Anachronism in Global Information Systems: the cases of Catalogue of Life and Unicode

Robert D. Montoya[1], Seth R. Erickson[1]
[1]University of California, Los Angeles

Abstract
The Catalogue of Life (CoL) and the Unicode Standard are examples of information systems that aim toward universals: the goal of the CoL is to provide a "single integrated species checklist and taxonomic hierarchy"; the goal of Unicode is to be a "universal character set" covering the world's writing systems. In this preliminary research paper we present *anachronism* as a key obstacle in the design, expansion, and evolution of such systems. We highlight the *preservation* of concepts (of species and of writing systems) through their inclusion in these systems as an example of how such anachronisms materialize. The goal in this piece is to present a more nuanced understanding of how information and documentary systems (viz-á-viz, indexes, taxonomies, knowledge organization systems, etc.) create new, multiplicitous temporal spaces as part of their construction—knowledge that can then be applied as information professionals build these systems and subsequently evaluate their functionality and efficacy.

Keywords: Temporality; Knowledge Organization; Standardization; Catalogue of Life; Unicode

Citation: Montoya, R. D. & Erickson, S. R. (2017). Anachronism in Global Information Systems: the cases of Catalogue of Life and Unicode. In iConference Proceedings 2017, Vol. 2 (pp. 123-128). `https://doi.org/10.9776/17232`

Contact: Robert D. Montoya <rmontoya@ucla.edu>

1 Introduction

The central concern of this paper is to examine how global information systems constitute new temporal spaces that typically reformulate the temporal identities of the documents they contain. Knowledge organization systems and technical standards radiate deeply in academic and professional practices, often facilitating the global communication of data and knowledge across scientific labs, scholarly domains, country boundaries, and societies. Global information collaboration now depends on distributed computational systems that facilitate what Elizabeth Eisenstein called the "cumulative change" Eisenstein (2009, p. 113) that is essential to the progression of, not only scientific discovery, but the evolution of humanity's collective production of knowledge in all domains. Information systems establish *standards* that are fundamental to the successful collection of information and its subsequent browsability and usefulness. In particular, for the purposes of this paper, the standardization of time as both an *impetus for* and *emergent principle of* the systems will be the analytic focus.

One problem with information aggregation is identifying how systems are used to stabilize temporally emergent processes (e.g., evolution of species and elements of writing systems) in order make useable the entities constituted within those processes. Such stabilization involves a degree of temporal flattening that is problematic because some of the attributes relevant to the 'use' of these entities concern their temporal position (e.g., phylogenetic relations and whether they are extinct or not matters). This creates a tension that we understand as a kind of anachronism. We present two case studies, The Catalogue of Life (CoL) and Unicode to illustrate how temporality functions in anachronistic ways in knowledge organization systems and standards, melding numerous qualities of time into one "pleated" (Bowker, 2009) interwoven fabric. We use approaches from infrastructure studies—specifically the analytical method of infrastructural inversion—to foreground these systems' temporal effects (Bowker, Baker, Millerand, & Ribes, 2009; Bowker & Star, 2000).

This paper isn't merely asserting that information systems need to (or, as entities, *do*) evolve over time (we take this to be a given quality that defines them). This paper *is* speaking to the temporal effects—composed and unavoidable—inherent in the construction of these systems (i.e., by describing or standardizing entities through them). We describe such effects as *anachronisms* in the sense used by Christopher S. Wood and Alexander Nagel (2010) in their monograph, *Anachronic Renaissance*. Our framing of anachronism is meant to highlight how we can study and disentangle these temporal effects (which are to be understood as both positive and negative, as much as they are inevitable). Working with and around anachronism both defines a fundamental purpose of these systems, as well as one of the key challenges in creating and implementing them.

2 Temporality and Anachronism: A Framework

In their study of temporality in Renaissance art, Nagel and Wood (2010) present the Ship of Theseus as a paradigmatic example of the conflict between temporal modes. This paradox from classical literature hinges on the question of whether a ship used by Athenians in an annual ritual continues to be the *same* ship even as its original planks and timbers, rotting over time, are replaced with new ones. Is the ship's "ontological stability across time" rooted in material continuity or structural relationships? Nagel and Wood argue that these twin modes of conceiving sameness over time form a dialectic. For *some* objects at *some* times, "substitution"—replacing rotting planks with new ones—is grounds for ontological rupture; in other cases it is not. In a third category are objects that fluctuate between these poles, manifesting what the authors describe as a dual historicity (ibid., p. 31). Anachronisms are fluctuations within this third category—occasions in which an object's relationship to the past oscillates between material and structural concerns.

Continuing in this vein, global universal information systems also manifest a kind of double historicity, in that they are generally seen as both fabricated artifacts (in the sense that Tennis, below, uses the term), as well as artifacts that are meant to somehow retain the historical and narrative qualities of the documents they mean to organize and *fix* within the frame of an information structure. Like the new planks on the Ship of Theseus, we understand the "preservation" of endangered entities through classification and standardization (in order to retain some capacity to use them) as a temporal reframing in which those entities enter into new, potentially anachronistic relationships with other entities/concepts. Metatheoretical examinations of knowledge organization (KO) systems have been especially sensitive to the temporal dimensions of information systems. Joseph Tennis understands classification systems to be artifacts that evolve as an "intellectual pursuit" and offers "subject ontogeny" as a method of tracing this temporal change as classification schedules change punctuatedly over time (Tennis, 2002). Tennis has more recently articulated the notion of "second-order classification" (2015, p. 246), a metatheoretical "contour" of KO systems concerned with the changes that classification systems undergo over time. The temporal anachronisms described here are an inherent and inevitable byproduct of the "work of maintaining" these infrastructures "over time" (2015, p. 246). It is within this discourse that this study situates itself, but broadens the theory to encompass more general standardizing infrastructures, and focuses more directly on the temporal change instantiated by including concepts in the system *at all*.

3 Case 1: Catalogue of Life

The CoL is a global initiative with the primary goals of creating the most comprehensive species checklist and taxonomic hierarchy of the world's biological organisms, estimated to be upwards of 1.7 million species (Species 2000, 2015). The CoL merges biodiversity data from nearly 160 Global Species Databases (GSDs), representing all major taxon groups (including animals, microorganisms, fungi, plants, and viruses). The CoL's "management hierarchy" serves as the organizing mechanism for contributed GSD data. The hierarchy includes taxon relationships, distribution information, and nomenclatural variants that effectively can be used to centralize the knowledge output of the numerous taxonomies produced by the biodiversity community (Ruggiero et al., 2015). This structure facilitates the ready use of species data within other online infrastructures, as well as browsing of the established taxonomic tree. The CoL checklist and taxonomy is integrated in many online data aggregators, including the Global Biodiversity Information Facility, Encyclopedia of Life, and the Barcode of Life Database, to map data to accepted species nomenclature. As David Remsen indicates, "scientific names have become the primary means for referencing a taxon whenever a piece of information is intended to refer unambiguously to a particular type of organism or group of organisms" (2010). Given the importance of nomenclature as the "label" for taxa (Remsen, 2010), the CoL has dedicated itself to providing the most authoritative listing of species names for use in the globe's biodiversity data ecology. Quite separate from its function as an authoritative communication and aggregative tool, however, this examination critiques the CoL taxonomy's internal ontological "consistency" (Furner, 2009, p. 12), and asks how, as a practical knowledge organization system, it transforms the temporal qualities of the concepts it seeks to document.

3.1 Preservation of Concepts in Catalogue of Life

Comprehensive species lists are closely aligned with biodiversity conservation activities. Stemming in part from early meetings of the Convention on Biological Diversity (CBD)—an international body dedicated to promoting sustainable development—nomenclature and taxonomy have been articulated as a core and necessary practice to understanding what species exist and how such identification information can be used "for effective decision-making about conservation and sustainable use" (Convention on Biological Diversity, 2016). The CoL strives to meet this pressing need by creating a "universal and complete reference" in order to "monitor, manage and protect biodiversity resources" (Species 2000, 2015).

Mapping data about species to *names* in the CoL, however, is no trivial matter. Despite rules dictating the application of names to taxa, nomenclature is hardly simple or consistent in practice. What defines a species (its circumscription) changes over time—one species can be given many names independently, multiple species can be merged under one species name after evaluation, etc. Each circumscription is, by definition, an approximation "equivalent to generating a new hypothesis in other branches of biology," and such a hypothesis is always open to new interpretations as new forms of evidence or new modes of data analysis (computational or otherwise) are introduced (Gaston & Mound, 1993, p. 139). Without a direct link to the species concept it represents at a given *place and time*, it is difficult to know *exactly* what species concept a name refers to. Franz, Peet, and Weakley (2008, p. 64), highlight this disjoint between "names and taxonomy" and the problems it causes when a "name and its meaning evolve independently."

One of the goals of the CoL is to impose a sense of order within a nomenclatural and taxonomic landscape that is *defined* by such conceptual fluctuation and divergent practices. Names (and thus the species concepts that these names are associated with) are *preserved* and *fixed* within the system and mapped accordingly to nomenclatural synonyms, misspellings, and other variants. Names are then embedded within a management taxonomy that is used to both map species concepts together in some manner consistent with taxonomic opinion.

3.2 Temporality and Anachronism in the Catalogue of Life

How then does temporality function as an emergent property of the CoL, and how do such temporalities perform anachronistically within the system? The *management hierarchy* can serve as one example, particularly since it is the primary vehicle by which users navigate and browse the system's nomenclature. Once the CoL *stabilizes* nomenclature, those names are then embedded within a "consensus higher level classification" (Ruggiero et al., 2015, p. 2). With no "consensus among the world's taxonomists concerning which classification scheme to use for the overall hierarchy of life" (Ruggiero et al., 2015, p. 2), the management hierarchy presents an alternative taxonomic schema agreed upon by a group of taxonomic experts. The management hierarchy alphabetizes taxa "below the rank of infra-kingdom" for "easier searching by those not familiar with the phylogenies of the many taxa therein" (Ruggiero et al., 2015, pp. 9,65). This rearrangement is due to the CoL function as a *standard* more focused on discovery rather than representing a cohesive taxonomic argument. Individual GSDs (each covering different taxonomic groups) are then attached to this taxonomic framework to facilitate a shareable 'tree of life' that is 'glued' together by this central taxonomic backbone (Species 2000, 2016).

The CoL becomes an amalgam of multiple taxonomic scales and analytic approaches (phylogenetic, evolutionary, etc.), exhibiting what Nagel and Wood term a "clash of temporalities" (Nagel & Wood, 2010, p. 37). Internally consistent temporal scales, however, are foundational to biological taxonomies *used in practice*, such as those represented by GSDs brought into the CoL. "Phylogenetic frameworks" are the typical "basis for … biological classifications" (Ruggiero et al., 2015), and inherent in these estimations of the "evolutionary past" (Baldauf, 2003, p. 345) is the understanding that the species on the tree of life evolve according to some quantifiable temporal timeframe. As an example, for trees formulated with molecular data, the "lengths of the branches correspond to the amount of evolution" (Baldauf, 2003, p. 346) between two species nodes. As a temporal melange, the CoL is not an internally consistent system, limiting its use in practice. Further, alphabetizing taxa in the CoL compresses the phylogeny-specific information that is a unique aspect of interpreting a GSDs ontology. In the CoL's case, anachronisms are *concealed* by the management hierarchy implemented across the system to form a functional whole. Something extremely valuable is gained in this process of temporal compression, however, and that is the effective management of data that is essential to building biodiversity knowledge as part of a global, collective endeavor.

4 Case 2: Unicode

The Unicode Standard defines a "universal character set"—a database of letters, symbols, ideograms, and other types of characters that can be used in plain text data (Unicode Consortium, 2016). Unicode is a critical technical infrastructure as it enables consistent handling and interpretation of textual data across disparate computer systems. Over the last decade the adoption of Unicode has steadily increased; as of September 2016, close to 88% of all websites are encoded with UTF-8, a character encoding format that implements Unicode (W3Techs, 2016). Unicode is an ongoing project in the sense that the Unicode Consortium—the standard's governing body—continues to accept proposals for new characters and scripts (i.e., collections of characters forming whole writing systems). While much of the current public interest in the standardization process concerns the addition of emojis (the pictographic characters frequently used in mobile messaging apps), most additions to Unicode begin as proposals from a small but dedicated group of linguists, typographic experts, and historians committed to the project of encoding the world's writing systems (Warzel, 2016). A notable project in this respect is the Script Encoding Initiative (SEI) at the University of California, Berkeley, which facilitates the preparation of proposals for as-yet unencoded scripts (SEI, n.d.). Although Unicode version 9.0 (the most recent) includes 128,172 characters covering 135 modern and historical scripts (West, 2016; Unicode Consortium, 2016), the SEI website notes over a hundred scripts that are not yet part of the standard (SEI, n.d.).

4.1 Preservation of Concepts in Unicode

In what sense are Unicode characters concepts and how might such concepts become endangered and deserving of preservation? The answers to these questions are perhaps less apparent than in the case of the Catalogue of Life. The conceptual nature of Unicode characters is best demonstrated in the standard's formal distinction between characters and glyphs. Characters are defined as "*abstract* representations of the smallest components of written language that have semantic value" (emphasis added), whereas glyphs "represent the shapes that characters can have when they are rendered or displayed" (Unicode Consortium, 2016, p. 15). Unicode is concerned exclusively with abstract characters, not glyphs. Although the published standard includes character glyphs, they are provided for reference only; Unicode characters are not *defined* by their graphical properties. Formally, each Unicode character—or code point—is defined by a unique numerical value (expressed in hexadecimal) and a canonical name meant to convey the character's "semantic value" (e.g., the code point for the letter "i" is: U+0069 → LATIN SMALL LETTER I).

With the distinction between characters and glyphs in mind, Walsh and Hopper aptly describe Unicode as "an idealist, abstract, Platonic standard" (2012, p. 68). As the authors note, however, Unicode's idealism isn't an ontological argument about characters or the "significance or insignificance of material aspects of a document" (p. 69). Unicode defines characters as it does—as code points—in order to stabilize them as non-ambiguous, fixed entities, well insulated from the messy business of font rendering and bit-level representation. It creates a space of *usable* abstractions, not *true* abstractions. This insight is nicely put in one of the standard's accompanying technical notes: "Because the Unicode Standard is a character encoding standard and not the Universal Encyclopedia of Writing Systems and Character Identity, the stability and uniqueness of published character names [e.g., LATIN SMALL LETTER I] is far more important than the correctness of the name" (Freytag et al., 2006).

Preservation in Unicode is, thus, about stabilizing the elements of writing as computationally tractable entities—a project that becomes more urgent as our relationship to text is increasingly digitally mediated. Deborah Anderson, head of the SEI, equates standardization with preservation as follows: "While the popular media has focused on the effort to save biological diversity and endangered languages, the case for preserving the writing systems of languages is largely unnoticed. Saving scripts by including them in

Unicode will help document the variety of writing systems while also enabling their study, appreciation, and use" (2005, p. 27).

4.2 Temporality and Anachronism in Unicode

The temporal rifts and foldings—the anachronisms—of Unicode become apparent by adapting an analytic approach from infrastructure studies, by asking *when* is a character? (Ruhleder & Star, 1996). On the one hand, the formal definition of a Unicode character can be dated precisely to the publication of the Unicode version in which it first appeared. (These dates are important because, once assigned, code points cannot be re- or un-assigned). On the other hand, characters are enmeshed in historical processes that both precede and follow from the standardization process. The inclusion of a character in Unicode does not bring an end to its graphical or semantic evolution; rather, it reconstitutes the structural and material conditions within which this evolution occurs.

Consider two quite distinct cases: the Han ideograph ([chinese character]) meaning "tool" in Chinese and Japanese was introduced in version 1.0.1, in June, 1992; similarly, the pistol emoji () was officially adopted in version 6.0, in October, 2010. The temporal flattening introduced by the standardization process belies these characters' historical complexity. In the first instance, regionally-specific variants of Han ideographs like ([chinese character])—variants that are the product of centuries of use and evolution in Chinese, Japanese, Korean, and Vietnamese writing systems—are often merged in Unicode through a process known as Han Unification (Lunde, 2008). The character ([chinese character]) is seen here in a simplified Japanese form, an effect of the font rather than the underlying code point, which has no regional specificity. On a much more compressed time scale, Apple's recent decision to change the way the pistol emoji appears on its operating systems—from an image of a lethal weapon () to that of a non-lethal water pistol ()—demonstrates the fragility of Unicode's standardization efforts. Apple's change generated significant outcry as an unwelcome intervention in the emoji's semantic value (Zittrain, 2016). These brief examples illustrate that the standardization process is constitutive, limited, and a reconfiguration of the relationships within which writing systems evolve.

5 Conclusion

In this discussion we have introduced *anachronism* as a potentially useful way of understanding how the *standardization* and *fixation* of concepts within information systems creates new temporal spaces that are infrastructurally specific. We are moving toward an argument that temporal changes are not just challenges for designers and users of information systems; information systems also constitute temporal structures that merit closer examination. This project suggests several avenues for future work. Most pressingly, the themes of materiality and structural relationships suggested by Nagel and Wood need to be more thoroughly explored in relation to the Catalog of Life and Unicode. Within Information Studies (and Knowledge Organization, more specifically), work is often (rightly) focused on interconcept and intraconcept entity relationships, but such work must be more attuned to how temporality functions within these ontological spaces, and how such activity redefines concepts by their alignment with other concepts articulated under discreet and separate circumstances. Furthermore there are important distinctions between forms of "stability" suggested in these systems that should be elaborated. Such lessons can then be used to create an analytic framework that can support similar examinations of temporality within the field of Information Studies and the STS community.

References

Anderson, D. (2005, January). Global Linguistic Diversity for the Internet. *Commun. ACM*, *48*(1), 27–28. Retrieved 2016-03-27, from http://doi.acm.org/10.1145/1039539.1039562 doi: 10.1145/1039539.1039562

Baldauf, S. L. (2003, June). Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, *19*(6), 345–351. Retrieved 2015-10-07, from http://www.sciencedirect.com/science/article/pii/S0168952503001124 doi: 10.1016/S0168-9525(03)00112-4

Bowker, G. C. (2009, November). A Plea for Pleats. In *Deleuzian Intersections: Science, Technology, Anthropology.* Berghahn Books.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2009). Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International Handbook of Internet Research* (pp. 97–117). Springer Netherlands. (DOI: 10.1007/978-1-4020-9789-8_5)

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences.* MIT Press.

Convention on Biological Diversity. (2016). *Why is Taxonomy Important?* Retrieved 2016-09-15, from https://www.cbd.int/gti/importance.shtml

Eisenstein, E. L. (2009). *The printing press as an agent of change communications and cultural transformations in early-modern Europe ; volumes I and II* (14th printing ed.). Cambridge [u. a.]: Cambridge Univ. Press.

Franz, N., Peet, R. K., & Weakley, A. S. (2008). On the Use of Taxonomic Concepts in Support of Biodiversity Reseach and Taxonomy. In Q. Wheeler (Ed.), *The New taxonomy.* Boca Raton: CRC Press.

Furner, J. (2009). Interrogating Identity : A Philosophical Approach to an Enduring Issue in. *Knowledge organization*, *36*(1), 3–16. Retrieved 2015-10-02, from http://cat.inist.fr/?aModele=afficheN&cpsidt=24455069

Gaston, K. J., & Mound, L. A. (1993, February). Taxonomy, Hypothesis Testing and the Biodiversity Crisis. *Proceedings of the Royal Society B: Biological Sciences*, *251*(1331), 139–142. Retrieved 2015-11-15, from http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.1993.0020 doi: 10.1098/rspb.1993.0020

Lunde, K. (2008). *CJKV Information Processing, 2nd Edition* (2nd ed.). O'Reilly Media, Inc.

Nagel, A., & Wood, C. S. (2010). *Anachronic renaissance.* New York: Zone Books.

Remsen, D. P. (2010, February). The All Genera Index Strategies for Managing the BIG Index of All Scientific Names. In A. Polaszek (Ed.), *Systema Naturae 250 - The Linnaean Ark.* CRC Press. Retrieved 2015-06-23, from http://www.crcnetbase.com/doi/book/10.1201/EBK1420095012

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., … Kirk, P. M. (2015, April). A Higher Level Classification of All Living Organisms. *PLOS ONE*, *10*(4), e0119248. Retrieved 2016-04-24, from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0119248 doi: 10.1371/journal.pone.0119248

Ruhleder, K., & Star, S. L. (1996). Steps toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information technology and organizational transformation: history, rhetoric, and practice/Joanne Yates, John Van Maanen, editors*, 305.

SEI. (n.d.). *website.* Retrieved 2016-09-17, from http://linguistics.berkeley.edu/sei/

Species 2000. (2015). *Species 2000–Home.* Retrieved 2015-09-01, from sp2000.org

Species 2000. (2016). *Catalogue of Life - 2016 Annual Checklist : Search all names.* Retrieved 2016-09-15, from http://www.catalogueoflife.org/annual-checklist/2016/

Tennis, J. T. (2002). Subject Ontogeny: Subject Access through Time and the Dimensionality of Classification. *Challenges in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge across Boundaries: Proceedings of the Seventh International ISKO Conference*, *8*, 54–59.

Tennis, J. T. (2015). Foundational, First-Order, and Second-Order Classification Theory. *Knowledge Organization*, *42*(4), 244–249.

Unicode Consortium. (2016). *The Unicode Standard, Version 9.0.0.* Mountain View, CA: The Unicode Consortium. Retrieved 2016-09-09, from http://www.unicode.org/versions/Unicode9.0.0/

W3Techs. (2016). *Historical trends in the usage of character encodings, June 2016.* Retrieved 2016-06-29, from https://w3techs.com/technologies/history_overview/character_encoding

Walsh, J. A., & Hooper, W. E. (2012, April). The liberty of invention: alchemical discourse and information technology standardization. *Literary and Linguistic Computing*, *27*(1), 55–79. Retrieved 2016-03-25, from http://llc.oxfordjournals.org/content/27/1/55 doi: 10.1093/llc/fqr038

Warzel, C. (2016, April). *Inside "Emojigeddon": The Fight Over The Future Of The Unicode Consortium.* Retrieved 2016-06-19, from https://www.buzzfeed.com/charliewarzel/inside-emojigeddon-the-fight-over-the-future-of-the-unicode

West, A. (2016). *BabelStone: What's new in Unicode 9.0?* Retrieved 2016-09-09, from http://babelstone.blogspot.com/2016/01/whats-new-in-unicode-90.html

Zittrain, J. (2016, August). Apple's Emoji Gun Control. *The New York Times.* Retrieved 2016-09-15, from http://www.nytimes.com/2016/08/16/opinion/get-out-of-gun-control-apple.html