

## XSEARCH<sup>TM</sup> BIETET VOLLSTÄNDIGE UNICODE UNTERSTÜTZUNG

### MIT ASCII FIND ALLES AN

Zu Beginn des Computerzeitalters war der ASCII Zeichensatz die einzige Möglichkeit, um jedes Zeichen jeder Sprache darzustellen. Der ASCII Zeichensatz umfasst 256 Zeichen. Hiervon sind die ersten 128 Zeichen reserviert für das Englische inklusive Punctuation, Zahlen und einiger Sonderzeichen. Somit verbleiben lediglich 128 Zeichen für die Darstellung von Nicht-Englischen Zeichen.

### BABYLON MIT HUNDERTEN VON ZEICHENSÄTZEN

Abhilfe schaffte die Industrie durch die Definition hunderter neuer, individueller Zeichensätze. Doch diese Zeichensätze sind untereinander nicht kompatibel, da unterschiedliche Kodierungssysteme für ein- und denselben Buchstaben jeweils unterschiedliche Zahlenwerte benutzen. Somit müssen Rechner und Software viele verschiedene Zeichensätze unterstützen.

### UNICODE SCHAFFT EINDEUTIGE KODIERUNG

Unicode ist der universale Standard für eindeutige Kodierung. Unicode gibt jedem Zeichen seine eigene Nummer und ist plattform-, programm- und sprachenunabhängig. Alle modernen Betriebssysteme und Softwareanwendungen wie Web Browser oder Office Anwendungen unterstützen bereits Unicode.

### 100.000 ZEICHEN AUS ALLER WELT

Im aktuellen Unicode Standard 4.0 finden sich 96.447 Zeichen und Symbole aus der ganzen Welt.

Hierbei wird unterschieden nach Groß- und Kleinbuchstaben, aber auch diakretische Buchstaben (Akzente) repräsentieren jeweils ein eigenständiges Zeichen. Hierdurch lassen sich Zeichen wie A, Å, Ä, Å, Å, Å, Å, Å, Æ, a, à, á, â, ã, ä, å darstellen und weiterverarbeiten.

### DREI KOMPATIBLE STANDARDS MIT UNICODE

Der Unicode Standard definiert 3 Codierungen (8, 16 und 32-bit Codierung pro Zeichen), die untereinander vollständig kompatibel sind.

**UTF-8** ist sehr gebräuchlich für HTML Text und transformiert die Unicode Zeichen durch eine variable Anzahl von Bytes, viele nur mit einem Byte.

**UTF-16** stellt die meisten Zeichen durch eine Informationseinheit mit einer Länge von 16 bit dar; weniger gebräuchliche Zeichen mit Hilfe eines 16-bit Paares.

**UTF-32** codiert jedes Zeichen als 32-bit Einheit.

### INDEXIEREN UND SUCHEN MIT UNICODE

Während Unicode also ein enormer Vorteil ist, um Text jeglicher Sprache darzustellen, ergeben sich für die Indexierung und Suche einige Schwierigkeiten. So sollen bei der Suche nach „a“ eben auch alle Akzente des gleichen Typs, also à, á, â, ã, ä, å und A, Å, Ä, Å, Å, Å, Å etc. gefunden werden. Gleichzeitig muss es natürlich auch möglich sein, explizit nur nach تطوير suchen zu können.

### XSEARCH<sup>TM</sup> UND UNICODE

XSEARCH erfüllt diese Kriterien dank seiner integrierten Verity Technologie bravourös. Durch seine vollständige Unicode Unterstützung und mit Hilfe einer speziell hierfür entwickelten toleranten Suchtechnik, die zusätzlich noch justiert werden kann, ist die Bewältigung dieser Anforderungen möglich.

### CHINESISCH TRADITIONELL UND Co.

Beispiele für Veritys internationale Sprachunterstützung sind ChinaByte, die erste Volltextsucheseite in China unter [www.chinabyte.com](http://www.chinabyte.com) oder Homeway, Chinas größter Online Publisher für Börsen- und Finanzinformationen unter [www.homeway.com.cn](http://www.homeway.com.cn).

Beide Portale beinhalten mehr als 5 Millionen indexierte Dokumente in chinesischer und anderen Sprachen. Ein weiterer Kunde ist [yf.sh.edu.cn](http://yf.sh.edu.cn), der rund 5 Millionen Studenten und Lehrern in China das erste Portal für Lerninformationen zur Verfügung stellte.

Daneben sind z. B. folgende bekannten internationalen Internet Seiten mit Veritys Unicode Technik realisiert:

Tokyo Restaurant Guide unter [www.asku.com/rgi](http://www.asku.com/rgi)

Cisco Connection Online Japan

CNN Italien unter [www.cnnitalia.it](http://www.cnnitalia.it)

Fujitsu Japan unter [www.fujitsu.co.jp](http://www.fujitsu.co.jp)

Kyobobook Korea unter [www.kyobobook.co.kr](http://www.kyobobook.co.kr)

NEC Corporation Japan unter [www.nec.co.jp](http://www.nec.co.jp)

Bank La Caixa, Barcelona

[www.senado.es](http://www.senado.es) des spanischen Senats

UNISYS Japan unter [www.unisys.co.jp](http://www.unisys.co.jp)

ZDNet Japan unter [www.zdnet.co.jp](http://www.zdnet.co.jp)

### WENDEN SIE SICH FÜR WEITERE INFORMATIONEN BITTE AN:

Weitkämper Technology GmbH  
Reindlweg 3  
D-82418 Seehausen a. Staffelsee

Telefon +49 (0) 8841 48824-0  
Telefax +49 (0) 8841 48824-88  
E-Mail: [info@weitkamper.de](mailto:info@weitkamper.de)  
<http://www.weitkamper.de>

