Between Text and Language: Unicode and the Rise of Emojis

by

S.E. “Shack” Hackney

BA, New York University, 2009

MSLIS, Pratt Institute, 2016

Submitted to the Graduate Faculty of the

School of Computing and Information in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

S. E. Hackney

It was defended on

December 14, 2021

and approved by

Eleanor Mattern, Teaching Assistant Professor, Department of Information Culture and Data Stewardship

Annette Vee, Associate Professor, Department of English

Martin B. H. Weiss, Professor, Department of Informatics and Networked Systems

Dissertation Director: Alison Langmead, Associate Professor, Department of Information Culture and Data Stewardship

Between Text and Language: Unicode and the Rise of Emojis

S.E. Hackney, PhD

University of Pittsburgh, 2022

The Unicode Standard is the de facto "universal" standard for character-encoding in nearly all modern computing systems. Unicode is what makes it possible for the order and appearance of characters within digital documents to remain consistent across time, operating system, and software. It was developed during the late 1980s as a replacement for previous-generation encoding standards such as ASCII and EBCDIC, which limited the number of possible unique characters to no more than 256. As personal and business computing expanded worldwide and the nascent public internet was looming on the horizon, these 256 character slots proved to be far too small for the orthographies of all human languages, and so Unicode—a much more capacious standard—became the dominant player.

The Unicode Standard is maintained and governed by the Unicode Consortium, a body that was formed concurrently with the development of the Standard in order to prepare the computing industry for its launch. This study situates Unicode within the history of character encoding, and reveals how the Unicode Consortium has been historically organized, why its members participate in its governance, and how the public presentation of the work of Unicode has changed over time. The original design of the Consortium, and the values of its members, reflected an assumption that these encoded characters should be understood as objects or goods in a marketplace, and the environment in which they exist as a limited resource. This research proposes a productive reframing of the Unicode Standard as a classification system rather than a "good," and uses the unique semantic nature of emojis as a focal point to examine the limits of the current paradigm.

Table of Contents

1.0 Introduction 1

1.1 Emojis 3

1.2 Purpose and Process 5

1.3 Methodological Approaches 12

1.4 Scope of this Study 14

1.5 Rationale and Approach 16

1.6 Structure of the Dissertation 18

1.7 Target Audience 19

1.8 The Researcher’s Perspective 19

1.9 Conclusion 21

2.0 Methods 22

2.1 Sense Making and Grounded Theory 23

2.2 Ethnography 25

2.3 Facet Analysis 26

2.4 Methodologies 31

2.4.1 Historiography 31

2.4.2 Network Analysis 32

2.4.3 Formal Analysis + Content Analysis 33

2.4.4 Telling the Story 34

3.0 Character-Encoding in Context 35

3.1 Organization of this Literature Review 36

3.2 Standards 38

3.2.1 What does Standardizing do? 38

3.2.2 Politics and Power of Standards 39

3.3 Economy of Standards 42

3.3.1 Locating the Unicode Standard 43

3.3.2 Consortia 45

3.3.2.1 The Unicode Consortium 48

3.3.3 Goods, Commons, Markets 52

3.3.3.1 Tragedy of the Digital Commons 54

3.3.3.2 Knowledge Commons 55

3.4 Text 56

3.4.1 What is Text? 57

3.4.1.1 Digital Text 57

3.4.1.2 Global Languages and Digital Text 59

3.4.2 Politics and Power in Computing Systems 61

3.4.3 Translation 63

3.4.4 Documents 65

3.5 Classification for Retrieval and Access 67

3.5.1 MARC and FRBR 68

3.5.2 Cataloging and Classification 72

4.0 History of Character-Encoding 79

4.1 Machine-mediated communication 80

4.2 Character Encoding Before Unicode 83

4.2.1 Mechanical Codes 83

4.2.2 ITA/ITA2 85

4.2.3 FIELDATA 88

4.2.4 ASCII........90

4.2.5 EBCDIC and the P-Bit 92

4.3 Creating Unicode 94

4.3.1 Compatibility........95

4.3.2 Linguistic Considerations........96

4.3.3 Technical Considerations 98

4.3.4 Market Expansion........101

4.4 Conclusion........ 101

5.0 Emojis 💯 104

5.1 The Birth of Emojis........ 104

5.1.1 What are Emojis and Where Did They Come From? 105

5.1.2 The first emojis........107

5.1.3 Bringing Emojis to Unicode 111

5.2 Emojis as Objects 116

5.2.1 Regulating Meaning........118

5.2.2 Emojis as Visual Objects 121

5.2.3 Emojis as Commercial Objects 128

6.0 Unicode and the Public 133

6.1 Unicode.org 134

6.1.1 Unicode.org Homepage..134

6.1.2 Unicode.org as Networked Document 140

6.2 Adopt A Character 148

7.0 Discussion, Findings, and Future Research... 151

7.1 A Catalog of Characters 151

7.2 Changing Linguistic Environment... 154

7.3 Decolonizing an Information Commons... 155

7.4 ALA Encoding Standard 156

7.5 Conclusion.. 159

References.. 161

List of Tables

Table 1 Data sources and types collected from Unicode.org website 10

Table 2 Data types and sources for Unicode.org data 11

Table 3 Data types and sources for non-Unicode.org data 11

Table 4 Research Facets 29

Table 5 David and Greenstein's taxonomy of standards (1990) 42

Table 6 Voting Members of the Unicode Consortium 50

Table 7 Four basic types of goods, adopted from Ostrom 2005, p. 24 53

Table 8 Smiling Face Emoji Presentation 124

Table 9 Characters necessary to create the "Woman Mechanic: Medium-Dark Skin Tone" emoji 128

List of Figures

Figure 1 Shannon and Weaver's Mathematical Model of Communication (Shannon and Weaver 1949, p. 34).. 64

Figure 2 Diagram of the visual telegraphy system developed by Claude and René Chappe85

Figure 3 Code table as presented by Baudot in the patent for his teletype machine........... 86

Figure 4 A reference chart for ITA2, showing how each character would appear on the punch tape output.. 88

Figure 5 FIELDATA Code Table, as presented by Tom Jennings 90

Figure 6 Unpacking UTF-8 codes... 100

Figure 7 Venn Diagram of Types of Encoded Characters Within Unicode 106

Figure 8 SoftBank's 1997 set of monochromatic emojis.. 108

Figure 9 Whale emoji on J-Phone 109

Figure 10 Kurita Emojis for DoCoMo 1999 110

Figure 11 (https://twitter.com/sigekun/status/1080848236653334529) 110

Figure 12 LATIN SMALL LETTER J 117

Figure 13 PILE OF POO... 117

Figure 14 Evolution of Pistol Emoji 2013 – 2018 120

Figure 15 Slightly Smiling Face 🙂 122

Figure 16 White Smiling Face ☺ 122

Figure 17 Variation Selector-16, U-263A ☺ + U-FE0F = 😊... 123

Figure 18 Hashtags and Accompanying Hashflags for #WorldAIDsDay2021 and #AskSnoopDogg. 129

Figure 19 Unicode.org homepage as of January 26, 1998 136

Figure 20 Unicode.org homepage, July 1, 2008. Retrieved from the Internet Archive 04-22-21 137

Figure 21 The Unicode.org homepage ten years later, on August 1, 2018. 137

Figure 22 The Unicode.org homepage (now using the home.unicode.org subdomain) 4-22-21 138

Figure 23 Basic Layout of the Unicode.org Website. 142

Figure 24 Basic Layout of Unicode.org Website, with Emoji pages highlighted in pink ... 143

Figure 25 Network Visualization of the Unicode.org Website, with Nodes colored by Page Type 145

Figure 26 Unicode.org Page Content Type Counts 147

Figure 27 Tweet announcing the Oakland Athletic's sponsorship of several emojis 149

1.0 Introduction

What I am about to tell you is a story about language. It is a story about the technology which allows us to tell stories, and about the people who developed one of those technologies: The Unicode Standard, which has become essential to how we write and communicate in the contemporary digital world. It is also a story about corporations, markets, and goods and the dangerous intersection where language becomes a commodity to be bought and sold. And it's also about what counts as language, what does not, and who gets to decide where that line is drawn.

Unicode is a digital character encoding standard— it is the system which currently ensures that when I type the letter "a" on my computer, it will show up as an "a" on your computer, regardless of operating system, software, or font. Fundamentally, the promise of character consistency via encoding standard is what has allowed global digital communication to flourish. By ensuring that messages can be typed, sent, and read using the exact same characters, the loss or undesirable transformation of characters is prevented. This is an essential part of ensuring that the meaning of the communication arrives intact and is able to be interpreted correctly by the receiver (Shannon and Weaver, 1949). The implementation of Unicode, its consortium, and its governance practices are major pieces of digital infrastructure history, and as such, this research seeks to examine it with a critical eye. Updating and maintaining character encoding standards— like any piece of infrastructure-- is essential to our ability to exist in the world that infrastructure is designed to uphold: one of documents, texts, emails, tweets, and myriad other forms of digital communication.

Over the course of this dissertation, I will tell you the story of how character encoding as we know it today was formed, and who manages it today. And I will introduce you to the X factor which has driven a(n as-of-yet largely imperceptible) wedge into the fabric of Unicode's infrastructure, emojis. Perhaps you've heard of them. Little thought has been put, I am afraid, into the theoretical consideration of the infrastructure for digital writing from a linguistic or human-information-seeking perspective. This work seeks to correct that oversight, and stake a claim for LIS, for classification and information infrastructure studies in particular, to lend their expertise to the bit of a pickle we seem to be headed towards, both linguistically and sociotechnically. It all stems from, as it often does, the difference between what someone says they can do, and what they are actually [capable of] doing.

I seek to present Unicode broadly as a lexicon of semantic objects, a contextual born-linked document, and an entity for the allocation and regulation of digital goods. I demonstrate the ways in which the Unicode Consortium treats its encoded characters as digital goods, existing in the finite marketplace of the Unicode namespace, and how this market-focused treatment creates a seeming scarcity within the character encoding system, both in terms of available space within the standard as well as access to and influence during the character creation process. This dissertation explores the history of digital character encoding, from its inception at the beginning of digital computing, through the launch of Unicode in 1990, and into the present day, where it oversees an ever-expanding lexicon of scripts and characters—including the set of pictographic characters known as emojis.

The original programmers and designers of Unicode and its predecessor encoding standards for text transmission placed their primary focus on the ability for "text packets" to be moved, stored, and translated as efficiently as possible by the computer (Bemer 2018). Only as these systems expanded to different orthographies and use contexts did considerations such as semantic reliability—and human readability-- come into play. When this technology was being invented, the digital computer didn't exist, wasn't considered something everyone would have access to the way we do today, there was always an 'operator' to mitigate between the machine and the public. Character encoding moved into the digital age as a means of efficient wartime communication, and within a specific linguistic, historical, and sociotechnical environment, which has ever since required everyone else to morph and stretch to its constraints. I track this evolution through the history of character encoding systems, with a special emphasis on emoji characters as the liminal objects which have opened the Pandora's box, so to speak, of semantic encoding, as these characters rely on a subjective interpretation of their visual representations, and are not associated with any pre-existing language or script.

1.1 Emojis

Emojis are "picture characters" and "are often pictographs […] or icons that represent emotions, feelings or activities" (Emojis and Pictographs, 2018) that can be typed in any compatible text entry field, as they are *classified* as textual objects by the computer software which interprets each character's unique identifier from pieces of binary code. While emojis are treated identically to other Unicode characters in terms of how the rules of the Standard are applied to them, and (for the most part) their object function within computer processes, the image-like appearance of emojis to human viewers causes a significant deviation in the ways that they are treated by us. They are treated as affective markers and nuanced emotional reaction images and therefore carry a great deal more, and a great deal more subjective, semantic information than a ‘normal’ encoded character—that is to say, we use them to indicate or clarify the emotional tone of a text-based communication (Alshenqeeti and Hamzda 2016, Barbieri and Camacho-Collados 2018, Gawne and McCulloch 2019).[1] The current system of understanding Unicode characters and namespace is simply not equipped for the documentation and regulation of these types of characters, due to the fact that Unicode intentionally shies away from attaching prescribed meanings to the characters they encode. However, since their earliest introduction on Japanese cell phones, emojis have seen massive worldwide popularity among users, and have been actively pursued as potential characters to be included within the Unicode Standard. The tension between emojis and the ‘intended purpose’ of the Unicode Standard as a digital repository of unique identifiers is the point of entry for this research, and the gaps between the stated scope of Unicode and the range of its contemporary application serve as a means to observe and analyze the power structures which undergird character encoding and all digital standards.

[1] Emojis can, of course, also be used on their own, independent of contextualizing words. See Chapter 2 for a discussion of some of the works with and about emojis in this area. For the purposes of this work, we will focus on emojis’ usage as affective markers in conjunction with the use of traditional orthographies, as we are concerned with the convergence of emojis and text in digital spaces where humans and machines are both reading messages written with them.

1.2 Purpose and Process

I'll begin by examining the underlying historical and social elements at play in the creation and adoption of digital character-encoding standards-- focusing on the development and implementation which led up to the creation and development of Unicode as a character encoding standard over the past 80 years. My aim here is to present this history in light of what we know about the linguistic landscape today, and point out and pick apart the assumptions, practices, and histories deeply embedded in the structure of Unicode as it has evolved. I analyze Unicode as an essential piece of digital infrastructure, and demonstrate how that structure draws boundaries around what is and is not a meaningful piece of textual information in digital space. By working to connect Unicode to the larger infrastructural network of contemporary digital society, I seek to better understand and define the relationship between the Unicode Standard, its governing Consortium, the characters it contains, and the public who uses them.   I believe that by reframing and explicitly defining Unicode's underlying structure as classification system, we will be better able to handle the expansion both in terms of scope and mass of Unicode which faces us in the coming years (Stone 2003, John 2013, Mueller 2010).

The formalization and standardization of writing and its related forms dates back far longer than Unicode, or even the printing press. We may often think of language and writing as a means of self-expression or artistry, and in so many instances it is—but the origins of writing were decidedly more practical. The earliest known preserved writing, Ancient Sumerian tablets, show that humans began writing things down as a memory aid, specifically with regards to the buying, selling, and transportation of livestock and other goods (Bonvillain 1993, Robinson 2002). This economic framing is important in considering the commodification of language through the record-keeping practices involved in building an infrastructural standard such as Unicode, and we will touch on how it continues to pervade the ways we conceptualize text throughout this writing.

Communication via text relies on a shared understanding of the relationship between form and meaning. Systems of formalization create prescriptive meanings for their constituent parts, thus standardizing their use across larger communities, streamlining and creating interoperability for previously disparate systems (Busch, 2011). Personally, I am interested in digital character encoding and the establishment of Unicode as the standard for such encoding on the web because of my own background in linguistics and ongoing research focus on information infrastructure and classification systems. The addition of emoji characters to the Unicode mix appeals to my interest at the intersection of the visual and the textual, and the meaning-making processes that separate (or merge) the two. I seek to make visible the infrastructure of one of the fundamental aspects of how human interaction happens via computers—the (type)written word-- and propose ways to enrich this structure towards the goals which Unicode has set out for itself regarding language preservation and documentation. Because so much of electronic text is governed by Unicode, it is important to know how Unicode works and how it conceptualizes text, both in theory and in practice. It’s also important to know not only how text is transmitted and interpreted by human readers, but by the computer parsers, programs, and many other processes that also must understand the format (if not the content) of that text.

Unicode says that its goals are to document, preserve, and allow the use of “all human languages,” but its current structure and character-expansion policies explicitly prevent this from being possible, as there are a limited number of spaces within the Standard while languages and their documentation continue to expand. If, as they state, Unicode wishes to preserve digital text for future generations via its promise of endlessly extensible backwards-compatibility, the definition of what they govern (text characters), and what aspects of those characters they have control over (their visual presentation on compatible devices), must be in line with both these goals and their ongoing practices.

Considering text as a commodity—as I argue is the current paradigm under which Unicode manages the character encoding process-- and using it as a means of affective data-mining in the digital world is incompatible with these goals, and serves the economic interests of the Consortium's membership, rather than the global user base for Unicode characters. This is particularly evident, as I will demonstrate, in the realm of emojis, and suggests that their addition to the Unicode Standard was less about language preservation and compatibility, and much more about cornering the market on text-based emotional expression for the purposes of seemingly endless economic growth. The written word is a tool that (nearly) the entire world relies upon, and Unicode holds a monopoly on its existence in the digital realm, with little to no outside regulation or oversight of its practices. Consider the following a warning to all of us about the precarious linguistic situation we find ourselves on the verge of today, whether we realize it or not. But I will not leave you empty-handed: I propose an alternative paradigm of organizing and understanding digital character-encoding—that of a classification system.

While this research focuses largely on 20th and 21st century developments in character encoding, I refer to the "interoperability" of all historical writing and communication systems using the language of contemporary technical standards, because being able to read and write the same language, using the same alphabet, is creating social interoperability—alphabets, characters, and text—are all technologies. Likewise, today's technical standards, such as the Unicode character encoding standard, are also social infrastructure, describing, stabilizing, and creating boundaries for what counts as socially meaningful forms of communication. While Unicode is the result of a long, contiguous line of textual technological innovations, the insertion of sprawling, diverse, and contradictory human writing systems into the rigid organization structure of the computer necessarily creates difficulties in the adaptation of languages and writing systems to the digital form, as well as in the long-term digital preservation of the documents they create.

As we have seen with many digital infrastructure systems, the exponential growth of computing power and the subsequent ubiquitous sociotechnical embeddedness of digital text in our lives has created shortcomings and contradictions within the Unicode Standard that could not have been predicted. The purpose of this dissertation is to examine how Unicode has come to be in a position of unparalleled power over the ways that text is produced in contemporary culture.

This dissertation, however, does not focus solely on the character-based minutia of the contents of the Unicode Standard, but also on its place as a largely invisible but extremely powerful part of the ideology which has shaped the modern web—both via the availability of certain characters within the Standard itself, but also as an active structure- and boundary-building presence within the shared virtual space of the internet. This requires several levels of abstraction from characters, text, and their use by individuals, to the rhetorical and semantic role of Unicode as an entity in and of itself, which both shapes and is shaped by the cultural foundations and development of the sociotechnical system of the internet. Like other forms of language-based communication, it is nearly impossible to describe Unicode without the use of Unicode—I would have to write this dissertation by hand or on a typewriter in order to escape it! Unicode's governing body, the Unicode Consortium is likewise dependent on the Unicode Standard in order to create, document, and share the means of governance of its own digital character standard. Therefore, this research cannot focus solely on the content of the Standard or its everyday usages, it also needs to analyze the ways in which Unicode itself is presented to its audiences and users across an international digital universe because that presentation itself is a part of the “good” that Unicode creates and promotes.

With this purpose in mind, I began my research by collecting data on the structure of the Unicode.org website over the history of its existence. A website is a born-digital document and doing historical archival research on born-digital documents has some distinct advantages. Websites are a series of hyperlinked pages, forming a structured network of documents, and the relationships between these documents can be just as meaningful, and often more complex, than the order in which pages appear in a book (Raley 2016). Additionally, the metadata, markup, and record of changes to born-digital documents are retrievable across time, again adding to the structural and contextual information available from a website. I have made use of a combination of data sources from the Unicode.org website, including the content of the webpages, the structure of the site as determined by links between pages, and the historical context of each page as it has developed over time. The hyperlink-based network of the Unicode.org site also serves as a source for the explicit record of the pages of the site, and their connections to one another are treated as a *text* that is able to be read both for its explicit content, as well as for the implicit reflections of its creators’ goals and assumptions during the standard-making process.

Additionally, the Unicode Consortium has made substantial use of the Unicode.org website as a repository for official internal documents relating to the Standard and its development. Documents such as meeting agendas, committee listservs, press releases, and even the remnants of a virtual “Unicode Museum” are available from the Unicode.org site, and this research makes ample use of these sources to document the history of Unicode as well as to provide contextual information about the people and policies creating Unicode and presenting it to the public. See Table 1, below, for a full list of research materials gleaned from the Unicode.org website.

Table 1 Data sources and types collected from Unicode.org website

Data	Source	Type	Method
Networked hyperlink data	Unicode.org, Internet Archive	Node/edge pairings, CSV	Web scraping
Page content data	Unicode.org	Text	Web scraping, hand-selected pages
Consortium listserv data	Unicode.org	Text	Available for download from site
Character tables	Unicode.org	Text, CSV	Available for download from site
Meeting agendas, internal memos, press releases, etc	Unicode.org	Text, PDF	Available for download from site

In addition to data retrieved directly from the unicode.org website, I have also collected primary and secondary historical sources relating to the development of character encoding standards in the 1980s as a part of a historiographic review of the character encoding landscape before Unicode. Specific sources associated with each category of data are listed below in Tables 0.2 and 0.3.

Table 2 Data types and sources for Unicode.org data

Data	Source	Type	
Networked hyperlink data	Unicode.org	CSV	Retrieved via webscraping, used to create network visualization
Page content data	Unicode.org	Text	Selected based on location within networked site map and page subject (relation to emojis)
Consortium listserv data	Unicode.org	Text	Available for download from site, full-text searched for threads concerning important emoji releases
Emoji committee meeting minutes	Unicode.org	Text, PDF	Retrieved based on links and references in listserv data and page content data
Nameslist	Unicode.org	Text, CSV	Human-readable version available on site, provides formal names and designations of Unicode characters
Emojis	Unicode.org	Text	Selection based on mentions in Unicode listserv and meeting minutes, as well as presence of discourse in popular media

Table 3 Data types and sources for non-Unicode.org data

Data	Source	Type	
Historical documents	https://sr-ix.com, https://bobbemer.com, Internet Archive	Text, PDF, Archived website	Contains commentary interview text, annotation of diagrams for encoding models

The data from outside sources also tries to get as close to primary source documents as possible, but much of the documentation of character encoding development before Unicode was formalized in 1990 has been lost. For this reason, I make extensive use of the documents and commentary retained on the personal websites of Tom Jennings and Bob Bemer, both of whom were intimately involved with the development of modern character encoding.

This dissertation, therefore, approaches the question of how the historical governance of digital text culminated in the creation of Unicode and its goods-based approach to character encoding by drawing as much as possible on primary- and secondary- source documents from Unicode itself, and analyzing not only what Unicode has to say for itself, but also how, when, and to whom it says it, as well as who it associates with. What results is a demonstration of the insidiousness of not only our embedded infrastructural technologies, but also of our economic paradigms, and our understanding of how knowledge is created through the circulation of ideas.

1.3 Methodological Approaches

The methodologies for this project are intentionally cross-disciplinary, and seek to analyze the Unicode Standard and its supporting documentation through both humanistic and information-focused lenses. My home discipline being the meta-field of Library and Information Sciences, I naturally look for relationships between and across methodologies, and many different strands of research methods are braided together to make this dissertation. All of them used together in this context have come together to form an approach I am calling Faceted Methods Analysis, because of the multiple methodological stances I take both to the material object of my research as well as in considering its theoretical place within its larger infrastructural context. Faceted Methods Analysis is not an entirely new methodology, but rather an umbrella term which I deploy here to describe a confluence of similar sets of methods that have been developed and applied across separate disciplines from Organization Studies to Anthropology.

Faceted Methods Analysis draws its name from seminal LIS scholar Ranganathan's faceted analysis approach to classification systems, which seeks to organize information objects according to a defined set of facets, which can then be related to similar works via each facet horizontally, rather than through the top-down perspective seen in other popular classification systems such as the Library of Congress Subject Headings (LCSH) or the Dewey Decimal system (Frank and Paynter 2003, Mitchell 2001). In practice, Faceted Methods Analysis serves as a framework for the ethnographic study of digital infrastructural systems and seeks to take the object of research (in this case, the Unicode Standard) and approach its analysis from a variety of methodological and theoretical perspectives, in order to develop a richer analysis of the subject at hand, and importantly, its relative position and relationship to its environment and peers across facets. I consider FMA especially apt for the study of infrastructural digital objects, such as the Unicode Standard, because of the innate embeddedness of such infrastructure, and the 'born-linked' nature of digital information objects.

The research actions I have undertaken under the auspices of Faceted Methods Analysis include historical analysis, systems analysis, and multiple digital humanities methods, including web scraping, network analysis, and content analysis. The historical analysis contributes contextual information about the creation and development of Unicode, as well as being a means of identifying and collecting source materials for analysis. The systems analysis provides technical, structural information about the functioning of Unicode as a system for character encoding and standardization, and the digital humanities methods offer direct insight into the technological structure of the Standard, and its instantiation in the digital world.

Faceted Methods Analysis and its related methods are intended here to contribute a non-technical, humanistic understanding to the historical and systems analyses, in order to better understand the choices made by the creators of the Standard, and members of its Consortium. These methods are designed to be able to answer big picture questions about Unicode and the sociotechnical implications of digital standards generally through careful and focused examination of specific documents, their structure, and the audience who interact with them.

1.4 Scope of this Study

As previously stated, the purpose of this research is to create a clearer and more holistic view of Unicode, as a standard, as infrastructure, and as a piece of contemporary information governance. I use emojis as the entry point into the world of Unicode because not only do they represent, as I show, a shift in the way that the Unicode Consortium presents itself to the public, but also because emojis bring semantically- and contextually-rich visual characters (back) into the conversation around language and communication within and via digital computing.

This research focuses heavily on the ways that emoji characters differ from the other types of orthographic characters documented by Unicode, and the implications of them for our conceptions of text; it is not a case-by-case analysis of the meaning(s), use, or appearance of individual emoji characters. There is a great deal of popular journalism and other writing about the explicit as well as hidden meanings of individual emojis, and their visual representations across platforms. While I will refer to specific emojis and their meanings as examples or as case studies to demonstrate a larger point about the role of emojis within the Standard, I leave more in-depth analysis of specific characters and their uses to linguists and rhetoricians.

While focusing on emojis, this research also draws upon and points to the long struggle of other languages and character sets which do not translate so easily to the Western alphabet character-based encoding system of not only Unicode but its predecessors. This is particularly true for Chinese, Japanese and Korean (CJK), and while this project draws on the findings of this body of research, a thorough historical analysis of CJK languages and digital character encoding is beyond the scope of this research.

Finally, this dissertation sets out a theoretical proposal for (re)considering how the digital object of text is treated by humans as they interact with computers. It does not, however, present guidelines or direct suggestions for changes to how computer hardware or software functions. As my historiography shows, the long road from early analog encoding to the Unicode Standard of today has built a firm foundation of technical and conceptual standards addressing not only text, but the overall nature of how information is stored and accessed within a computing system. And while part of the purpose of this research is to explore the assumptions, biases, and contradictions of the palimpsest which has become the de facto global character encoding standard for digital text, I have not, in this document, created a replacement for Unicode, nor do I wish to place blame on solely on Unicode for the shortcomings in today’s digital text environment.

1.5 Rationale and Approach

I begin by outlining the history of character encoding to follow the lineage of the first character-level mass communication codes, through to the encoding standards for US military machine operation, and the direct route from there to the more familiar standards of the modern computing environment, EBCDIC, ASCII, and eventually Unicode. Using this historical perspective, I discuss the ways that these early types of character encoding focus on the communication of characters as objects to be transferred from computer to computer efficiently as possible, rather than as semantically-valuable objects on their own, both for reasons of technical streamlining, but also as the narrow-sighted result of a Western-centric perspective on both writing systems and languages across computing, linguistics, and communication studies at the time. I situate the Unicode Standard within the larger contexts of digital computing and communication during the years leading up to and during the Standard's founding. By comparing Unicode to its competitors in the late 1980s we can see the spectrum of options at a critical crossroads in the development of digital technology, as well as pinpoint the priorities, interests, and biases which made Unicode so much more successful than its competitors.

Continuing this narrative, I introduce emoji characters; their original development history, and eventual relationship to the Unicode Standard. Using specific character examples, backed up with supporting documents from the publicly-available Unicode Consortium records held on the Unicode.org website, I demonstrate some of the challenges of adopting emojis as a part of the Standard, and how these technical and rhetorical nuances influence Unicode's ability to live up to their own stated mission. I show how the addition of emojis to the Standard introduced a new level of visual and semantic complexity to the nature of what exactly it is that Unicode encodes, and demonstrate that the increasing popularity and demand for emoji characters has pushed Unicode further into the public eye.

I then discuss how this objectification of encoded characters which has lent itself to the ongoing treatment of the Unicode Standard and its characters as a limited resource, governed by a consortium using that resource as a potential source of profit—this includes Adopt a Character, a long-running effort to commodify Unicode characters, notably taking advantage of the ‘first wave’ of emoji popularity, as well more contemporary integration of emojis into all aspects of digital life. Having done this, I discuss some of the contradictions and fallacies necessary to maintain Unicode’s current structure, with examples of languages and scripts that contradict or are otherwise left out of the current intellectual paradigm of character encoding— what was lost by focusing on the rapid transmission of characters as “marks” as opposed to other aspects of the written word, and what languages and other potential character uses are inconvenienced by the way that the Standard has been organized.

Subsequently, I propose an alternate paradigm for the organizational understanding of digitally encoded characters: that of a classification system. Specifically, I propose that Unicode is better understood as a born-networked expansible classification system for the orthographic systems of human language in digital environments. This alleviates some of the difficulty with the backwards compatibility, permanence, and ‘space’ issues which Unicode and its predecessors have had to deal with in constructing character-encoding standards. I also speak to a *cataloging mindset* towards character-encoding, and this perspective’s ability to undermine the application of a top-down semantic perspective on encoded characters. The born-networked nature of digital characters, especially emojis, allows for this kind of multi-faceted approach to both the assignment of meaning as well as the building of relationships between these objects.

Finally, I look forward to the future of emojis and of digital text, anticipating future bottlenecks and particular areas of expansion within the current structure of Unicode. I also suggest ways to expand the conceptual paradigm of what digital text is and how it is regulated. I do this with a particular focus on the ever-changing nature of human language usage, and potential methods to avoid a primarily-prescriptive encoding of characters and their meanings.

1.6 Structure of the Dissertation

From here on out, this dissertation proceeds in the following way: Chapter 1 discusses my methodological approach to this research, and introduces Faceted Methods Analysis in depth. Chapter 2 places character-encoding (and Unicode in particular) within the context of the larger worlds of digital infrastructure, text, code and encoding, and standardization. This literature is connected with documentalist and classification theory work within LIS, situating the reader within the world of language and standards we will explore together. Chapter 3 begins the in-depth history and formal analysis of character encoding standards before Unicode, as well as the state of the field as Unicode began to be developed. Chapter 4 introduces emojis, outlines how they became a part of the Unicode Standard, and discusses their unique semantic role within character-encoding and digital communication. Chapter 5 considers Unicode as a public entity, examining the ways that the Standard is presented to the public, and how its governing consortium positions itself relative to its audience and the contents of the Standard itself. Finally, Chapter 6 pulls together all of the threads from the previous chapters to consider exactly how Unicode "sees itself" within the digital universe, and proposes an alternative intellectual paradigm which moves Unicode out of the marketplace and more firmly into the realm of language and digital preservation.

1.7 Target Audience

This dissertation addresses the following audiences: Academics in LIS and STS, and to a lesser extent, academics and practitioners in Computer Science. It is my hope, however, that this research will also be relevant to non-academic readers and users of Unicode characters. The presence of character encoding standards remains largely invisible to people typing on their computers and other devices every day, however, the introduction of emojis to Unicode has already led to a more heightened popular awareness of Unicode. With this research, I intend to contribute to this understanding by providing a clear and thorough description of what Unicode is, how it functions in our everyday lives, and why emojis feel like (and actually are) such a big deal.

1.8 The Researcher's Perspective

I come to this research via linguistics, writing, and LIS, as I have mentioned before. I see this project as a synthesis of how principles from all of these fields of practice function in technological, digital space. While the text that digital character encoding allows us to create may appear to be the least material of the ways that humans have discovered to share writing with one another, that materiality is not gone, simply displaced-- the bit has not yet replaced the atom as the building block of our universe (Blanchette 2011, Gunn 2020). The headfirst sprint into machine-mediated communication over the last century has done so with little concern for the problem of material storage for digital entities, and this has already caused important provenance, authority, and control records from the boom of encoding standards in the 1950s and 1960s to be disconnected from points of access, or lost entirely. There is much to be said about the preservation of early digital documents, and so much has already been lost. I hope that through the historical aspects of this work to create a more complete picture of the thought process and environment around the development of character encoding standards for computing, as well as to preserve the histories that are at risk of being lost in the slagheap of 'old internet stuff.'

I have focused on emojis as the point of rupture within the established character-encoding paradigm, as well as within sociolinguistic language change for two reasons. The first is that emojis blur the line between text and image, and binary computing does not deal well with blurry spaces. This, I believe, allows for the opportunity to reassess and reevaluate the processes which have defined "text" and "image" in the sociotechnical world which we live in today. The second reason is that the sudden and overwhelming popularity of emojis in popular culture since the release of Emojis 1.0 in 2015 has caused the Unicode Standard and its Consortium to address directly for the first time the needs of human beings who were raised in a digitally-literate environment. This includes functional needs, such as which scripts are supported, and also cultural needs (which is also a matter of which scripts are supported), but importantly for this work, the economic needs of the tech companies who have invested their time and money into making sure that Unicode is a standard which serves not only their users and their technology, but also the long-term business and research and development goals of its constituent members.

1.9 Conclusion

In conclusion, this dissertation is meant to be as holistic a contextual analysis of the Unicode Standard's place in the larger world of digital infrastructure as possible. By examining Unicode as a lexicon of semantic objects, a contextual born-linked document, and as an entity for the allocation and regulation of digital goods, I aim not only to elucidate the ways that Unicode has come to define itself in contemporary digital culture, but also seek to raise larger questions about the culturally-situated nature of technology, and the often-invisible power structures built into the infrastructure of our everyday digital communications. My LIS-centered approach to this analysis, using Faceted Methods Analysis as a mixed-methods modality serves to highlight the historical, economic, and political factors at play with regards to the creation and development of digital character-encoding standards and reveals a significant shift in how the Unicode Consortium and its membership of tech industry leaders presents the Standard to its users because of the introduction and subsequent wild popularity of emoji characters.

2.0 Methods

This work is best described as “mixed-methods” as it makes use of frameworks and techniques from a variety of disciplines and methodological paradigms (Kaplan and Maxwell, 2005). Beginning with Sense Making and Grounded Theory, I identified a multi-pronged, faceted approach as being best suited to the large-scale, discovery-focused study of a complex sociotechnical system, such as Unicode. This approach draws on established and parallel methods across a variety of academic disciplines and serves the goal of a nuanced and multi-angle take on a single subject of research. As a means of organizing this methodology, I take the principles from S.R. Ranganathan’s faceted approach to classifying information objects and apply it to a subject of research, subsequently approaching these different facets with different methods appropriate to each (Ranganathan, 1950). In this chapter, I begin by locating my methods generally in the world of Grounded Theory and Sense Making, and establish the need for a faceted approach to a system of the size and complexity of Unicode. Next, I discuss S.R. Ranganathan’s facet analysis, and how that paradigm is particularly relevant to the study of Unicode as a large infrastructural system at work in the world today. I then review my process for identifying methodologies to best analyze the various facets of Unicode which this research addresses, and the appropriateness of each methodology for its requisite facet.

2.1 Sense Making and Grounded Theory

Grounded Theory and Sense Making are cross-disciplinary methods used widely across fields such as Library and Information Science, Nursing, Sociology and others, though sometimes with differing names and details (Strauss and Corbin, 1994; Dervin 1999). Both take an iterative and exploratory approach to research materials and theory, and are well suited to the research of multi-dimensional complex systems. Brenda Dervin writes about Sense-Making, the methodology she developed, as such:

> "Sense-Making Methodology from the beginning has mandated itself to the design of methodology for the communicative study of communication. Information seeking and use are defined as communicative practices. So too are the practices of researching information needs and seeking." (Dervin, 1999, p. 729)

This is particularly relevant in the case of this research, where specific linguistic communication—writing—is used as a way to explore the information structure of a system which itself regulates writing. Rather than seeking to separate or neutralize the subject of research from the practice of researching it, Sense-Making serves as a meta-methodology, acknowledging and reminding us that the process of meaning-making also requires the deconstruction and analysis of already-made meanings. Writing about writing systems is a necessarily meta practice, and the sense that we are able to make from this exercise remains contextually centered within the understood paradigm of text and writing as a meaningful and valid way of exploring complex ideas.

Likewise, Grounded Theory is a methodology focused on identifying what exists in the data or area of research, and repeatedly updating, refining, and refocusing the direction of the research based on the results of the previous iterations of the work. Like Sense-Making, Grounded Theory is a qualitative methodology used across a wide variety of humanities and social sciences fields. Grounded Theory was developed by Barney Glaser and Anselm Strauss, and was first published in their book, *The Discovery of Grounded Theory: Strategies for Qualitative Research* (1967). Strauss and later collaborator Juliet Corbin define GT as such:

> "Grounded theory is a general methodology for developing theory that is grounded in data systematically gathered and analyzed. Theory evolves during actual research, and it does this through continuous interplay between analysis and data collection." Strauss and Corbin (1994, p. 273)

Critical to the process of doing grounded theory research is the concept of iteration. Iteration is the idea that the researcher must return again and again to their sources throughout the research process to assess how well their process of collecting and creating data reflects the phenomena which they are studying. One difficulty with this approach is that

> "the researcher who has carried out the research knows more than anyone else about the phenomenon under study. The researcher knows every single detail of the research and it is not easy to share all of this knowledge with others." (Mansourian 2006).

An iterative process not only broadens and deepens the researchers understanding of their subject, but also provides insights into how the knowledge they are gaining might best be synthesized and shared with others (Glaser and Strauss 1967, Strauss 1998). Grounded theory is a methodology, therefore, which positions knowledge as inherently subjective, requiring careful observation and documentation to create as much contextually supportive information about the research process as possible (Timmermans and Tavory 2012, Star 1998).

Documentation and metadata creation naturally mesh with the work of information scientists, and “The fields of library and information science have no shortage of research questions and phenomena needing thorough exploration and continue to need more well-founded theories, so there is certainly a need for more grounded theory research” (Powell 1999, p. 103). The process and principles of Sense-Making and Grounded Theory have informed the construction of the present research, and are methodologies with which I am familiar (Hackney et al., 2018, Sula, Hackney and Cunningham 2017).

2.2 Ethnography

In many senses, however, this work is an ethnography of digital text—a “thick description” of its environment, how it came to be that way, and how the humans moving through it are shaped (Hine, 2015). Ethnography has precedent in infrastructure studies, providing the researcher a methodology to study the ecological effect of infrastructure models, which are often difficult to parse in action (Star 1999, p. 379). Scholar of infrastructure and advocate for information infrastructure studies, Susan Leigh Star, argues that ethnographic field work is often a type second-hand study of infrastructural systems, as culture itself is an infrastructure which shapes the way that human lives are lived (Star 1999, Jewett and Kling 1991, Neumann and Star 1996). Scheffer (2002) proposes that ethnography of systems can be accomplished via the use of analytical ethnography, in which systems of behavior are observed and analyzed with regard to different “fields” – or facets—which alter the researcher’s perspective on what exactly the boundaries of what the subject of study are. Analytical ethnography, defined as “an ongoing dialogue of empirical and theoretical perspectives," provides a way for ongoing theorizing by the researcher during the process of conducting research in the "field." (Lofland 1995, Scheffer 2007) Large infrastructural systems provide a rich research site for analytical ethnography, because of their broad scope and influence across many industries, cultural milieus, and intellectual paradigms (Starr 1999) .

Additionally, scholars in media studies, community dynamics, and informatics have found ethnography a useful method for study of internet-based groups of people, and the digital spaces they occupy. Unicode is different from the environments of many of these other ethnographies, which cover topics such as the social norms of instant messaging, gender presentation in massively-multiplayer online roleplaying games (MMORPGs), in that it is not a space or a service which users opt-in to, or are necessarily able to opt-out of (Hine 2016, Lewis and Fabos, 2005, Kendall 1999, 2000). By using an ethnographic framework towards applying the cross-disciplinary nature of a faceted approach to a complex system, I then allow for the creation of connections between disciplines that perhaps have yet to be made by bringing interdisciplinary scholarship to play in the expansive infrastructural environment of digital character-encoding, or places under the same umbrella similar methods and theories using different terminology across disciplines, creating a wholistic cross-disciplinary view of the subject at hand.

2.3 Facet Analysis

Facet analysis is the process by which an object gets labeled within a classification system, and how that labelling enables it to relate to other objects within the same system, according to its creator, S.R. Ranganathan, Indian mathematician and LIS scholar (Ranganathan 1950). Broadly defined, facet analysis is the process of dividing a complex subject into its constituent parts by relating them to a set of five fundamental categories: personality, energy, matter, space, and time (Ranganathan 1951). These are categories that Ranganathan determined based on his own research and experience, but does not necessarily represent a set-in-stone paradigm for facet analysis, but rather Ranganathan's implementation of its principles (Ranganathan 1950, 1951).

Other classification systems such as Library of Congress, use nested systems where each 'level' of classification is a subcategory of its parent category, creating a hierarchy, which like so many top-down approaches end up with unevenness in object distribution, and flattens nuanced works into a single category (Adler, 2017, Billey et. al 2014, Olson 2001). Alternatively, facet analysis relates objects within a classification system horizontally, allowing for multiple labels associated with the different facets of the object, preserving some of the inherent complexity of the information object at hand, and allowing for more metadata about an item to be stored, which in turn creates more points of access for users and more nuanced connections between items. For instance, in the Library of Congress Classification, information objects must fall under one of 26 top-level categories (labeled A-Z) before any further classification must happen. This means that in practice, catalogers must look at a work and decide from the very beginning of the classification process whether it belongs under "B -- PHILOSOPHY. PSYCHOLOGY. RELIGION" or "P -- LANGUAGE AND LITERATURE" and it cannot fall under both, regardless of the actual subject content of the work (Library of Congress, 2021). This can result in not only the intellectual silo-ing of multidisciplinary works, but can also cause a single author's work to be distributed widely across classifications, resulting in the physical dispersion of their work in the library as well (Adler

2017). Colon Classification makes use of non-hierarchical facets in order to allow for multiple subject headings and categories to be applied to a single work without reducing the work to a single classification (Ranganathan 1951, Joudrey et al., 2015).[2]

I have adapted and recontextualized the principles of facet analysis that Ranganathan outlines for the purposes of this research, beginning by identifying facets of the research subject/area that are of direct relation to the research questions at hand, and then selecting and applying the most appropriate method to each of these facets based on the research questions, data, and goals of the researcher. This *Faceted Methods Analysis* (FMA) is specifically designed for the study of large, complex systems and datasets, and works best when clear boundaries are drawn around what is and is not included in the definition of that system. Identifying facets is a part of this process and is intended to not only focus the researcher on the relevant areas of inquiry for their research, but also to identify what is not covered by their research. This creates clarity from the beginning of the project around the scope of research on complex systems and large datasets.

FMA consists of four steps:

1.0 Identifying facets

2.0 Associating – data with facets, facets with methods

3.0 Analysis

4.0 Synthesis

[2] Though, of course, it does not solve the fundamental problem of potentially shelving a single book in two physical locations. That is a problem for another scholar, on another day.

Once a body of research has been identified, the first step in FMA is (a) identifying facets of the subject that coordinate best with the research questions and; (b) identifying the particular disciplinary interests of the researchers. This is a bidirectional process, meaning that facets can be identified for study from the data and particular facets of interest to the researcher can be used to identify particular areas of the subject to focus on. In the case of this dissertation project, the original focus of my research was "The Unicode Standard," and I approached it from a Library and Information Sciences perspective. Beginning from there, I identified facets in order to refine the focus of my research and hone in on the aspects of my subject and data that could best answer my research questions about the infrastructural nature of Unicode in the contemporary digital world.

Table 4 Research Facets

Facets of Unicode →	Facets of Research Interest ↓	Digital Infrastructure	Standardization of Character-Encoding	Documentation Practices
	Language Change		X	
	History of Technology		X	X
	Digital Governance and Communication	X		X

Looking at Table 4 (above), the facets of Unicode that I identified as being of greatest interest to my practice/research/work were "digital infrastructure," "standardization of character encoding," and "documentation practices." Similarly, the facets of my own research interests as being relevant to Unicode are “language change,” “history of technology,” and “digital governance and communication.” The points where the rows and columns intersect are marked by shading and an X, and represent areas I identify as relevant to one another in the context of this research, and well-suited to examination as a distinct facet of this work. Having identified these overlaps, I next need to determine what kinds of methods are most appropriate for each facet—aka Step 2: Associating.

Because both language change and history of technology are related to the standardization of character-encoding, and both deal with the development and long past of Unicode, I chose historiographic research into the origins of character encoding and development of Unicode as the appropriate method. Likewise, as digital governance and communication intersects with digital infrastructure and documentation practices, I chose network analysis as a means to explore the structural makeup of Unicode and its associated documents. Finally, history of technology intersects with both the standardization of character-encoding and documentation practices, I have chosen a combination of close reading and the more ‘distant reading’ practice of content analysis to apply to the documentation and forms of character-encoding.

2.4 Methodologies

2.4.1 Historiography

Gavin Andrews has described historiography as being “concerned with historical interpretations and representations of the past—put another way, the writing of history as opposed to history itself” (p. 399). Foundational work in the shaping of historiography defines it as a method of “notation” on earlier historical works, assessing the accuracy, sources, and viewpoint of the author (Becker, 1938). Like so much of this dissertation, historiography takes a meta view of already existing histories and sources, and attempts to contextualize the labor that has gone into creating already existing histories. The history of Unicode from the beginning of digital character encoding to the present day has never been thoroughly documented, and those who worked on Unicode’s predecessor, ASCII, bemoan the loss of much of the official documentation for that standard (Jennings 2020). I drew largely on documents archived casually (that is to say, saved intentionally, but outside of traditional archives settings and standards) by persons who worked with character-encoding throughout the 20th century, and are presented with commentary, contextual information, and amendments from their collectors. These sources are ideal for historiography, and through them I am able to reconstruct a more thorough history of Unicode, both in terms of temporality, as well as with regards to the attitudes, opinions, and biases of the creators of that history.

2.4.2 Network Analysis

Creating networks from data allows us to analyze the underlying structure of the networked entities, and draw conclusions about the nature of their relationships (Painter, Daniels, Jost, 2019). Network Analysis is particularly suited for the study and mapping of hyperlinked pages in a website, as these links serve as explicit records of a relationship between pages (Adamic 2009). Additionally, internet search algorithms make use of these connections to determine which pages are the most linked between one another, with links being weighted by the relative connectivity of the linked pages, and display search results based partially on those findings (Noble 2018, Altman and Tennenholtz, 2005). Within this research, I use Network Analysis as a way to create visual, mapped representations of the Unicode.org website and its associated documents in order to better conceptualize ideas and entities, and to emphasize and highlight aspects of the data which may not be readily apparent. (Card, Schneiderman, et al. 1999) In this case, using network visualization to create an overall view of Unicode's website allows us to see the "space" that is taken up by its constituent pages, while retaining the valuable relational information about the relationships between those pages. Running centrality and ranking algorithms on that network provides additional important information about which pages are the most connected to others. To extend the map metaphor, this data is used to determine which "towns" (pages) are most frequently travelled to, and by which "roads" (hyperlinks).

2.4.3 Formal Analysis + Content Analysis

Formal Analysis, a method frequently employed in Art History and Literary Studies is “a careful and methodical examination of the physical components” of an object as a way to “decode” its meaning (Turnbull et al. 2021) It produces a precise and detailed description and evaluation of the object in question, and should serve to enhance the viewer’s understanding of the object and its purpose (Smith 2016). Formal Analysis in this work focuses on two areas. First, in the appearance of certain Unicode characters, comparing their appearance in text to their entry descriptions in the Unicode nameslist. Second, on the presentation of character encoding charts from the generations before Unicode, which, due to their much smaller size, are able to be displayed as a single chart or graphic, and often were produced materially to serve as a reference for users of a given standard (Standage 1998, Jennings 2006). This was done in order to better understand and make clear the rationale behind each code’s creation and unique layout.

This same type of formal analysis via close reading is nigh impossible for a standard as large and broad as Unicode, and therefore the distant reading techniques of content analysis and network analysis were employed to help situate Unicode within a larger world of standards, their consortia, and the public. Content Analysis is “any technique for making inferences by objectively and systematically identifying specified characteristics of messages" (Holsti 1969, p. 14). It is especially well-suited to large corpuses of text, and is used to determine themes and focuses of the analyzed documents by “coding” the text via agreed-upon and stable categories (Stemler 2000). By using content analysis on a corpus, it becomes possible to determine and highlight themes of the documents’ content allowing the researcher to “infer from symbolic data what would be either too costly, no longer possible, or too obtrusive by the use of other techniques” (Krippendorf 1980,

p. 51). Content analysis is especially suited to the analysis of the Unicode.org website due to its large scale and non-linear nature, which would make reading and analysis by other methods prohibitively time-consuming. While at first it may seem contradictory to group a close reading method such as formal analysis and content analysis, a more ‘distant’ form of analysis, within this work, their connection reflects the meta- nature of my approach to this research. Linking them together emphasizes the vast scale of the Unicode Standard, from the bit level all the way to the institutional level.

2.4.4 Telling the Story

I outline all these methods here, and note their history and relevance in the field of LIS and infrastructure studies, but beyond that it is my hope that the rest of this work reads less as an academic accounting of research done on a particular subject, and more as a narrative history of just one of the technologies foundational to creating the digital world we live in today. By telling you this story, I hope to enable you to make your own sense of Unicode’s role in the history and future of text and technology, while presenting the current state of affairs within the larger sociotechnical environment from a LIS/classification-focused linguistic perspective.

3.0 Character-Encoding in Context

The amount of research that has been done about emojis has been rising dramatically, since the formal launch of Emoji 1.0 in 2015. They talk about affect, they talk about gesture, about race and gender, about what order the ingredients on a cheeseburger go (Alshenqeeti and Hamza 2016, Gawne and McCulloch 2019, Korn 2021, Riorden 2017, Brogan 2017). The majority of this work accepts emojis as they are presented to us by Unicode—images that are encoded as text—without any further exploration into the infrastructural system which houses and governs them. There are of course, notable exceptions, and more work being done in this area as time goes on. But Unicode still has not be thoroughly dissected and theorized, and while I do not claim to be able to do that in its entirety within this document, I do want to stake a claim for the field of LIS, and encourage further discourse in this area. For this reason, I use this literature review to take the broadest possible lens on the theoretical frameworks underpinning Unicode as a piece of sociotechnical infrastructure.

Unicode is a digital organization system which deals with communicating human language through the regulation of the unique appearance and IDs of orthographic characters. Because Unicode lives within the larger sociotechnical environment of networked digital computing, the foundations of how communication happens in these spaces is essential to understanding what exactly it is that Unicode does, and how it differs from other forms of digital information encoding.

One of the purposes of this work is to examine and problematize exactly what it is that the Unicode Standard regulates, and some of the friction that the ambiguity surrounding that question raises in trying to apply it to more conventional frameworks of standardization. By working to define specifically the goods which Unicode provides and regulates, as per Elinor Ostrom's 2005 description of the four main types of goods, which are determined by the difficulty of exclusion and subtractability of the good in question, I reconsider the understanding of character-encoding as per the paradigm of the Knowledge Commons, highlighting how this reframing of the Standard's "space" rejects the notions of scarcity which threatens traditional common pool resources.

3.1 Organization of this Literature Review

This research deals with the transfer of coded information-- specifically text-- from one location to another in digital environments, it is therefore essential to understand how such transfer has been conceptualized, but also how it has been implemented practically since its inception in the mid-20th century. I begin this review of the literature by reviewing and pinpointing exactly what is meant by "text" in the context of this paper. This definition is in many ways phenomenologically tied to the material technologies which make producing text possible. Since Unicode is a standard for digital text, I will review understandings of text in digital environments as well, and specifically consider their relation to materiality. After considering broader implications of "text" I will then move to discuss documents and documentalism as a perspective on understanding information objects in a broader sense than traditional orthographic texts.

In the next section, I will cover what standards are and what standardization does for our collective understanding of the spaces it creates. I will review the role of standards and their governance in the field of economics, and how power is accumulated by the ability to uphold and maintain a standard in a given market. This leads to the next section wherein I discuss the economy of standards which has built up around large socio-technical infrastructural systems such as Unicode, and where text-encoding falls within accepted taxonomies of standards. With their introduction to a market then comes considerations of the materiality of standards, “where” they live, and how they build information organization systems through their instantiation in digital space.

Subsequently, expanding on the economic paradigm of standards, I review the types of goods which Elinor Ostrom has defined as circulating within these markets, and the concept of a “commons” from which they are drawn. This leads to a discussion of the “Tragedy of the Commons,” a thought problem in the field wherein shared resources are over-extracted, creating a manufactured shortcoming. I will discuss instances of this occurring within digital computing spaces, and then consider Ostrom’s model of the “Knowledge Commons” as an alternative conceptualization of a shared resource pool which does not produce scarcity in the same way that a traditional commons does.

Finally, I will discuss the theory and practice of classification and cataloging with LIS literature, and how it is particularly equipped to maintain and interpret standards (especially humanistic ones, such as that of a language) because of its focus on retrieval and access. I will introduce two major metadata constructs within the LIS world: MARC and FRBR, and discuss how they each deal with issues of text, materiality, and document when organizing information objects. This presents an alternative perspective on standardization, and specifically highlights a different set of needs for users of digital character-encoding in this area as opposed to in a market environment.

3.2 Standards

In the broadest of terms, Lawrence Busch describes the four types of standards which shape human experience as Counting, Shape, Weight, and Time (Busch, p 77). These standards remained generalized and vague, he says, until “The advent of written language brought standards to the fore in a new way. Written language appears to have had its origins not so much in the recounting of great exploits or in the retelling of epic poems but in the rather prosaic need to account for things and people” (p.79). This required a much more accurate shared understanding not only of the characters and glyphs of written language, but of what needs to be standardized and how. Writing’s origin in commerce and trade links it inexorably to accounting, regulation, recordkeeping—that is to say, to standards and standardization. This association remains strong today, and while Unicode as a standard for digital writing may at first seem to be an academic or technical pursuit, its roots in commerce remain visible in who governs it and how it has developed within its own digital marketplace.

3.2.1 What does Standardizing do?

Research regarding the creation and maintenance of standards spans several academic fields. In fact, the term “standard” has a variety of meaning and implications, depending on application and context. Lawrence Busch outlines these nuances in *Standards: Recipes for Reality* (2011), and sums up by saying:

> “Standards may imply that something is the best, or that it may be used as an exemplary measure or weight; or they may emphasize the moral character of someone

> or the superb qualities of something. Standards may also refer to rules or norms that embody the ideal or merely the average. Finally, standards may refer to tolerances permitted for both people and things. These various meanings are inextricably linked together. All say something about moral, political, economic, and technical authority." (p. 25)

Standards are entities that are imposed by humans, whether individually or collectively, and they serve to make navigation through the world easier. They "become embedded in particular technological forms, and then diffused or integrated into local practices" (Lischer-Katz, 2016). The creation of and compliance with standards is a form of knowledge creation, and a performance of knowledge for social purposes, and can serve and phenomenological evidence of knowledge or expertise in the standardized area (Timmermans and Epstein, 2010).

3.2.2 Politics and Power of Standards

The ubiquity and interoperability of a well-designed and fully-adopted standard can give the impression that standards and standard-making are apolitical, but the process of adoption as well as the control exerted by a standard are shaped by power dynamics and political negotiation. Kindleberger (1983) writes that, "it is hard to think of international standard that did not start out as the public good of some particular country, usually one with high international standing because of its economic and/or military power." (p. 392) That is, entities with a great deal of preexisting power use that power (intentionally or not) to create and adopt standards which are in line with their needs, goals, and established norms.

Just as the history of technology is tied up in military innovation[2] the standards around computing technology are as well. While Unicode itself was created and adopted for use in the business and then public sectors, the need for character-encoding standards at all would be impossible without the development of computers for military purposes. This history is very difficult to remove from later instantiations of technology, and Greenstein (1992) observes that “there are well-documented historical links between military demand and later civilian development of computing equipment. […] In that case, the standards imbedded in military equipment may strongly determine later design choices.” (p. 542).

Even separate from the military-industrial complex, the existence of standards can be a means to maintain or rebalance power dynamics, as is evidenced in the prioritization of the Roman alphabet in early character-encoding models, as well as in the script- and emoji-adoption process in place today (Berard 2018). Janet Abbate writes in Inventing the Internet,

> “Standards are a political issue because they represent a form of control over technology. Interface standards, for instance, can be empowering to users of a technology. If all manufacturers of a device use the same interface (for example, the touch-tone keypad of a telephone), users need learn how to operate the device only once. Standards also ensure that components from different manufacturers will work together. When standard interfaces make products interchangeable, consumers can choose products on the basis of price or performance, rather than just compatibility. This increases consumers’ power in the marketplace relative to producers” (Abbate, 2015 p. 147)

This power exchange is a balancing act for producers, who want to capture as much of the market as possible, while also maintaining a certain amount of compatibility with their competitors (David 1987). Achieving a profitable market share, as Abbate notes, is often a matter of acquiescing to the users' desire for interoperability. In the area of character-encoding, interoperability often manifests as 'lossless' communication between devices created by different manufacturers and is often only apparent when that interoperability fails.[3]

Scholars of standards and standardization agree that the input of users is necessary for a standard to meet the needs of those users (Foray, 1994), but note that perfect representation is extremely difficult, both because of potential conflicting needs from different user groups (Farrell and Saloner, 1986), but also because "future generations of standards users […] may not yet have representatives in the present market place, much less know what features they will desire in their product standards." (David and Greenstein, 1990, p.7) Likewise, in order to make participation possible, Foray notes that standards and means by which they are made must be intelligible to their users— actively working against the slipping into ubiquity and invisibility that so many infrastructure scholars note as a part of the universal adoptions of standards (Bowker & Star, 1999; Edwards, 2003). We can see this conflict of interests at play in the development of Unicode, especially once emojis were adopted—the creators and original user base for Unicode was professional and purely text-based, and there was no way to anticipate the widespread adoption and demand for visual emoji characters.

3.3 Economy of Standards

In economics, there is abundant literature discussing the role of standards in the marketplace, and widespread historical efforts for interoperability which in many cases lead to industry standardizations (Kindleberger, 1983; Weiss and Cargill, 1992). David and Greenstein (1987, 1990) have developed a typology of standards based on this literature which divides standards by their function in relation to the entity which they are standardizing (reference, compatibility, or process), and their method of implementation (*de facto*, *de jure*, or voluntary consensus). They begin by defining a standard as "a set of technical specifications adhered to by a producer, either tacitly or as a result of a formal agreement" (1990, p.4).

Table 5 David and Greenstein's taxonomy of standards (1990)

	Reference	Compatibility	Process
De facto			
De jure			
Voluntary Consensus			

Reference standards are those which set a "measure against which the relative extent of some quality dimension is compared (David 1987, p.213), and which include such examples as currencies, weights and measures, and accreditation standards. A process standard sets a scale for agreement with the standard, with bounded end points, where individuals adhering to the standard fall in between those points—these are quality and safety standards, and measures of competency. Compatibility standards are "dichotomous sets, one being compatible with a standard and the other not” (p.219), these assure that products and protocols from competing sources are able to interact with one another, and include things like electric plugs, signal frequencies, and diplomatic protocols.

The second vector along which David and Greenstein classify standards is that of the method in which they are instated and/or enforced. A de facto standard is one that has not been formally adopted by any governing organization but is accepted as the go-to standard in practice. A de jure standard is a standard which has been formally endorsed by a governing body as official, and deviations from it are considered to be incompatible. Voluntary consensus standards fall somewhat in the middle and are those standards which are agreed to be used by the main user base or vendors in a particular area, but are not endorsed from a top-down perspective by any larger governing body.

3.3.1 Locating the Unicode Standard

Character encoding standards are *compatibility* standards—that is, all of the standards discussed in this work, from Baudot’s code to Unicode serve the purpose of ensuring that the same codes are used to represent the same characters from place to place, whether those codes be the position of rotating wooded planks, or an eight-bit unique identifier corresponding to an indexed character. Compatibility standards ensure that objects (in this case the UTF-0000 codes corresponding to characters) from various and potentially competing sources are able to interact with one another in as efficient and lossless a way as possible. We can see this compatibility fail when a document saved with one encoding standard is opened in a program using a different standard, and characters are transformed or unable to be displayed at all, making the message illegible. While backwards-compatibility between Unicode and ASCII makes legibility possible in many cases, and similarities in character arrangement within a standard may give the appearance of partial compatibility, the assignment of characters to unique identifiers must be identical between the original encoding standard of the document's author and that of the reader in order for true compatibility to be achieved. This is in part why code assignments cannot be removed or changed once added to Unicode, to maintain the hard-won stability of its compatibility with previous standards.

Unicode is not governed by any entity outside of its own Consortium. It was created by a process of voluntary consensus among the manufacturers and vendors of technology that make use of character encoding standards, but because this consensus happened alongside the development of the technology to which it applied, it has become the *de facto* standard (David and Greenstein 1990). Unicode is mapped directly to ISO/EIC 10646, making it officially sanctioned by the International Standards Organization, making the ISO's particular implementation of 'Unicode' a *de jure* standard. However, this image of Unicode captured as the ISO standard is static, and the ISO governing bodies do not have the power to adapt and expand ISO/EIC 10646 beyond Unicode. For this reason, we can separate ISO/EIC 10646 as the *de jure* standard, and Unicode as the *de facto*.

Having identified Unicode as a *de facto* compatibility standard, it is much easier to recognize that its responsibilities as far as its own growth through the addition of characters as scripts is done largely reactively. While alternate encoding standards provide a way to type in languages not yet included in Unicode, the ability of such documents to traverse a digital environment is extremely constrained, making both the dissemination of knowledge from these linguistic and cultural groups difficult, and heightening the risk of deprecation via potential loss of backwards-compatibility as ‘niche’ standards are absorbed into Unicode.

3.3.2 Consortia

Weiss and Cargill (1992) have developed a taxonomy of standards consortia working complementarily with the taxonomy of standards discussed above, and focusing specifically on the Information Technology (IT) field. In their work they define a consortium as “a collection of like-minded interests that participate in the development of what may be a market accepted solution to what is perceived to be a user problem” (p. 560). The types of consortia they identify and define are implementation consortia, application consortia, and proof-of-technology consortia (p. 561).

Implementation consortia are established to aid with the adoption of existing standards, usually ones that are not intuitive or easily transitioned-to. Their focus is on use of the standard, its implementation and continued integration into its user base. Application consortia are focused on creating new users for a standard by repackaging or promoting certain aspects of the standard. This type of consortium is especially used by open systems or free and open software standards, which are available to all, but encounter a hurdle in usability. And finally, proof-of-technology standard consortia are designed to create and/or prepare a market for a standard which is in development. They serve as heralds for the standard, to encourage potential users to adopt it as ‘the’ standard once the technology is produced (Weiss and Cargill, 1992). Across all these types, the function of the consortia is based on the place that the standard in question has within the existing market. These consortia are advocates for their respective standards and are seeking to appeal to users (whether at the individual or collective level) and provide a needed good (the standard itself) to those users.

Anticipatory standards—standards created before the technology they are intended to standardize (Cargill 1989)—became popular with the rapid development of digital technology in the 1980s and 1990s, in an attempt to keep up with the broadening of the computer technology market (Byrne and Golder, 2002). However, as the number of users booms, the anticipated use of such a standard can be outpaced by its actual usage. As noted previously, the documentation of the Unicode Standard, created by the Consortium, aims to provide a unique encoding for “any abstract character that ever could be encoded” (UTR#17: Unicode Character Encoding Model). This is anticipatory in the grandest sense possible, and any defined linguistic standard’s codespace is necessarily too small for the infinite potential characters, present and future—especially when the removal of characters from the standard, or relocation of them within the codespace is explicitly forbidden in the name of backwards compatibility, as it is with Unicode.

It seems, however, despite this risk, there is no (or very little) desire for exclusion for the use of standards, but rather for compatibility across otherwise competing users of a particular standard. In fact, Kindleberger (1983) considers standards to be a public good, noting that even in the private sector, “there is strong pressure [… to be] compatible, leading to the private and collective good of standardization” (p. 387). As seen in Table 5, a public good requires a low subtractability of use, meaning that an individual making use of a public good does not detract from another’s ability to use it,18 however, subtractability becomes an issue in technological standards when the ‘space’ within a standard is finite, and the inclusion of one point of compatibility within the standard uses a ‘slot’ which cannot be reused or shared. While this does not affect the individual ability to make use of a standard, it does affect how universal a standard can be.

However, anticipatory standards are not the only option for establishing standards within a given area. The development of standards can arise in a variety of ways in order to create efficiency and interoperability within a marketplace. A diffuse market—one where there are many buyers and sellers—can lead to the development of multiple competing standards (Greenstein 1992), and we see that this is the case during the late 1980s – early 1990s with regards to character encoding standards. Users and developers agreed that ASCII and other 128- or 256-bit character sets were inadequate for modern computing needs, and several entities were working to develop replacements standards19. The factors that contribute to a particular standard being adopted over another are not always logical, but rather may result from a "bandwagon" effect due to a particularly well-connected producer of a standard, rather than due to optimal efficiency of the standard itself (Greenstein 1992, p. 539).20 Weiss (1991) applies political theory and game theory concepts to the creation of voluntary consensus standards. Voluntary consensus standards are those developed within committee by the industries to whom the standard applies. Weiss observes that this type of standard creation process often does not create an 'ideal' standard, due to conflicts interests among committee members, as well as the power dynamics between those members. This results in standards which, while not ideal, are approximately equally effective for all members (Weiss 1991). Like all pieces of infrastructure, standards are designed for use, and must be implemented in order to be effective, whether they are perfectly planned or not. That implementation, however, is dependent on what it is that a standard standardizes on a conceptual level, and the ability to identify what resources in the real world become standardized upon that implementation. There is not always a 1-to-1 relationship between these two things, as we see with Unicode, making it that much more important to clearly define the good that is provided, and the resource that is used up by it.

3.3.2.1 The Unicode Consortium

The Unicode Consortium, formed concurrently with the development of the Standard, appears to have begun its life as a proof-of-concept consortium, in order to make the forthcoming Unicode Standard the de facto character encoding standard for the next generation of computing, regardless of its members being otherwise in competition for the same user base. As time has passed, however, the Unicode Consortium has become more of an implementation consortium, working for universal encoding, integration, and consistent visual representation of the Standard across platforms.

The Unicode Consortium is a mix of public, private, government, and individual interests—this helps give the appearance of accessibility, as do features such as adopt-an-emoji, however, this is symbolic involvement at best, and the nexus of power within the consortium remains largely in line with that of the remainder of the digital world—Western based tech companies with large market shares of the hardware and or software that makes use of Unicode. Table 6 below shows the voting member organizations of the Unicode Consortium as of May 2018. For the purposes of this research, I broadly divide the "Type" of member into three categories: Technology, Government, and Education. The 14 Technology members, making up 77% of the voting members of the Consortium, are those which are any business or group which produces hardware and/or software on which Unicode is used. This includes producers of operating systems, such as Microsoft and Apple, as well as entertainment, social media, and visual design enterprises.

The four government-type members are representative of their respective governments, whether at a national level, or as a particular committee or subgroup within the government and make up 22% of the voting members. These members, as of 2018, are all representatives from the Mid-East and Central Asian regions of the world, and represent culturally- and linguistically-rich areas, and can be assumed to have a vested interest in the preservation of their local languages, especially those utilizing non-Roman scripts. Finally, there are two Educational voting members of the Consortium, making up 11% of the voting members. First is the Linguistics Department at the University of California, Berkeley, a traditional academic department in a field closely linked to the mission and work of Unicode. The second is Emojipedia.org, a website where visitors can look up emojis to see how they appear across platforms, changes to their designs over time, and other information related to each of the emoji characters. While not an ‘educational institution’ in the same way that UC Berkeley Linguistics is, I characterize Emojipedia as an Educational-type member of the Consortium as it exists to serve as a publicly-accessible reference resource for emojis and their usage.

Table 6 Voting Members of the Unicode Consortium

Member	Ranking	Type
Adobe	Full Members (Voting)	Tech
Apple		Tech
Facebook		Tech
Google		Tech
Huawei		Tech
IBM		Tech
Microsoft		Tech
Ministry of Aqwaf and Religious Affairs,		Gov
Sultanate of Oman		Tech
Netflix		Tech
Oracle		Tech
SAP		Tech
Shopify		
Government of Bangladesh	Institutional Members (Voting)	Gov
Government of India		Gov
Government of Tamil Nadu		Gov
University of California, Berkeley		Edu
Linguistics Department		
Monotype Imaging	Supporting Members (Voting)	Tech
Emojipedia		Edu

The overwhelming numbers of global tech giants at the highest level (91.6% of Full Members) within the Consortium speaks to the influence both that digital text has on how contemporary life is lived, but also to the global standardization of digital technology. It seems clear that the Governments of India and Bangladesh have an interest in their writing systems being incorporated into Unicode, for both communication and cultural preservation reasons. What interest, then, does Facebook or Netflix have in Unicode? Bethany Berard speaks to the social influence that can be had from exerting control over text, writing "standardization not only facilitates circulation and increases the potential of rapid and widespread adoption, but demonstrates a particular form of control over technology." (Berard, 2018). Having their name associated with the 'official' standard for character encoding ensures a continuing piece of the market share for Consortium members, as well as the continued opportunity to influence the direction of Unicode going forward.

Looking only at the membership of its Consortium, it would be easy to say that Unicode itself is also a therefore a tech company focused on the development of standardized digital text. Unicode itself describes its purpose as providing "a unique number for every character, no matter what the platform, no matter what the program, no matter what the language" (What is Unicode, 2017). However, Unicode also positions itself as a cultural heritage institution, noting that they "[include] popular languages like English and Mandarin, but also endangered languages like Navajo. […] Help us preserve the world's heritage by Adopting a Character or becoming a Member." (Overview – Unicode, 2019). These two things—standards development and linguistic documentation, appear to go hand in hand, and together they position the 'space' of the Standard as a repository for the archiving and preservation of the characters of every human writing system. Importantly here, is the conflation of "every character" in those writing systems with the full sum of a 'language.' As I discussed in Chapter 3, this equivalency becomes unstable when we consider languages which deviate from the 'norm,' such as Mandarin Chinese, whose writing system ascribes meaning at the character level, and currently makes use of more than 50,000 unique characters.

Indeed, the development of written text was itself a technological innovation. Unicode, though, does not invent or even reinvent text. Instead, it creates a uniform and finite resource pool of potential character codes for users to draw from in order for writing to happen digitally. Therefore, a more accurate assessment might be that Unicode is a resource pool for the assignment of digital textual characters, governed by its Consortium, which works to formalize and regulate how that resource is used. In this light, Unicode becomes an index of codes to be assigned, rather than a library of letters that users “check out” when they make use of them— making it a descriptive document of which character codes are available for use. The trouble with this, however, is that an approximate description of a world is at risk of becoming or being taken as a prescriptive outline of all that does and should exist in that world (Langmead, Newbery 2020). In other words, scripts (and the languages they represent) which are not encoded within Unicode for all intents and purposes do not exist in the digital world, and their addition comes as an ‘addendum’ to what is already seen as the ‘normal’ or ‘standard’ pool of resources which are already in circulation (Berard 2018). This is already playing out in front of our eyes, with regards to the representation of human beings via emojis, in terms of gender and especially racial representation.

3.3.3 Goods, Commons, Markets

While the taxonomies discussed above set boundaries for what roles standards and their governing bodies play within their respective markets (or ‘spheres of influence,’ more broadly), they do not address directly the type of goods (or ‘sociocultural value’) that such standards create. Elinor Ostrom (2005) outlines four basic types of goods: toll/club goods, private goods, public goods, and common pool resources (see Table 7, below), which account for the difficulty of exclusion and subtractability of the good in question.

Table 7 Four basic types of goods, adopted from Ostrom 2005, p. 24

		Subtractability of Use	
		Low	High
Difficulty of Exclusion	Low	Toll/Club Goods	Private Goods
	High	Public Goods	Common Pool

Toll or Club goods are those with a low difficulty of exclusion and a low subtractability of use—things such as country clubs, gyms, or even movie theaters, where a user must be a member or pay a fee to have access to the good, but do not consume the good in a way that prevents other from also consuming it. Also with a low difficulty of exclusion, but a high subtractability of use are private goods—what we regularly think of as consumer goods such food and clothing—where there is a limited supply and when one user purchases or consumes the good, it is then no longer available for other users to access. Goods with a high difficulty of exclusion are generally accessible to everyone, or to all those who wish to access them. Public goods are those which are not subtractable, meaning that they can be accessed by multiple users without depleting the supply of the good—items such as public highways, spoken language, or sunshine. This is the category of good that Unicode is positioning itself as governing, based on the linguistic idea that characters are not consumed by users in a 'permanent' way when typed on a device—There are not a limited number of letter "e"s to go around. However, this framing addresses only the linguistic nature of the characters governed by the Standard, and not the material, technological ones. For this perspective, the final type of resource outlined by Ostrom is a perhaps a better fit: the Common Pool Resource

Common Pool Resources (CPRs) are goods which have high difficulty of exclusion and high subtractability of use. Classic examples of CPRs are wild fish stocks, and shared public grazing land. (Ostrom 1990, 2005). These resources are available to be used by anyone, but run the risk of being depleted through overuse, denying further access to the resource to anyone. What follows from this situation, then, is the "tragedy of the commons" which plays out when there is high incentive to make use of the resource, and low incentive to consider others' potential use[5] (Hardin, 1968). The traditional tragedy of the commons results in the overfishing or overgrazing of the shared resource, which then cannot be renewed. Hardin, the originator of the concept applies it to resources in the natural world, rather than constructed ones, such as those existing digitally, but as we now exist in a digital world with constructed limits, it becomes possible to 'overuse' these resources as well.

3.3.3.1 Tragedy of the Digital Commons

The classic example of the tragedy of the commons in contemporary IT standards is that of Internet Protocol version 4 (IPv4). Milton Mueller (2006, 2008, 2010) discusses the decades-long impending panic surrounding the limited number of IPv4 addresses—the unique device IDs used to identify devices on digital networks and provides insight into the nuanced difference between scarcity and consumption in the digital realm, noting how the technical standards regulating IP addresses created the current scarcity problem:

> "The Internet protocol creates a virtual resource, the IP address space, of finite dimensions. […] The address space size is fixed by the technical standards defining the Internet protocol. IP addresses are scarce in the strict sense that economic theory defines scarcity: it is not possible for all of us to have all of the addresses we would

> like at zero cost. As a virtual resource, they are not ‘consumed’ or used up when put into production; rather, they are occupied” (Mueller 2010, p. 406).

In the case of IPv4, the space set aside to be occupied within the standard anticipated a certain amount of usage, which greatly underestimated the actual usage of the standard space. Meaning, in this case, that there have become a great deal more devices connected to the internet than we ever anticipated there being. While this scarcity may at first seem purely manufactured and ephemeral—why not just add more codes, or redistribute those being held back from use by their respective licensees? —it is in these moments that the materiality of the digital world becomes real. Hardware is programmed for compatibility with one standard, and is rendered useless if those resources are reorganized; the functionality of software depends on the consistent, organized, and persistent location of its reference material both on a local hard drive, as well as on the hard drives and cloud servers with which they interact.

3.3.3.2 Knowledge Commons

However, Ostrom also proposes the concept of a ‘knowledge commons’ – a pool of resources that can be added to but not subtracted from, meaning that the tragedy of the commons becomes irrelevant as the stockpile of knowledge in the commons grows (Ostrom 1990). This works against the traditional understanding of the movement of goods in the market, and Ostrom points to the Library and Information professions as being well-suited to the management and organization of these types of spaces (Hess and Ostrom, 2005). Librarians and LIS scholars have asserted that conceptualizing the library as a commons is necessary to do justice to the library mission of equitable access (Halperin 2020, Fister 2014, ALA Key Action Areas 2022). Further, I assert that the particular realm of digital infrastructure which character-encoding occupies, and currently is governed by Unicode, is exactly the sort of “knowledge commons” which Ostrom and Hess describe, and is well within the purview of LIS as a field to take a serious interest in this particular infrastructural space. Librarianship speaks often of communities— but whither the *commons* it implies?

3.4 Text

Moving on to the realm which Unicode is standardizing, let us discuss for a moment what is considered “text”—especially within the context of this work and the larger digital computing environment. For the purposes of this research, I begin with my own definition of “text” as *an orthographically-based communication using culturally established characters, symbols, and their associated meanings.* This definition is drawn from and builds on both philosophical and intellectual imaginings of the term, but also on its colloquial uses and practical definitions. Importantly, I consider a text to be an object, composed of other, smaller objects. Characters and symbols also have culturally established meaning, whether purely lexicographic/phonetic, or encompassing larger semantic ideas on their own. There are certainly other definitions of “text” which take a broader scope, beyond that of orthography and linguistics altogether. Because, however, this project focuses on the standardized presentation of encoded characters for the purpose of digital writing, it will leave the textuality of art, gesture, or performance for other researchers. My definition also attempts to cut across philosophical and technical definitions of “text” and “texts” in order to create a functional hybrid that encompasses the materiality and ephemerality of human communication and technology.

3.4.1 What is Text?

Philosophers and rhetoricians have grappled with what compromises a text, and whether its textuality is defined situationally by its ability to be read, or objectively by its conformance to a particular model. I choose to begin with Paul Ricoeur because of his focus on a text's relationship to its audience. Ricoeur begins his essay "What is a Text?" by defining text as "every utterance or set of utterances fixed by writing" (135). He concludes this work by determining that the act of reading accomplishes the "destiny of text" (150). Throughout, text is linked to human acts and actors, whether through creation, dissemination or consumption. Text, therefore, he claims, is defined by its usability, and its usefulness comes from the information which it conveys. Additionally, text remains based in the act of writing, and therefore to the orthographies of human languages. A text, according to Ricoeur, is an object that we can hold, and from which we can extract meaning, without destroying said object—and a text is created by writing with characters and symbols which are part of the lexicon of a known language. Ricoeur's definitions read as fundamentally material, having come from a period before the creation of the World Wide Web, and the concomitant normalization of purely digital documents and communication.

3.4.1.1 Digital Text

Digital texts, however, have a fundamentally different relationship to materiality than texts on stone tablets, papyrus, or bound in books, which affects the ways in which they are circulated, consumed, and otherwise used. The idea that digital information is no longer material is one that has persisted since the dawn of digital storage and communication. Negroponte (1995) proclaimed a transition "from atoms to bits" in the digital era, effectively declaring that digital data, and the bits that make it up are excluded from the building blocks of the universe, the laws of physics, and are a purely ephemeral object. This is, of course, patently false—digital objects, text included, merely have their materiality displaced. In order to read a digital text, the observer must bring it up on a display, but the displayed characters are a series of translations from 1's and 0's represented by electrical pulses that are made at physical point on a hard drive, through the entire stack and the programming languages in them, to a recognizable form of writing on a screen (Ford 2016).

Media Studies scholar Rita Raley describes code as "a deep structure that instantiates a surface" (2006), which is not dissimilar to Ricoeur's definition of text, but working from the back-end to the front, while Ricoeur's texts begin at the surface and develop depth through their being read. For most users, the surface is the only part of the code that they see or interact with, and often do not recognize the depth behind that surface until something breaks, or as Raley discusses, in code art, deliberately repositions the visible and invisible aspects of digitality, giving the viewer a peek at the work being performed behind the scenes. In concrete terms, new software and tools use the rules, norms, and even pieces of code from the older ones, and are dependent on hardware than is less malleable/updateable than software is. Because the digital layers at the surface of our interactions with technology continually build on one another, whatever the final instantiation of code that reaches us has become an amalgamation of not only the interfaces before it, but also of the depths of code behind it, all the way down to the hardware. And the social norms, constructions and performances that were built into those depths regenerate themselves with each new layer. This means that character and text production remains deeply tied to the paradigms of the cultural within which it was constructed, even (and perhaps especially) in the completely contstructed digital environment of the computing world.

That is to say, infrastructures which support digital text— word processors, command lines, other places you might type on computers, cellphones, etc.— have also all been created under the auspices of a specific understanding of what text is. DeRose et al examine this “model of text” and the ways in which it limits what users are able to do with digital vs analog writing systems. They highlight five ways that computers treat text: “text as bitmap, text as a stream of characters, text as formatting instructions, text as page layout, and text as a stream of content objects” (p. 8). They define the essential feature of text documents as "’content objects,’ [which] are of many types, such as paragraphs, quotations, emphatic phrases, and attributions. Each type of content object usually has its own appearance when a document is printed or displayed, but that appearance is superficial and transient rather than essential” (p.5). From this definition, and the rules of use between how each of these content objects (quotations can span multiple sentences, but paragraphs cannot span multiple chapters, etc), they propose that text, then, is “an ordered hierarchy of content objects or OHCO” (p.6). This definition of text has been widely adopted across text-encoding standards such as XML (Extensible Markup Language) and TEI (Text Encoding Initiative), although there has been debate in the TEI community especially about the accuracy and rigidity of OHCO as a model (Renear, Mylonas, Durand 1993). However, these standards deal with text at the word- or sentence- level, taking for granted somewhat the accurate and meaningful encoding of individual characters.

3.4.1.2 Global Languages and Digital Text

There is a swath of literature about the ways that character-encoding systems fail the orthographies and cultural literary requirements of languages that use non-roman alphabets and much of this comes from the period before the more widespread adoption of Unicode as an international standard and look to Unicode and/or the ISO to work towards a viable option for multilingual computing (Kim 1992, Shapard 1993, Osborn 2010). Kyongsok Kim, in his (1992) discussion of Korean Hangul orthography, points out a major discrepancy in priorities between the designers of character-encoding standards and the users of a particular language/alphabet: Korean Hangul characters represented in digital character-encoding standards predating Unicode have been broken down in linguistically bizarre as well as technically inefficient ways, due to programmer's unwillingness to deviate from the status quo of the Roman alphabet. He warns against valuing corporate interest over the preexisting rules of a language's written script and suggests that readers and writers will be reticent to change their language use habits to accommodate what is easiest or cheapest for computer developers or manufacturers.

Thomas Mullaney (2017) documents this struggle between orthographic tradition and western technological development over the past two centuries in his book *The Chinese Typewriter*. In it, he documents the adoption and adaptation of the typewriter in China and the larger Southeast Asian community, and the ways that the "stifled imagination" of technologists and designers from the West made it essentially impossible to conceive of a typewriter for Chinese, or any machine that produced Chinese writing as constituting a "typewriter." Likewise, there have long been movements within China and from the West to abandon the Chinese character writing system and "modernize" by adopting an alphabet. The difficulty in creating a Chinese typewriter influenced these movements, and in some cases "Chinese Alphabets" or other means of breaking characters down into a 'manageable' number of strokes were paired with particular physical pieces of technology. This struggle within Chinese is an example of how technologists have attempted to control and change language to better suit the needs of the machines they have already created, long before digital character encoding became the purview within which it was happening.

3.4.2 Politics and Power in Computing Systems

These linguistic and cultural discrepancies within the computing universe demonstrate the Western-skewed perspective of the computer scientists and engineers developing text standards for digital use. Considering the readable text as the "surface" that Raley refers to, it stands to reason, then, that the power dynamics and inequalities of these is also present in the deeper levels of the code which instantiates them. Using the Linux kernel as an example, Adrian Mackenzie demonstrates the ability of computer software to circulate and proliferate in cultural contexts. He writes, "Software production and consumption are also the object of intensely embodied identifications and personal styles" (2005 p. 72) Identity becomes bound up in not only the act of production of code, but in the form of production itself, to the point that 'illegible' code can be seen as a demonstration of true mastery of a programming language, and therefore a marker of membership in the inner circle of 'real coders.' The argument here is that this illegibility ensures limited access to the recreation and creation of code, and results in digital texts that reflect the ideologies of the people who are allowed to create it. Mackenzie touches on the gendered issues of how code functions within technological subcultures, and uses feminist lenses to examine the narrow worldviews that went into creating not only the technical code at the heart of computing, but also the communities of practice that today are actively resisting diversification. Mar Hicks notes that "histories of computing sometimes reflexively and unconsciously privilege those with the most power and implicitly endorse an ahistorical fiction of technological meritocracy" (2018

p. 16). This perhaps comes somewhat from the binary environment in which computers exist, and for which a clear distinction must be made the two ends of the binary in order for the logic of the computer to function.

Tara McPherson (2013) writes about the development of this type of emphasis on the binary abilities of computers as they developed in the 1950s and 1960s, which she calls "lenticular thinking," and the its parallels to the civil rights movement happening in the United States during the same time period. The reconceptualization of race in combination with an increased fascination with binary computing may see disconnected at first, but McPherson calls on us to understand these paths as being parallel in terms of who we think of as being a part of each movement, and who can be included in technical systems, and who is excluded. Adrian Mackenzie (2005) writes additionally about the development of the UNIX kernel, and how the circulation of code is both a product of and constitutive of a community of "true" coders-- which he notes is made up primarily white men, and he critiques this type of performative belonging through feminist lenses, including that of Judith Butler, wherein membership is reinforced by the ability to read and respond to in-group signs and signals. This results in a socio-technical infrastructure where the socially outcast remain outcast and are unable to engage with the process by which information is delivered to them in any meaningful way. Safiya Noble (2018) writes about the results of searching for the term "black girls" and being confronted with pornography and describes this as a way in which non-dominant identity categories become commodities and unable to be reclaimed, because the systems which control them remain dominated by white men, to the detriment of us all.

3.4.3 Translation

All of these definitions of text and texts deal with aspects of human communication, done largely through language. Unicode as a digital character encoding standard is a part of the sociotechnical infrastructure which makes up what we understand as writing in the digital world, and spans beyond the scope of a single language or orthography to encompass the means not only for multi-lingual communication via the computer, but also for the written translation of one text into another in digital space. The process of communication is vital to understanding how information is able to be moved between people, places, and contexts. Claude Shannon and Warren Weaver's (1949) mathematical model of communication (Figure 1) outlines a process by which information is moved from one entity to another. Originally applied in engineering, this model of communication has become foundational for Information Science as well as machine learning models today. Shannon and Weaver state that the information that moves between two ends of a communication is separate from what either entity perceives, due to the introduction of noise at each step of the process of communicating. Clear (that is, semantically equal from start to finish) communication requires that both parties agree on which parts of the message are signal, and which are noise.

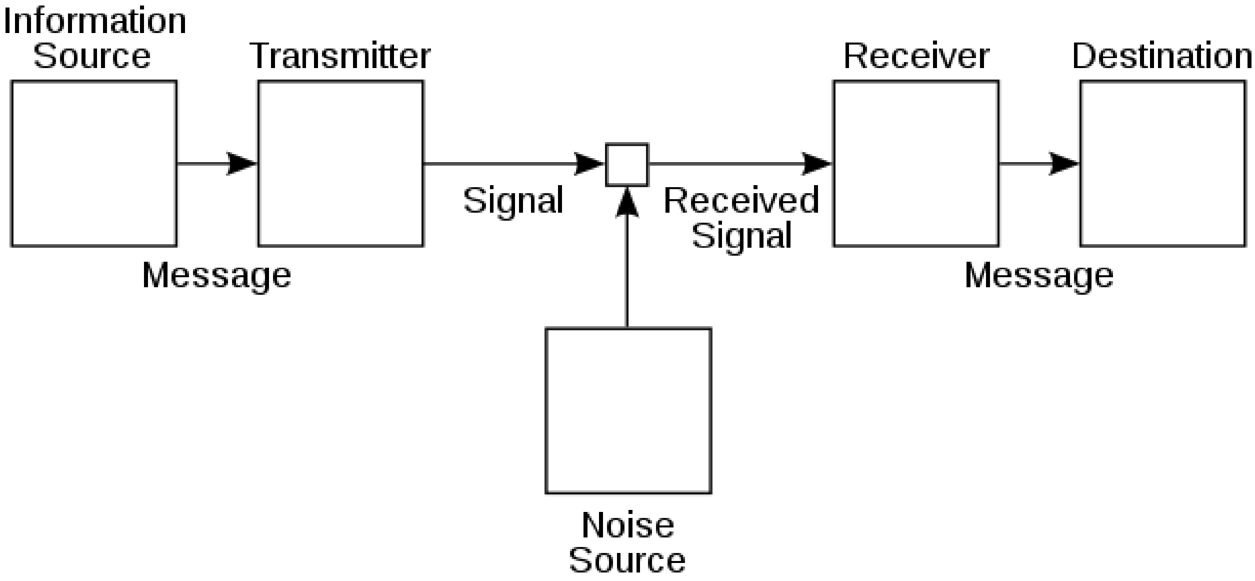


Figure 1 Shannon and Weaver's Mathematical Model of Communication (Shannon and Weaver 1949, p. 34)

The information at each step in this process is unique from its neighbors, and while each (hopefully) conveys the same message from start to finish, these multiple texts of a single piece of information can become records documenting the translation process. In dealing with speech or printed matter, we are often only left with the end result of a series of communications, a single document covering a single informational narrative. However, in digital spaces where iteration and reproducibility seem so much more expansive, there can be a proliferation of varying documents about the same information-communication in different stages and with different perspectives from one another—something we often refer to as 'versioning' today. This preserves some of the "noise" in the translation process that Shannon and Weaver describe, and can allow for it to be read itself as a meaningful piece of information. Which is exactly what this research does with documentation relating to Unicode and its development.

3.4.4 Documents

Considering the texts of web-based systems necessitates considering how documents exist online, and how we define a document, and its relationship to text. This question and those surrounding it has been the work of scholars since at least the 19th century; Paul Otlet, the Belgian academic and sometimes ‘father of information science,’ sought to organize all of existence, and developed a model for considering objects as documents-- a term that had previously only been applied to traditional written texts (Wright, 2014). He used this type of "document" and "documentation" to amass an indexed and cataloged museum-like collection and established a tradition of European documentalism that continued through to the 20th century.

Otlet outlined a series of ideals for the work of general classification, which included “the creation of documents in such a way that each item of information has its own identity and, in its relationships with those items comprising any collection, can be retrieved as necessary” (1934, p. 391) which places documents in between information and collections in a hierarchy of information-objects. Living in a pre-digital era, Otlet also defines a document in terms of materiality as “everything which expresses an element of knowledge directly (a known fact) by any graphic representation whatever (manuscripts or printed texts, inscriptions, epigraphy, drawings, iconography)” (Rayward 1990, p. 73). This is also placed within a hierarchy moving from the ephemeral “Knowledge or Understanding” through to the specific and material “Books.” The separation of documents from being specifically written texts, but rather defined by their ability to convey knowledge was revolutionary at the time, and Otlet’s taxonomies retain their influence in library classifications systems, and despite its analog origins, Otlet’s definition can apply seamlessly to digital documents.

After the second World War, Suzanne Briet took up the banner of documentation, and pushed the definition of a document even further. In her conceptualization, the defining feature of a document is not its text, or explicit ability to be "read" (in the way that both written texts and museum/art objects are), but its ability to be accessed in a controlled environment, and to provide evidence in support of facts (Briet & Martinet, 2006; Buckland, 1997). Notably, Briet speaks of “facts,” and defines documents by the ability to further some definite search for truth, but we think now of meaning and truth as socially constructed, with the implication then that something has the possibility to be a document in some contexts, and not in others. Contemporary document scholar Michael Buckland succinctly outlines Briet’s definition of a document thusly:

> 1. There is materiality: Physical objects and physical signs only;
>
> 2. There is intentionality: It is intended that the object be treated as evidence;
>
> 3. The objects have to be processed: They have to be made into documents; […]
>
> 4. There is a phenomenological position: The object is perceived to be a document. (1997, p. 504)

Like her contemporary, Vannevar Bush, who wrote that, "A record if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted," (1945), Briet considers that it is "indexicality—the quality of having been placed in an organized, meaningful relationship with other evidence—that gives an object its documentary status.” (Buckland, 1997, p. 806) Both Bush and Briet place heavy weight on contextualization, such as in Briet's famous antelope example, where an animal in the wild is not a document, but one held in a zoo and studied is.

Julian Warner (1990) attempts to unify the concept of the document and that of the computer, arguing that “At the level of discourse of logical operations, there is no distinction between a written expression, or program, and the computer which executes the transformation rules of that written expression” (p.19). This suggests that a well-functioning computer can transcend the map-territory problem, and create texts which are purely performative, in the semiotic tradition of J.L Austin (1962), and more broadly, Saussure’s (1919) concepts of signifier and signified. In Warner’s conceptualization, the computer is both the signifier, and the document itself.

3.5 Classification for Retrieval and Access

The discussion of organizing standards and regulation of texts is not, however, limited to the field of economics, and has a long history within the Library and Information Sciences as a part of the discussion of knowledge organization as well as infrastructure studies. Bowker et al note that the term ‘information infrastructure’ is “usually associated with the internet” (2009, p. 98), but information infrastructure is not only a digital phenomenon, and is made up of a “underlying concept...within a historical lineage.” (p.99). Its manifestation is dependent on its contemporary technologies but is by no means uniquely digital. Edwards defines information infrastructure as “ways to handle the functional problems of information storage, transfer, access, and retrieval” and points out that, “books and libraries remain our most important information infrastructures even today.” (Edwards, 2003, p. 197) However, knowledge organizations systems, taxonomies, and classification schemas are by no means the only forms of information infrastructure. Like any type of infrastructure which supports a society, there are physical and social aspects to the infrastructure, as well as conceptual or intellectual ones (Busch, 2011). While in “traditional” forms of infrastructure, such as roads or power grids, the physical aspects of these systems are most obvious— a power grid must have wires in the correct places, and individual homes must be connected correctly to that system— there are more ephemeral aspects to these systems as well— all generators on the grid must know and agree to where their power is distributed, and all homes must be built to accommodate the correct voltage, and comply with safety and usage monitoring standards (Hughes, 1987). This ratio of the visibility of physical to intellectual infrastructure is often flipped in computer-based infrastructure systems; digital standards exist and are applied entirely within the software of computers, but also require physical hardware to store and execute, as well as human labor to be maintained. There is a distinction between information about an infrastructural system, and infrastructure that is about information organization and dissemination—governing and scoping information is necessary for any type of standard, in order to communicate clearly to its users and handlers, but serves a largely supportive role, whereas an standard working with information infrastructure is governing the formatting, dissemination, and circulation of certain types of information. A classification system is the latter: a meta-information infrastructure. It describes what and where information is, and how it relates to the other information objects in its system.

3.5.1 MARC and FRBR

Philosophical definitions often conflict with practical ones, especially in the combination of physical and intellectual objects and their circulation—as in a library. The field of LIS is generally quite concerned with texts, but more specifically with the ability to define textual objects for the purpose of classifying, locating, and circulating those objects. The practice of cataloging, discussed above in infrastructural terms, involves the creation of texts about texts— metadata records—in order to document what is held in a particular collection, and provide meaningful descriptive data for other librarians as well as library patrons. The standard for creating catalog records which are machine-readable is MARC 21 (MAchine-Readable Cataloging), created and endorsed by the US Library of Congress (MARC Standards, 2006). MARC records today are born-digital records, and therefore depend on the presence of an inclusive character-encoding standard to create records which can be read consistently by people and machines across platforms and devices. The conflict between existing character-encoding systems and the needs of catalogers led the ALA to propose its own encoding scheme in 1990, which, while not adopted, is ample evidence that LIS has been concerned with the production of digital texts and the technologies that support them since the beginning of second-generation encoding standards. However, MARC records do not define categorically what is and is not a text, but rather focus on creating meaningful definitions for existing (usually text-based) objects. But this issue is not ignored entirely by LIS, in fact it is the subject of an ongoing intellectual project among the cataloging and LIS standards-creating community, which has led to the development of guidelines for defining texts at several conceptual and physical levels: FRBR.

Functional Requirements for Bibliographic Records (FRBR) is a “a model of the bibliographic universe” suggested by International Federation of Library Associations and Institutions (IFLA) in 1998, but is not itself implemented in any bibliographic practices (OCLC 2018, Floyd and Renear 2007). FRBR suggests that the intellectual content of a work is separate from the physical document which contains that content, with a spectrum of entities between a specific book and the story it tells. The Group One categories for works in FRBR are:

- The work, a distinct intellectual or artistic creation
- the expression, the intellectual or artistic realization of a work
- the manifestation, the physical embodiment of an expression of a work
- the item, a single exemplar of a manifestation. (OCLC 2018)

FRBR has been suggested since 1998 as a replacement for understanding information objects within the LIS framework, and was most recently updated in 2009. It currently consists of what is essentially a thought problem in the field. The division of a "work" into distinct sub-entities, with particular ownership and use rights for each level, proposes a radical shift in how documents are collected, stored, circulated and archived within libraries and other collecting institutions. This sort of paradigm is anything but unprecedented, with strong ties to the "idea/expression" division which comes into play with regards to copyright case law, as well as harkens back to ideas of signifier and sign with Saussere, and ideas and forms described by Plato (Saussure 1959, Plato, Lewis Cambell, and Benjamin Jowett, 1987). It seems that ever since humans could conceptualize information as separate from its material instantiation, we have been pondering how best to describe the relationship between the two.

But even considering FRBR as having a practical application for cataloging represents a philosophical shift within the practice in how documents are related to the information they contain. FRBR attempts to account for the multiplicity of documents which contain the same information (a library may have dozens of physical and digital copies of the latest bestseller), and make distinctions about what types of intellectual material the library trades in. In one sense FRBR firmly rejects the idea that libraries circulate ideas, or immaterial conceptions of ‘information,’ because that circulation functions at the “item” level, that of particular books.[3] However, by creating a hierarchy which distinguishes the “work” from the documents which manifest it, FRBR acknowledges that the value of a document or a text is more than just that of letters on a page. As stated above, the theoretical consideration of the distinctions between the types of objects which FRBR outlines is not new, or limited to LIS. Philosopher Roman Ingarden draws out these differences in his consideration of Chopin’s B Minor Sonata (1986), noting that there is no tangible “definitive” sonata, because a piece of music is instantiated through its performance, and while each performance may be an individual manifestation of the B Minor Sonata, there is in no perfect form to which it can be compared. Considering FRBR as a framework for an ontological model of understanding Unicode as a type of *classification commons* requires several additional levels of abstraction from the material—caves upon caves, with flickering shadows melting together, to put it in Platonic terms. The physical packets being created, managed, moved, and organized by Unicode consist of electrical pulses through wires, representing the raw binary series of ones and zeros that, in the end, makes up all the computer’s software, stored material, indices, files, photos, and so on—and working with the space of Unicode requires an almost surgical-like precision to distinguish pieces and levels of information from one another. It is not difficult to understand why

[3] This still applies to some forms of digital library media like ebooks as well, which have licenses limiting their ability to be accessed infinitely.

this area is so muddled, when man and machine are working from completely different ends of the stack to sort through the information the computer contains. It is for this reason that I believe that the FRBR framework is worth pursuing as a model for classifying and standardizing digital text (transmission). LIS scholars have reckoned with what FRBR implementation would look like, and in what ways the theoretical model would have to be stretched to accommodate the complexities of the digital world—to say nothing of the analog world it lives nested within. Renear and his colleague Floyd, while championing nearly opposite methods for doing so, both argue that in order for any practical implementation to be achieved, FRBR needs a dramatic remodeling of the relationship between its ontological structure and the way that names and roles are applied to objects within that ontology (Renear, Choi, Furner, Lagoze, and McDonough 2006, Floyd and Renear 2007). With these initial inquiries and known points of tension, I nevertheless suggest FRBR for examination as a potential starting point for information infrastructure modelling within the specific digital, textual world of digital character-encoding. We will return to this discussion in Chapter 6.

3.5.2 Cataloging and Classification

Classification systems are used by libraries and librarians to sort books and other library holdings into related groups. This practice is fundamental to what makes a library different from a storeroom of books, and while there are many classification systems used worldwide, the two most prevalent ones are the Library of Congress Classification/Subject Headings (LCC/LCSH) and the Dewey Decimal System (Olson, 2001, p.641). These classification systems divide all potential areas of knowledge into sub-groupings, and then provide descriptions and related terms for each grouping. Classification provides “a descriptive and explanatory framework for ideas and a structure of the relationships among [them]” (Kwasnik 1992, p. 63). The practice of applying these grouping to items in a library’s collection is called cataloging and translates the broadness of a classification system into the individual practice of a cataloger, and the item they are cataloging, the purpose of which is to facilitate access to users of the library holdings (Rubin 2010).

Classification systems are set up in order to make items more easily findable, and to group related items together. Any knowledge organization system—of which classification is a subset—is attempting to make sense of the world, whether merely through naming, or by creating complex taxonomies of categories and relationships for the various entities that make up the universe.

A classification system is a map of information and is subject to the socially-constructed nature of all maps (Crampton 2001). Classification and standardization are then, by definition, a means of flattening the complexities of the natural world in order to aid access and navigation. LIS scholar Emily Drabinski writes, “If social categories and names are understood as embedded in contingencies of space, time, and discourse, then bias is inextricable from the process of classification and cataloging” (2013, p.108). It is the responsibility of information organization professionals, and all those who work with the organizing infrastructures of our world, to tread carefully, and observe the ways that classification shapes and reshapes the lived reality of the world around us.

As Hope Olson observes, Western ideas of classification are built upon the idea of sameness and difference. The distinction between these two things is the fundamental structuring that makes all other classification possible. She writes,

> It is so ingrained that we do not even think of it as a "real" way of finding information. It is not uncommon to hear people deprecating their searching skills by admitting that in a library they just find a call number and then browse the shelves. They take the classification for granted as though it were a natural landscape rather than a well-manicured lawn that is the product of intellectual labor. Classification gathers things according to their commonalities. In doing so it demonstrates the effectiveness of this sameness/difference-principle duality.” (Olson, 2001, p. 115)

The invisibility of this type of meaning-making is evidence, then, of how fundamental it is to the way we organize information while navigating the world. Groundbreaking information infrastructure scholars Bowker and Star use their work to highlight the ways that infrastructures that we use to navigate the world often become invisible as they become indispensable (2000), and we therefore must consider systems of classification as foundational infrastructural parts of western knowledge-making standards.

However, as Olson goes on to point out, a classification schema must necessarily privilege some kinds of sameness or difference over others, and the process of doing so means the incorporation of a particular ideological foundation or worldview into the entire system. It also takes for granted the value of the sameness/difference distinction as one that is meaningful, and it has been so thoroughly incorporated into the way that we, as westerners, organize and describe the world that it is nearly impossible to conceive of a classification system that is not built from the idea of grouping similar things together, and separating them from dissimilar ones.

Susan Leigh Star discusses the loss that standardizing systems necessarily create when favoring one type of person or worldview over another. She uses the example that ‘people who are allergic to onions’ are not a powerful enough group to influence the systems of fast food chains, meaning that burgers will continue to come with onions by default, and even when ‘no onions’ are requested, the request may not be taken seriously, because the state in which someone is not allergic to onions is so much more frequent, and therefore powerful, that it exerts a great deal of control over the systems which produce fast food burgers in a timely and standardized fashion. This demonstrates how essential having to exclude some demographics in order to maintain functionality is, and the ways that it necessarily others those deemed ‘the exception’ (1990). While this example may seem frivolous, its frivolity is testament to how ingrained the norm of “not allergic to onions” is within our society. By making these choices, the world is necessarily flattened in ways that make the representation that the system is making of the world poorer, but also more functional for the majority (or favored set) of users.

This plays out in actual cataloging practice, wherein individual items are subject to the system, and may result in subject headings and shelf placements that make clear the biases of the systems creators. Several solutions to this problem have been proposed over the years, ranging from subject heading reform, to user-generated classification, to information literacy campaigns, and most acknowledge that the catalog and the shelf is a place where power must be continually negotiated and where ethical decisions must be made (Drabinski 2013, 2017). Activist catalogers have continually raised issues with the ethics of flattening a complex and ever-changing world into a functional classification system. Sanford Berman has called on the Library of Congress to erase or amend offensive subject headings, and the publication of his treatise Prejudices and Antipathies systematically outlines parts of the LCC that are outdated or offensive. He notes in the introduction to Prejudices that, “the LC list can only ‘satisfy’ parochial, jingoistic Europeans and North Americans, white-hued, at least nominally Christian (and preferably Protestant) in faith, comfortably situated in the middle- and higher-income brackets, largely domiciled in suburbia, fundamentally local to the Established Order, and heavily imbued with the transcendent, incomparable glory of Western civilizations” (Berman, 1971, p. 15). Berman’s work over the past 40+ years has resulted in changes to more than 140 of the headings that he suggested should be amended, and he continues to advocate for further additions and revisions (Knowlton, 2005, p.127-8).

Melissa Adler explores the difficulty of finding materials in the library catalog related to gender and sexuality, both in her own experience, and through the example of the work of queer theorist Eve Sedgwick, whose writings span literary criticism, poetry, and personal essay and whose subject headings fail to capture the nuance of her work (Adler, 2017). Adler raises questions of access, wondering “why wouldn’t this literary memoir [Sedgwick’s Dialogue on Love] be placed in the section that seems to be trying to collocate her work under her name?” (p.95). But in addition to collocation and access, Adler explores the potential censoring effects of classifying works by queer authors with such headings as “Sexual Perversion” and “Deviance.” How then, Adler argues, can a classification system claim to be unbiased and neutral when homosexuality continues to be associated with criminality? In this case, the LSCH itself is taking a moral stance through its application, and that stance is one that asserts that non-heterosexuality is immoral.

When homosexuality is associated with criminality of deviance, or a racial/ethnic group with outdated terminology or even slurs, its results not only in the exposure of the white, patriarchal heteronormative foundations of LC, but it can also confuse, alienate, or traumatize actual human beings who confront this classificatory violence on the shelf. Similar issues arise when discussing the experiences of people of color, whose language and self-descriptors are absent entirely from the vocabulary provided by the classification system and are likely to be cataloged by librarians who likewise lack access to the appropriate vernacular (Olson, 2001).

Catalogers working today, such as Barnard Zine Librarian Jenna Freedman, also struggle to apply headings to works that do not fit the standardized ideal of materials collected by libraries. Freedman's work with zines in particular, and her struggle to find appropriate headings, reflects the narrowness of scope that the authors of the LSCH had (and continue to have) in mind when creating subject headings. Freedman notes that "A typical LC excuse for its offensive headings is that their job is to serve members of Congress, so the headings they choose reflect Congressional language and culture," however, she continues, "The works I'm cataloging, zines, are usually created by women, and young women at that. They are often created by queer women, and in smaller numbers they're by women of color, people outside the gender binary, and women with disabilities. The zines are typically informed by an anarchopunk political and social ethos that I would venture to say is not highly represented in the House of Representatives." (Freedman, 2016).

While activists such as Berman and Freedman advocate for changes to the Library of Congress, other LIS scholars debate the ethics of erasing the controversy and complications within classification systems. Emily Drabinski considers cataloging and classification systems from the perspective of queer theory, which argues that categories are permeable and in continuous flux. She suggests, then, that no classification system will ever be perfect or unbiased, and that the work of librarians should be not to erase past offenses, but to highlight the gaps within the system, and to engage in “dialogue with patrons that will help them tell the troubles of those schemes. Users can be invited into the discursive work of both using and resisting standard schemes, developing a capacity for critical reflection about subject language and classification structure” (Drabinski, 2013, p. 107). The work of this research is in line with what Drabinski proposes, and by engaging with the creation, application, and continual negotiation of the standardized space of digital-character encoding, I intend to create a dialogue around the power structures at play within information structures, and the long-term effects of favoring certain human experiences and paths of use over others.

4.0 History of Character-Encoding

The history of digital character encoding runs in parallel, unsurprisingly, with the development of the digital computer. But codes and encoding processes have been in use much longer than the mid-twentieth century. In this chapter, I will outline the history of the tools we have used to encode alphabets, from the analog through to the development of the Unicode Standard which continues to reign supreme. First, I discuss pre-computer long distance communication tools, and the codes which developed around those technologies to streamline the transmission of language from once place to another. These technologies set a precedent for the modes in which messages are encoded into a transmittable format which still serves as a framework for how encoding happens today. Next, I return to the work of Warren Weaver, whose work ideologically if not physically, shaped how computers deal with language today. Weaver's conceptions of language and translation, specifically, have become the foundation for the architecture of modern computing, and I demonstrate how his assumptions about the nature of language have reverberated and amplified into a core part of how computers are able to interact with language today. Finally, I introduce ASCII and EBCDIC, the two most prevalent encoding standards prior to Unicode, and outline the conflict of interests and pursuit of market share which led to the split between the two, which in turn lead directly to the establishment of the Unicode Consortium and the development of the Unicode Standard.

4.1 Machine-mediated communication

The earliest of what we would call modern computers had to deal with encoding characters to and from binary, but not with the digital display of those characters; this lineage remains in the way that computers deal with text today, with ‘translation’ for human legibility being essentially secondary to creating and maintaining the bits and bytes which make the computer work. Throughout this chapter and elsewhere in this work we will see these choices being made, and some of the effects those choices have on how technology is able to be implemented and used today.

My focus here is not on whether or how computers understand language/text, but rather on how a particular definition of what text is has become hard-coded into our technological systems, and from there into social structures built on top of those systems. All modern computers, “operate only upon binary numbers. Letters, decimal numbers, and symbols can only be processed after being coded into binary numbers” (Brock 1975, p. 85). Gerald Brock, writing at the beginning of widespread computing in industry as well as the US government, frames the need for self-regulation in computers standards (not limited to character encoding), noting that “effective standards greatly facilitate the interchange of data and programs among the machines of different manufacturers and allow the user to combine equipment from several suppliers.” (p. 75) As Brock implies, this history is just as much dependent on hardware as it is on software. In the case of digital text, this is done through the use of character encoding standards, which, in the most basic sense, create a cipher linking certain pieces of coded information—in the case of digital computers, combinations of binary bits --to particular letters, numbers, and/or symbols used in writing. Such standards predate computing, and such pre-digital systems include Morse Code, International

Marine signal flags, and even the original standard typefaces of early printing presses, as alphabets themselves initially became formalized (Standage, 1998). Today, digital text must still be displayed via a piece of hardware, and that hardware retains systems linking it back to previous generations of computing and the infrastructures which upheld them. This is particularly evident in text input. The keyboard input system was taken directly from the typewriter and continues to emphasize the computer's focus on text even as it is moved into a purely-digital, on screen form.

In the earliest days of digital computing, Warren Weaver (1949) conceptualized machine translation of languages, which imagines that each language is a one-to-one representation of each other (or of English-- the implied "true" language), and has can thus be translated through a binary, machine-based translation process. Weaver suggests that machines could be used to translate one language to another effectively because there is a universal language. Weaver's suggestions are colonialist at best—he refers to Chinese as no more than "English translated into a Chinese code", a fundamental misunderstanding of not only Chinese as a language, but of human linguistic development as a whole (Weaver 1949, Dong 2014). While he admits to having only a layman's understanding of linguistic theory, his concept has proven sufficient to functionally *approximate* this type of translation, and in doing so has ended up creating the universal language that Weaver hints at: binary code. This is a language no human speaks, and no human reads, but is the interstitial point for the transmission of digital text (and all digital information) across networks today.

Well before the internet, home computing, or digital media storage was widespread, the human imagination conjured up wild ideas of machines might be able to do for us, and much of that imagining focused on making the production of meaningful messages to be transferred from human to human (Veness and Barbrook 2007, Taub 1961). Likewise, long before the invention of digital computing, humans have pursued technology to make writing easier-- it is impossible to unlink the development of text from the development of language. In this sense, it seems almost inevitable that computers should take on the problem of language. As the computing power of the original mainframes increased, so did their ability to process complex information, such as text.

One of the most famous early instances of the increasing capability for computers to work with texts is that of Father Roberto Busa, who via the labor of dozens of women-- at the time referred to as "computers" themselves-- was able to create a concordance of the works of St. Thomas Aquinas (Terras and Nyhan, 2016). This was accomplished by dividing the writing by word, encoding each word onto a punch card, and then using the computer to essentially house a physical database of the words Aquinas used in his writing. Busa could then ask questions about Aquinas' writing, and the computer (a machine this time) would be able to answer. It was possible to ask the computer ‘How many times was the word "Christ" used in Aquinas' writing’, and receive an answer via the computers processing. This is often cited as an origin point for the field of digital humanities, and while Busa’s experiment may not claim sole ownership of that title, it does certainly represent a turning point in conceptualizing how computers might aid or transform literary research (Underwood 2017).

It is important to note, however, that Father Busa was working at the word level of text, meaning that each word was encoded into the system as an indivisible unit, and the computer could tell you very little about how many times the letter "C" appeared in the data. And indeed, it remains the case that most textual analysis, is at the word level or higher, meaning that the complexities of text creation at the character level, and its analysis by computer specifically, remain largely unexplored.

4.2 Character Encoding Before Unicode

Despite the fantastic variation of languages and scripts in use today, the invention and development of language as a tool evolved from the same basic needs. From a phenomenological perspective, all human language and writing are devices for expanding our individual and collective memories. Storytelling practices passed down essential knowledge and cultural practices between generations, and the first written marks were used to track the passage of time, the movement of animals and goods, and to establish a shared identity between groups with shared literacies (Bonvillain 1993, Robinson 2002). Alongside the development of written language and its widespread adoption, came the development of technologies to streamline the production of written texts. Increasing manpower, such as in the case of entire monasteries of monks who dedicated their lives to translating and reproducing editions of the Bible, is one approach to increasing production. Another is to reduce the effort required to produce, transmit, and translate written messages through standardization practices and the adoption of standardized equipment to do so across time and place. And naturally, that is the route we will be tracing in the remainder of this chapter.

4.2.1 Mechanical Codes

Mechanical codes for transmitting language technically began with the invention of writing—but in the interest of time, we will skip forward to the development of codes for transmitting language in real-time across long distances—usually considered to have begun with the telegraph. The first telegraph put into wide-spread use was entirely visual, very different from what we think of as a “telegraph” today—a tachygraph or semaphore system. It was originally demonstrated by Claude Chappe in 1792, and eventually put into use across France (Standage 1998). It consisted of an antenna-like pole rising from the transmission station, with a large rotating crossbar, called the regulator, with a smaller rotating bar at each end, see Figure 2 (below). The regulator “could be aligned horizontally or vertically, and each of the small arms, called indicators, could be rotated into one of seven positions in 45-degree increments. This allowed for a total of 98 different combinations, four of which were used as control codes, leaving 94 codes to represent numbers, letters, and common syllables” (Standage 1998, p.11). This technology was developed by Chappe and his brothers, who also developed a 94 page code book, with 94 numbered words or common phrases on each page, allowing for nearly 9,000 predetermined messages to be sent with only four digits—the first two representing the page, and the second pair the numbered item on that page.

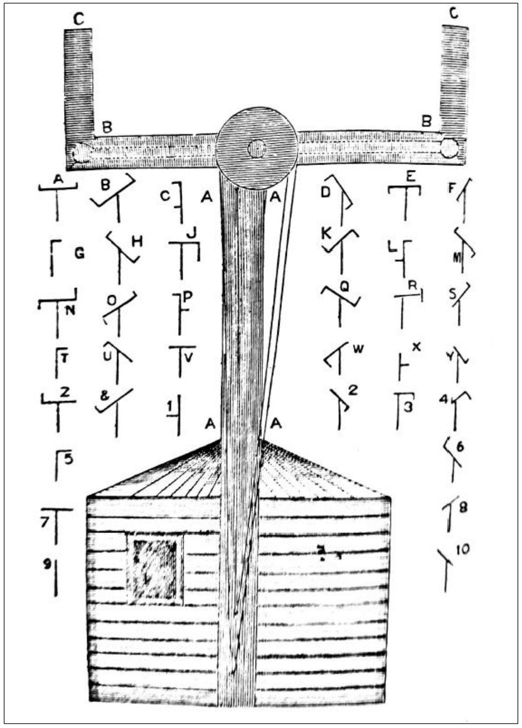
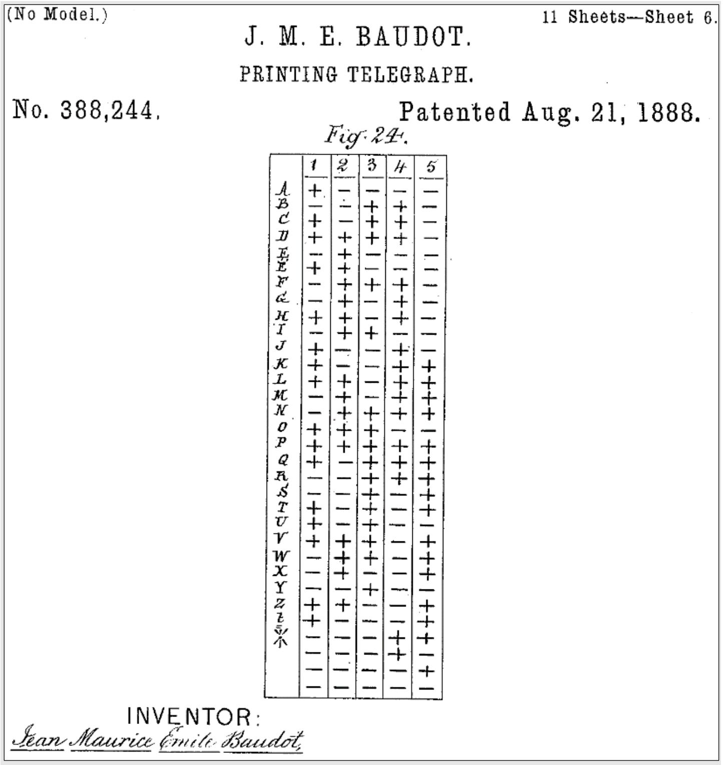


Figure 2 Diagram of the visual telegraphy system developed by Claude and René Chappe

While it took a significant amount of time for the concept of telegraphy to marry with electricity and create the modern telegram, the Chappes’ efforts, along with myriad scientists and inventors of the era focused on the same problem, helped to create and establish many of the norms we still make use of in transmitting linguistic messages over long distance, such as control codes, synchronization between stations, and the beginning of standardization of encoded messages.

4.2.2 ITA/ITA2

The International Telegraph Alphabet (ITA), also known as the Baudot code, is a character encoding standard designed in the 1870s for teleprinters, the hardware used to transmit and receive telegraphic messages (Baudot code, 2003). ITA was a five-bit code, consisting of 30 capital letters of the French alphabet, a set of 24 numbers, symbols and six control codes: Null, Figure, Blank, Erasure, Blank, and Letter. It was later adapted for use with the English alphabet, with fewer letters and an altered set of symbols, eschewing punctuation for more mathematical symbols.



(No Model.)

11 Sheets—Sheet 6.

J. M. E. BAUDOT.

PRINTING TELEGRAPH.

No. 388,244.

Patented Aug. 21, 1888.

Fig. 24.

	1	2	3	4	5
A	+	−	−	−	−
B	−	−	+	+	−
C	+	−	+	+	−
D	+	+	+	+	−
É	−	+	−	−	−
E	+	+	−	−	−
F	−	+	+	+	−
G	−	+	−	+	−
H	+	+	−	+	−
I	−	+	+	−	−
J	+	−	−	+	−
K	+	−	−	+	+
L	+	+	−	+	+
M	−	+	−	+	+
N	−	+	+	+	+
O	+	+	+	−	−
P	+	+	+	+	+
Q	+	−	+	+	+
R	−	−	+	+	+
S	−	−	+	−	+
T	+	−	+	−	+
U	+	−	+	−	−
V	+	+	+	−	+
W	−	+	+	−	+
X	−	+	−	−	+
Y	−	−	+	−	−
Z	+	+	−	−	+
t	+	−	−	−	+
*	−	−	−	+	+
	−	−	−	+	−
	−	−	−	−	+
	−	−	−	−	−

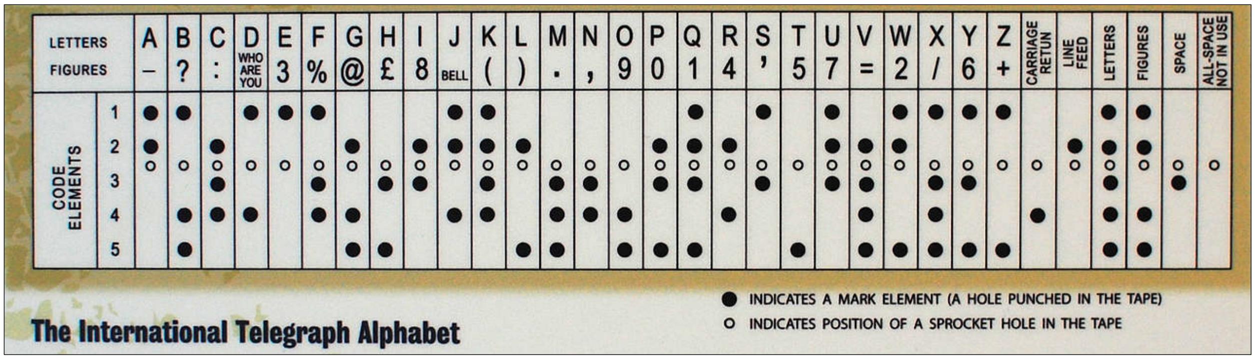
INVENTOR:
Jean Maurice Émile Baudot,

Figure 3 Code table as presented by Baudot in the patent for his teletype machine.

As seen in Figure 3, above, the Baudot code introduces the idea of using a binary code to represent letters and symbols. In each row above, we see the character (A, B, C…, and so on), followed by columns labelled 1 through 5. Each of these columns represents a "bit" within the code which can either be programmed as a "+", equivalent to a 1 in computing binary, or a "-", equivalent to a 0. Each character, therefore, is made up of the combination of values for these five columns. A is “+ - - - -“ and Z is “+ + - - +”, or 10000 and 11001, respectively.

The success of the Baudot code, and the patented machine it was used with, caused the rapid dissemination of the code worldwide, and several altered versions, adaptations, and spin-offs appeared (Standage 1998). The most successful of these was the Murray code, created in 1901 by Donald Murray which adapted Baudot’s code to be used with paper tape and a keyboard perforator, where a punched mark indicated a +/1 and unpunched a -/0. Murray’s code was then adopted by Western Union, which they used until the 1950s (Fischer 2012).

In 1924, Murray’s code was adapted and expanded into ITA2. ITA2 characters were also 5 bits, but made use of two shift keys (Letter Shift, and Figure Shift), for indicating which of two possible symbols (the aforementioned Letters and Figures) on the code table to refer to. In this modality, “11000” would refer to the letter “A” when used I n the Letter Shift mode, and the symbol “-“ in Figure Shift mode. ITA2 was adapted for the Russian alphabet in RTK-2, and with only a few minor changes for American English users (Cyzborra 1998, Fischer 2012). ITA2 was the encoding standard most used globally until the introduction of ASCII.



LETTERS		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	CARRIAGE RETUN	LINE FEED	LETTERS	FIGURES	SPACE	ALL-SPACE NOT IN USE
FIGURES		–	?	:	WHO ARE YOU	3	%	@	£	8	BELL	(	)	.	,	9	0	1	4	'	5	7	=	2	/	6	+						
CODE ELEMENTS	1	●	●		●	●	●				●	●						●		●		●		●	●	●	●			●	●		
	2	●		●				●		●	●	●	●				●	●	●			●	●	●					●	●	●		
		○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
	3			●			●		●	●		●		●	●		●	●		●		●	●		●	●				●		●	
	4		●	●	●		●	●			●	●		●	●	●			●				●		●			●		●	●		
	5		●					●	●				●	●		●	●	●			●		●	●	●	●	●			●	●		

● INDICATES A MARK ELEMENT (A HOLE PUNCHED IN THE TAPE)
○ INDICATES POSITION OF A SPROCKET HOLE IN THE TAPE

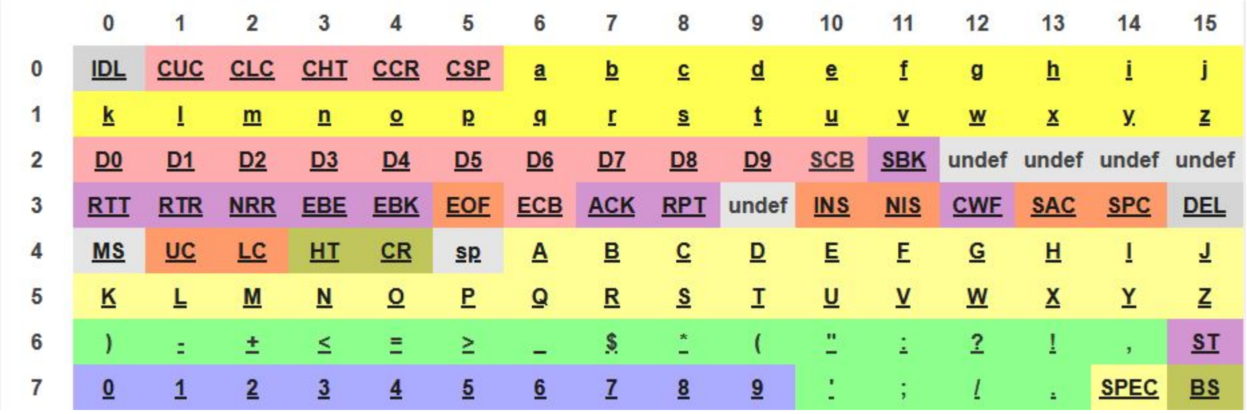
The International Telegraph Alphabet

Figure 4 A reference chart for ITA2, showing how each character would appear on the punch tape output.

4.2.3 FIELDATA

While ITA2 was being used in the business and public sectors, the US government and military were looking for alternatives for use in the field. FIELDATA was a project of the US Army, started in 1956, to unify, standardize, and streamline communications on the battlefield. It is an “integrated family of data processing and data transmission equipment being developed for Army use. A unique feature of this family is the almost complete disappearance of conventional distinctions between communications and data processing.” (Luebbert 1959, p. 189) Unique to FIELDATA at the time was its emphasis on data formatting. Both ASCII and FIELDATA were both intended as standards for hardware reading, with as much information packed into the code itself as possible, to minimize the need for software ‘translation.’ The goal here was not to produce documents (printed or displayed) for the human reader, but to streamline the reading and writing process for the computer itself. Early computing historian Tom Jennings notes that while “ASCII’s design was well underway when FIELDATA was deployed, at least one person worked on both standards.” (Jennings, 2020)

The FIELDATA encoding system uses 7 bits; 6 bits for the character code itself, and the seventh as a control bit indicating which of two 64-character tables to refer to. Consider this seventh bit something like a 'shift key' for FIELDATA, effectively doubling the number of characters possible with a 'six-bit' code from 64 to 128. This was important because "computers of the time considered characters to be six bits in width, and […] had register and memory widths of 18 and 36 bits." (Ibid.) This causes FIELDATA to function quite differently from encoding standards today, where a character code is thought of as an "atomic unit" (ibid) which loses its meaning if the bits are separated from one another. Rather, in order to create extra space 'within' the six-bit format, the seventh bit, called a tag, gives the other six bits a different meaning based on which set of characters it 'points' to. In FIELDATA, there are two tables: Alphabetic (indicated by a seventh bit value of 1) and Supervisory (with a value of 0). Supervisory codes were (presumably) standardized codes and functions related to specific bit slots within FIELDATA which conveyed information about the transmission itself, including message formatting functions, and basic error/flow control correction. In other words, Supervisory codes were pieces of metadata in the FIELDATA format to describe the context of the 'text' contained within the alphabetically-coded portion of a FIELDATA message. Contextually, this makes sense for a standard developed by the Army for use in the field, and with a focus on hardware and software efficiency. These Supervisory codes allowed the computers to unpack and translate FIELDATA-encoded messages without specialty software or human intervention, while also mechanically indicating which table a character in a message should refer to. FIELDATA notably also first introduces separate encodings for capital and lower case letters.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	IDL	CUC	CLC	CHT	CCR	CSP	a	b	c	d	e	f	g	h	i	j
1	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
2	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	SCB	SBK	undef	undef	undef	undef
3	RTT	RTR	NRR	EBE	EBK	EOF	ECB	ACK	RPT	undef	INS	NIS	CWF	SAC	SPC	DEL
4	MS	UC	LC	HT	CR	sp	A	B	C	D	E	F	G	H	I	J
5	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
6	)	-	±	≤	=	≥	_	$	*	(	"	:	?	!	,	ST
7	0	1	2	3	4	5	6	7	8	9	'	;	/	.	SPEC	BS

Figure 5 FIELDATA Code Table, as presented by Tom Jennings

FIELDATA, while mechanically obsolete today, is a godfather of contemporary character encoding, and remains present in UNIVAC computers running legacy COBOL software It also lives on in its sibling ASCII, which itself remains persevered within the Unicode standard.

4.2.4 ASCII

In March of 1968, President Lyndon B. Johnson mandated that all US computers support the character encoding standard ASCII – the American Standard for Information Interchange (Johnson 1968). ASCII was first developed in the early 1960s and made its debut on commercial computers in 1963. It used 7 bits to encode characters, allowing for a total of 128 characters. While FIELDATA, discussed above, also technically used 7 bits and accommodated 128 characters, ASCII does not use the seventh bit as a 'tag' for 6-bit encoding charts, but rather incorporates all 7 bits into the numbering of their character codes. ASCII does support backwards compatibility with FIELDATA, ITA2 and other 6-bit standards, by reserving the first 64 character slots in ASCII for those compatible with 6 bit standards. Additionally, characters 1-32 (excluding the 0 slot, which is a null character) are control characters, similar to the Supervisory codes present in FIELDATA, and interoperable with them. These characters speak directly to the hardware of the computer and specify formatting and print settings for the surrounding text.

It is important at this point to note, that while ASCII is often considered the first ‘modern’ character encoding scheme, and has far more name recognition today than any of its contemporaries, it is still a computer-facing standard. That is, it is designed to encode information in a way that is efficient and meaningful for the computer, rather than human users. Its function is more complex than a simple cipher where A=1, B=2, and so forth. By encoding the control characters into the same namespace as the alphabetical, numerical, and assorted other characters, ASCII comes with all the tools to translate itself, and an ASCII-compatible computer requires no outside interference or assistance to read and process ASCII-encoded information. Previous standards, including FIELDATA also made use of extra code spaces within their standard to convey metadata about their content, but ASCII was the first to have its system of such codes formalized, standardized, and endorsed by the US government.

Between its first release in 1963, and ASCII-67, lowercase English characters were added to the standard, filling out the table, and largely taking away the possibility for individual implementations of the standard to add non-English characters to the unused code slots, such as country-specific currency symbols, diacritics, or non-Roman characters. YUSCII (Yugoslav Standard Code for Information Exchange) and its variants reformatted the ASCII code space to work with both the Roman and Cyrillic alphabets, depending on software to interpret the encoded text and determine which alphabet to use (Cyzborra 1998)

The alphanumeric characters and symbols within ASCII include the 26 upper- and lower-case characters of the English alphabet (52 total), numerals 0-9, and from 11 to 25 special graphic symbols already in common use among other standards. Add these to the 32 control characters, and ASCII contains 105 to 119 standardized characters, leaving 9 to 23 empty slots for expansion or localization. Initially ASCII took a typewriter approach to diacritics, meaning that each such character takes up not only a unique code space in within the standard, but a unique space in print/display as well, resulting in accents being written after the letter they modify, such as "cafe'."

4.2.5 EBCDIC and the P-Bit

The tech industry was more or less ready to roll over to the use of ASCII in the early 1960s, with its first official documentation being released in 1963. However, IBM, one of the main sponsors in developing ASCII, faced a crossroads at the same time: They had created the IBM 360 mainframe computer, with the intention of it being compatible with ASCII, as well as backward compatible with their own in-house encoding standard, Binary Coded Decimal (BCD). Bob Bemer, head of Logical Systems Standards for IBM at the time, describes the situation thus:

> "IBM was going to announce the 360 in 1964 April as an ASCII machine, but their printers and punches were not ready to handle ASCII, and IBM just HAD [sic] to announce. So T.V. Learson (my boss's boss) decided to do both, as IBM had a store of spendable money. They put in the P-bit. Set one way, it ran in EBCDIC. Set the other way, it ran in ASCII. But nobody told the programmers! […] They spent this huge amount of money to make software in which EBCDIC encodings were used in

> the logic. Reverse the P-bit, to work in ASCII, and it died. And they just could not spend that much money again to redo it (Bemer, 1999)."

The hardware was designed with an internal switch, called the P-bit, that was supposed to change the encoding system back and forth between ASCII and EBCDIC—with EBCDIC being the updated 7-bit version of IBM's BCD standard. In the rush to push the hardware to market, however, there was a miscommunication between the higher-ups and the thousands of programmers working on the ground to build the machines and their operating systems. Bob Bemer, referred to as "the father of ASCII" and the "father of COBOL", blames this mistake directly on his boss, T.V. Learson, and IBM's refusal to wait for the final updates to the ASCII standard. EBCDIC was built in-house at IBM, specifically for IBM computers, which was not that uncommon at the time, as the process of technical standardization for computer software was in its infancy. However, as I mentioned earlier, IBM was actively involved in the process of developing a 'universal standard' for character encoding, which became ASCII. IBM's undermining of ASCII's efforts set back attempts for standard consolidation by 30 years.

The problem was that the order in which characters are encoded in ASCII as opposed to in EBCDIC. ASCII lists the capital letter alphabet A-Z, first and consecutively in its table, while EBCDIC integrates the upper- and lower-case alphabets: A, a, B, b... Z, z. Additionally, the programmers used EBCDIC encoding in the programming of the IBM 360, meaning that if the hardware was switched to be ASCII compatible, the computer would not function. The reason this happened appears to have been largely financial. IBM did not want to wait for the formal release of ASCII, nor spend the time to reprogram the thousands of computers they were waiting to push to the market. As a result, the 'universality' to which ASCII aspired was undermined from the very beginning, and EBCDIC continued to be present as an encoding standard for IBM machines until the official adoption of Unicode, which is expansive enough to maintain backwards compatibility with both ASCII and EBCIDC (Jennings 2020, Bemer 1999).

4.3 Creating Unicode

The initial proposal for what would become Unicode appeared in August of 1988, from Dr. Joseph Becker, at the time employed by the Xerox Corporation, though the system he proposed was not yet endorsed by Xerox, who was using their own proprietary standard (Becker 1988). It lays out a system of encoding and namespace assignment that uses 16 bits, resulting in a total of ~65,500 possible characters. In a section titled "Sufficiency of 16 Bits?" he writes,

> "Are 16 bits, providing at most 65,536 distinct codes, sufficient to encode all characters of all the world's scripts? Since the definition of a "character" is itself part of the design of a text encoding scheme, the question is meaningless unless it is restated as: *Is it possible to engineer a reasonable definition of "character" such that all the world's scripts contain fewer than 65,536 of them?*
>
> The answer to this is Yes" (Ibid, emphasis added.)

Future troubles with this assumption aside for the moment, we can see in plain English the prioritization of efficient engineering and compact code for the machines that will use Unicode over any kind of realistic assessment of the breadth of human writing systems, and fails to consider altogether the needs of the human user in relation to the system being built. This harkens back to Warren Weaver's assessment of machine translation nearly forty years earlier, and reflects the general attitude that the efficiency of the computer, and data’s legibility to the computer was tantamount, and that human behavior could (and should) adapt to accommodate the code, as well as that all human scripts could (and should) be reduced to fit within the original 16-bit framework of Unicode (Weaver 1945, Mackenzie 1980, Mackenzie 2005).

4.3.1 Compatibility

In considering the motivations of the founding members of the Unicode Consortium who created the Standard and the Consortium itself, we must look back again to the series of events which took place surrounding the previous attempt to launch a unified character-encoding standard in the mid 1960s. As discussed previously, ASCII was already in use during this time, but as computing grew more widespread and communication between humans via computers became more common, designers and engineers at Bell Labs and IBM realized that fundamental linguistic agreement on the back end was necessary to make sure messages could be created, transferred, and stored without translation or file degradation issues. The flawed hardware switch on the IBM 360 computers which would have allowed for the use of both ASCII and EBCDIC encoding standard proved damning enough to undermine the entire operation. The technical failure of the P bit was a pivotal moment in the history of character encoding, but was by no means the only technical shortcoming faced by 7-bit encoding systems (ASCII). The concept of cross-compatibility in computer hardware and software was nascent at the time, and the ability to universally update the shared components (including encoding standards) of computers was almost entirely absent. This resulted in an environment where each individual computer or group of computers working together could ‘evolve’ on its own, outside of standardized updates applied from the top down.

In order to successfully create a universal standard, the early developers of Unicode determined that the new standard would need to be backwards compatible with ASCII-encoded files, at the very least. The failure of the universal hardware roll out of ASCII machines, and lack of ability to push updates to computers universally made this necessary in order to preserve as much of the ASCII-and-earlier information extant in the world’s computers. This marks the beginning of (what would become) Unicode’s interest in preservation of digital textual materials, which remains an important part of its mission today, but is notably distinct from the idea of “language preservation” with which Unicode promotes itself today.

4.3.2 Linguistic Considerations

Encoding standards such as Unicode permanently assign unique identifiers to defined, constituent pieces of the larger whole. In Unicode’s case, those pieces are characters. The amount of semantic information contained within a single character, however, varies dramatically among orthographies. When these semantic and other linguistic variations are not accounted for during the development of the systems designed to standardize them, trouble inevitably occurs.

As we have seen, ASCII was developed for largely US - UK (English speaking) communication, building on the legacy of wartime communication standards created by the US military’s DARPA, and later, ARPAnet. The use of other scripts or special characters during this time were mutations of ASCII, rather than ever becoming incorporated into the ASCII standard formally. These mutations vary from the addition or substitution of some [accent] characters, to a complete remapping of the standard to support different alphabets, and often continued down their own branch of the evolutionary tree, morphing into language/alphabet-specific standards. The countries of the former USSR, in particular, formalized and maintained ASCII-variant standards for Cyrillic languages and keyboards developed during Soviet isolation (Czyborra 1998).

Dozens of global ASCII variations, as well as bespoke standards for languages that did not map well to the 128-character layout of ASCII (and QWERTY keyboards) were popping up, and this had the potential to prevent clear communication between PCs with different encoding standards. As Unicode began its development, making sure that as many of these standards could be comfortably ported into the Unicode Standard as possible was a high priority. Like with ASCII, efforts were made to arrange characters within the expanded namespace to make backwards compatibility as universal as possible. As a result of this, many languages' full orthographies were dumped into Unicode as-is. This aides ease-of-access from an encoding perspective, as all the characters for a single language could be captured via a range of encoded IDs, for instance, the Greek alphabet is encoded together under the UTF U-03XX encodings, preceded by Latin in the U-02XXs, and followed by Cyrillic in the U-04XXs. This organization system, however also created points where similar or identical characters appeared more than once within the Standard.

In many places within the Standard, characters are cross-referenced with their sibling characters, much in the same way that library catalogs cross-reference items listed under more than one subject heading and point the user to similar categories within the classification system. This solves the problem of multiple identical characters for the computer—for which it was not really a problem in the first place—by creating built-in connections between characters which would normally be siloed by language. It also provides some assistance to humans attempting to trawl the Unicode nameslist, but Unicode itself makes it quite clear that the publically available nameslist is not truly intended for human use, but is rather an adaptation of the machine-readable version, providing a text-searchable, but not semantically linked in any way, list of all the characters Unicode supports. That is to say, while the beginnings of a "cataloging mindset" appear within this structure, the infrastructure necessary to support its use by humans remains missing.

Taking this approach allowed developers and programmers to work quickly on creating and implementing the new standard by passing over the specifically language-based issues with ASCII and focusing on building the infrastructure of Unicode, which was to be big enough to accommodate the occasional double-encoding of a character. In fact, "get bigger" seems to have been the primary solution the Unicode Standard developers came up with for dealing with the growing multilingualism of computing environments.

4.3.3 Technical Considerations

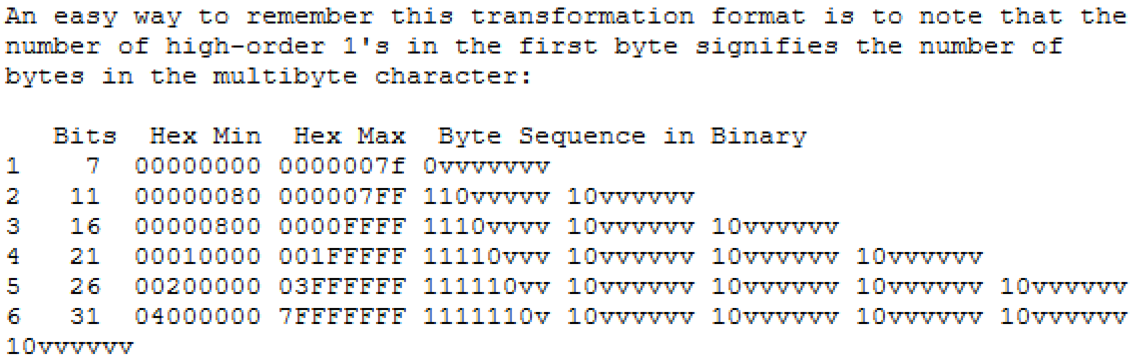
The number of characters possible on a typewriter were limited to the number of keys (though doubled by the use of the Shift key), and the number of keys were limited by a vaguer idea of what would be considered 'efficient' for the process of typing, and 'superior' to the previous generation's technology. The number of characters in digital text is not limited by the number of keys (though quick access to characters remains bounded to that physicality), but by the number of bits available for characters to be assigned to. ASCII, being a 7-bit standard had a limit of 128 unique identifiers to be assigned to the letters and characters of the English alphabet.

The Unicode Standard intended to increase that limit by several orders of magnitude, first by increasing the basic namespace to 8-bits, doubling the number of character slots to 256 and then by the implementation of a very clever bit of numbering trickery, which allows Unicode to function as a 32-bit standard, while still limiting the length of each character’s unique identifier to no more than 8 characters.

The standard and format developed by IBM and originally set to be used as Unicode had a few issues, especially in regards to backward compatibility. This first draft universal coded standard (UCS) “provides the capability to encode multi-lingual text within a single coded character set. However, UCS and its UTF [Universal Character Set Transformation Format, or later Unicode Transformation Format] variant do not protect null bytes and/or the ASCII slash ("/") making these character encodings incompatible with existing Unix implementations” (Pike, 2003). These issues were identified by researchers working on Plan 9 from Bell Labs, an operating system based on the Unix kernel, and therefore with a specific investment in compatibility with Unix implementations. One developer, Rob Pike, recalls feeling frustrated with IBM’s UTF, and along with Ken Thompson set out to create an alternative version which maintained encoding efficiency while solving some of the problems they had encountered.

> “UTF-8 was designed, in front of my eyes, on a placemat in a New Jersey diner one night in September or so 1992. What happened was this. We had used the original UTF from ISO 10646 to make Plan 9 support 16-bit characters, but we hated it. […] I received a call from some folks […] who were in an X/Open committee meeting. They wanted Ken and me to vet their FSS/UTF design. […] Ken and I suddenly realized there was an opportunity to use our experience to design a really good standard and get the X/Open guys to push it out.” (Ibid.)

What they essentially created was a way to encode the character encodings so that when processed by the computer they ‘unpacked’ themselves, turning a single 8-bit string into as many as 6 bytes of data, with the first byte in the sequence indicating how many bytes were encoded in total. This creates efficiency by not only allowing the code to remain packed up until needed, but also by creating patterns within the byte sequence which made it easier to find the start of a character from anywhere in the byte stream, one of the major issues that Pike and Thompson had with IBM’s initial UCS.



```
An easy way to remember this transformation format is to note that the
number of high-order 1's in the first byte signifies the number of
bytes in the multibyte character:

   Bits  Hex Min  Hex Max  Byte Sequence in Binary
1    7  00000000 0000007f 0vvvvvvv
2   11  00000080 000007FF 110vvvvv 10vvvvvv
3   16  00000800 0000FFFF 1110vvvv 10vvvvvv 10vvvvvv
4   21  00010000 001FFFFF 11110vvv 10vvvvvv 10vvvvvv 10vvvvvv
5   26  00200000 03FFFFFF 111110vv 10vvvvvv 10vvvvvv 10vvvvvv 10vvvvvv
6   31  04000000 7FFFFFFF 1111110v 10vvvvvv 10vvvvvv 10vvvvvv 10vvvvvv
10vvvvvv
```

Figure 6 Unpacking UTF-8 codes

This bit of code packing and unpacking means that UTF-8 and UTF-32 can both support as many as 2^{32}, or 4,294,967,296 individual codes. However, Unicode is based on UTF-16 encoding, the limitations of which mean that in practice Unicode is limited to 17 planes of 65,536 characters each, and with 2048 code points reserved as surrogates and 66 as non-characters, that leaves just over 1.1 million total unique codes available to be filled within the Unicode standard. As of Unicode 13.0, 143,859 of these spaces have a character permanently associated with them.

This represents 12% of the total available space, and with over a thousand new characters added with each update, that total will continue to rise.

4.3.4 Market Expansion

Working off of the ways that ASCII failed to enter the market in the 1960s as a universal standard, despite being endorsed by the American government, IBM, and Bell Labs created an anticipatory consortium to ensure cross-compatibility and to prepare the market for the launch of Unicode as a replacement for ASCII and all its variants. In this context, an “anticipatory consortium” is one that is formed deliberately, before the release or formalization of a standard, in order to ensure that all the major players in the area are in agreement as to how the final standard itself will be structured and applied (Weiss and Cargill, 1991).

So, it was not out of a sense of moral obligation, or a desire to preserve the world’s languages that Unicode was produced, but rather to ensure that the global tech market remained united under one standard, preferably under the control of the corporations (often descendants of what were originally academic or military operations) who were already dominating the market.

4.4 Conclusion

The way that the ‘space’ of Unicode is structured currently means that the Standard has a limited number of slots, which then must be curated by the Consortium to best make use of that finite space. Anyone can apply to add a new character/character set to the Standard, which involves a process of proving the audience for and necessity of a new addition to Unicode. This process is vetted by the voting members of the Consortium, and final say on design standards, official ‘meaning’ and adoption/rollout remains with many of the same corporations which include Unicode as a part of their product. The implementation of the Adopt a Character program in 2015 provides those without institutional power a chance to engage with the Unicode ‘brand,’ rather than with the Standard or Consortium in any meaningful way. Notably this program was implemented shortly after the first integration of emojis as permanent part of the Standard.

While Unicode assures users that they have enough space for all human languages, the namespace of Unicode is currently limited to just over 1.1 million potential codes, 143,000 of which are being occupied as of March 2020 and fails to consider the ever-developing and expansive nature of human communication, while retaining its strict policy stating that once a character is encoded it cannot be moved or removed, creating a clear bottleneck in a supposedly spacious system (Unicode ® Statistics 2021). While new languages may not be springing forth fully formed left and right, the addition and rapid expansion of emojis to the Standard demonstrates the fine and quickly blurring line between preserving language and creating an incubator for future linguistic evolution—something which Unicode has stated is not a part of their intended scope.

When approaching the creation of Unicode in the 1980s, the descendants of Bell Labs and IBM were fundamentally picking up where they left off in the 1960s, by attempting to solve a hardware compatibility problem between their devices. At the engineering level, a new character-encoding standard was being approached as an information management issue, at the corporate level, a new standard would help solve a problem with their product, and at an industry level, facilitate the opening of the global technology sector. Despite acknowledging that any new character-encoding standard must incorporate a larger namespace in order to allow for the many extra characters that were already critical to global communication, neither of these groups approached the encoding problem as a linguistic one. The priorities of the Consortium members and Standard developers at this juncture have since gone on to shape the digital world which we currently inhabit in untold countless ways.

5.0 Emojis 💯

In this chapter I explore the relationship between Emojis and the Unicode Standard, looking at the origins of these visual characters in digital text, and examining the treatment given to them by the Unicode Consortium and its constituent members in creating and extending the Unicode Standard. First, I give an overview of the history of emojis, from their origin on Japanese mobile devices in the 1990s, their role as hardware-specific features on those devices, and their subsequent integration into the Unicode Standard concurrent with their adoption into the Western market. Secondly, I probe the nuances of what makes a character an emoji versus any other sort of visual symbol or glyph in Unicode. This leads to a discussion on how meaning is assigned and regulated (or not) within Unicode, and via a close reading of a few selected emojis, I will discuss the gaps within Unicode's object-based system of organization, and the added pressure that emojis' semantically complex nature puts on how we identify and use the objects housed within the encoding standard.

5.1 The Birth of Emojis

In the past decade, emojis have taken the internet and the wider digital world by storm. There have been novels, movies, and innumerable think pieces about these little pictures that have made their way into our everyday vocabularies (Mukerjee 2017, Roger 2015, Schwedel 2018, Lenton 2018). Unicode existed before emojis, and even had their own set of 'pictographic characters' which predates the original emoji set (UTS #51: Unicode Emoji, 2021). Despite this, it was the addition of emojis specifically which prompted larger conversations around how these types of characters are used, stored, and perceived in digital space. What makes emojis so different from any other character that Unicode encodes? Part of the answer lies in the origin of emojis, and the way that they came to be added to Unicode, and at the request of and to the benefit of whom.

5.1.1 What are Emojis and Where Did They Come From?

The term "emoji" comes from the Japanese characters meaning "picture character:" 絵 (e ≅ picture) + 文字 (moji ≅ written character), and is not, as is sometimes believed, a spin-off of the English term "emoticons" used to describe pictures (usually faces) created in-line with text via regular alphabetical, numerical, and punctuation characters (FAQ – Emoji & Pictographs, 2021). Unicode defines emojis thusly: "Emoji are often pictographs—images of things such as faces, weather, vehicles and buildings, food and drink, animals and plants—or icons that represent emotions, feelings, or activities." (UTR-17 Unicode Character Encoding Model, 2008). The Emoji Report goes on to explain,

> "Emoji may be represented internally as graphics or they may be represented by normal glyphs encoded in fonts like other characters. These latter are called emoji characters for clarity. Some Unicode characters are normally displayed as emoji; some are normally displayed as ordinary text, and some can be displayed both ways." (Ibid.)

The relationship between emojis "emoji characters" and the other glyph characters within Unicode is illustrated in Figure 7, below. The nuance in this differentiation between "normal glyphs" and emoji characters is crucial to understanding emojis' place in the world of digital text. Unlike pre-existing pictographs, dingbats, and symbols, emojis are the first characters in Unicode to be permanently (and sometimes exclusively) encoded with a prescribed graphical representation, rather than with language- or orthography-based descriptors, which rely on reference back to a previous analog set of meaningful characters, such as an alphabet, and which can vary visually based on the typeface, font, or other environmental factors.

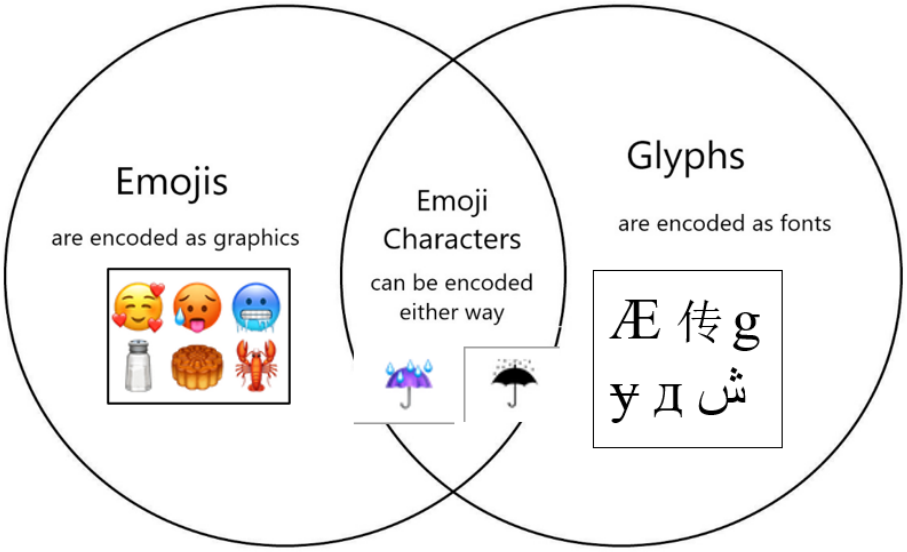


Figure 7 Venn Diagram of Types of Encoded Characters Within Unicode

Emojis formally encode for the first time versions of smileys, emoticons, and other makeshift affective symbols developed on the typewriter and with the characters of previous encoding standards, marking a shift in the way that tone and emotion are able to be conveyed via formalized writing systems (Highfield 2018; Freedman 2018). Emojis did not, however, pop fully-formed from the brow of Unicode, but were the result of a nearly decade-long process of global integration and standardization.

5.1.2 The first emojis

The first emojis appeared on Japanese cellphones in 1997, following some initial use of pictographs in Japanese pagers (UTS #51: Unicode Emoji, 2021). Initially they were carrier-exclusive to J-Phone, now known as SoftBank, and premiered with their SkyWalker DP-211SW model phone (Hånberg Alonso, 2021). The set included 90 characters designed in black-only pixel art and containing such future classics as “thumb’s up,” “beating heart,” and “smiling poo.” See Figure 8 below for the entire character set.

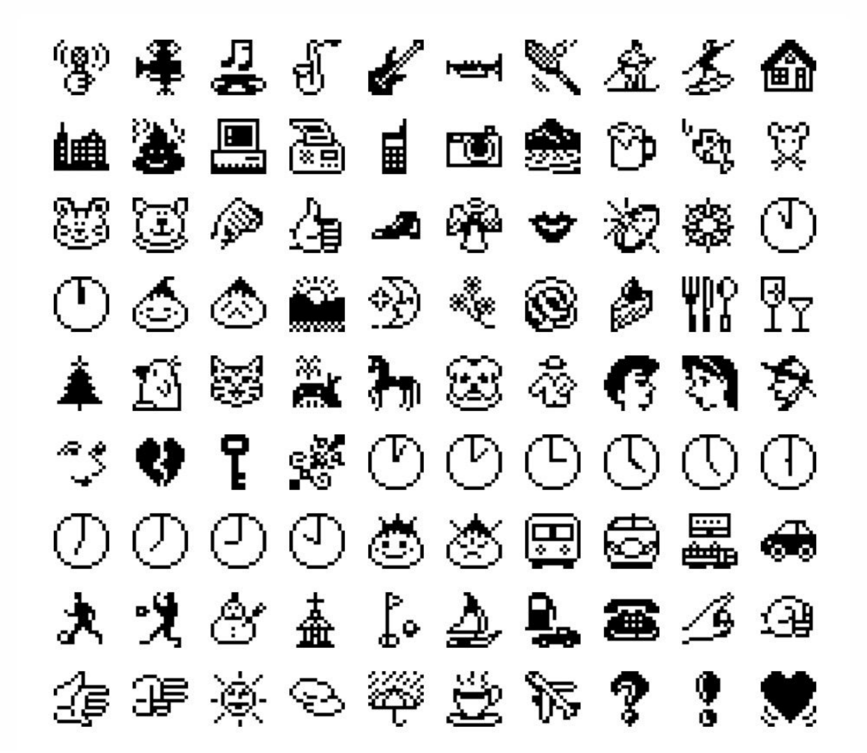
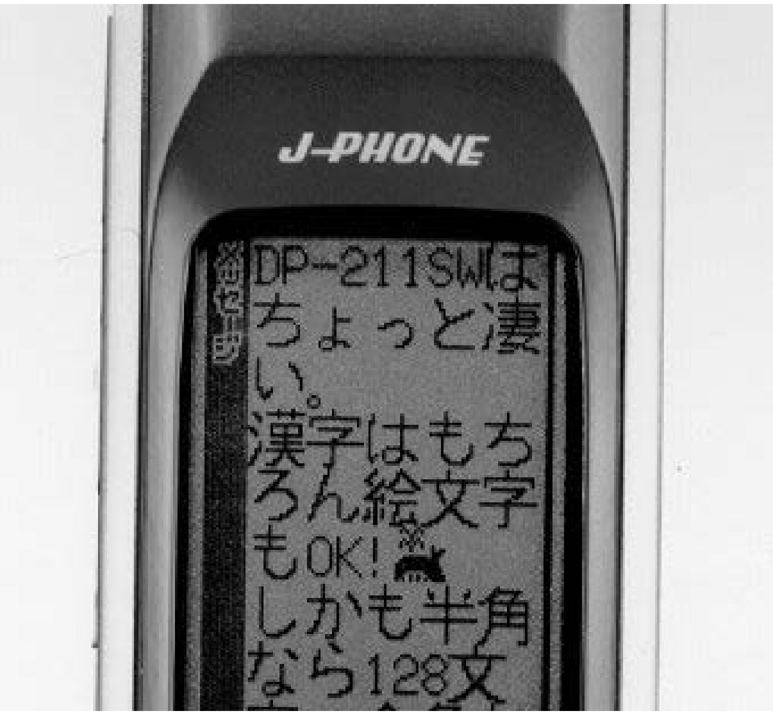


Figure 8 SoftBank's 1997 set of monochromatic emojis

Following the J-Phone/SoftBank set, Japanese carrier NTT DoCoMo introduced a set of 176 emojis for use within their proprietary internet service called “i-mode.” This set of pixel-art emojis, available in a rainbow of colors, but with each individual character being monochromatic, debuted in February 1999, and were designed by Shigetaka Kurita as a way to expand the communicative capacity of i-mode, which only allowed for 250 characters per message (Hånberg Alonso, 2021). The DoCoMo set is widely credited with launching the emoji phenomenon, as this was the emoji set which gained popularity in Japan initially, and reached a wider market than the J-Phone, whose touch-screen hardware placed it enough outside the mainstream mobile phone market at the time that the J-Phone set of emojis did not gain the audience or popularity of the following emoji generations. Prior to comments made on Twitter by Kurita himself in 2019, it was believed that the DoCoMo emojis were the first set ever produced for cellphones.



DP-211SWはちょっと凄い。漢字はもちろん絵文字もOK!🐋しかも半角なら128文

Figure 9 Whale emoji on J-Phone

Kurita’s DoCoMo emoji set remained the standard for emoji characters in Japan, and throughout the Southeast Asian markets for nearly a decade. Little appears to be written in English about the development of these character sets during that decade, and emojis would remain largely unknown to the West until American technology companies attempted to expand into the Japanese mobile phone market.

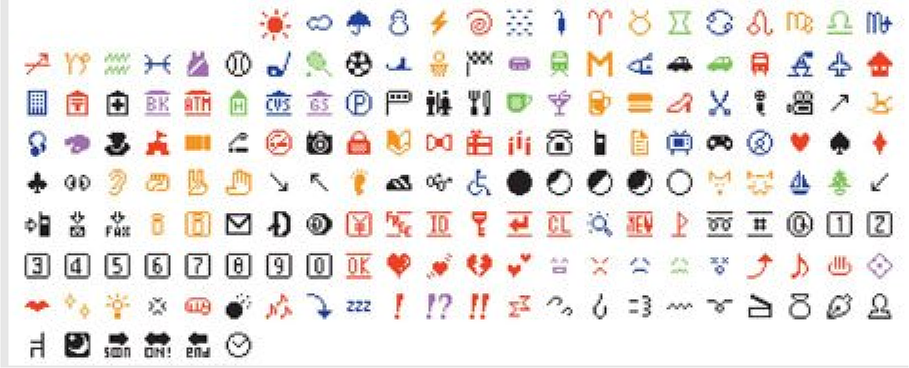
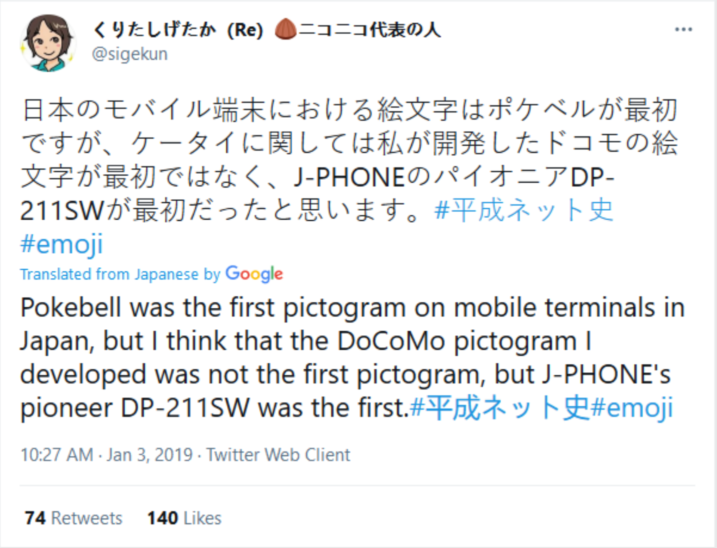


Figure 10 Kurita Emojis for DoCoMo 1999



くりたしげたか（Re）🌰ニコニコ代表の人
@sigekun

日本のモバイル端末における絵文字はポケベルが最初ですが、ケータイに関しては私が開発したドコモの絵文字が最初ではなく、J-PHONEのパイオニアDP-211SWが最初だったと思います。#平成ネット史 #emoji

Translated from Japanese by Google

Pokebell was the first pictogram on mobile terminals in Japan, but I think that the DoCoMo pictogram I developed was not the first pictogram, but J-PHONE's pioneer DP-211SW was the first.#平成ネット史#emoji

10:27 AM · Jan 3, 2019 · Twitter Web Client

74 Retweets **140** Likes

Figure 11 (https://twitter.com/sigekun/status/1080848236653334529)

Apple created a partnership with DoCoMo in 2008 to integrate the two platforms. This was a strategic move on Apple's part to gain a foothold in the Japanese and wider Asian mobile phone market. Apple's adoption of emojis as a part of this merger was a reflection of their enormous popularity already in Japan. They recognized that popular features, such as emojis, on Japanese phones must be brought over/compatible with Apple products entering the market, to entice customers away from established Japanese brands, as well as to prevent iPhone users from being isolated or restricted in their communications with older / different Japanese DoCoMo phone models (iPhone: About using the Emoji keyboard 2009). It was Kurita’s emoji set on DoCoMo which had the greatest aesthetic influence on Apple when they began the process of adding emoji characters to the iPhone’s repertoire, especially since the two phones would be interoperable and needed to maintain visual (and presumably semantic) consistency between how the characters appeared across the different devices. Prior to this, all emoji or pictogram characters available for use on mobile devices were carrier-specific. This was possible because the emoji characters were not yet part of any standardized text or character system, and were essentially being created as a ‘feature’ of the hardware they appeared on.

5.1.3 Bringing Emojis to Unicode

In 2006, Mark Davis, on behalf of Google, created an internal memo for consideration by the Unicode Consortium. It read in part, “There are a number of symbol sets that are in widespread use, but currently can only be mapped to private use characters on input. The UTC should consider whether or not it would be useful to encode these, or some subset. (Davis 2006) He makes reference to widely-available symbol fonts such as Webdings and Wingdings, then links to the DoCoMo character sets, as well as symbol sets being used on Vodafone devices in Japan, as well as on KDDI “au” brand devices. This was the first official Unicode documentation to take into consideration the encoding of visual symbols of this type, and also acknowledges that Western tech leaders such as Google have had a close eye on the development of the Japanese cellphone market. Following this, Google began encoding some of the Japanese emojis to private-use codes, separate from Davis' suggestion to Unicode. (Hånberg Alonso, 2021). Private-use codes are sets of 'blank' spots for characters within an encoding system, for which individual developers can then create their own proprietary characters. These characters will only appear on devices / in programs where they have been programmed to appear, and are not cross-platform compatible. The set of emojis characters encoded to these private use codes became available exclusively for use within the Google Gmail platform.

In response to Davis' proposal to Unicode, Andreas Stoetzner, a type designer and editor of the sign and symbol magazine SIGNA, and Dr. Deborah Anderson, a scholar of Signography—the study of signs and symbols—urged Unicode to accept their own proposal for the encoding of approximately 1,000 symbols "dealing with the orientational signage in public space", recommending that all the proposed symbols be accepted as a single batch of related characters, in the same manner that a language's alphabet might be encoded. (Stoetzner 2006). Stoetzner and Anderson's proposal focused on the integration of *already common* signs with established cultural meanings in to the Unicode Standard, and proposes that they be treated as its own orthographical set. This is something that has come to be a part of the Unicode Standard today, with sets of alchemical and astrological signs encoded together, but was met with some resistance at the time (Scherer 2008, Davis 2014).

Perhaps more importantly, Stoetzner and Anderson, "[S]trongly recommend NOT to randomly encode some picked-up industry fonts in use by any company. Remind "Zapf Dingbats".

For the proposal mentioned above we embarked on extensive collecting and research work in order to destillate [sic] the TYPES from the TOKENS.” (Stoetzner 2006.) This appears to be in direct response to Davis’ suggestion that symbol fonts, a group of which Zapf Dingbats is a prominent member, be potentially included for direct mapping as Unicode symbols (ITC 1978, p. 36). Stoetzner’s concern appears to be in regards to eating up valuable encoding space with ‘meaningless’ symbols designed solely for the adornment of text, evidence of the scarcity mindset which has surrounded the Unicode-as-Good model from its inception. This proposal was largely ignored, and can be considered the beginning of the debate within Unicode regarding which types of symbols, both from pre-existing sources and those developed specifically for Unicode, should be included within the Standard, and which should remain delegated to fonts or private-use codes. This is a debate which continues to this day, as evidenced by its frequent presence in meeting agendas, notes, and listserv threads.

While Google was quietly adapting Japanese emoji to private-use characters, Apple completed its deal with DoCoMo to bring the iPhone to Japan in 2008. The subsequent release of the iPhone (OS 2.2) in Japan brought with it the first emojis to appear on an Android or Apple mobile device (iPhone: About using the Emoji keyboard 2008). Emojis were still limited to phones using Japan’s SoftBank as their carrier, they required specific keyboard setup, were limited to communications between SoftBank devices, and were not backwards compatible. Emojis were not available on Western phones at the time without a hack. (iPhone 2008, Hånberg Alonso 2021). Despite this, emojis quickly became a phenomenon both in Asia and North America, prompting both the first wave of cultural critique of the characters, as well as their permanent codification into Unicode.

The addition of emoji characters to the Unicode Standard was put to an informal vote via the Unicode Consortium listserv in December 2008 (Davis, 2008). While there was debate about the usefulness of creating an entirely separate category of character for these visual characters, the first emojis became an official part of Unicode with the release of Unicode 5.2.0 on October 1, 2009. Perhaps as a concession to the dissenters, the set of 114 characters added—already popularly referred to as emojis by the public, consisted of many of which had already existed within the Standard as dingbats and other pictographs and the set was not yet formally called emoji by Unicode, but rather fell under the heading of "Miscellaneous Symbols." The Unicode 5.2.0 release notes describes them as such;

> The Miscellaneous Symbols block consists of a very heterogeneous collection of symbols that do not fit in any other Unicode character block and that tend to be rather pictographic in nature. These symbols are typically used for text decorations, but they may also be treated as normal text characters in applications such as typesetting chess books, card game manuals, and horoscopes. (Unicode 5.2.0, p 482)

The closest thing to a direct reference to emojis themselves comes just a few paragraphs later, as the report describes how,

> The Miscellaneous Symbols are derived from a large range of national and vendor character standards. Among them, characters from the Japanese Association of Radio Industries and Business (ARIB) standard STD-B24 are widely represented in this block. The symbols from ARIB were initially used in the context of digital broadcasting, but in many cases their usage has evolved to more generic purposes (ibid).

This acknowledges that these symbols which originated with Japanese cell phone characters had grown to wide enough usage that their permanent encoding became necessary. Prior to this, while characters were encoded as sets, these sets usually represented a distinct and complete orthography, essentially “adding a language” to those that Unicode could support. The addition of these visual characters, and the acknowledgement by Unicode that they came to be via a different route than traditional orthographic symbols left the door open to the set to be continually built upon, essentially growing the new emoji vocabulary within the environment of Unicode. We have seen this in practice in the years since, as petitions and advocacy groups have developed with the purpose of getting a particular emoji or another added to the Standard—something that we do not see, and in fact makes little sense, in the context of a traditional ‘language’ (Glaser 2019).

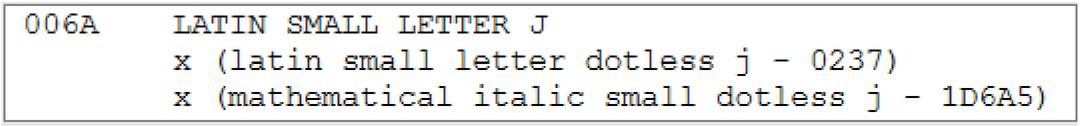
The western media and public fascination with the characters began in earnest once they had been encoded in Unicode and were widely available in the United States. At first this was only on Apple products, as Apple’s had gained exclusive access to the software integration of emojis into their devices (iPhone: About using the Emoji keyboard 2009). Apple enthusiastically advertised the new characters and their potential uses to its userbase. Google then began its integration of their own email smileys, which used private-use character code, with the Unicode Standard emoji sets, with coverage eventually expanding across all operating systems using the latest Unicode version. Apple and Google in particular, because of the huge amount of market control and influence they wielded, were invested in a cross-platform solution to character / symbol compatibility, and were able to take hold of both ends of the emoji release and distribution process—both companies have been represented as Consortium members since 1998 and 2006, respectively (The Unicode Consortium Members 1998, 2006). And as emojis continued to sell mobile devices, the growing influence of emojis on the Standard and the Consortium became more apparent within the Unicode website, both its content as well as its structure.

5.2 Emojis as Objects

Let us take a moment to examine the idea of emojis as ‘semantically complex’ objects, relative to the types of characters which Unicode was initially designed to encode. Complexity is, of course, relative and languages such as Chinese, Japanese, and Korean (CJK) assign meaning at a character level in a way that languages such as English do not.

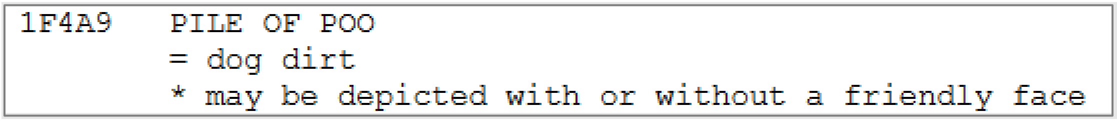
The letter "j" for example can have many contextual meanings, as a name or initial, and a variety of other things, but just the letter "j" on its own and out of context does not convey a culturally relevant piece of information on its own-- not even how it is pronounced, considering that "j" is vocalized differently across languages which use it in their alphabets. Emojis, however, can serve as meaningful linguistic objects on their own, without the necessity of context provided from other characters, functioning more on the level of a word or morpheme than a phoneme within the linguistic structures where they are used. Many CJK languages also struggle with the Unicode framework, as each character encoded essentially represents a “word” in the language, meaning that for “complete” Unicode compatibility, hundreds of thousands of characters would need to be encoded. This issue predates emojis, and even Unicode, and is the direct result of Warren Weaver’s assumptions about how languages are translated (Weaver 1949, Mullaney 2017). However, because of their born-digital nature, emojis are not associated with a particular language or its alphabet, and it therefore becomes easier to see how they vary from Unicode’s expected norm, allowing us to learn about those norms and how they play out in real world use.

A morpheme in formal linguistics is the smallest unit of semantically meaningful text—smaller than a word, as it includes things such as prefixes, suffixes, abbreviations, and other similar parts of speech. A phoneme is similarly the “minimal unit in the sound system of a language” which translate roughly but not directly to the letters in a given language’s alphabet (Glossary of Linguistic Terms 2003). This distinction is at the crux of what makes emojis so difficult for Unicode to manage both within their own organizational structure, but also as textual objects which circulate in digital space. The morphemic quality of emojis is evident through the way that emojis are labeled both within Unicode as well as within popular culture. The "pile of poo" emoji (💩 U+1F4A9) is able to stand on its own representing such a pile of poo, but requires us as readers, and Unicode as record-keepers, to string together the letters "smiling pile of poo" to form a descriptor. Let’s compare the entries of “j” to that of “💩” within the official Unicode nameslist—the Standard itself, insomuch as it can be reduced to a single .txt document.



```
006A    LATIN SMALL LETTER J
        x (latin small letter dotless j - 0237)
        x (mathematical italic small dotless j - 1D6A5)
```

Figure 12 LATIN SMALL LETTER J



```
1F4A9   PILE OF POO
        = dog dirt
        * may be depicted with or without a friendly face
```

Figure 13 PILE OF POO

In this format, we can see the first inkling of how a ‘cataloging mindset’ defines how Unicode characters are documented. To the left is the unique ID which tells a computer which set of code needs to be activated to display the associated character. On the same line, in all caps, is the formal title of the character, here “Latin Small Letter J” and “Pile of Poo.” Note that in both these cases, the character being defined is not present in its own name, but this does not hold true with other characters, such as 004A LATIN CAPITAL LETTER J. Conversely, it is currently impossible for emoji characters to appear in the Unicode nameslist, as it remains encoded in ASCII, which does not allow for such characters. This is done as a compatibility measure, and we will discuss its impact on the Standard in a later chapter.

Beneath the UID and name line, indented once, appear such pieces of metadata as alternate names (indicated by “=”), related terms (“x”), and notes on the appearance or categorization of the character. Another possible metadata items are <control>, indicating that the character is a control character which performs a function within the text, i.e. a line or page break, or does not appear as a displayable character on its own. And finally, is the #, which cross indexes where visually identical but encoding-unique characters exist within the Standard. Each item of this nature appears on its own line.

5.2.1 Regulating Meaning

The method that Unicode uses to define a character tries not to assign a fixed meaning to them, but rather reinforce their equality at the level of the descriptor, as each receives a single unique identifier within the Unicode namespace. From the computer’s perspective, all Unicode characters are indivisible, atomic pieces of code. Emojis in particular, however, are stylized in their presentation, and humans perceive them as images.

Even emojis which 'contain' text such as 💯, 🈯, or 🆒 are stylized to emphasize that they are not a part of the 'regular' text. The red coloration and two emphatic underlines on the “Hundred Points Symbol”, and the green and blue boxes on the “Squared CJK Unified Ideograph-6307” and “Squared Cool” respectively, serve to immediately separate these characters from those around them and emphasize their symbolic sociocultural meanings, such as doing well on an exam, noting a reserved spot or the designated hitter spot in a baseball lineup, or associating a message or person with the cultural vernacular meaning of *cool*. These characters can all be typed easily without emojis-- “100”, “指”, “cool” –but lose much of their contextual meaning when doing so.

Because we see and categorize emojis differently than 'regular' characters, we treat them differently as well— both in terms of the linguistic role which they play in our communication, and in terms of how meaning is applied to them. Because emojis do not draw on an already existing alphabet/language, and are developed out of whole cloth based on user proposals and Consortium research/agendas, Unicode is sidled with the weight of standardizing, normalizing, and reifying the meanings of each emoji-- despite this being explicitly outside the scope of their stated goals (UTS #51 Unicode Emoji 2021, Overview—Unicode 2021). Unicode has handled this by attempting to divorce themselves as much as possible from the standardization of the visual representation of emojis, while still prescribing standardized interpretations on essentially 'invisible' objects.

The Unicode Standard indicates some guidelines for the appearance of emojis, and in recent years has worked to resolve issues with characters that have appeared inconsistently across platforms, notably in the case of the pistol emoji, which transitioned from a realistic-appearing firearm to a brightly colored toy gun across all major platforms after Apple changed their own representation in response to the high rate of gun-based violence in the United States (see figure below).

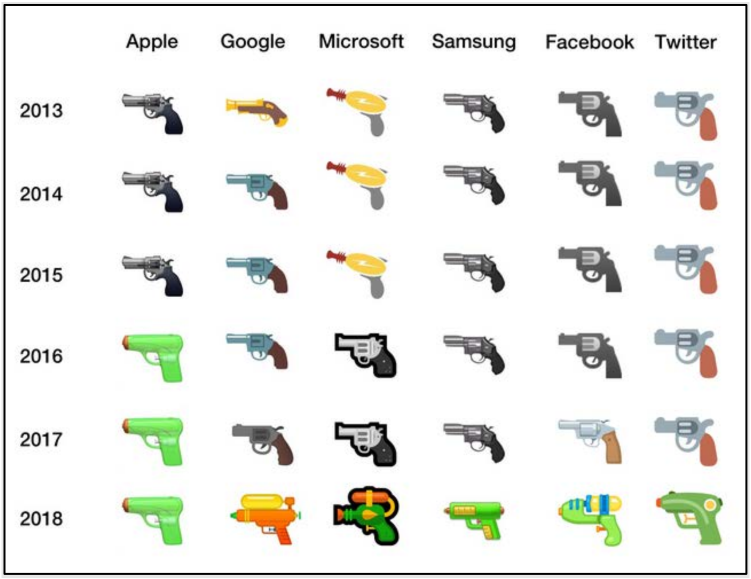


Figure 14 Evolution of Pistol Emoji 2013 – 2018

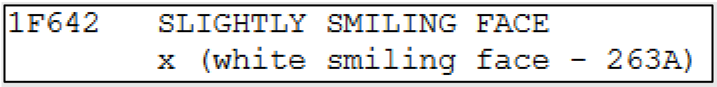
Because Apple is a part of the Consortium, and has been intimately involved with the adoption and popularization of emojis, they made use of their power behind the scenes at the Consortium as well as within the marketplace to essentially peer pressure the other major emoji representation platforms (Google, Twitter, Samsung, Microsoft, Facebook, WhatsApp) into matching their representation of the pistol (Berard, 2016, Burge 2018). This shift is not reflected in the Unicode nameslist, where the emoji continues to be listed only as “pistol” with the synonyms “handgun” and “revolver” associated with it. The majority of the less popular platforms’ representations of this character remain that of a realistic-looking handgun, highlighting the public relations aspect of the change for the major carriers seeking to remain compatible with the industry leader—Apple—and prevent potentially disastrous cross-platform miscommunications (Burge 2016, 2018).

A look at the meeting minutes and listserv discussions around the adoption of new emojis and the tweaking of current ones reveals a distinct increase in the amount of visual control Unicode is attempting to exert over emojis, an implicit acknowledgement from Unicode that emojis are different from the other characters that Unicode encodes. I also interpret this as an attempt to regulate the meaning of emojis, as much of the metadata associated with non-emoji Unicode characters is in regards to the long linguistic heritages from which they derive—a history that emojis do not have. It is also important to point out that while the discussion on future emojis is open for any member of the consortium to join, the number of actual individual participants (as opposed to organization-level participation) engaged with the emoji approval process is tiny— the most recent meeting minutes record comments from three individuals regarding their thoughts on the proposed emojis. Each of their contributions, while detailed and thorough, represents a tiny fraction of the members who are eligible to so comment.

5.2.2 Emojis as Visual Objects

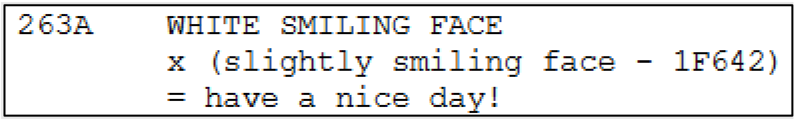
But just the written name or description of an emoji is not the end of its meanings, and only serves to unlock another level of semantic complexity these objects have as compared to other characters encoded by Unicode. One aspect of this is culturally-specific interpretations of individual emojis-- much like how "j" has multiple pronunciations in different languages, emojis can have different meanings depending on the audience and context. I'll give an example:

The "smiling face" emoji appears at first to be a straightforward adaptation of the classic "smiley", consisting of a circular yellow face containing black eyes and smiling mouth. U-1F642 🙂 is "Slightly Smiling Face" in Unicode, and its description from Emojipedia reads, "A yellow face with simple, open eyes and a thin, closed smile. Conveys a wide range of positive, happy, and friendly sentiments. Its tone can also be patronizing, passive-aggressive, or ironic, as if saying *This is fine* when it’s really not (Slightly Smiling Face Emoji, 2021)."



```
1F642    SLIGHTLY SMILING FACE
         x (white smiling face - 263A)
```

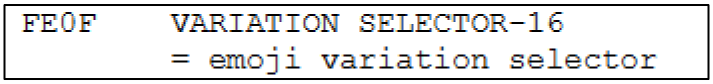
Figure 15 Slightly Smiling Face 🙂



```
263A     WHITE SMILING FACE
         x (slightly smiling face - 1F642)
         = have a nice day!
```

Figure 16 White Smiling Face ☺

As we can see from their entries in the Unicode nameslist, U-1F642 Slightly Smiling Face is cross-referenced with U-263A White Smiling Face. In this case, “White” does not refer to the race or ethnicity of the depicted character, but rather is used to indicate that “when shown with monochrome fonts, this character is intended to be displayed using a hollow or outlined appearance” (Emoji Glossary, 2021). U-263A is a pre-emoji symbol, and falls into the category of characters which can be displayed either graphically as an emoji, or non-graphically as a part of a character font set. This character defaults to being displayed textually, and must be combined with U-FE0F Variation Selector-16 in order to show a smiling yellow symbol in the same style as other emojis. This is done automatically in most mobile keyboards, meaning that the white U-263A ☺ *emoji character* presentation is rarely seen in practice, and users must deliberately seek out that version of the character should they desire to use it.



```
FE0F     VARIATION SELECTOR-16
         = emoji variation selector
```

Figure 17 Variation Selector-16, U-263A ☺ + U-FE0F = 😊

However, upon combining these two characters, the resulting emoji gains several new affective visual features. In addition to now being a colored character in shades of yellow and pink with black lines, the U-263A + U-FE0F emoji has closed eyes, and gains rosy cheeks and relaxed eyebrows, see Table 8, below. This presentation of the character is known simply as Smiling Face Emoji, and is, according to Emojipedia, “A classic smiley.” (White Smiling Face Emoji, 2021). Significantly, Smiling Face Emoji does not carry with it any of the disingenuousness of the U-1F642 🙂 Slightly Smiling Face, but rather represents a truly happy state for the character, and presumably its human user.

Table 8 Smiling Face Emoji Presentation

Emoji Presentation	Character Code	Name	Date Added to Unicode
☺	U-263A	White Smiling Face	1993, Unicode 1.1, as a pictographic symbol
🙂	U-1F642	Slightly Smiling Face	2014, Unicode 7.0, incorporated into Emoji 1.0 in 2015
☺️	U-263A + U-FE0F	Smiling Face	U-FE0F modifier added in 2015, to create emoji presentation

This is interesting because while U-263A ☺ + U-FE0F ☺️ and U-1F642 🙂 are visually distinct, making determining the difference in their intended meaning easy for the reader, the underlying character of U-263A ☺ appears almost identical visually to U-1F642 🙂, aside from the color of the smiling face. This would not appear to be a problem as modern technology automatically converts U-263A ☺ to a different presentation, however, the two characters remain cross-referenced within the official Unicode documentation. What this means is that these two glyphs are being marked by Unicode as being semantically similar, if not identical to one another.

So why not simply wipe the slate clean and “rewire” U-263A ☺ to be the emoji character version of U-1F642 🙂, and establish Smiling Face Emoji ☺️ as a single character encoding, rather than the result of a character-plus-modifier set up? Unicode does not allow for the deletion of characters, and is also strictly bound to the permanence of its encodings as a part of backwards compatibility. The timeline of these characters’ addition to the Standard provides some insight into why these characters are related to one another in the way that they are, and the way that the Consortium chose to treat emojis in general differently from other encoded sets.

White Smiling Face U-263A ☺came first, encoded with Unicode 1.1 in 1993, six years before emojis were introduced on Japanese cellphones, indicating the smiley’s pre-emoji heritage, as well as pointing at the long history of people using mechanical devices to create the appearance of faces amidst text (Typographical Art, 1881). Next came U-1F642 🙂 in 2014, under a completely different encoding. Following this release, both of these characters were grouped under the initial Emoji 1.0 release in 2015, with the emoji characters such as U-263A ☺then being linked to the emoji variation selector U-FE0F, which as seen above, automatically alters that character’s visual presentation to create a consistent aesthetic look to the emoji set. However, in the case of U-263A ☺ specifically, a simple paint job would have only confused things more, creating a U-263A ☺ character colored in yellow like the rest of the emoji face set—making it then almost visually identical to U-1F642 🙂. This resulted in the makeover of the U-263A + U-FE0F emoji presentation to the softer, more detailed visual representation persistent to this day. (Scherer 2008, White Smiling Face Emoji 2021). However, the two characters and its three presentations remain cross-referenced within the Unicode nameslist, despite the divergence in meaning and affect between the two ‘final’ emoji presentations. This preserves the historical relationships between these characters, and ensures greater stability for Unicode in the future, but poses a risk of semantic obsolescence as the characters' meanings diverge.

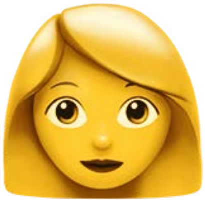
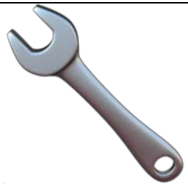
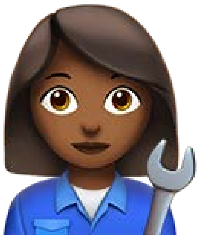
This seeming redundancy makes more sense when viewed from the lens of the Unicode Standard being treated as a discrete, finite set of code "slots," essentially creating a controlled vocabulary of U+0000 namespace codes. Characters cannot be moved from spot to spot, or consolidated into fewer spots, because the long-term functionality of Unicode depends on the reliability of those namespace codes, resulting in the type of work-around demonstrated above. This is also what results in features such as gender and skin tone becoming add-ons to the baseline yellow emoji people, causing the character 👩🏾‍🔧 "Woman Mechanic: Medium-Dark Skin Tone" to only be possible through a combination of the 👩 Woman, □ Medium-Dark Skin Tone, Zero Width Joiner (ZWJ) and 🔧 Wrench emojis, see Table 9, below. As a result, 👩🏾‍🔧 as a textual object to the computer is four characters long, despite displaying as a single character from Emoji 4.0 (2016) and onward.[4] Note here that the "Woman Mechanic: Medium-Dark Skin Tone" combination character is added to the Emoji 4.0 set, but not to the Unicode Standard directly. This means our mechanic friend is not a character in and of herself, but rather a combined presentation of four existing encoded characters. On the back end, this means that each non-yellow person emoji does not need to be encoded individually into the Unicode Standard, which would increase the space that emojis take up in the Standard by at least fivefold, with each human character being

[4] Previous Unicode versions will simply display 👩 🔧 and the Medium-Dark Skin Tone modifier, which ironically will not display in Microsoft Word as a standalone character.

encoded individually for each of the five skin tones Unicode supports.[5] Rather, that inefficiency is skirted at the code end, and passed down to the user end where it then requires four characters to represent a single Black woman. Scholars have written about the implications of this as an Othering of non-white persons, while the default yellow emojis are assumed to represented "regular" (read: white) people (Barbieri and Camacho-Collados 2018, Berard 2018, Sweeney and Whaley 2019).

[5] To say nothing of the many other characters that are unrelated to skin tone variation that this type of combination allows.

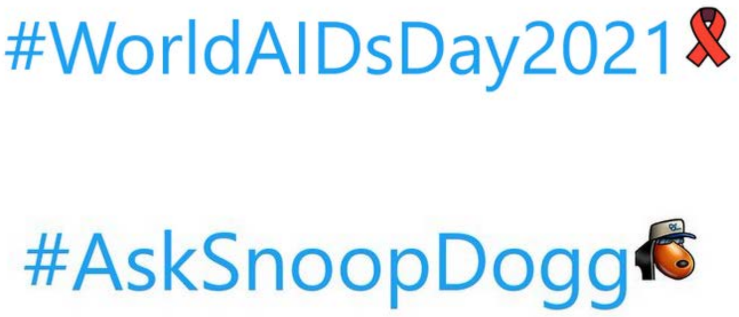
Table 9 Characters necessary to create the "Woman Mechanic: Medium-Dark Skin Tone" emoji

Emoji Presentation	Character Code	Name	Date Added to Unicode
	U+1F469	Woman	Unicode 6.0 in 2010, incorporated into Emoji 1.0 in 2015
🟫	U+1F3FE	Medium-Dark Skin Tone (Previously Emoji Modifier Fitzpatrick Type-5)	Unicode 8.0 and Emoji 1.0 in 2015
	U+1F527		Unicode 6.0 in 2010, incorporated into Emoji 1.0 in 2015
[invisible when used alone]	U+200D	Zero Width Joiner	Unicode 1.1 in 1993, incorporated into Emoji 1.0 in 2015
	U+1F469 U+1F3FE U+1F527 U+200D	Woman Mechanic: Medium-Dark Skin Tone	Emoji 4.0 in 2016

5.2.3 Emojis as Commercial Objects

On the flip side, platforms such as Twitter and Facebook, themselves prominent members of the Unicode Consortium, are able to then use the ZWJ characters within their proprietary environments to create customized "emoji-like" characters. This is similar to how unassigned code slots in previous character-encoding standards would be customized based on local language, hardware requirements, and other orthographic needs (Fischer 2012).

Twitter is perhaps best known for its use of these characters, called “hashflags” which are then associated with hashtags, Twitter’s own crowdsourced subject-based classification system (Highfield 2018). This results in such hashtag and hashflag combinations such as “#WorldAIDsDay2021” followed by a red ribbon icon, and “#AskSnoopDogg” followed by a cartoon dog head (see images below).



#WorldAIDsDay2021

#AskSnoopDogg

Figure 18 Hashtags and Accompanying Hashflags for #WorldAIDsDay2021 and #AskSnoopDogg.

Twitter users can use a hashflag simply by typing in the associated hashtag, and the accompanying image will appear with the tagged tweet(s), adding visual flair to text-based messages, and encouraging communication within the Twitter environment, as these images are not visible elsewhere. We can see here how Twitter, despite the hashtag’s natural inclination towards information organization through cross-referencing, intentionally makes use of emojis as marketable *visual* objects with cultural cache in order to entice users to its platform. In fact, Twitter states up front that hashflags are products which are for sale. Highfield writes, “Indeed, Twitter describes hashflags as ‘commercial products’: a hashflag is part of an advertising campaign encompassing promoted tweets and trends (Highfield 2018, Twitter Ads 2017). The exact costs and potential income these hashflags provide for Twitter and their sponsors are difficult to estimate, but as of 2016, approximately $1 million USD was considered the going rate (Johnson, 2016). This stands as solid evidence of emoji characters being considered by tech companies—including those with significant power within the Unicode Consortium-- a good to be bought and sold. Despite hashflags themselves working via an encoding loophole on Unicode’s end and not *technically* being emojis, they are in fact visual objects encoded as pieces of text by the computer, just as emojis are. The public prominence, visual similarity, and known status as #brandedcontent of hashflags associates all emoji characters with this market-based paradigm. Additionally, characters which began life as proprietary hashflags on Twitter have since gone on to become fully-integrated emojis within the Unicode Standard. The most prominent example of is the rainbow flag emoji, which was introduced as a part of the #Pride hashtag in, and met with such resounding positive feedback that it was later added as an official emoji (Highfield 2018, Johnson, 2016).

Making this type of analysis possible requires a great deal of work by the Unicode Consortium, but the members of the committees and working groups dedicated to the adoption and integration of emojis into the digital text environment also have a vested economic interest in their widespread dissemination and use. Emojis are added to Unicode one character at a time, based on individual proposals for characters, which can be submitted by persons, businesses, governments, or other organizations. The proposals are then reviewed for relevance, necessity of addition, visual appeal, and a variety of other guidelines and then passed up the line for further review (Unicode n.d., Berard 2016).[6] This process is not dissimilar to that by which other characters are accepted into Unicode, except for one major difference: that of scale. Emoji additions to the Unicode Standard are considered on a case-by-case basis, while other alphabets, scripts, or orthographic sets can contain thousands of individual characters. This makes the relative cost of adding emojis to Unicode extremely high—and gives one pause to wonder what exactly makes these semantically, technically, and visually cumbersome characters worth the effort?

Put simply, emojis “solve” the problem of conveying tone and emotion in text-based settings, allowing users to express their feelings with a combination of emojis and words, or without words entirely. This is certainly not lost on the Unicode Consortium, and while the Consortium publicly tries to distance itself from the job of prescribing fixed meanings to individual emojis, the businesses who make up the majority of the Consortium’s membership have jumped on this opportunity to give users additional options to show affective responses to media—such as the addition of the thumb’s down, angry face, and others to the repertoire of Facebook reactions. This not only responds directly to the needs of the market (more emojis!), but can then provide Facebook with important data points about which kinds of posts elicit which types of reactions, while at the same time encoding the affective meaning of those characters within the Facebook environment and beyond (thumb’s up = I like it, thumb’s down = I dislike it) which allows

[6] The Republic of North Korea has submitted several times a proposal for an emoji glyph of the national seal of the DNP’s ruling party. It has been repeatedly rejected by the committee on the grounds that political symbols are not permitted. The North Korean contingent, however, makes the argument that the hammer and sickle glyph ☭ has been encoded since Unicode 1.1 in 1993 (Emojipedia 2021, Unicode n.d.)

sentiment analysis to then be able to determine the emotional tone of a post based on the reactions it receives, and the emojis related to it. The Unicode Consortium needs user engagement with emojis to make their time and money investments in the technology worthwhile. Fortunately for them, this has never proved difficult to do, and we will discuss some of their major engagement tactics in the upcoming chapter.

6.0 Unicode and the Public

Thus far, we have followed the history of mechanical character-encoding from its early analog days, through the formation and launch of Unicode, to the adoption of emojis into the Standard, and have delved into some of the ways that emoji characters bend or break the guidelines of the Standard. Throughout, I have endeavored to highlight the ways in which the engineers who constructed the Unicode Standard were building upon previous conceptualizations of both computing hardware and software, but also of language and translation. This has led, I have argued, to the Unicode Standard, and the characters it represents, being treated as goods existing in a digital marketplace, and able to be consumed by its users. Building on this, I now turn to the ways that Unicode presents itself to the public, examining the choices that the Consortium has made in how they allow engagement with the Standard.

In this chapter I look at several of the largest and longest-lasting ways that Unicode has acted in order to engage with their audience: through the Unicode.org website, and through the Adopt a Character feature. First, I take a look at the Unicode.org website, and through content and network analysis of the site's pages, I reveal the ways that the structure of this document signal information about the Standard to visitors, particularly as relates to emojis, concurrently providing information about the value of emojis to Unicode and its members. Next, I conduct an analysis of the Unicode Standard's "Adopt a Character" feature, the longest-running way for members of the public to interact with the Standard directly. The interesting history of this engagement, and the records related to it, reveal the value of the public to the Consortium as one of *perceived* rather than actual support. Combined, these analyses support the narrative of Unicode as a product being marketed to the public, framed by the Consortium’s stated desire to preserve human language, and asks users to “invest” in individual characters on a superficial level, while broader policy change remains in the hands of the Consortium’s voting members—the makeup of which we will discuss in depth later in this chapter.

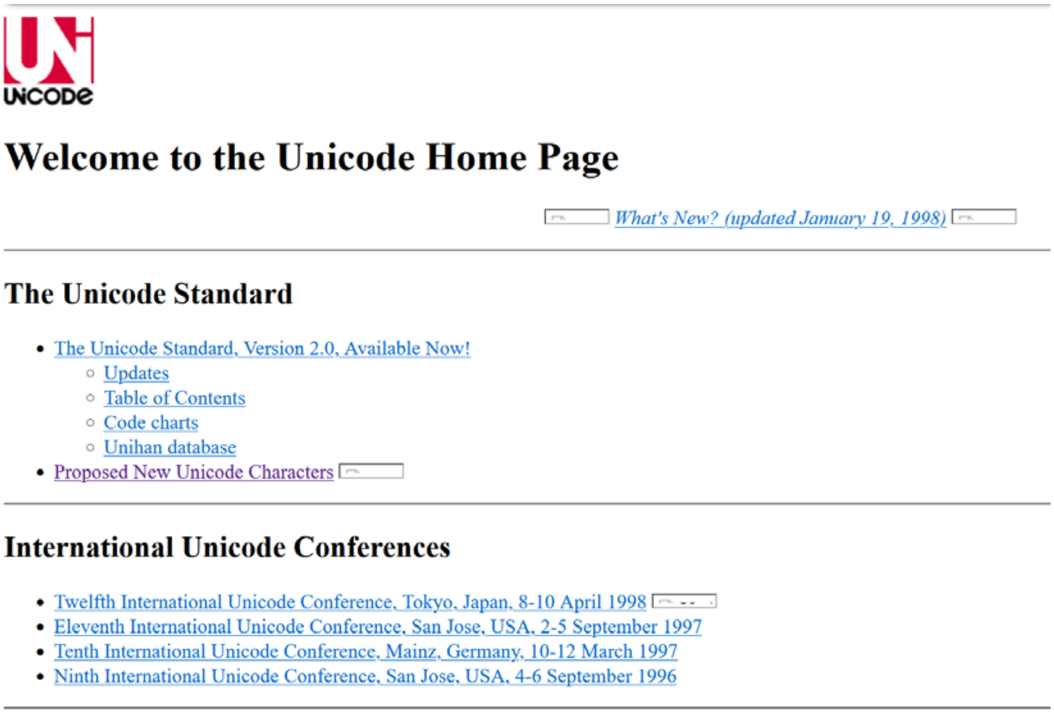
6.1 Unicode.org

As the example of the pistol emoji discussed in Chapter 4 shows, Unicode and its constituent members weld a huge amount of power over not only which languages are digitally reproducible, but also over the appearance and cultural context of individual characters in the digital world. Unicode takes great care in how their work as a Standard and Consortium is presented to the public, and the major forum through which Unicode does this is their website, http://unicode.org.[15] Through content and network analysis of the site, I show that Unicode has made deliberate choices about how to organize the information on their website in order to present a particular narrative of their work to the public, and that a great deal of emphasis is placed on emojis. This, in turn, indicates the high value of emojis to the Standard, and its mostly-corporate membership.

6.1.1 Unicode.org Homepage

The Unicode.org URL was first captured by the Internet Archive’s Wayback Machine on January 26, 1998, and that image of the site says it was last updated January 19 of the same year, see Figure 19, below. (Unicode Homepage, 1998). It also informs us that the site was designed by Glenn Adams, then the Technical Director of the Unicode Consortium, who announced the site's official launch to the Unicode listserv December 20, 1994 (Adams, 1994). Adams notes, however, that http://www.unicode.org redirects to http://www.stonehand.com/unicode.html, where the database of Unicode's information was held at the time, and he indicates that the Unicode.org URL should be considered the primary domain name in case of data migration at some point during the site's life, and in fact, though the earliest capture of the stonehand.com URL by the Wayback Machine was January 5, 1997, by August of that same year, all of the site's data and pages were held under the Unicode.org name (Wayback Machine 2021). This distinction is important for a few reasons: first, it indicates that Unicode anticipated being around on the web for a while. Second, the Unicode.org domain secures the .org domain name, signaling that Unicode is an *organization* rather than a *commercial entity*. While there have never been requirements for an organization to qualify for the .org domain name, it is commonly associated with nonprofits and other philanthropic groups, as opposed to more commercial entities on the .com domain (Pope et al 2012). This association is emphasized in internal Unicode communications as an important means of establishing legitimacy as a standard (Adams, 1994).

Since its establishment in the mid-90s, the Unicode.org home page has gone through several visual overhauls as online style and technology has developed. By 2008, ten years after the first Wayback Machine capture of the URL, the content of Unicode.org had become stabilized, and received updates as new versions of the Standard were developed, without any major reorganization to the home page / site map itself, and remained as such for the following decade, as well, see Figures 19-21, below. This can be seen in the persistence of the top-level menu items on the homepage over this time such as “New to Unicode?”, “General Information”, “The Consortium”, and “The Unicode Standard”, each of which links to the same URL over the course of this decade of evolution, and again, while visually the pages changed over that time, their content remained consistent with Unicode’s general policy of only adding and never subtracting content.



UNICODE

Welcome to the Unicode Home Page

What's New? (updated January 19, 1998)

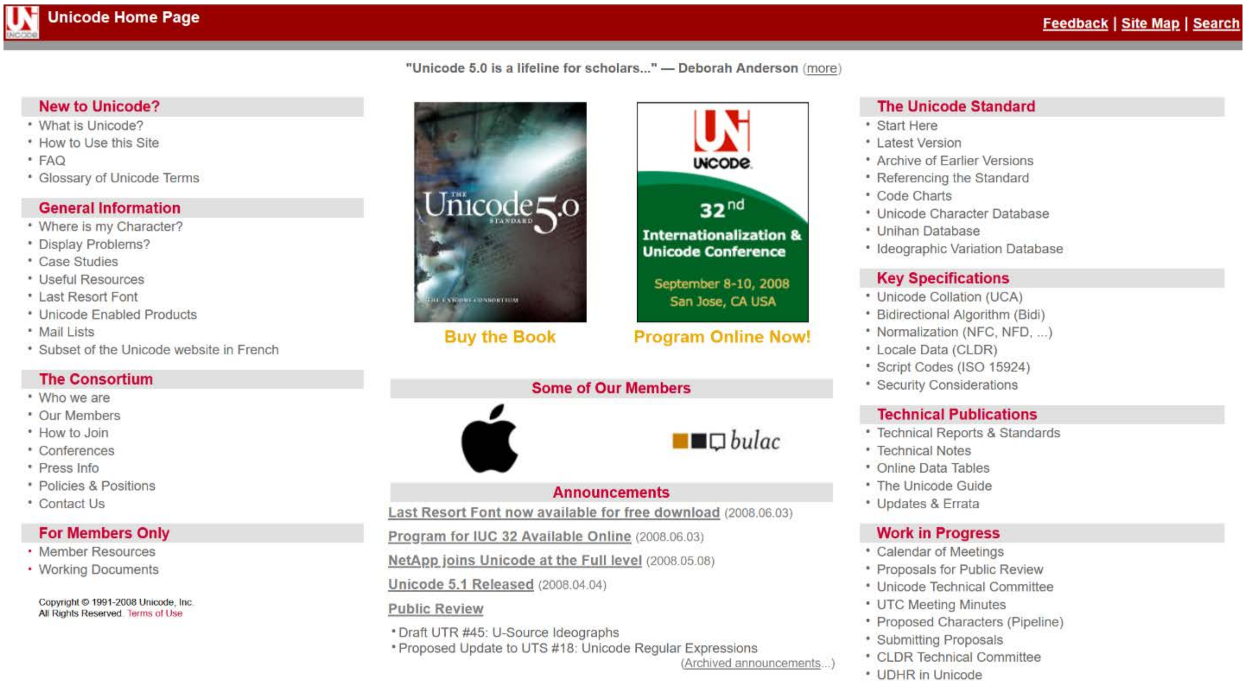
The Unicode Standard

- The Unicode Standard, Version 2.0, Available Now!
  - Updates
  - Table of Contents
  - Code charts
  - Unihan database
- Proposed New Unicode Characters

International Unicode Conferences

- Twelfth International Unicode Conference, Tokyo, Japan, 8-10 April 1998
- Eleventh International Unicode Conference, San Jose, USA, 2-5 September 1997
- Tenth International Unicode Conference, Mainz, Germany, 10-12 March 1997
- Ninth International Unicode Conference, San Jose, USA, 4-6 September 1996

Figure 19 Unicode.org homepage as of January 26, 1998



Unicode Home Page

Feedback | Site Map | Search

"Unicode 5.0 is a lifeline for scholars..." — Deborah Anderson (more)

New to Unicode?
- What is Unicode?
- How to Use this Site
- FAQ
- Glossary of Unicode Terms

General Information
- Where is my Character?
- Display Problems?
- Case Studies
- Useful Resources
- Last Resort Font
- Unicode Enabled Products
- Mail Lists
- Subset of the Unicode website in French

The Consortium
- Who we are
- Our Members
- How to Join
- Conferences
- Press Info
- Policies & Positions
- Contact Us

For Members Only
- Member Resources
- Working Documents

Copyright © 1991-2008 Unicode, Inc.
All Rights Reserved. Terms of Use

Buy the Book

Program Online Now!

Some of Our Members

Announcements

Last Resort Font now available for free download (2008.06.03)

Program for IUC 32 Available Online (2008.06.03)

NetApp joins Unicode at the Full level (2008.05.08)

Unicode 5.1 Released (2008.04.04)

Public Review
- Draft UTR #45: U-Source Ideographs
- Proposed Update to UTS #18: Unicode Regular Expressions

(Archived announcements...)

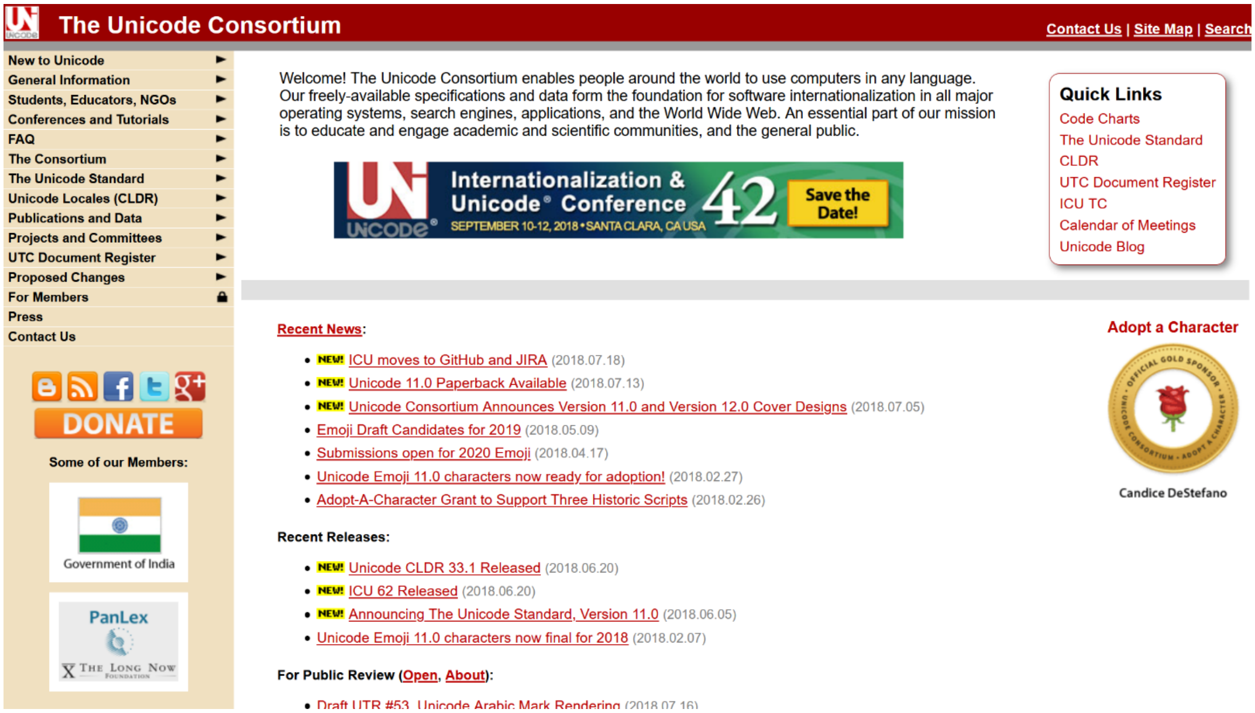
The Unicode Standard
- Start Here
- Latest Version
- Archive of Earlier Versions
- Referencing the Standard
- Code Charts
- Unicode Character Database
- Unihan Database
- Ideographic Variation Database

Key Specifications
- Unicode Collation (UCA)
- Bidirectional Algorithm (Bidi)
- Normalization (NFC, NFD, ...)
- Locale Data (CLDR)
- Script Codes (ISO 15924)
- Security Considerations

Technical Publications
- Technical Reports & Standards
- Technical Notes
- Online Data Tables
- The Unicode Guide
- Updates & Errata

Work in Progress
- Calendar of Meetings
- Proposals for Public Review
- Unicode Technical Committee
- UTC Meeting Minutes
- Proposed Characters (Pipeline)
- Submitting Proposals
- CLDR Technical Committee
- UDHR in Unicode

Figure 20 Unicode.org homepage, July 1, 2008. Retrieved from the Internet Archive 04-22-21



The Unicode Consortium

Contact Us | Site Map | Search

- New to Unicode ▶
- General Information ▶
- Students, Educators, NGOs ▶
- Conferences and Tutorials ▶
- FAQ ▶
- The Consortium ▶
- The Unicode Standard ▶
- Unicode Locales (CLDR) ▶
- Publications and Data ▶
- Projects and Committees ▶
- UTC Document Register ▶
- Proposed Changes ▶
- For Members 🔒
- Press
- Contact Us

DONATE

Some of our Members:

Welcome! The Unicode Consortium enables people around the world to use computers in any language. Our freely-available specifications and data form the foundation for software internationalization in all major operating systems, search engines, applications, and the World Wide Web. An essential part of our mission is to educate and engage academic and scientific communities, and the general public.

Internationalization & Unicode® Conference 42
SEPTEMBER 10-12, 2018 • SANTA CLARA, CA USA
Save the Date!

Quick Links
- Code Charts
- The Unicode Standard
- CLDR
- UTC Document Register
- ICU TC
- Calendar of Meetings
- Unicode Blog

Recent News:

- NEW! ICU moves to GitHub and JIRA (2018.07.18)
- NEW! Unicode 11.0 Paperback Available (2018.07.13)
- NEW! Unicode Consortium Announces Version 11.0 and Version 12.0 Cover Designs (2018.07.05)
- Emoji Draft Candidates for 2019 (2018.05.09)
- Submissions open for 2020 Emoji (2018.04.17)
- Unicode Emoji 11.0 characters now ready for adoption! (2018.02.27)
- Adopt-A-Character Grant to Support Three Historic Scripts (2018.02.26)

Recent Releases:

- NEW! Unicode CLDR 33.1 Released (2018.06.20)
- NEW! ICU 62 Released (2018.06.20)
- NEW! Announcing The Unicode Standard, Version 11.0 (2018.06.05)
- Unicode Emoji 11.0 characters now final for 2018 (2018.02.07)

For Public Review (Open, About):

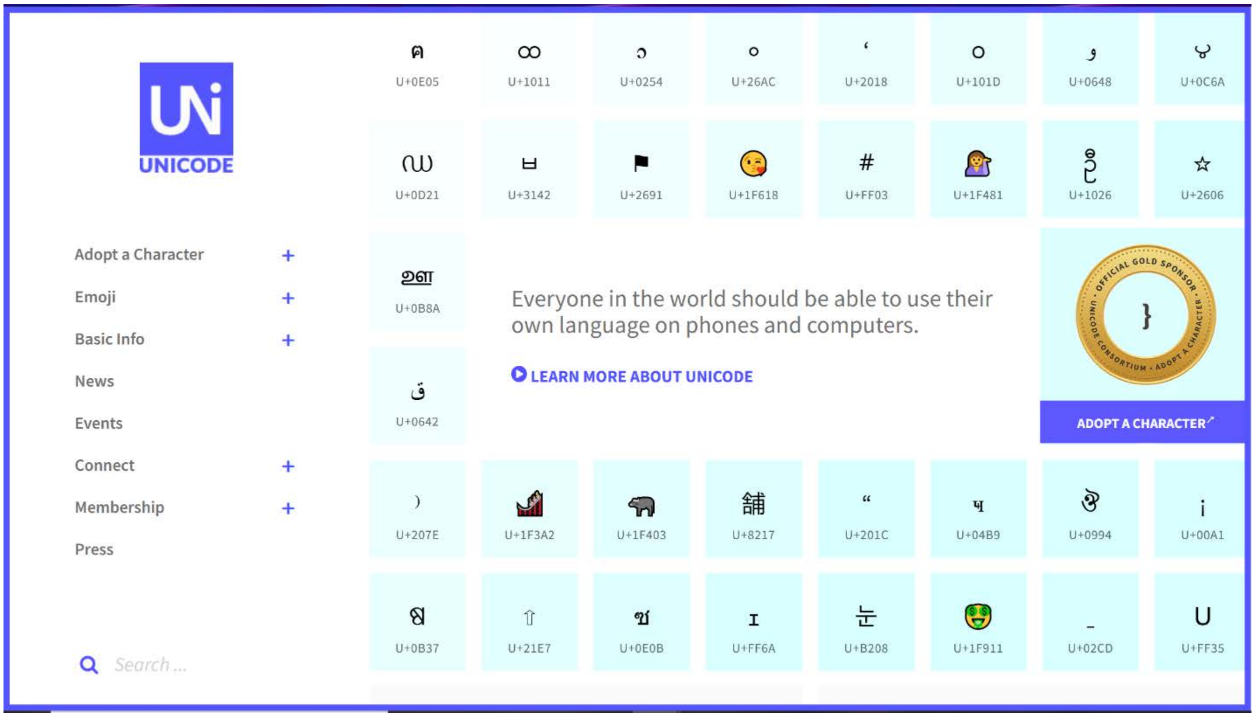
- Draft UTR #53, Unicode Arabic Mark Rendering (2018.07.16)

Adopt a Character

Candice DeStefano

Figure 21 The Unicode.org homepage ten years later, on August 1, 2018

As seen in Figure 21, we do see the addition of the Adopt a Character feature and icon as a persistent part of the home page, beginning in early 2016. We will discuss this feature and the timing of its appearance as a part of the Unicode.org website later in this chapter. By the latter half of 2019, however, the Unicode.org site saw its first major visual reorganization in over a decade, see Figure 22, below. This homepage now uses the http://home.unicode.org URL, and for several months after its initial launch, visitors to the site were greeted with a pop-up message reading, “Welcome to the re-designed Unicode Consortium home page. We hope you like it. If you want to access our traditional, technical site, click here: unicode.org/main.html” (Unicode – The World Standard for Text and Emoji, 2019).



Everyone in the world should be able to use their own language on phones and computers.

LEARN MORE ABOUT UNICODE

Figure 22 The Unicode.org homepage (now using the home.unicode.org subdomain) 4-22-21

This new homepage greatly condenses the number of navigation links on the page, and notably, the “New to Unicode?” menu which had maintained its top spot on the site for more than ten years is gone, being usurped by “Adopt a Character” and “Emoji”, before appearing as the more toned-down “Basic Info” link, which continues to redirect to the same URL as the “New to Unicode” link did previously. Likewise, news about the goings on of the Consortium and academic work related to Unicode have been wiped from the home page altogether in favor of a more visual presentation, emphasizing the global nature of the Standard, showing characters and symbols from a broad spectrum of orthographies and symbol sets, including several emojis, as well as the Adopt a Character badge, which serves as a rotating gallery of adopted characters, many of which are emojis. And, in the middle of the page, there is a new slogan, “Everyone in the world should be able to use their own language on phones and computers” (Ibid). This represents a major departure in the original mission of Unicode, which as we have seen, was to create a character-encoding standard compatible across global hardware and software, and in its original forming document asks " Is it possible to engineer a reasonable definition of "character" such that all the world's scripts contain fewer than 65,536 of them? The answer to this is Yes”—a far cry from all languages and scripts as implied by the new slogan.

Clearly, the importance of emojis as a part of their public image is not lost on the Unicode Standard, and they have gone out of their way with this latest redesign to center the public’s relationship with emoji characters as key to their work. Of note is the redesign pop-up, indicating that Unicode’s technical site remains intact, and signaling that *this* site is not a technical one, but rather for general engagement—particularly via emojis in general and Adopt a Character specifically. This homepage, in fact, is not much more than a façade, and travelling further into the website reverts back to the previous site structure and visual design. Next, I will explore that structure, and discuss further how Unicode has positioned its publicly-available documentation to support its role as a universal public good.

6.1.2 Unicode.org as Networked Document

For a more in-depth exploration of the Unicode.org website, and the way its constituent documents have been linked together to form various paths of navigation through the site, I turn to network analysis. My interest and approach to creating this network is a geographical one: I can safely assume that the pages of the Unicode.org website are already linked together under the Unicode.org domain, and are connected to one another via hyperlinks between pages, and that the resultant network visualization can be then read as a "map" of the site. This map, combined with metadata collected about the contents of each of the site's individual pages indicates to us the ways that Unicode has organized itself across a variety of subject and content types, and further demonstrates the ways that Unicode intentionally positions itself relative to its public in its ongoing attempt to be seen as a public good.

This analysis was conducted by first using Python to scrape the Unicode.org website, creating a list of all of the links on the home page, and then from each successive page linked from the previous. For the sake of limiting the scope of the URL collection, this was limited to URLs with the Unicode.org domain, meaning that while links to external pages are indicated, once encountering a non-unicode.org URL, the code collecting data stopped its process. The result of this is a table listing each unique URL under the top-level Unicode.org domain name, and the pages it is linked directly to. This type of structured data can then be imported into network visualization software, in this case Gephi, to create the “maps” of the site, as seen in the figures below. A limitation of this type of visualization is that captures a static image of the links between pages, when in reality websites are updated and deprecated all the time. All of the network visualizations for Unicode.org which follow are based on a capture of the website as it existed on June 1, 2018.

In these network visualizations, each circle represents a unique URL within the Unicode.org domain— http://unicode.org/news, http://unicode.org/emojis, and so on. The lines between the circles indicate that a direct link exists between two pages. Within the environment of a website, these links are *directional*, meaning that a link from Page A to Page B does not necessarily indicate the presence of a reciprocating link from Page B to Page A. In these network visualizations, the size of each page’s circle is directly correlated to the number of both in- and out-links to that page. That is, better-connected pages with more links to and from them are depicted as larger in the diagram. The Unicode.org homepage has been identified in the center of the map, and in Figure 24 the pages within the site which contain the word “emoji” in either the URL or page title are highlighted in pink.

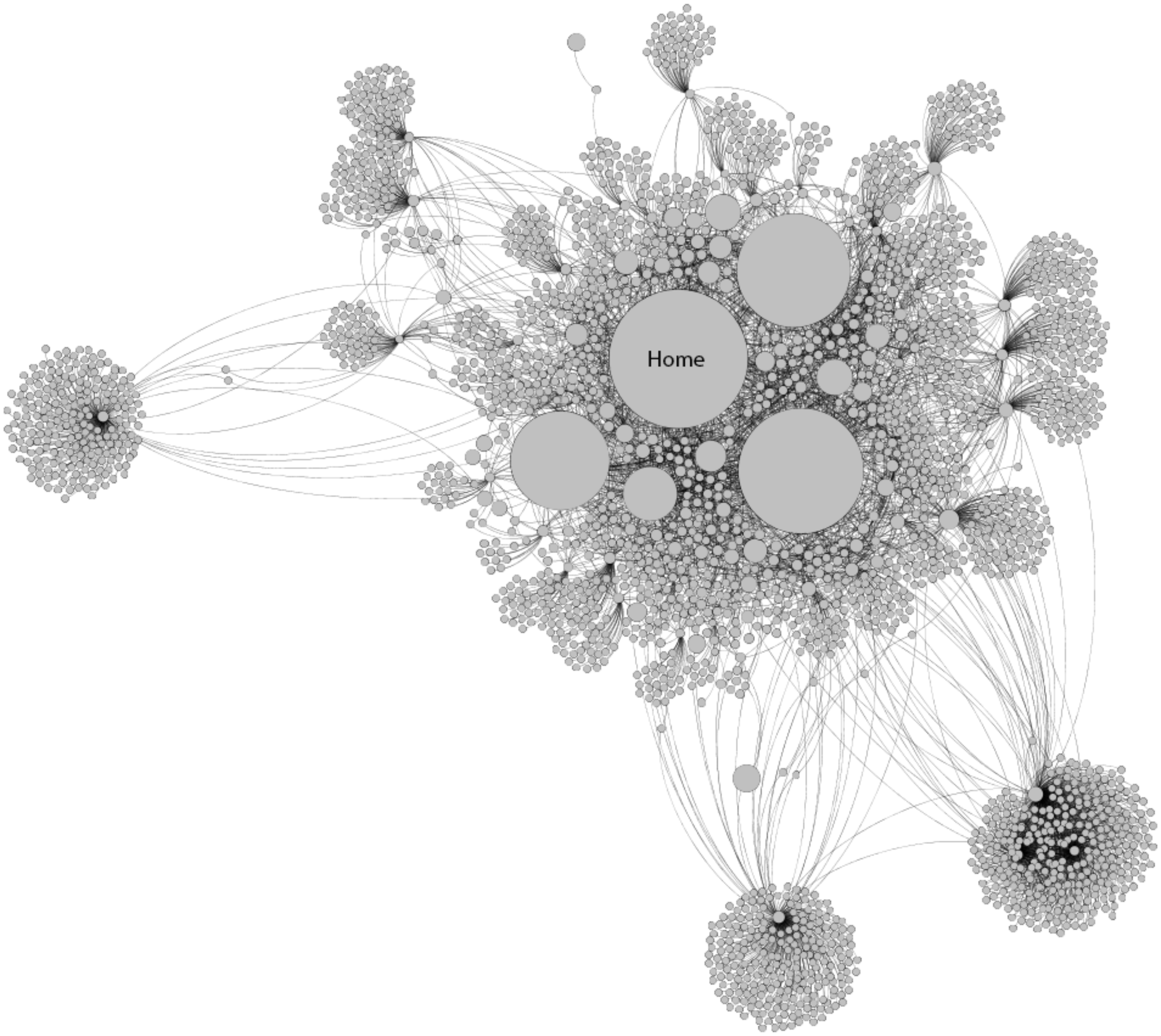


Figure 23 Basic Layout of the Unicode.org Website.

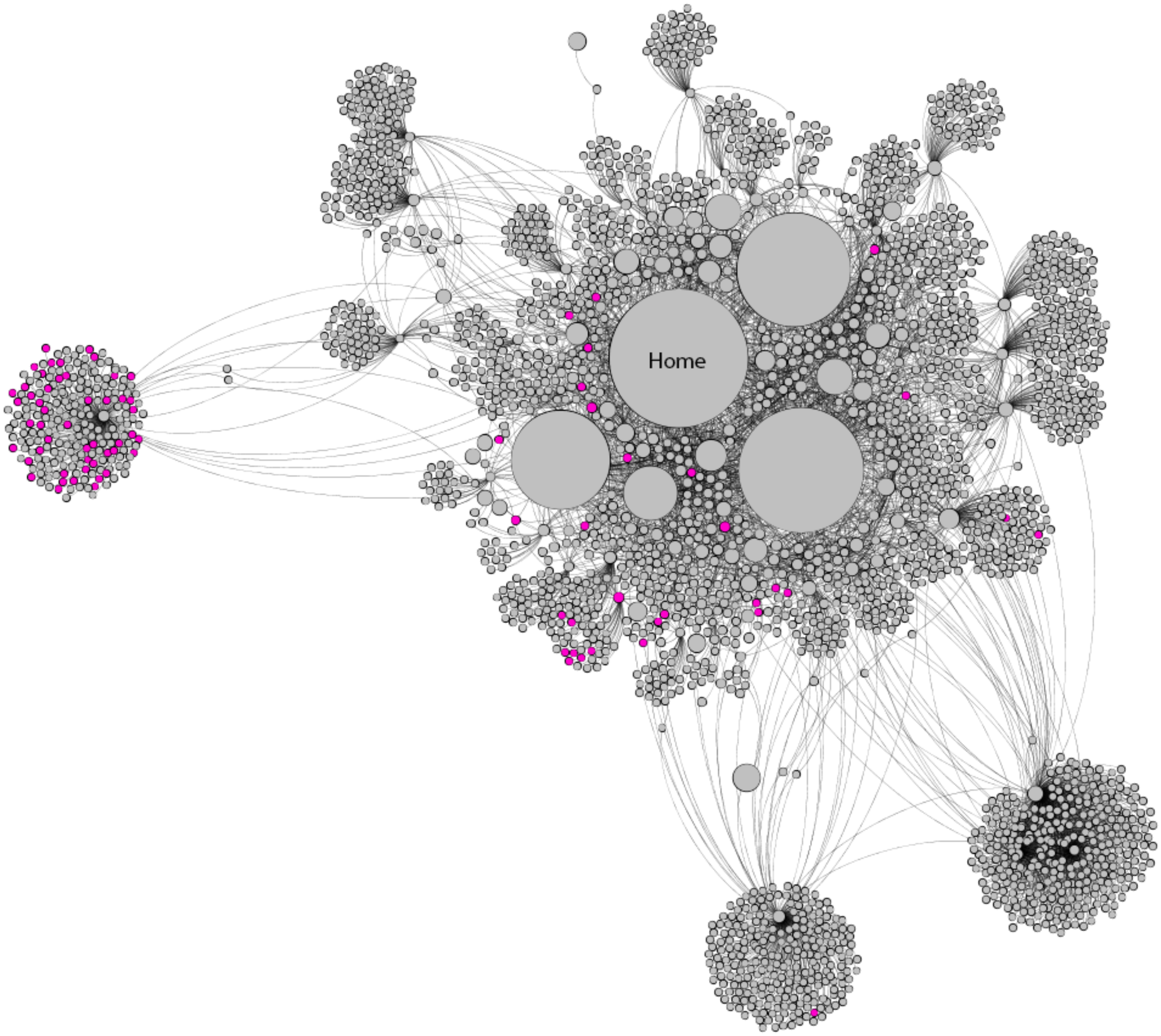


Figure 24 Basic Layout of Unicode.org Website, with Emoji pages highlighted in pink

From these network maps of the Unicode.org website, we can learn several things about how Unicode organizes its information, and how it chooses to present it to its online audience. Starting with the central node of the website, the home page, and working outwards we can see five larger pages crowded with many smaller pages. From these top pages, the relative connectivity of the remaining pages drops precipitously, indicating that these main pages are far more likely to receive traffic from visitors, as any given link on a page of the website is more likely to direct to one of these top pages than any other page.

Additionally, we see the presence of hubs appearing around the edge of the diagram, indicated by pages with few in-links, and a large number of out-links leading to a digital ‘dead-end’ page, a page which itself does not link out to other pages. What would create such hubs? That information is revealed when we color the individual page nodes by the type of content they contain (see Figure 25, below). These hubs can be best defined as topic areas within the Unicode.org website, where the hub page represents an HTML document with descriptive data about the linked documents, which in many cases are PDF documents containing meeting minutes, descriptive metadata for various encoded scripts, or other documents relating to the technical aspects of the Standard or managerial aspects of the Consortium. This metadata about the Unicode.org webpages was collected through keywords present in page URLs, page titles, and file-type extensions such as .pdf, .txt, or .mov.

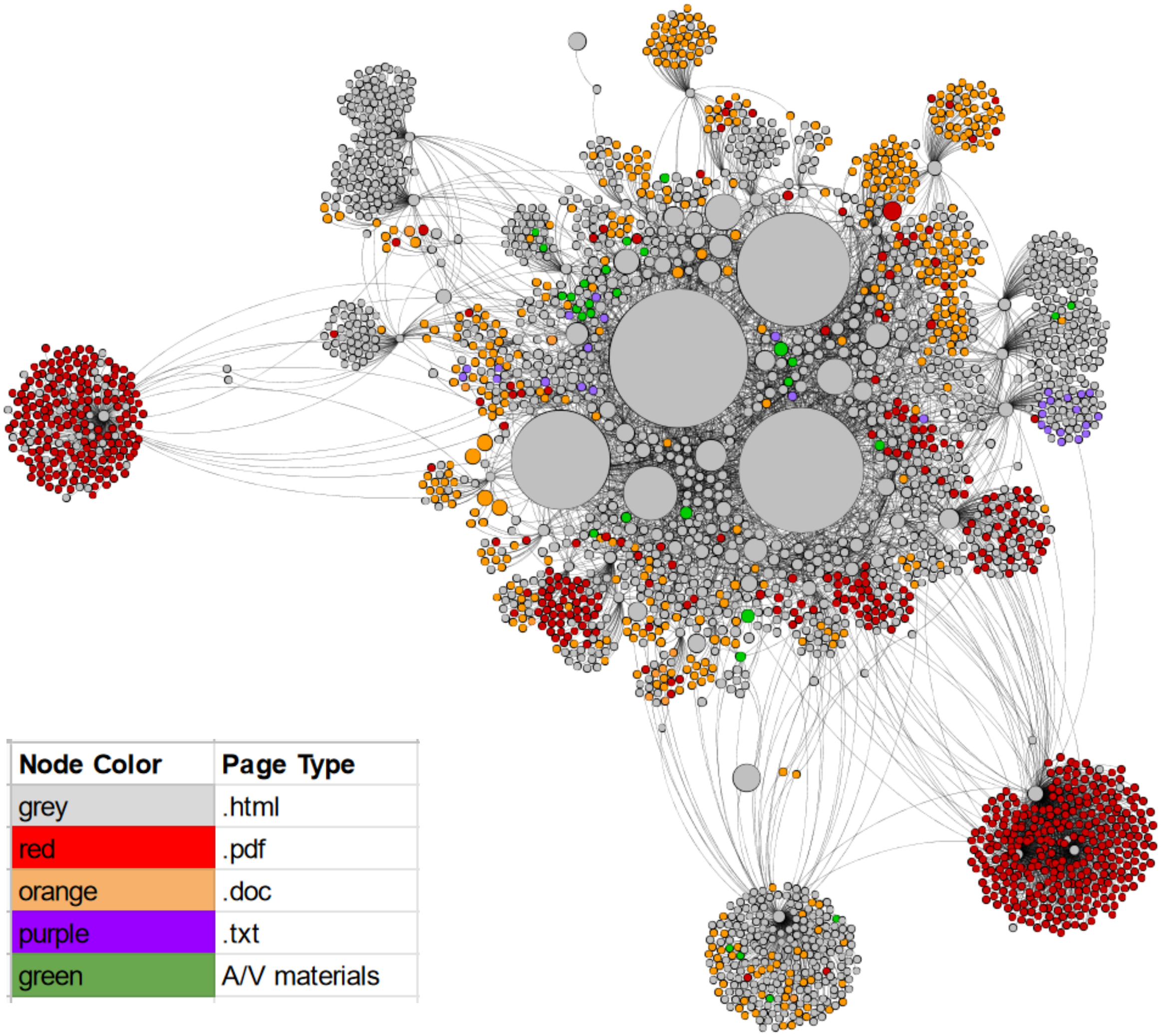


Figure 25 Network Visualization of the Unicode.org Website, with Nodes colored by Page Type

Once equipped with the full list of Unicode.org domain URLs and their links, I was able to fill in additional metadata about the majority of the pages across three categories: Page Type, Content Type, and Audience. Page Type is indicated by page format ending (.html, .pdf, .doc, etc), Content Type is based on page title and subdomains (all pages under http://cldr.unicode/org/...

have the Content Type CLDR, for example). For the Audience category, I labelled each individual node as being part of one of three major audience categories: Internal, Public, and Emojis. Emoji pages, in this case, are drawn from pages that would otherwise be label either Internal or Public, but I have chosen to highlight them separately to demonstrate the relative amount of emoji vs non-emoji content on the site. These final categories were developed based on an iterative process of coding, visualizing, and re-evaluating the data in order to incorporate as many of the scraped URLs under as few categories as possible. The full list of categories and labels is presented in Table 5.1, below:

Table 5.1: Unicode.org page categories and labels

Page Type	Content Type	Audience
.html	documentation	Internal
.pdf	versions	Public
.doc	reports	Emojis
.txt	CLDR	
A/V materials	consortium	
	emoji	
	FAQ	
	announcements	
	ISO	
	public	
	notes	
	history	
	policy	
	charts	

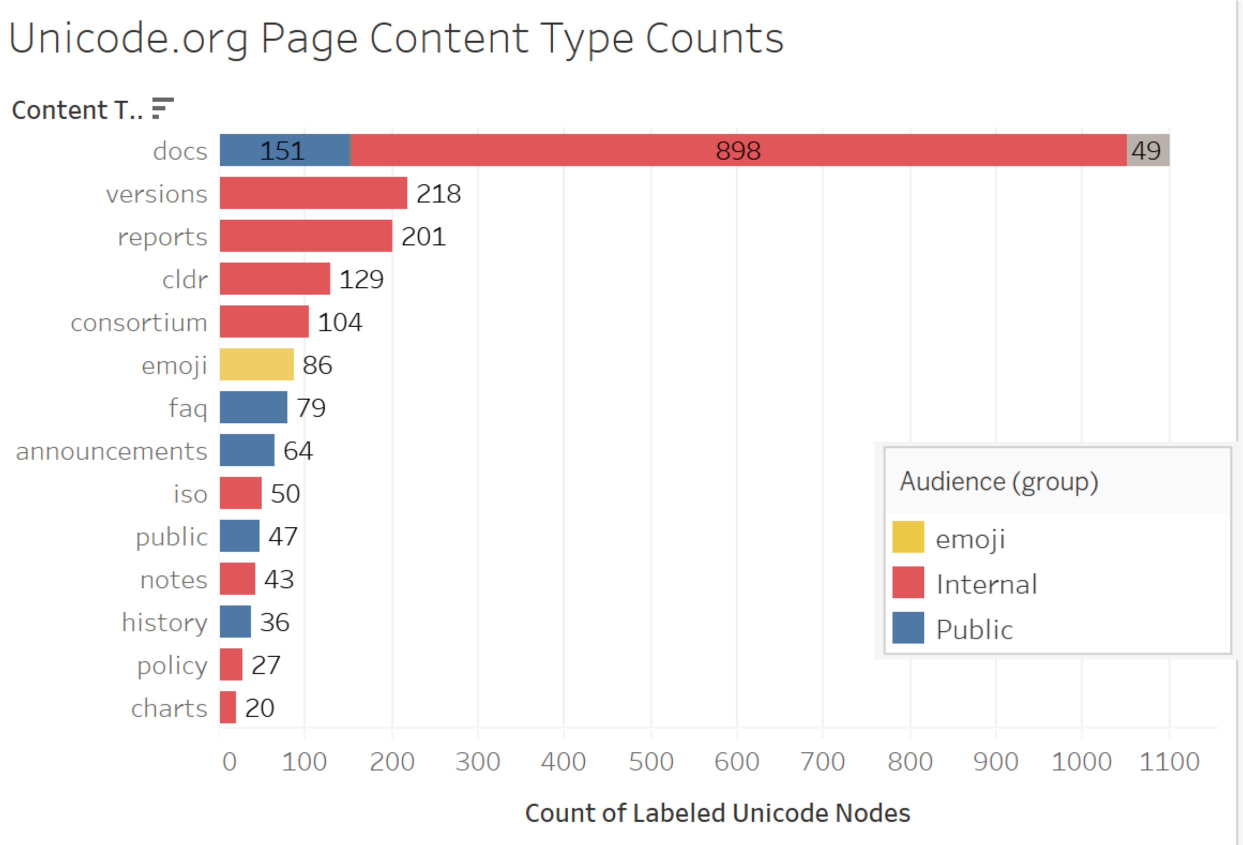


Figure 26 Unicode.org Page Content Type Counts

From this data, we can see that the majority of the content on the Unicode.org website is internally-focused (see Figure 26, above). In fact, only public documentation pages (The "docs" Content Type being distinct from Page Type: .doc), outnumber all emoji pages across the website. Documentation pages, perhaps understandably, make up the vast majority of the categorizable pages on the website, and represents the only Content Type that spans several audiences. What this shows us is that Unicode.org serves primarily as a repository for organizational documentation about the Standard and the Consortium, despite the home page's emphasis on public engagement.

6.2 Adopt A Character

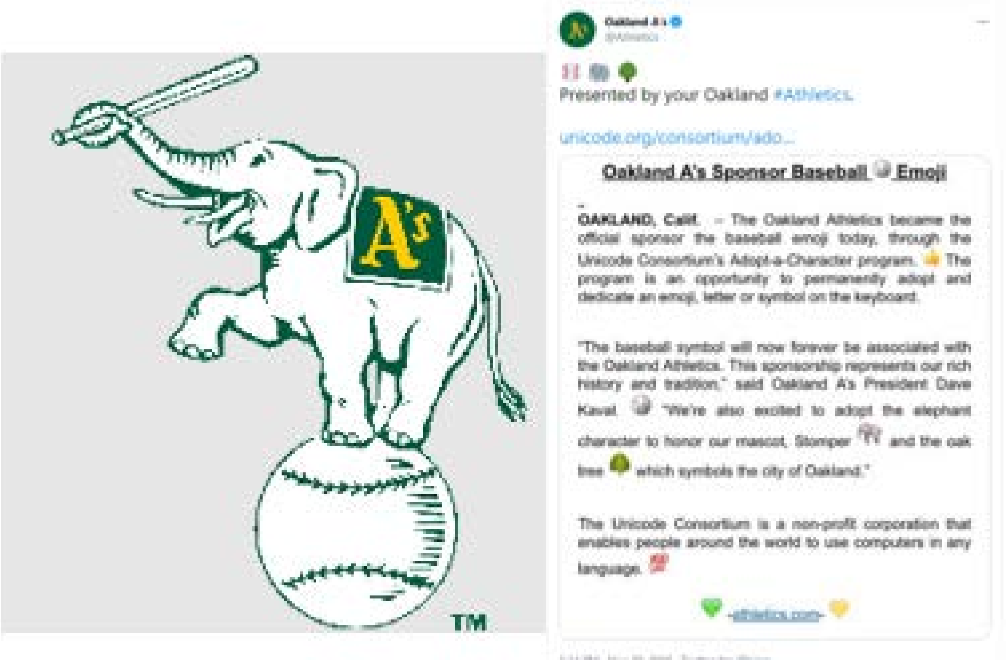
As discussed briefly above, Adopt a Character is a feature which Unicode promotes prominently on the website. The service functions much like “adopt a tiger” or other sponsorship-based fundraising schemes. For a fee, individuals, groups, and businesses are able to associate their name (or someone else’s of their choosing) with a particular character of the Unicode set. While this is open to most of the entire set of Unicode characters (with some exceptions for control and other special characters), the majority of the emphasis in the promotion and consequently, in the set of adopted characters, are emojis.

> Do you want your company to be associated with the 🍔 emoji? Do you feel like the poor semicolon and the equals sign never get any respect? Do you want to declare your love with a 💍 or 💘 dedication for your partner?[7] (About Adopt a Character, 2019)

The page advertises over 136,000 characters that can be adopted, all in the service of helping “the non-profit Unicode Consortium in its goal to support the world’s languages. (2019). Sponsorship can happen at the Bronze ($100 - $999), Silver ($1000 - $4999), and Gold ($5000+) levels. Multiple entities may sponsor the same character at the Bronze level, but adoption at the Silver and Gold tier is exclusive on a first come, first served basis. Sponsorships at the Gold level

[7] Ironically, on the Unicode.org site, this statement does not make use of the emoji characters it depicts, instead images of each character are inserted into the sentences’ formatting.

include the Oakland A’s, who adopted the baseball, elephant, and deciduous tree emojis, as a representation of their team logo in 2016.



Oakland A's
@Athletics

⚾ 🐘 🌳
Presented by your Oakland #Athletics.

unicode.org/consortium/ado...

Oakland A's Sponsor Baseball ⚾ Emoji

OAKLAND, Calif. – The Oakland Athletics became the official sponsor the baseball emoji today, through the Unicode Consortium's Adopt-a-Character program. 👍 The program is an opportunity to permanently adopt and dedicate an emoji, letter or symbol on the keyboard.

"The baseball symbol will now forever be associated with the Oakland Athletics. This sponsorship represents our rich history and tradition," said Oakland A's President Dave Kaval. ⚾ "We're also excited to adopt the elephant character to honor our mascot, Stomper 🐘 and the oak tree 🌳 which symbols the city of Oakland."

The Unicode Consortium is a non-profit corporation that enables people around the world to use computers in any language. 💯

💚 -athletics.com- 💛

Figure 27 Tweet announcing the Oakland Athletic’s sponsorship of several emojis

As of June 2021, prior to the upcoming Unicode 14.0 release, there are currently 60 Gold Sponsors, 49 of which (81.6%) are emojis or emoji characters, 69 Silver Sponsors, 55 of which are emojis (79.7%), and 816 Bronze Sponsors, 571 of which are emojis (70%). This comes to a total of 945 adopted characters out of the advertised 136,000 possibilities (~0.7% sponsorship rate), and of those 945, 675 are emojis (71.4%).

Assuming the minimal donation amount per tier, this outreach program has generated at least $300,000 for the Unicode Consortium, and likely a great deal more, as the donations are tax deductible any many at the Gold Tier level come from multi-billion dollar corporations, including members of the Consortium themselves, such as IBM and Adobe. However, fundraising is not the only purpose of having Adopt a Character as a prominent part of the public presentation of the Standard—by being able to associate an individual’s name with a specific character, the character, and the Standard as a whole, are humanized. This is consistent with the Standard’s ongoing efforts to position itself as providing a public good, one that the consumer can support through the Adopt a Character feature.

7.0 Discussion, Findings, and Future Research

Where do characters, as visual/written objects, live within Unicode? Do they? The visual appearance of characters varies between operating system, program, and chosen font or typeface. The nameslist of Unicode makes use of a cross-referencing system, as described in Chapter 4, to indicate which characters share visual similarities, or are derivatives of other characters, and occasionally indicates best practices regarding the use of one character over another. This reflects the ‘cataloging mindset’ necessary when managing and organizing data at this volume. While Unicode continues to treat its characters as a commodity, and its namespace as a limited resource with which to hold those characters, the presence of this small cross-referencing practice indicates the necessity of such a cataloging mindset when dealing with data at such volume.

7.1 A Catalog of Characters

The distinction between the ‘space’ in the Standard and its occupants-- the characters which we all use-- is about consumption. A slot in the Standard is used up when a new character is assigned that spot and given an associated number U+XXXX, but the use of any given character by someone typing a letter does not detract from the total pool of characters and does not prevent other users from typing with that same letter. In this sense, a character added to the Unicode Standard is made into a piece of information, in the same way that Suzanne Briet describes how the antelope in a zoo become a ‘document’—it has been placed into a meaningful context where it can be observed, regulated, and made use of ‘safely’—in the case of the antelope by viewing it from outside the zoom enclosure, and in the case of Unicode, by assuring cross-compatibility and code stability (Briet and Martinet 2006).

This process seems to be in place within Unicode regardless of whether its contents are considered goods or classifications, but the abstraction of a ‘true form’ of a character which is made by associating it with its unique U+XXXX identifier elevates its place in Unicode to that of “the work,” in terms of the FRBR classification system, or the “Idea,” in Platonic terms. The “j” that lives in Unicode is the ideal representation of “j” and its essence is not changed whether it is typed “j” “𝐣” or “𝒿” or whether it is pronounced dʒ or like the English letter y.

However, because Unicode takes a monopolistic approach to the construction of language, making sweeping assumptions and generalizations about the pieces used to build human communication, and because the Unicode Consortium is invested in the continued expansion and use of its standard, the ability for Unicode, as it currently exists, to serve as an orthographic classification system is severely limited. For instance, Unicode assumes that languages are made of discrete characters which can be combined to create words and phrases with distinct linguistic meaning, which is then conveyed to others solely via their viewing those characters—that is that languages can be written and read in the way that English is written and read. This philosophical understanding and application of ‘what language is’ has been applied from a top-down perspective not only to the Unicode Standard and its characters, but also to all of the hardware and software which makes use of it. To create a technological object which is Unicode-compatible is to implicitly endorse and reinforce a discrete-character-based understanding of language. However, there are ample examples of known human languages which do not conform to this understanding.

From all this, we can say conclusively that Unicode treats its namespace as a limited resource existing in a marketplace of digital goods, and the corporations involved in the Consortium as having a major power advantage in dictating how that space gets used up. This is consistent with my earlier labelling of the Standard's accession practices as being largely reactionary—reactionary to the submissions of its users and their perception of the Unicode "brand" via surface-level changes to the Unicode.org website, as well as reactionary to the goals of its constituent members, tech giants intent on collecting user data and recognizing that the emoji boom has provided access to a heretofore difficult to access part of digital written communication—how users are *feeling* about what they say and interact with.

However, we can reframe Unicode and its 'permanent collection' as descriptive representations of character-objects, documented in a structured manner. This interpretation falls much more in line with Unicode's own stated mission to make permanently digitally accessible the complete breadth of human written communication forms. In this light, Unicode becomes a repository with a controlled vocabulary of objects, connected by cultural, historical, technical and other characteristics, aka a classification system. However, for such a system to be meaningful, records within it must be findable. This is a situation that remains difficult for Unicode characters, whether it be because of the physical limits of a QWERTY keyboard, inconsistent metadata fields and application or the Standard's inherent incompleteness.

Emojipedia, one of only two educational voting members of the Consortium has taken on this task on behalf of emojis, providing a searchable database of all emojis with information on their official and unofficial titles, meanings, and related characters. It also provides examples of how each emoji appears across different platforms, including examples dating as far back as the

DoCoMo set of characters. Because Emojipedia is an official member of Unicode, it is reasonable to say that this is an accurate, up-to-date representation of the emoji content of the Unicode Standard. However, no such database exists for the rest of Unicode's non-emoji characters—instead depending on yearly print releases of the latest Standard documentation as a human-readable reference, and the .txt nameslist which Unicode itself indicates is a resource intended for machine use, but rendered legible for human readers.

7.2 Changing Linguistic Environment

Emojis are directly descended from the use of mobile phone keyboards, and while they can be used in other settings more and more easily, it focuses affective communicative value on the mobile device/keyboard. This is reenforced by the fact that mobile devices, with their on-screen digital keyboards, greatly outnumber computers with analog keyboards in terms of general use worldwide, and often provide access to people and places that would not normally be able to make use of such technology (Graham, Hale, Stevens 2012). I could write a whole other dissertation about the visual evolution of the digital keyboard, and the stubborn refusal of QWERTY to die, but that is indeed best saved for another time. Suffice to say at this moment that while the bare-bones alphabetic structure of the keyboard remains the same, methods for inserting emojis, gifs, images, or stickers into regular conversation are increasing in number and the barrier to creating such affective objects is also dropping precipitously, with users on iMessage, Signal, WhatsApp, and other messaging platforms being able to create and share their own stickers with other users. Similarly, the phenomenon of bitmoji allows users to model themselves in the style of an emoji and send messages with their bitmoji avatar actually ‘performing’ the emoji for the receiver. This is evidence of the expansion of emojis beyond the fixed realm of Unicode, and allows them to exist as meaning-making objects independently, unlike most of the gylphs in Unicode, which function only as a part of an existing script or alphabet. This frees users of social obligations or pressures to use emojis in a prescriptive way, but allows them to modify, remix, or otherwise manipulate images of emojis to expand the affective vocabulary they create.

7.3 Decolonizing an Information Commons

The work of the Unicode Consortium on expanding Unicode is good, but still treats non-roman alphabet scripts as “additions” to the foundational understanding of text. Barring a complete descent into a new dark age, followed by the opportunity to “start over” with computing technology, we will never be able to undo the biases and assumptions which are built into our technological systems. My work here has been to point out exactly where and how these slips and false assumptions made it into the logic of our technology, and how such frameworks limit what we are able to do with that technology. People have been advocating for the inclusion of their language/script in Unicode since the beginning of Unicode, but it was with the addition of Emoji 1.0 in 2015 that real discussion in the public sphere began about who was being represented by these characters. This, if nothing else, I would argue, has made the addition of emojis to the Standard worthwhile, but emojis also no doubt represent the bleeding edge of both technological and linguistic innovation.

Work happening in LIS has begun the conversation around the role of libraries as places for infrastructural abundance, using the language and framework of the commons. (Halperin 2020, Mattern 2014, Fister 2014) FRBR stands as just the latest iteration of US librarianship's attempt to grapple with some of the conceptual depth of organizing information objects, one in a long line of such practical and philosophical proposals that spans disciplines so broadly that it might be meaningless to even invoke "interdisciplinary" practices. I, along with others in this profession, believe that global, alternative, and indigenous knowledge paradigms have much to offer the world of classification and ontology, and in particular I urge LIS to take up the work, at least theoretically, of imagining a linguistic knowledge commons based on abundance, rich network embeddedness, and equitable access for all (Joranson, 2013, Bonnand and Donahue, 2010).

7.4 ALA Encoding Standard

Unicode, developed and endorsed by the same groups which created and adopted ASCII, seemed like the obvious and easiest replacement for the previous generation's standards, but the ways that the Consortium chose to address each of the issues they hoped to resolve with a UCS are just that—choices. We can see in many places throughout the history of character-encoding the individual opinions, preferences, and styles of a system's designer manifest in the standard itself. But Unicode was not the only option, and the nascent Unicode Consortium was not the only organization working to establish an encoding system which best met their needs. In fact, so many alternate encoding systems were in use or being proposed for adoption during this time period that it is beyond the scope of this dissertation to discuss them all. I would like to focus for a moment on one alternative, proposed by the American Library Association in 1989, which considers the classification of characters from a perspective outside the traditional grouping by language or script, as we see Unicode organized today.

The field of library classification has had a vested interest in a universal encoding standard since the 1970s, when catalog records began to be created digitally. In the midst of this transition the American Library Association (ALA) was in the process of developing its own character encoding standard (Peruginelli et al 1992). Because library collections often hold books in a variety of languages, and search and retrieval rely on the accuracy of title and author transcription in to the catalog, librarians struggled with ways around the 128- or 256- character limit of early encoding standards. While Unicode seemed to be the best possible solution to this problem— not only would every possible character be available, they would be able to be moved between platforms and devices losslessly— the need for efficiency and multiple paths of use on the developer end resulted access and normalization difficulties on the user end.

A major goal of library classification is to make search and retrieval more accurate and efficient. As MARC records have become the standard method of creating such classification, and text- or keyword-search the user's means of reaching those records, consistency in spelling, capitalization, and romanization or translation of foreign-language texts has become of tantamount importance. Perguinelli et al point out, however, that the lax control and multiple entry points for visually similar characters in Unicode creates problems for record normalization and character-matching retrieval. In Unicode, the letter "ñ" can be composed of a single character, "Latin Small Letter N with Tilde", (U+00F1), or with two characters combined, "Latin Small Letter N" (U+006E), and "Combining Tilde" (U+0303). These two possibilities provide identical visual outputs, but different results when sorted by machine, with the second being sorted either under "n" or "~" rather than "ñ." This raises obvious problems for librarians attempting to sort materials, and for patrons attempting to find them.

This issue was, in part, the reason that ALA created their own character set in the late 1980s, and attempted to standardize the hardware and built-in encoding software in library-use computer terminals (Library technology report, 1989). The ALA report on this projects identifies three major "searching problems" for text-based library cataloging systems (especially in non-English or mixed-language collections): Upper-lower case equivalence, stop words, and "noise" characters (pp. 270-271). Additionally, sorting and consistency problems are addressed. Sorting problems are largely character diversity as addressed above, and consistency addresses human-related input errors or variations in character usage.

The three searching problems boil down to which characters are considered significant to the searching algorithm, which can vary widely between languages, and be further complicated through cross-language cataloging practices. What makes the ALA proposal so very different from Unicode or any of its predecessors is focus on human usability. Unicode is a computer-facing standard, as was ASCII before it, and it is designed with the computer's efficiency in mind, and can lead to problems where human users think they are using one character when actually using another, visually identical one—resulting in the library search and retrieval issues mentioned above, among other issues.

In dealing with the treatments of texts, in both a physical and ideological sense, the struggles of the ALA in the 1980s highlights the chicken-or-the-egg problems of attempted linguistic uniformity, especially when a variety of different technological standards are involved.

An ALA encoding standard would have made library cataloging practice easier to standardize, and made text-searching a more feasible access option for library users much earlier in the history of LIS technological innovations. However, the ALA standard would have also required a fundamental restructuring of how characters were associated with one another—prioritizing visual consistency over traditional orthographic grouping in an effort to prevent duplicates.

Based on a cross-disciplinary understanding of “text”, it is impossible to create or circulate a text outside of a cultural paradigm, and the existence of “texts” in the first place is deeply dependent on the technologies which support them. That is to say, that we are not in a position to be able to dig ourselves out of the character-encoding hole that we are currently in, as it has become foundational to the way that communication happens in contemporary society, but as critical LIS scholar Emily Drabinski (2013) proposes with regard to *any* classification system, the goal should be not to suppress or sweep aside the deviations and mutations which appear to threaten the stability of such systems, but to treat these difficult-to-categorize items as evidence of the breadth of human experience and an invitation to critically examine the systems ordering our world which we often take for granted.

7.5 Conclusion

The Unicode Standard releases a single major annual update, and since beginning this dissertation, we have moved from Unicode 11.0 to 14.0. These updates have added 6,782 characters, including 153 emojis. In that time, emojis have been formally integrated as their own subset of Unicode, and labeled with their own versioning system. Likewise, emojis have become an everyday item in the lives of millions of people worldwide, and have proven their staying power as a visual vocabulary which bypasses—I hesitate to say ‘transcends’-- individual languages, and which has allowed its own artistic and creative subcultures to flourish in the digital realm. Early indications suggest that changes in communication due to the COVID-19 global pandemic are shifting the value of affective indicators in written text, and we are now seeing the worldwide boom of emojis as such semantic markers.

The Unicode Consortium continues to review applications for new characters, both emoji and non-emoji, and its constituent members continue to make use of the affective and visual possibilities of emojis to frame our digital conversations and make extant within the computer what has heretofore been so ephemeral about language—the subtle signals that we as humans use to convey more than the writ meaning of our words, and to place ourselves within the emotional context of our communications. The unregulated cycle of tech corporations both creating and exploiting the systems of digital writing for profit hangs heavy over the digital world as we know it, and risks linguistic privatization or even dissolution in the wake of potential technological collapse, with the entire world largely dependent on the continued growth and success of the (Western) companies which make up the Unicode Consortium to guarantee that the language preservation the Standard promises is maintained.

References

"😊 White Smiling Face Emoji." n.d. Emojipedia. Accessed November 20, 2021. https://emojipedia.org/smiling-face/.

197805 Interface Age V 03 I 05. n.d. Accessed September 23, 2020. http://archive.org/details/197805InterfaceAgeV03I05.

197806 Interface Age V 03 I 06. n.d. Accessed September 23, 2020. http://archive.org/details/197806InterfaceAgeV03I06.

197807 Interface Age V 03 I 07. n.d. Accessed September 23, 2020. http://archive.org/details/197807InterfaceAgeV03I07.

11319660. n.d. "Emoji by WOMANZINE." Issuu. Accessed April 8, 2021. https://issuu.com/lindseyweber5/docs/emoji_by_womanzine.

"A Companion to Digital Literary Studies 'Ss1-6-12.'" n.d. Accessed February 4, 2022. http://digitalhumanities.org:3030/companion/view?docId=blackwell/9781405148641/9781405148641.xml&doc.view=content&chunk.id=ss1-6-12&toc.depth=1&brand=9781405148641_brand&anchor.id=0.

Abbate, Janet. 2000. *Inventing the Internet*. MIT press.

"About the Unicode Standard." n.d. Accessed November 16, 2018. https://www.unicode.org/standard/standard.html.

Adamic, Lada A. 2009. "The Social Hyperlink." In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, 1–2. HT '09. New York, NY, USA: ACM. https://doi.org/10.1145/1557914.1557916.

Adler, Melissa. 2017. *Cruising the Library: Perversities in the Organization of Knowledge*. New York: Fordham University Press.

Albrechtsen, Hanne, and Elin K. Jacob. 1998. "The Dynamics of Classification Systems as Boundary Objects for Cooperation in the Electronic Library." https://www.ideals.illinois.edu/handle/2142/8212.

Aliprand, Joan M. 2011. "The Unicode Standard." *Library Resources & Technical Services* 44 (3): 160–67. https://doi.org/10.5860/lrts.44n3.160.

"All Major Vendors Commit to Gun Redesign." 2018. Emojipedia. April 27, 2018. https://blog.emojipedia.org/all-major-vendors-commit-to-gun-redesign/.

Alshenqeeti, and Hamza. 2016. "Are Emojis Creating a New or Old Visual Language for New Generations? A Socio-Semiotic Study." SSRN Scholarly Paper ID 3709343. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3709343.

"Alternate ASCII Registry." 2018. April 2, 2018. https://web.archive.org/web/20180402215348/http://www.bobbemer.com/REGISTRY.HTM.

Altman, Alon, and Moshe Tennenholtz. 2005. "Ranking Systems: The PageRank Axioms." In *Proceedings of the 6th ACM Conference on Electronic Commerce*, 1–8. EC '05. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1064009.1064010.

Amy R. Krosch and David M. Amodio. 2014. "Economic Scarcity Alters the Perception of Race." *Proceedings of the National Academy of Sciences - PNAS* 111 (25): 9079–84. https://doi.org/10.1073/pnas.1404448111.

Andrews, Gavin J. 2008. “Historiography.” In *The SAGE Encyclopedia of Qualitative Research Methods*, 400–400. Thousand Oaks: SAGE Publications, Inc. https://doi.org/10.4135/9781412963909.

“ASCII Articles in Interface Age Magazine.” 2018. April 2, 2018. https://web.archive.org/web/20180402220906/http://www.bobbemer.com/INSIDE-A.HTM.

Austin, J. L. 1962. *How to Do Things with Words*. Cambridge: Harvard University Press.

Ayotte, Kenneth, and Henry Smith. 2011. *Research Handbook on the Economics of Property Law*. Edward Elgar Publishing. https://doi.org/10.4337/9781849808972.

Backhouse, James, Carol W. Hsu, and Leiser Silva. 2006. “Circuits of Power in Creating de Jure Standards: Shaping an International Information Systems Security Standard.” *MIS Quarterly* 30: 413–38. https://doi.org/10.2307/25148767.

Barbieri, Francesco, Miguel Ballesteros, and Horacio Saggion. 2017. “Are Emojis Predictable?” *ArXiv:1702.07285 [Cs]*, February. http://arxiv.org/abs/1702.07285.

Barbieri, Francesco, and Jose Camacho-Collados. 2018. “How Gender and Skin Tone Modifiers Affect Emoji Semantics in Twitter.” http://repositori.upf.edu/handle/10230/34971.

Bartley, Tim. 2007. “Institutional Emergence in an Era of Globalization: The Rise of Transnational Private Regulation of Labor and Environmental Conditions.” *American Journal of Sociology* 113 (2): 297–351. https://doi.org/10.1086/518871.

Becker, Carl. 1938. “What Is Historiography?” *The American Historical Review* 44 (1): 20–28. https://doi.org/10.2307/1840848.

BEGHTOL, CLARE. 1995. "'FACETS' AS INTERDISCIPLINARY UNDISCOVERED PUBLIC KNOWLEDGE: S.R. RANGANATHAN IN INDIA AND L. GUTTMAN IN ISRAEL." *Journal of Documentation* 51 (3): 194–224. https://doi.org/10.1108/eb026948.

Bemer, R. W. 1960a. "A Proposal for Character Code Compatability." *Communications of the ACM* 3 (2): 71–72. https://doi.org/10.1145/366959.366961.

———. 1960b. "Survey of Coded Character Representation." *Communications of the ACM* 3 (12): 639–41. https://doi-org.pitt.idm.oclc.org/10.1145/367487.367493.

Berard, Bethany. 2018. "I Second That Emoji: The Standards, Structures, and Social Production of Emoji." *First Monday* 23 (9). https://doi.org/10.5210/fm.v23i9.9381.

Bergamin, Giovanni, Susanna Peruginelli, and Pino Ammendola. 1992. "Character Sets: Towards a Standard Solution?" *Program* 26 (3): 215–23. https://doi.org/10.1108/eb047115.

Berman, Sanford. 1993. *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People*. McFarland.

Blanchette, Jean-François. 2011. "A Material History of Bits." *Journal of the American Society for Information Science and Technology* 62 (6): 1042–57. https://doi.org/10.1002/asi.21542.

Bonvillain, Nancy. 1993. *Language, Culture, and Communication*. Englewood Cliffs, N.J: Prentice Hall.

Borgman, Christine L. 2003. *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*. MIT Press.

Bowker, Geoffrey C., Karen Baker, Florence Millerand, and David Ribes. 2009. "Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment." In

International Handbook of Internet Research, 97–117. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-9789-8_5.

Bowker, Geoffrey C., and Susan Leigh Star. 1998. "Library Trends 47 (2) 1998: How Classifications Work: Problems and Challenges in an Electronic Age." https://www.ideals.illinois.edu/handle/2142/8213.

Bowker, Geoffrey C, and Susan Leigh Star. 2008. *Sorting Things out: Classification and Its Consequences*. Cambridge, Mass. [u.a.: MIT Press.

Bray, David, and Vinton Cerf. n.d. *The Unfinished Work of the Internet*. *Society and the Internet*. Oxford University Press. Accessed August 21, 2020. https://oxford-universitypressscholarship-com.pitt.idm.oclc.org/view/10.1093/oso/9780198843498.001.0001/oso-9780198843498-chapter-25.

Briet, Suzanne, and Laurent Martinet. 2006. *What Is Documentation?: English Translation of the Classic French Text*. Scarecrow Press.

Bromwich, Jonah Engel. 2017. "How Emojis Find Their Way to Phones." *The New York Times*, December 21, 2017, sec. Technology. https://www.nytimes.com/2015/10/21/technology/how-emojis-find-their-way-to-phones.html.

Brunsson, Nils, and Bengt Jacobsson. n.d. *A World of Standards*. *A World of Standards*. Oxford University Press. Accessed October 22, 2020. http://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199256952.001.0001/acprof-9780199256952.

Brunsson, Nils, Andreas Rasche, and David Seidl. 2012. “The Dynamics of Standardization: Three Perspectives on Standards in Organization Studies.” *Organization Studies* 33 (5–6): 613–32. https://doi.org/10.1177/0170840612450120.

Buckland, Michael K. 1997. “What Is a ‘Document’?” *Journal of the American Society for Information Science* 48 (9): 804–9.

———. n.d. “What Is a Digital Document?” Accessed December 28, 2020. https://people.ischool.berkeley.edu/~buckland/digdoc.html.

Busch, Lawrence. 2011. *Standards: Recipes for Reality*. Mit Press.

Bush, Vannevar. 1945. “As We May Think.” *The Atlantic*, July 1945. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Busha, Charles H., and Stephen P. Harter. 1980. *Research Methods in Librarianship: Techniques and Interpretation*. Academic press.

Butler, Eamonn. 2012. “Public Choice - A Primer.” SSRN Scholarly Paper ID 2028989. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2028989.

Byrne, Bernadette M., and Paul A. Golder. 2002. “The Diffusion of Anticipatory Standards with Particular Reference to the ISO/IEC Information Resource Dictionary System Framework Standard.” *Computer Standards & Interfaces* 24 (5): 369–79. https://doi.org/10.1016/S0920-5489(02)00057-0.

Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman. 1999. *Readings in Information Visualization: Using Vision to Think*. Book, Whole. San Francisco, Calif: Morgan Kaufmann Publishers.

http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwfV27jsIwEFwduYoKCCc4HkpDCbLs-JEagfgA-igYW6IJBYLvv93E4RFxdNlYWsUjZyexPWMAwVds2aoJimUGucVbbbPC0eKT50dmsTQKbbzlrU2W95O_WiZI1w8rM_idIzj-vHewLteK8WZcEWUK8aBZilP5GOcUK2boBCi8NIrkqVmwg2pi9WJ-WpHPtgcRCRL68OXKAXSfDARjWIQ98JfkVCbBA5WQTm6nC6kla43lECbbzX69W1LqPEzY5KEr_Aei8ly6ESTqoFTlzG4ZPnkhDdNKFqmUzqSFd3oM8bsMY5i93G5gy5HntTbpv-3Yaaw77Pd92gniVJkX0ETEFL49vihuVuMzr5D_A0gMi_E.

Cargill, Carl F. 1989. *Information Technology Standardization: Theory, Process, and Organizations*. Newton, MA, USA: Digital Press.

Caswell, Michelle, and Marika Cifor. 2016. "From Human Rights to Feminist Ethics: Radical Empathy in the Archives." *Archivaria* 81 (1): 23–43.

Certeau, Michel de. 2008. *The Practice of Everyday Life*. Berkeley, Calif.: Univ. of California Press.

"Choosing Characters." n.d. Accessed October 20, 2021. https://www.unicode.org/consortium/choosing.html.

"CiteSeerX — Discovery of Grounded Theory." n.d. Accessed January 5, 2021. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.461.6630.

"Cloud Computing: From Scarcity to Abundance | SpringerLink." n.d. Accessed August 19, 2020. https://link-springer-com.pitt.idm.oclc.org/article/10.1007/s10842-014-0188-y.

"CNN - 1963: The Debut of ASCII - July 6, 1999." n.d. Accessed September 11, 2020. http://edition.cnn.com/TECH/computing/9907/06/1963.idg/.

Crampton, Jeremy W. 2001. “Maps as Social Constructions: Power, Communication and Visualization.” *Progress in Human Geography* 25 (2): 235–52. https://doi.org/10.1191/030913201678580494.

Crystal, David. 2006. *Language and the Internet*. 2nd ed. Cambridge, UK ; Cambridge University Press.

———. n.d. “Morpheme.” In *Dictionary of Linguistics and Phonetics*. Accessed June 1, 2021a. https://search-credoreference-com.pitt.idm.oclc.org/content/title/bkdictling?institutionId=1425&tab=entry_view&heading=phoneme&sequence=0.

———. n.d. “Phoneme.” In *Dictionary of Linguistics and Phonetics*. Accessed June 1, 2021b. https://search-credoreference-com.pitt.idm.oclc.org/content/title/bkdictling?institutionId=1425&tab=entry_view&heading=phoneme&sequence=0.

Cunningham, Andrew. 2014. “Unicode 7.0 Introduces 2,834 New Characters, Including 250 Emoji.” Ars Technica. June 16, 2014. https://arstechnica.com/gadgets/2014/06/unicode-7-0-introduces-2834-new-characters-including-250-emoji/.

David, Paul A., and Shane Greenstein. 1990. “The Economics Of Compatibility Standards: An Introduction To Recent Research.” *Economics of Innovation and New Technology* 1 (1–2): 3–41. https://doi.org/10.1080/10438599000000002.

Davis, Katie. 2012. “Tensions of Identity in a Networked Era: Young People’s Perspectives on the Risks and Rewards of Online Self-Expression.” *New Media & Society* 14 (4): 634–51. https://doi.org/10.1177/1461444811422430.

Davis, Mark. n.d. “The Rapid Evolution of a Wordless Tongue.” Accessed November 27, 2021. https://www.unicode.org/mail-arch/unicode-ml/y2014-m11/0064.html#replies.

Dervin, Brenda. 1999. “On Studying Information Seeking Methodologically: The Implications of Connecting Metatheory to Method.” *Information Processing & Management* 35 (6): 727–50. https://doi.org/10.1016/S0306-4573(99)00023-0.

Diana. n.d. “About Adopt a Character.” *Unicode* (blog). Accessed June 10, 2021a. https://home.unicode.org/adopt-a-character/about-adopt-a-character/.

———. n.d. “Overview.” *Unicode* (blog). Accessed October 18, 2021b. https://home.unicode.org/basic-info/overview/.

Dietz, Thomas, Elinor Ostrom, and Paul C Stern. 2003. “The Struggle to Govern the Commons.” *Science* 302 (5652): 1907–12.

Dodge, Martin, Rob Kitchin, and C. R. Perkins. 2011. *The Map Reader: Theories of Mapping Practice and Cartographic Representation*. Book, Whole. Chichester, West Sussex, UK;Hoboken, NJ; Wiley-Blackwell. http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwdR1JCsIwcHA5KCioVdwq_UCl2Rp7FsUHeJekTfBiEdGDvzdJW8FSIZdsQzJZZsnMBIDgbRTW7gSeJVxrxnSG0pRLzAUTiSBW3IjTRNCakeX3569aEKTX_5eZHYkZI7QNbUJJ4TH-1bVElJszHzuHdG6FQdO4Cr5T5pOf4KaOuBxH0LEOB2NoqXwCvfJX8uvbg6FZw-Am7sHD2RtPYXU8nPen0EK4lHqXSzkiPIOBsNbq-dN5tWVzCDAlWklJpGXnMVI7wzaksVSCGn5JI7YArwnUAvyf4goNpt7QoQgtm7utoF9oQm1aQ1ebja38Yr4bh6kPe_J0UQ.

Drabinski, Emily. 2013. “Queering the Catalog: Queer Theory and the Politics of Correction.” *The Library Quarterly* 83 (2): 94–111. https://doi.org/10.1086/669547.

———. n.d. “Standard Practice: Libraries as Structuring Machines.” *Parameters* (blog). Accessed February 8, 2018. http://parameters.ssrc.org/2017/07/standard-practice-libraries-as-structuring-machines/.

Dvorak, John. 1992. “Kiss Your ASCII Goodbye.” *PC Magazine*, 1992. https://unicode.org/announcements/kissascii.pdf.

“Easy Principles of Computer Character Sets.” 2018. April 2, 2018. https://web.archive.org/web/20180402221037/http://www.bobbemer.com/CODESETS.HTM.

Edwards, Paul N. 2003. “Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems.” *Modernity and Technology* 1: 185–226.

“Emoji, Emoji, What for Art Thou? | Lebduska | Harlot: A Revealing Look at the Arts of Persuasion.” n.d. Accessed April 8, 2021. http://harlotofthearts.org/index.php/harlot/article/view/186/157.

Erickson, Seth. 2021. “Plain Text & Character Encoding: A Primer for Data Curators.” *Journal of EScience Librarianship* 10 (3). https://doi.org/10.7191/jeslib.2021.1211.

“Facet Analysis | Theory of Ranganathan.” n.d. Encyclopedia Britannica. Accessed October 12, 2021. https://www.britannica.com/science/facet-analysis.

“FAQ - Basic Questions.” n.d. Accessed November 29, 2021. https://unicode.org/faq/basic_q.html.

"FAQ - Emoji & Pictographs." n.d. Accessed November 19, 2018. https://www.unicode.org/faq/emoji_dingbats.html.

"FAQ - UTF-8, UTF-16, UTF-32 & BOM." n.d. Accessed November 16, 2018. https://unicode.org/faq/utf_bom.html.

Farokhmanesh, Megan. 2017. "Is There a More Fundamental Human Question than 'Why Isn't This an Emoji?'" The Verge. May 11, 2017. https://www.theverge.com/2017/5/11/15623432/why-isnt-this-an-emoji-questions-unicode.

Farrell, Joseph, and Garth Saloner. 1986. "Installed Base and Compatibility: Innovation, Product Preannouncements, and Predation." *The American Economic Review*, 940–55.

FEBRUARY 2, Lauren Johnson and 2016. n.d. "Twitter's Branded Emojis Come With a Million-Dollar Commitment." Accessed November 27, 2021. https://www.adweek.com/performance-marketing/twitters-branded-emojis-come-million-dollar-commitment-169327/.

Fennell, Lee Anne. 2011. "Commons, Anticommons, Semicommons." Chapters. Edward Elgar Publishing. https://econpapers.repec.org/bookchap/elgeechap/13202_5f2.htm.

Fischer, Eric. n.d. "The Evolution of Character Codes, 1874-1968," 33.

Flammia, Madelyn, and Carol Saunders. 2007. "Language as Power on the Internet." *Journal of the American Society for Information Science and Technology* 58 (12): 1899–1903. https://doi.org/10.1002/asi.20659.

Floyd, Ingbert R., and Allen H. Renear. 2007. "What Exactly Is an Item in the Digital World?" *Proceedings of the American Society for Information Science and Technology* 44 (1): 1–7. https://doi.org/10.1002/meet.1450440374.

Foray, Dominique. 1994. “Users, Standards and the Economics of Coalitions and Committees.” *Information Economics and Policy*, Special Issue on “The Economics of Standards,” 6 (3): 269–93. https://doi.org/10.1016/0167-6245(94)90005-1.

Ford, Paul. n.d. “What Is Code? If You Don’t Know, You Need to Read This.” *Bloomberg.Com*. Accessed October 4, 2016. http://www.bloomberg.com/graphics/2015-paul-ford-what-is-code/.

“FORMAL OR CRITICAL ANALYSIS.” 2019. Humanities LibreTexts. May 7, 2019. https://human.libretexts.org/Bookshelves/Art/Book%3A_Introduction_to_Art_-_Design_Context_and_Meaning_(Sachant_et_al.)/04%3A_Describing_Art/4.02%3A_FORMAL_OR_CRITICAL_ANALYSIS.

Foskett, Anthony Charles. 1971. “Misogynists All; A Study in Critical Classification.” *Library Resources and Technical Services*.

Frank, Eibe, and Gordon W. Paynter. 2004. “Predicting Library of Congress Classifications from Library of Congress Subject Headings.” *Journal of the American Society for Information Science and Technology* 55 (3): 214–27. https://doi.org/10.1002/asi.10360.

Franklin, Ursula M. 1992. *The Real World of Technology*. Book, Whole. Concord, Ont: House of Anansi. http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwdV3BTsMwDLUYXEA7QAei24p8QnAY2hqnyc6IiQ_gXqVJe0LbYeP_idO06qbtmESJEieOI9vvBUDkH8vFyZ1A7K9wip-v1lBFxiklZK20WFpjQi75MMmy__nrhATp73JkhrwhVpJGMBIkesQ465Amf87yvlCsChmZd7pGGpZ9uz5iOg2WZnMP14weICrepvA3YAtMIEsYgzwFSOIiIWKUTsTSFuobVexx7fIKf0-gbk_Dujfh78YKFJx1-Chd6o_wmzz9fP5veDZlNGhU8al5k8wNpwGvz0EuJx7BiRdaVGolbSWTXWxlus6d1IZZyuhmiqFybmhUsiOqjv5lhwWFJKm57vN4LbNXmWPxBxuGq8xddbK7iVswT-dTo-f.

Freedman, Alisa. 2018. “Cultural Literacy in the Empire of Emoji Signs: Who Is Crying with Joy?” *First Monday*, September. https://doi.org/10.5210/fm.v23i9.9395.

———. n.d. “View of Cultural Literacy in the Empire of Emoji Signs: Who Is Crying with Joy? | First Monday.” Accessed April 8, 2021. https://firstmonday.org/article/view/9395/7567.

Freedman, Jenna. 2016. “Can I Quit You, LC?” *Lower East Side Librarian* (blog). March 27, 2016. http://lowereastsidelibrarian.info/lcsh/quityou.

Garcia, Megan. 2016. “Racist in the Machine: The Disturbing Implications of Algorithmic Bias.” *World Policy Journal* 33 (4): 111–17. https://doi.org/10.1215/07402775-3813015.

Gawne, Lauren, and Gretchen McCulloch. 2019. “Emoji as Digital Gestures.” *Language@Internet* 17 (2). https://www.languageatinternet.org/articles/2019/gawne.

Gillam, Richard. 2003. *Unicode Demystified: A Practical Programmer’s Guide to the Encoding Standard*. Boston, MA: Addison-Wesley.

Glaser, April. 2019. “About Dam Time.” *Slate*, May 2, 2019. https://slate.com/technology/2019/05/beaver-emoji-proposal-is-hilarious-and-extremely-correct.html.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. Book, Whole. Garden City, N.Y: Doubleday. http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwfV1Nb8IwDLVGd9ltwBDlQ_JpGgdQSVDTnhGIH7B71cWOhIQ6JE799zhLtgkEXBMnshLHlp28FwCtFtn8yifomp

SxdbksCpKgI2FK5eKe_d9I1im698jyigPpYcaoV8aYPO9AR1KWABm_ICz9CRjbV0g8iKALT9z0IA0YWIzn6IQfkex51oPRH2IE3zHKBeqOtg8z2UQ8_gOEGvx2eOKDw32DLEbYUt3iYe_4Dcbbzed6N_eaVLEmU0Vl1QASyfJ5CJir0pA1dmWtZGg1F0Z_UcaOjKKMWKfQvzVDCtOL5t87K99fisjo9rAxvHhWmVBLmMCzE1vnaViuM_3rgLQ.

Graham, Mark, and William H. Dutton, eds. 2014. *Society and the Internet: How Networks of Information and Communication Are Changing Our Lives*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199661992.001.0001.

Graham, Mark, Scott Hale, and Monica Stephens. 2012. “Digital Divide: The Geography of Internet Access.” *Environment and Planning A* 44 (5): 1009–10.

Grahame, Peter R. 1998. “Ethnography, Institutions, and the Problematic of the Everyday World.” *Human Studies* 21 (4): 347–60.

Granovetter, Mark S. 1977. “The Strength of Weak Ties1.” In *Social Networks*, edited by Samuel Leinhardt, 347–67. Academic Press. https://doi.org/10.1016/B978-0-12-442450-0.50025-0.

Gray, Kellie, and Steve Holmes. 2020. “Tracing Ecologies of Code Literacy and Constraint in Emojis as Multimodal Public Pedagogy.” *Computers and Composition* 55 (March): 102552. https://doi.org/10.1016/j.compcom.2020.102552.

Greenhow, Christine, and Beth Robelia. 2009. “Informal Learning and Identity Formation in Online Social Networks.” *Learning, Media and Technology* 34 (2): 119–40. https://doi.org/10.1080/17439880902923580.

Greenstein, Shane M. 1992. “Invisible Hands and Visible Advisors: An Economic Interpretation of Standardization.” *Journal of the American Society for Information Science* 43 (8): 538–49. https://doi.org/10.1002/(SICI)1097-4571(199209)43:8<538::AID-ASI4>3.0.CO;2-2.

Gunn, Chelsea. n.d. “Unruly Records: Personal Archives, Sociotechnical Infrastructure, and Archival Practice.” Ph.D., United States -- Pennsylvania: University of Pittsburgh. Accessed October 13, 2021. https://www.proquest.com/docview/2499078842/abstract/814FEDCD75D64B85PQ/1.

Hammersley, Martyn. 2018. “What Is Ethnography? Can It Survive? Should It?” *Ethnography and Education* 13 (1): 1–17. https://doi.org/10.1080/17457823.2017.1298458.

Hånberg Alonso, Daniel. n.d. “Emoji Timeline.” Emoji Timeline. Accessed June 7, 2021. https://emojitimeline.com/.

Hardin, Garrett. 1968. “The Tragedy of the Commons.” *Science* 162 (3859): 1243–48.

Haythornthwaite, Caroline. 2002. “Strong, Weak, and Latent Ties and the Impact of New Media.” *The Information Society* 18 (5): 385–401. https://doi.org/10.1080/01972240290108195.

Heilmann, Till A. 2015. “Reciprocal Materiality and the Body of Code. A Close Reading of the American Standard Code for Information Interchange (ASCII).” *Digital Culture & Society* 1 (1): 39–52. https://doi.org/10.25969/mediarep/678.

“Hello World or Καλημέρα Κόσμε or こんにちは 世界.” n.d. Accessed July 29, 2021. http://doc.cat-v.org/plan_9/4th_edition/papers/utf.

Hess, Charlotte, and Elinor Ostrom. 2005. “A Framework for Analyzing the Knowledge Commons : A Chapter from Understanding Knowledge as a Commons: From Theory to Practice.” *Libraries’ and Librarians’ Publications*, January. https://surface.syr.edu/sul/21.

Hicks, Marie. 2018. *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*. History of Computing. Cambridge, MA London, UK: MIT Press.

Highfield, Tim. 2018. “Emoji Hashtags // Hashtag Emoji: Of Platforms, Visual Affect, and Discursive Flexibility.” *First Monday* 23 (9). https://doi.org/10.5210/fm.v23i9.9398.

Hine, Christine. 2020. *Ethnography for the Internet: Embedded, Embodied and Everyday*. Routledge.

“Historical Analysis.” 2006. In *The SAGE Dictionary of Social Research Methods*, by Victor Jupp. 1 Oliver’s Yard, 55 City Road, London England EC1Y 1SP United Kingdom: SAGE Publications, Ltd. https://doi.org/10.4135/9780857020116.n91.

Holsti, OR. 1969. *Content Analysis for the Social Sciences and Humanities*.

“How ASCII Came About.” 2018. April 2, 2018. https://web.archive.org/web/20180402200104/http://www.bobbemer.com/ASCII.HTM.

“How ASCII Got Its Backslash.” 2018. April 2, 2018. https://web.archive.org/web/20180402195951/http://www.bobbemer.com/BACSLASH.HTM.

“How Will Software Clean Up Its Present Mess?” 2018. April 2, 2018. https://web.archive.org/web/20180402200014/http://www.bobbemer.com/SLCLUNCH.HTM.

Hughes, Thomas P. 1987. “The Evolution of Large Technological Systems.” In *The Social Construction of Technological Systems*, edited by W.E Bijker, T.P. Hughes, and T.J. Pinch, 51–82. Cambridge Mass.: MIT Press. http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwtV07T8QwDLa4sgALTx0ciEwsUJQ2udION6FDLEwc86lKXUCC9gQ9fj92H-lDuoGBJapSKW7sqLEtf58BlH8r3cE_IQpRx37oURynyWWPDLnZdwHKWE3TNCwB350iS9t9rJ37V8PTHJmegbR_ML5dlCbomY4AjXQIaBz4x71MbGmo1zfsFAE1SK72eOBPLZo9xg8uCb8umkx7hZbs0JnXmYGyxKLJDDQBIkVYzEkWVJ1l-3TTg2vAFudRkEFOhfbkFbOPfybvpphh5r48j2AURtKBbbox5082kyUDxotz80MrTdfURlY6E7GafLX-7lzei33YY0CHYKQFfcABbGF2CLsdIsYjuCF1CKsOkaeiVIfoqUPU6jiGy4f54v7RrWQtVxVhx9LuSJ2Ak-UZjkEESJGm9NFDbbRSSazNVCZRymw1ARp9CuNNq5xtfjWBndYS5-AUX2u8qGhzfwGSvvAH.

Hunter, Dan. 2003. “Cyberspace as Place and the Tragedy of the Digital Anticommons.” *California Law Review* 91 (2): 439–519. https://doi.org/10.2307/3481336.

“IBM - EBCDIC and the P-Bit.” 2018. May 13, 2018. https://web.archive.org/web/20180513204153/http://www.bobbemer.com/P-BIT.HTM.

Ingarden, Roman. 1986. *The Work of Music and the Problem of Its Identity*. Edited by Jean G. Harrell. Translated by Adam Czerniawski. Book, Whole. Berkeley: University of California Press. http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwfV3BbsIwDLXQuEziMNjQymDKDzA1cdLW5wnEB3Cv0jaROKw7jAt_PyctCFDFMY5kR44Sx479DIDqK13f3Qm

OzVTtdU3OImpvG42VVM4xzVgZMZOukyzPfS_vMJAeeoyoTR5QJUdYXArGI6iQMYpUj7VzGd9gmUZbsn2Bp1BfMIWRa2cwjn2WX0HydomQIyV-vfgJNMEuvuDXmehbvoSJw_FPHLrK2tMbLLab_fduHQSUfRSm7Jan5jCxIXW9PcYSt-YdRMbqcJR5dhxIZ40ib03lKCW0ufQaE5gNcEpgeUM9f1aVSKizIoHV8HRhDMvJF4NMP-BZUtGHGJYw9nwE3KpT1WdU6j88WYWi.

International Typeface Corporation. 1978. “ITC Zapf Dingbats (Signs, Symbols & Ornaments).” *Upper & Lower Case, The International Journal of Typographics* 5 (2): 36–39.

“Interview with Bob Bemer.” 2012. October 26, 2012. https://web.archive.org/web/20121026023854/http://www.bobbemer.com/arranga.htm.

“IP in a World Without Scarcity - University of Pittsburgh.” n.d. Accessed August 19, 2020. https://pitt.primo.exlibrisgroup.com/discovery/fulldisplay?docid=cdi_gale_infotracacademiconefile_A414091276&context=PC&vid=01PITT_INST:01PITT_INST&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central.

“IPhone: About Using the Emoji Keyboard.” 2009. July 3, 2009. https://web.archive.org/web/20090703152413/http://support.apple.com/kb/TS2404.

Jennings, Tom. 2018. “WPS:Projects.” December 7, 2018. https://web.archive.org/web/20181207045913/http://worldpowersystems.com/PROJECTS/computer-numbers.html.

———. n.d. “Wayback Machine.” Accessed September 17, 2020a. https://web.archive.org/web/20160812085410/http://worldpowersystems.com/projects/codes/X3.4-1963/page4.JPG.

———. n.d. “World Power Systems:Texts:Annotated History of Charactercodes.” Accessed September 17, 2020b. https://www.sr-ix.com/Archive/CharCodeHist/index.html.

———. n.d. “World Power Systems:Texts:Annotated History of Charactercodes.” Accessed October 1, 2020c. https://www.sr-ix.com/Archive/CharCodeHist/index.html#SUP4.

John, Nicholas A. 2013. “The Construction of the Multilingual Internet: Unicode, Hebrew, and Globalization.” *Journal of Computer-Mediated Communication* 18 (3): 321–38. https://doi.org/10.1111/jcc4.12015.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47 (2): 263–91. https://doi.org/10.2307/1914185.

Kendall, Lori. 1998. “Meaning and Identity in ‘Cyberspace’: The Performance of Gender, Class, and Race Online.” *Symbolic Interaction* 21 (2): 129–53. https://doi.org/10.1525/si.1998.21.2.129.

Kim, Kyongsok. 1992. “A Future Direction in Standardizing International Character Codes — with a Special Reference to ISO/IEC 10646 and Unicode.” *Computer Standards & Interfaces* 14 (3): 209–21. https://doi.org/10.1016/0920-5489(92)90020-E.

Kindleberger, Charles P. 1983. “Standards as Public, Collective and Private Goods.” *Kyklos* 36 (3): 377–96. https://doi.org/10.1111/j.1467-6435.1983.tb02705.x.

Köhler, Heinz. 1968. *Scarcity Challenged; an Introduction to Economics.* xxviii, 660 p. New York: Holt, Rinehart and Winston. //catalog.hathitrust.org/Record/009908674.

Korn, Jenny Ungbha. 2021. “Connecting Race to Ethics Related to Technology: A Call for Critical Tech Ethics.” *Journal of Social Computing* 2 (4): 357–64. https://doi.org/10.23919/JSC.2021.0026.

Korpela, Jukka K. 2006. *Unicode Explained*. Sebastopol, CA: O'Reilly.

Krechmer, Ken. n.d. "OPEN STANDARDS REQUIREMENTS," 34.

Kwasnik, Barbara H. 1992a. "The Role of Classification Structures in Reflecting and Building Theory." *Advances in Classification Research Online* 3 (1): 63–82.

Kwasnik, Barbara H. 1992b. "The Role of Classification Structures in Reflecting and Building Theory." *Advances in Classification Research Online* 3 (1): 63–82. https://doi.org/10.7152/acro.v3i1.12597.

Lambrecht, Maxime. 2017. "The Time Limit on Copyright: An Unlikely Tragedy of the Intellectual Commons." *European Journal of Law and Economics* 43 (3): 475–94. https://doi.org/10.1007/s10657-016-9538-z.

Lewis, Cynthia, and Bettina Fabos. 2005. "Instant Messaging, Literacies, and Social Identities." *Reading Research Quarterly* 40 (4): 470–501. https://doi.org/10.1598/RRQ.40.4.5.

"Library - The Dewey Decimal System." n.d. Encyclopedia Britannica. Accessed June 13, 2021. https://www.britannica.com/topic/library.

Licklider. 1960. "Man-Computer Symbiosis," 8.

Luckerson, Victor. 2016. "Meet the 63-Year-Old in Charge of Approving New Emojis." Time. 2016. http://time.com/4244795/emoji-consortium-mark-davis/.

Mackenzie, Adrian. 2005. "The Performativity of Code Software and Cultures of Circulation." *Theory, Culture & Society* 22 (1): 71–92. https://doi.org/10.1177/0263276405048436.

Mackenzie, Charles E. 1980. *Coded Character Sets: History and Development*. Reading, Mass: Addison-Wesley Pub. Co.

Manning, Dale T., J. Edward Taylor, and James E. Wilen. 2018. “General Equilibrium Tragedy of the Commons.” *Environmental and Resource Economics* 69 (1): 75–101. https://doi.org/10.1007/s10640-016-0066-7.

Manovich, Lev. 2009. “The Practice of Everyday (Media) Life: From Mass Consumption to Mass Cultural Production?” *Critical Inquiry* 35 (2): 319–31. https://doi.org/10.1086/596645.

Mansourian, Yazdan. 2006. “Adoption of Grounded Theory in LIS Research.” *New Library World* 107 (9/10): 386–402. https://doi.org/10.1108/03074800610702589.

Marino, Mark. 2018. “Critical Code Studies | Electronic Book Review.” August 2, 2018. https://web.archive.org/web/20180802025238/http://www.electronicbookreview.com:80/thread/electropoetics/codology.

Marwick, Alice E. n.d. “The Public Domain: Social Surveillance in Everyday Life.” *Surveillance & Society*. Accessed January 31, 2018. https://search.proquest.com/openview/44c209545c4b499f790342486af023b9/1?pq-origsite=gscholar&cbl=396354.

Marzluff, John, Eric Shulenberger, Wilfried Endlicher, Marina Alberti, Gordon Bradley, Clare Ryan, Craig ZumBrunnen, Ute Simon, and John Marzluff. 2008. *Urban Ecology: An International Perspective on the Interaction Between Humans and Nature : An International Perspective on the Interaction Between Humans and Nature*. New York, NY, UNITED STATES: Springer. http://ebookcentral.proquest.com/lib/pitt-ebooks/detail.action?docID=337050.

"Masterpiece Engineering." 2018. April 2, 2018. https://web.archive.org/web/20180402221131/http://www.bobbemer.com/666.HTM.

McEnery, A. M., and R. Z. Xiao. 2005. "Character Encoding in Corpus Construction." In , edited by M. Wynne. Oxford, UK: AHDS. https://eprints.lancs.ac.uk/id/eprint/60/.

McPherson, Tara. 2013. "U.S. Operating Systems at Mid-Century: The Intertwining of Race and UNIX." In *Race after the Internet*, edited by Lisa Nakamura and Peter Chow-White. Routledge. https://books.google.com/books?id=aOiSAgAAQBAJ&lpg=PA21&ots=Pxxotr3cjB&dq=Tara%20McPherson%2C%20%E2%80%9CU.S.%20Operating%20Systems%20at%20Mid-Century%3A%20The%20Intertwining%20of%20Race%20and%20UNIX%2C%E2%80%9D%20in%20Race%20After%20the%20Internet%20(New%20York%3A%20Routledge%2C%202012)%2C%2021-37.&lr&pg=PA21#v=onepage&q&f=false.

Mehta, Lyla. 2010. *The Limits to Scarcity : Contesting the Politics of Allocation*. Florence, UNITED KINGDOM: Taylor & Francis Group. http://ebookcentral.proquest.com/lib/pitt-ebooks/detail.action?docID=1144665.

Miller, H., Daniel Kluver, Jacob Thebault-Spieker, L. Terveen, and Brent J. Hecht. 2017. "Understanding Emoji Ambiguity in Context: The Role of Text in Emoji-Related Miscommunication." In *ICWSM*.

Milner, Ryan M. 2013. "Hacking the Social: Internet Memes, Identity Antagonism, and the Logic of Lulz. | The Fibreculture Journal : 22." *The Fibreculture Journal*, no. 22. http://twentytwo.fibreculturejournal.org/fcj-156-hacking-the-social-internet-memes-identity-antagonism-and-the-logic-of-lulz/.

Miltner, Kate M. 2021. "'One Part Politics, One Part Technology, One Part History': Racial Representation in the Unicode 7.0 Emoji Set." *New Media & Society* 23 (3): 515–34. https://doi.org/10.1177/1461444819899623.

Mitchell, Joan S. 2001. "Relationships in the Dewey Decimal Classification System." In *Relationships in the Organization of Knowledge*, edited by Carol A. Bean and Rebecca Green, 211–26. Information Science and Knowledge Management. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-9696-1_14.

MLIS, Steven A. Knowlton. 2005. "Three Decades Since Prejudices and Antipathies: A Study of Changes in the Library of Congress Subject Headings." *Cataloging & Classification Quarterly* 40 (2): 123–45. https://doi.org/10.1300/J104v40n02_08.

Morrissey, Sheila M. 2011. ""More What You'd Call 'Guidelines' Than Actual Rules"[1]: Variation in the Use of Standards." *Journal of Electronic Publishing* 14 (1). http://dx.doi.org/10.3998/3336451.0014.104.

"Moving to Unicode 5.1." n.d. *Official Google Blog* (blog). Accessed September 18, 2018. https://googleblog.blogspot.com/2008/05/moving-to-unicode-51.html.

Mueller, Milton. 2010. "Critical Resource: An Institutional Economics of the Internet Addressing-Routing Space." *Telecommunications Policy* 34 (8): 405–16.

Mueller, Milton L. 2006. "IP Addressing: The next Frontier of Internet Governance Debate." *Info* 8 (5): 3–12. https://doi.org/10.1108/14636690610688051.

Mueller, Milton L. 2008. "Scarcity in IP Addresses: IPv4 Address Transfer Markets and the Regional Internet Address Registries." *Internet Governance Project* 20.

Mukerjee, Aditya. n.d. “I Can Text You A Pile of Poo, But I Can’t Write My Name.” *Model View Culture* (blog). Accessed October 10, 2017. https://modelviewculture.com/pieces/i-can-text-you-a-pile-of-poo-but-i-cant-write-my-name.

Mullaney, Thomas S. 2017. *The Chinese Typewriter: A History*. MIT Press.

Nichols, David M., Ian H. Witten, Te Taka Keegan, David Bainbridge, and Michael Dewsnip. 2005. “Digital Libraries and Minority Languages.” *New Review of Hypermedia and Multimedia* 11 (2): 139–55. https://doi.org/10.1080/13614560500351071.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Novak, Petra Kralj, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. “Sentiment of Emojis.” *PLOS ONE* 10 (12): e0144296. https://doi.org/10.1371/journal.pone.0144296.

“Oakland A’s Sponsor Baseball Emoji.” n.d. MLB.Com. Accessed June 21, 2021. https://www.mlb.com/press-release/oakland-a-s-sponsor-baseball-emoji-209950970.

“OCLC Research Activities and IFLA’s Functional Requirements for Bibliographic Records.” 2018. June 21, 2018. https://www.oclc.org/research/activities/frbr.html.

“OECD Glossary of Statistical Terms - Controlled Vocabulary Definition.” n.d. Accessed October 18, 2018. https://stats.oecd.org/glossary/detail.asp?ID=6260.

Olson, Hope A. 2001. “Sameness and Difference.” *Library Resources & Technical Services* 45 (3): 115–22. https://doi.org/10.5860/lrts.45n3.115.

Osborn, Don. 2010. *African Languages in a Digital Age: Challenges and Opportunities for Indigenous Language Computing*. Cape Town: HSRC Press.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Canto Classics. Cambridge, United Kingdom: Cambridge Univ Press.

Otlet, Paul. 1934. *Traité de Documentation: Le Livre Sur Le Livre, Théorie et Pratique*. Editiones mundaneum.

———. 1990. *International Organisation and Dissemination of Knowledge: Selected Essays of Paul Otlet*. Edited by W. B. Rayward. Elsevier for the International Federation of Documentation. https://www.ideals.illinois.edu/handle/2142/4004.

Painter, Deryc T., Bryan C. Daniels, and Jürgen Jost. 2019. "Network Analysis for the Digital Humanities: Principles, Problems, Extensions." *Isis* 110 (3): 538–54. https://doi.org/10.1086/705532.

Papacharissi, Zizi. 2002. "The Presentation of Self in Virtual Life: Characteristics of Personal Home Pages." *Journalism & Mass Communication Quarterly* 79 (3): 643–60. https://doi.org/10.1177/107769900207900307.

———. 2010. *A Networked Self: Identity, Community, and Culture on Social Network Sites*. Routledge.

Patton, Tracey Owens. 2020. "Visual Rhetoric: Theory, Method, and Application in the Modern World." In *Handbook of Visual Communication*. Routledge.

Pike, Rob. 2003a. "UTF History." 2003. https://www.cl.cam.ac.uk/~mgk25/ucs/utf-8-history.txt.

———. 2003b. "UTF-8 History," April 30, 2003. https://www.cl.cam.ac.uk/~mgk25/ucs/utf-8-history.txt.

Poirier, Lindsay. 2017. "Devious Design: Digital Infrastructure Challenges for Experimental Ethnography." *Design Issues* 33 (2): 70–83. https://doi.org/10.1162/DESI_a_00440.

Pope, Michael Brian, Merrill Warkentin, Leigh A Mutchler, and Xin Robert Luo. 2012. "The Domain Name System—Past, Present, and Future." *Communications of the Association for Information Systems* 30 (1): 21.

Powell, Ronald R. 1999. "Recent Trends in Research: A Methodological Essay." *Library & Information Science Research* 21 (1): 91–119.

Puck. 1881. "Typographical Art," March 30, 1881.

Raley. n.d. "Code.Surface || Code.Depth." Accessed November 21, 2016. http://www.dichtung-digital.org/2006/01/Raley/index.htm#_ednref16.

Ranganathan, S. R. (Shiyali Ramamrita). 1951. *Classification and Communication*. Delhi University Publications. Library Science Series ; 3. Delhi: University of Delhi.

Ranganathan, Shiyali Ramamrita. 1950. *Classification, Coding and Machinery for Search*. Unesco.

Reilly, E.D., A. Ralston, and D. Hemmendinger, eds. 2003. "Baudot Code." In *Encyclopedia of Computer Science*, 4th ed. Credo Reference: Wiley. http://pitt.idm.oclc.org/login?url=https://search.credoreference.com/content/entry/encyccs/baudot_code/0?institutionId=1425.

Renear, Allen H, Elli Mylonas, and David Durand. 1993. "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies."

Reuters. 2019. "As Wildfire Rages, LA's 'fire Proof' Getty Museum Sees No Risk to Art," October 29, 2019, sec. Environment. https://www.reuters.com/article/us-california-wildfire-gettycenter-idUSKBN1X82P2.

"Rfc1591." n.d. Accessed November 21, 2021. https://datatracker.ietf.org/doc/html/rfc1591.

Ricoeur, Paul. 1981. “What Is a Text? Explanation and Understanding.” *Hermeneutics and the Human Sciences: Essays on Language, Action and Interpretation*.

Riordan, Monica A. 2017. “Emojis as Tools for Emotion Work: Communicating Affect in Text Messages.” *Journal of Language and Social Psychology* 36 (5): 549–67. https://doi.org/10.1177/0261927X17704238.

Riva, Massimo, and Vika Zafrin. 2005. “Extending the Text: Digital Editions and the Hypertextual Paradigm.” In *Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia*, 205–7. HYPERTEXT ’05. New York, NY, USA: ACM. https://doi.org/10.1145/1083356.1083396.

Roberts. n.d. “Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media’s Waste – Wi Journal.” Accessed April 11, 2018. http://wi.mobilities.ca/digitalrefuse/.

Robinson, Andrew. 2002. *Lost Languages : The Enigma of the World’s Undeciphered Scripts*. New York: McGraw-Hill.

Rubin, Richard. 2010. *Foundations of Library and Information Science*. 3rd ed. Book, Whole. New York: Neal-Schuman Publishers. http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwfV09D8IgEL0YXdz8jLXWMLlpsFB6zsbGH-DeQIHERQf_fyJUGm3TOEJyQC58vYP3DoClB7rv7AncMzhlrlEISbWpND_iyQjlplNFqzr3RN8ny44G0l_EKFLq8LND7wOGX8Z45iWmHALEIO_UlEVLzLQ-TIoJDD3BYAoD85hBEogDZEcCM8h7ioQlN4fDN-_RizwtCVEXIh-a3H8Mwkm2gLi43M7Xve-

1DLGZMgw6XcLQoX2zAqI0WsuZwqOQXCKqXFmeC55n7m4mkUcw72shgqRV3bxdlYJ5hRm67jeLYfx5DvcxhQ2MrJvzJvm4Zls78Q3Rz3-3.

Scheffer, Thomas. 2007. “Event and Process: An Exercise in Analytical Ethnography.” *Human Studies* 30 (3): 167–97.

Scherer, Markus. n.d. “Unicode Mail List Archive: Emoji: Public Review December 2008.” Accessed November 27, 2021. https://unicode.org/mail-arch/unicode-ml/y2008-m12/0063.html.

Schoechle, T. 2003. “Digital Enclosure: The Privatization of Standards and Standardization.” In *ESSDERC 2003. Proceedings of the 33rd European Solid-State Device Research - ESSDERC ’03 (IEEE Cat. No. 03EX704)*, 229–40. https://doi.org/10.1109/SIIT.2003.1251210.

Schuster, Kristen, and Stuart Dunn, eds. 2020. *Routledge International Handbook of Research Methods in Digital Humanities*. London: Routledge. https://doi.org/10.4324/9780429777028.

Schwedel, Heather. 2018. “Actually, We Don’t Really Need a White-Wine Emoji.” Slate Magazine. August 2, 2018. https://slate.com/technology/2018/08/we-dont-need-a-white-wine-emoji.html.

Shannon, CE, and W Weaver. 1949. “A Mathematical Model of Communication.” *)(IL: University of Illinois Press, 1949).*

Shapard, Jeffrey. 1993. “Islands in the (Data)Stream: Language, Character Codes, and Electronic Isolation in Japan.” In *Global Networks: Computers and International Communication*, edited by Linda M Harasim.

https://symbiose.uqo.ca/apps/LoginSigparb/LoginPourRessources.aspx?url=http://ieeexplore.ieee.org/servlet/opac?bknumber=6267399.

Smith, Barbara Herrnstein. 2016. “What Was ‘Close Reading’?: A Century of Method in Literary Studies.” *The Minnesota Review* 2016 (87): 57–75. https://doi.org/10.1215/00265667-3630844.

Smith, Fred W. 1964. “New American Standard Code for Information Exchange.” *Western Union Technical Review*, April. https://www.sr-ix.com/Archive/CharCodeHist/New-ASCII/index.html.

Smith, H. J., and F. A. Williams. 1960. “Survey of Punched Card Codes.” *Communications of the ACM* 3 (12): 639. https://doi.org/10.1145/367487.367491.

SPERBERG-McQUEEN, C. M. 1991. “Text in the Electronic Age: Texual Study and Textual Study and Text Encoding, with Examples from Medieval Texts.” *Literary and Linguistic Computing* 6 (1): 34–46. https://doi.org/10.1093/llc/6.1.34.

Standage, Tom. 1998. *The Victorian Internet: The Remarkable Story of the Telegraph and the Nineteenth Century’s Onlline Pioneers*. London: Weidenfeld & Nicolson.

Star, Susan Leigh. 1990. “Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions.” *The Sociological Review* 38 (1_suppl): 26–56. https://doi.org/10.1111/j.1467-954X.1990.tb03347.x.

———. 1998. “Grounded Classification: Grounded Theory and Faceted Classification.” https://www.ideals.illinois.edu/handle/2142/8215.

STAR, SUSAN LEIGH. 1999. “The Ethnography of Infrastructure.” *American Behavioral Scientist* 43 (3): 377–91. https://doi.org/10.1177/00027649921955326.

Star, Susan Leigh, and Karen Ruhleder. 1994. “Steps towards an Ecology of Infrastructure: Complex Problems in Design and Access for Large-Scale Collaborative Systems.” In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 253–64. CSCW ’94. Chapel Hill, North Carolina, USA: Association for Computing Machinery. https://doi.org/10.1145/192844.193021.

Stark, Luke, and Kate Crawford. 2015. “The Conservatism of Emoji: Work, Affect, and Communication.” *Social Media + Society* 1 (2): 2056305115604853. https://doi.org/10.1177/2056305115604853.

Stemler, Steve. 2019. “An Overview of Content Analysis.” *Practical Assessment, Research, and Evaluation* 7 (1). https://doi.org/10.7275/z6fm-2e34.

Stone, A. 2003. “Internationalizing the Internet.” *IEEE Internet Computing* 7 (3): 11–12. https://doi.org/10.1109/MIC.2003.1200294.

Strauss, Anselm, and Juliet Corbin. 1994. “Grounded Theory Methodology: An Overview.” In *Handbook of Qualitative Research*, 273–85. Thousand Oaks, CA, US: Sage Publications, Inc.

Strauss, Anselm L. 1998. *Basics of Qualitative Research : Techniques and Procedures for Developing Grounded Theory*. 2nd ed. Thousand Oaks: Sage Publications.

Sweeney, Miriam E., and Kelsea Whaley. 2019. “Technically White: Emoji Skin-Tone Modifiers as American Technoculture.” *First Monday* 24 (7). https://doi.org/10.5210/fm.v24i7.10060.

“SwiftKey Emoji Report | The United States | Data.” n.d. Accessed March 22, 2019. https://www.scribd.com/doc/262594751/SwiftKey-Emoji-Report.

Taube, Mortimer. 1961. *Computers and Common Sense. The Myth of Thinking Machines*. Columbia university press.

Taylor, Greg. n.d. *Scarcity of Attention for a Medium of Abundance: An Economic Perspective*. *Society and the Internet*. Oxford University Press. Accessed August 21, 2020. https://oxford-universitypressscholarship-com.pitt.idm.oclc.org/view/10.1093/oso/9780198843498.001.0001/oso-9780198843498-chapter-19.

Taylor, Hugh A. 1991. "Chip Monks at the Gate: The Impact of Technology on Archives, Libraries and the User." *Archivaria* 33 (0). https://archivaria.ca/archivar/index.php/archivaria/article/view/11808.

"The ALA Character Set and Other Solutions for Processing the World's Information." 1989. *Library Technology Reports* 25 (2): 255–73.

"The Cyrillic Charset Soup." 2007. January 21, 2007. https://web.archive.org/web/20070121043016/http://czyborra.com/charsets/cyrillic.html.

"The Library Commons: An Imagination and an Invocation – In the Library with the Lead Pipe." n.d. Accessed February 24, 2022. https://www.inthelibrarywiththeleadpipe.org/2020/the-library-commons/.

Timmermans, Stefan, and Marc Berg. 2016. "Standardization in Action: Achieving Local Universality through Medical Protocols." *Social Studies of Science* 27 (2): 273–305. https://doi.org/10.1177/030631297027002003.

Timmermans, Stefan, and Iddo Tavory. 2012. "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis." *Sociological Theory* 30 (3): 167–86.

Trashi, N., L. Z. Shu, Q. Nuo, P. Dun, C. Z. Yang, and Y. Tso. n.d. “A Tibetan Mobile Phone Based on CDMA System.” In . ACTA Press. Accessed February 18, 2020. http://www.actapress.com/PaperInfo.aspx?PaperID=30325.

Turnbull, Michelle, Paul Ricciardi, Matthew Forman, Maria Rosario, Monica Walker, and Andrew Wilder. 2021. “HUM1: Modern Humanities: Arts & Ideas.” *Open Educational Resources*, January. https://academicworks.cuny.edu/kb_oers/21.

Underwood, Ted. 2017. “A Genealogy of Distant Reading.” *Digital Humanities Quarterly* 011 (2).

“Unicode – The World Standard for Text and Emoji.” 2019. November 30, 2019. https://web.archive.org/web/20191130223544/http://home.unicode.org/.

“Unicode 11.0.0.” n.d. Accessed November 19, 2018. https://www.unicode.org/versions/Unicode11.0.0/.

“Unicode 11.0.0 Final Names List.” n.d. Accessed October 18, 2018. https://www.unicode.org/Public/UCD/latest/ucd/NamesList.txt.

“Unicode Character Encoding Stability Policies.” n.d. Accessed October 18, 2018. https://www.unicode.org/policies/stability_policy.html.

“Unicode Consortium.” n.d. Accessed November 16, 2018. https://www.unicode.org/consortium/consort.html.

“Unicode Home Page.” 1998. January 26, 1998. https://web.archive.org/web/19980126155923/http://www.unicode.org/.

“Unicode Mail List Archive: Unicode Web Page Now Available!” n.d. Accessed November 7, 2021a. https://www.unicode.org/mail-arch/unicode-ml/Archives-Old/UML004/0045.html.

"———." n.d. Accessed November 15, 2021b. https://www.unicode.org/mail-arch/unicode-ml/Archives-Old/UML004/0045.html.

"Unicode Statistics." n.d. Accessed October 20, 2021. https://www.unicode.org/versions/stats/.

"Usage Survey of Character Encodings Broken down by Ranking." n.d. Accessed November 16, 2018. https://w3techs.com/technologies/cross/character_encoding/ranking.

Üstün, Ahmet, Murathan Kurfalı, and Burcu Can. 2018. *Characters or Morphemes: How to Represent Words?* Association for Computational Linguistics. https://doi.org/10.18653/v1/w18-3019.

"UTF-8 Encoding." n.d. Accessed June 20, 2021. https://www.fileformat.info/info/unicode/utf8.htm.

"UTR#17: Unicode Character Encoding Model." n.d. Accessed September 24, 2018. https://www.unicode.org/reports/tr17/.

"UTS #10: Unicode Collation Algorithm." n.d. Accessed October 18, 2018. https://www.unicode.org/reports/tr10/.

"UTS #51: Unicode Emoji." n.d. Accessed June 7, 2021. https://www.unicode.org/reports/tr51/#Introduction.

Van Dijck, José. 2013. *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press.

Veinot, Tiffany C, and Kate Williams. 2012. "Following the 'Community' Thread from Sociology to Information Behavior and Informatics: Uncovering Theoretical Continuities and Research Opportunities." *Journal of the American Society for Information Science and Technology* 63 (5): 847–64.

Veness, Alex, and Richard Barbrook. 2007. “Imaginary Futures: From Thinking Machines to the Global Village.”

Wajcman, Judy. 2002. “Addressing Technological Change: The Challenge to Social Theory.” *Current Sociology* 50 (3): 347–63. https://doi.org/10.1177/0011392102050003004.

Warner, M. 2002. *Publics and Counterpublics*. Zone Books. https://books.google.com/books?id=fZfaAAAAMAAJ.

“Wayback Machine.” n.d. Http://Www.Stonehand.Com/Unicode.Html. Accessed November 27, 2021. https://web.archive.org/web/19970301000000*/http://www.stonehand.com/unicode.html.

Weaver, Warren. 1955. “Translation.” In *Machine Translation of Languages: Fourteen Essays*, edited by William Nash Locke and Andrew Donald Booth. Published jointly by Technology Press of the Massachusetts Institute of Technology and Wiley, New York.

Weiss, Martin B. H. 1991. “The Standards Development Process: A View from Political Theory.” Monograph. June 1991. http://d-scholarship.pitt.edu/18250/.

Weiss, Martin B. H., and Marvin Sirbu. 1989. “Technological Choice in Voluntary Standards Committees: An Empirical Analysis.” School of Library and Information Science, University of Pittsburgh. http://pitt.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2AwNtIz0EUrE0wsTZLSQJNaoNM-0kCJyDAlJc0iNTnNxDA52SwFbZEl7P67FF1YFy8jswB2ElKpPjAZWlgyM7AaGZubglK1RaQxytmk4LrBTZCBNSCxILVIiIEpNU-EwRc-Vg3yvYJzRj4wJypk5imEAeM4rySxqFIhGNp7L1YA7c4AGpeaWmyl4Jin4JpbkAksUMBdlSIKIOlm2uIs4cuyNb4VCRnxoMObE6JRxaBqEkpjQc73EiMgTcRtHw9rwS8zS1FgkHBMM08xTwp1dQgMc3AxDDNCNRqSEtNMjVMS0tKS0xOk2QwIt0iKXI0STNwQZZRgVZsyDCwpgFzQaosJHTlwEENALvMlec.

Weiss, Martin, and Carl Cargill. 1992. “Consortia in the Standards Development Process.” *Journal of the American Society for Information Science* 43 (8): 559–65. https://doi.org/10.1002/(SICI)1097-4571(199209)43:8<559::AID-ASI7>3.0.CO;2-P.

“What Is Unicode?” n.d. Accessed March 22, 2019. https://www.unicode.org/standard/WhatIsUnicode.html.

Wilson, Samuel M, and Leighton C Peterson. 2002. “The Anthropology of Online Communities.” *Annual Review of Anthropology* 31 (1): 449–67.

“Wireless Application Protocol: WAP Pictogram Specification.” 2001. Wireless Application Protocol Forum.

Wolske, Martin, and Colin Rhinesmith. 2016. “Critical Questions for Community Informatics in Practice.” *The Journal of Community Informatics* 12 (3). https://doi.org/10.15353/joci.v12i3.3289.

Wright, Alex. 2014. *Cataloging the World : Paul Otlet and the Birth of the Information Age*. Oxford: Oxford University Press.

“WWW-Talk Jan-Mar 1993: Re: Telecom Digest Archives.” n.d. Accessed November 10, 2020. http://1997.webhistory.org/www.lists/www-talk.1993q1/0241.html.