Unicode, Text Search and You

Article contributed
by dtSearch®

Designed to encompass the world's languages, Unicode is a product of the Unicode Consortium, Unicode.org. Unicode text is what fills your files and emails.

Unicode mappings cover English as well as other Western European, Eastern European and Nordic languages. Separate mappings cover double-byte characters like Chinese, Japanese and Korean. Still other mappings encompass right to left languages like Arabic and Hebrew. Even historical languages like Egyptian hieroglyphics have Unicode mappings. Along with international languages, Unicode covers a wide range of numbers and symbols as well as the world's emojis 😊

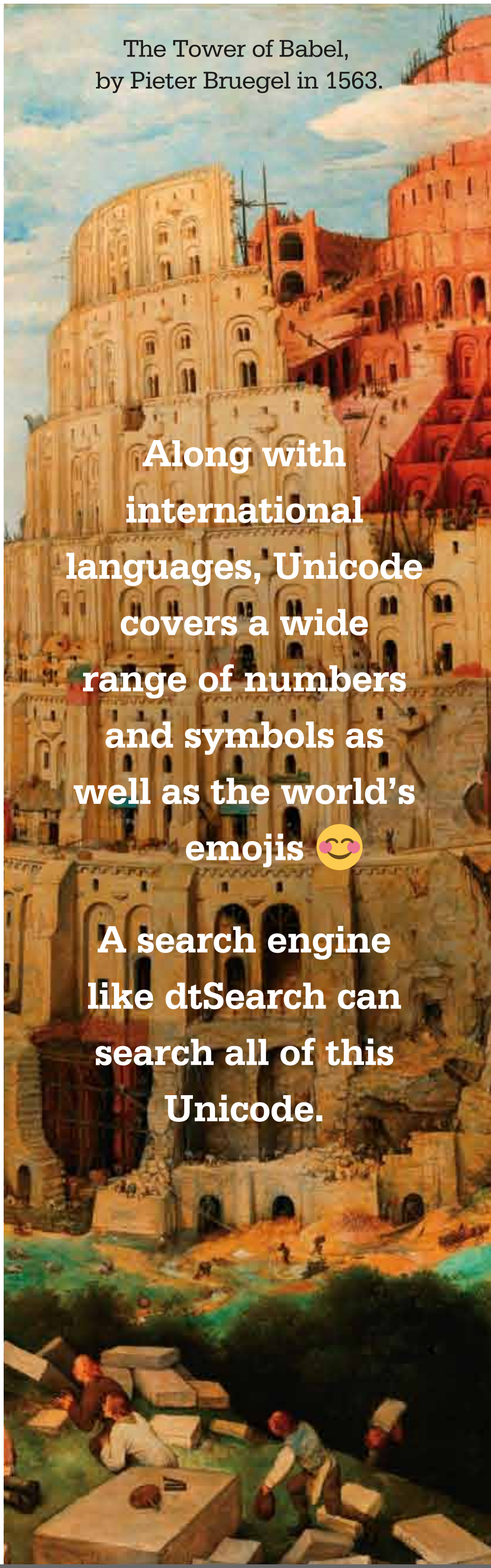
A search engine like dtSearch can search all of this Unicode. For background, dtSearch has enterprise and developer products that run "on premises" or on cloud platforms to instantly search terabytes of "Office" files, PDFs, emails along with nested attachments, databases and online data. Because dtSearch instantly search terabytes, many dtSearch customers are Fortune 100 companies and government agencies. But anyone with lots of data to search can go to dtSearch.com and download a fully-functional 30-day evaluation copy.

A search engine instantly searches terabytes of Unicode after first building one or more terabyte-size search indexes. Indexing takes no effort at all for you the end-user. Just point to all of the file directories, email repositories and the like you want to index, and the search engine will do the rest.

The search engine approaches each file, email, etc. in its binary format. The binary format view of a file or email is very different from how you would normally see that file or email inside its native application. The whole point of looking at an email in Outlook is that you can easily read it. But in its binary format, the same email might look like complete gibberish to the naked eye.

The search engine has to sift through the mass of binary codes to find the relevant Unicode text to index. In doing so, the search engine first has to figure out the correct file parsing specification to apply. You might think that the file extension would work for this purpose. So, if a file ends in .PDF, you might think: PDF file.

But what if someone renames a PowerPoint with a PDF extension? Just looking at the file extension would result in a completely wrong parsing specification and likely miss the contents of the entire file. So a search engine needs a more foolproof method for determining file type. Instead of looking to the file extension, dtSearch looks inside the binary file itself to figure out the relevant file type and then uses that to apply the correct parsing specification.



The Tower of Babel,
by Pieter Bruegel in 1563.

Along with international languages, Unicode covers a wide range of numbers and symbols as well as the world's emojis 😊

A search engine like dtSearch can search all of this Unicode.

Reprinted with permission of USA Daily Chronicles, usadailychronicles.com

After applying the correct parsing specification, the search engine can obtain the full-text and metadata Unicode information it needs for its search index. A search index is not like a “back of book” index. Rather, it is strictly an internal tool enabling either a single end-user or multiple concurrent users to instantly sift through terabytes using over 25 different search features.

dtSearch search features enabled by indexing include Boolean and/or/not Unicode searches, proximity Unicode searches, fuzzy matching Unicode searches and numeric range Unicode searches. The index also supports advanced search options like the ability to identify credit card numbers in Unicode text and developer-enabled faceted or drill-down metadata search and other data classification. And all of this works regardless of the underlying international language or potentially multiple different languages in each file.

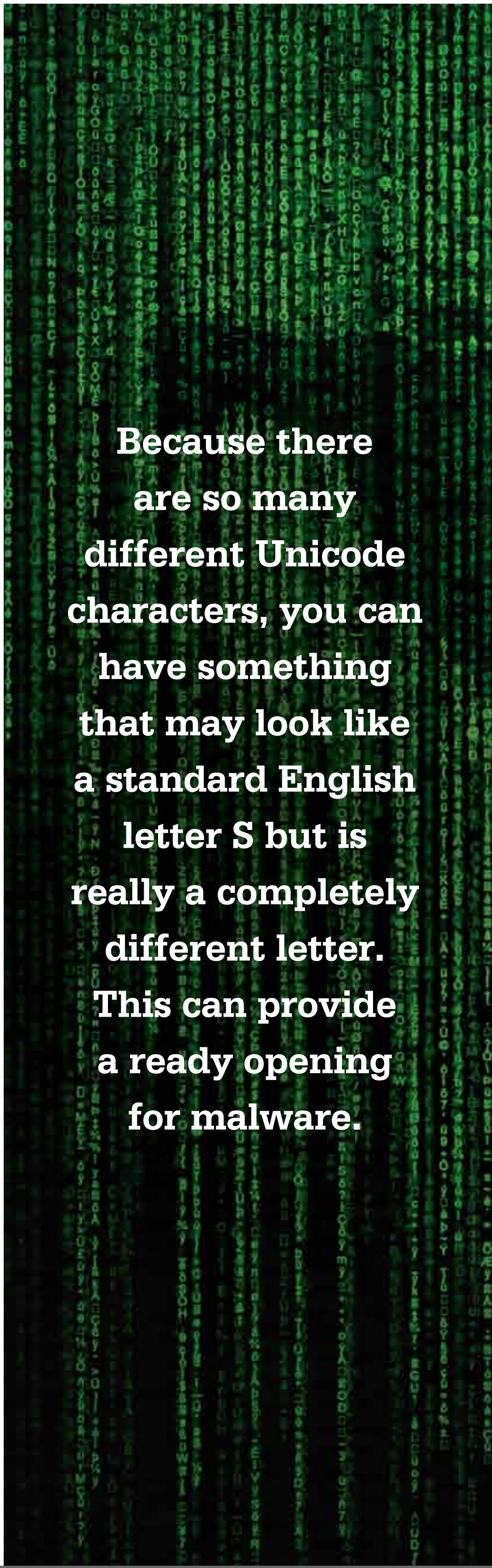
But along with search avenues opened up by Unicode, there is a major caution. Because there are so many different Unicode characters, you can have something that may look like a standard English letter S but is really a completely different letter. This can provide a ready opening for malware. Suppose you have an email promising a 30% off sale with a tempting link to what looks like a “name brand” website. If even one letter of the link—S or something else—represents a different Unicode mapping, the link might go to a very different URL than you would expect.

The easy answer is to ignore the email with the link. But suppose you really want to check out the sale. In that case, the safe way to proceed is to type the “name brand” web address directly into your own web browser. Copy and paste will just repeat the same misleading Unicode. But if you retype from scratch where you want to go, you are in charge of the destination.

Another Unicode caution: sometimes when you are working with PDFs, for example, you see what may look like standard Unicode. But when you try to copy and paste the text, nothing happens. In that case, you are probably working with an “image only” PDF. You need to apply some type of OCR application like Adobe Acrobat to turn that into Unicode that you can work with and that a search engine can search. dtSearch has an option to bulk identify image-only PDFs. And then you can send them all through an OCR program like Adobe Acrobat to turn them into regular Unicode PDFs.

Please go to dtSearch.com and download a fully-functional evaluation version to instantly search terabytes of your own Unicode.

Article contributed
by dtSearch®



Because there are so many different Unicode characters, you can have something that may look like a standard English letter S but is really a completely different letter. This can provide a ready opening for malware.

Reprinted with permission of USA Daily Chronicles, usadailychronicles.com