, D., & Zrzavý, J. (2003). Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. Cladistics, 19(2), 120-127.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4), 406-425.

Steward, J. H., & Shimkin, D. B. (1961). Some mechanisms of sociocultural evolution. Daedalus, 90(3), 477-497.

Swofford, D. (2014). PAUP 4.0: phylogenetic analysis using parsimony. Smithsonian Institute.

Tehrani, J. J. (2013). The Phylogeny of Little Red Riding Hood. PloS one, 8(11), e78871.

Tëmkin, I., & Eldredge, N. (2007). Phylogenetics and material cultural evolution. Current Anthropology, 48(1), 146-154.

Thomer, A. K., & Weber, N. M. (2013). Dispatches, digests and doodles: Exploring the significant properties of field notebooks. iConference 2013 Proceedings (pp. 937-941). doi:10.9776/13483

Wagner, G. P. (1989). The biological homology concept. Annual Review of Ecology and Systematics, 51-69.

Weber, N., Mayernik, M and Worley, S. (2014) The metric roots of ICOADS cooperation. Proceedings of the Fourth JCOMM Workshop on Advances in Marine Climatology (CLIMAR).

Wiley, E. 0. (1981). Phylogenetics: The Theory and Practice of Phylogenetic Systematics. New York: Wiley.

Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Eric Freeman, J., Berry, D. I., ... & Wilkinson, C. (2011). ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. International Journal of Climatology, 31(7), 951-967.

Worley, S. J., Woodruff, S. D., Reynolds, R. W., Lubker, S. J., & Lott, N. (2005). ICOADS release 2.1 data and products. International Journal of Climatology, 25(7), 823-842.