

## Thema: Unicode-Unterstützung

Stand Juni 2011

### Einführung: Dupehunter Professional nutzt Unicode

Nach einer mehrmonatigen Entwicklungsarbeit können Sie nun das neue Dupehunter Professional in der Version 9 benutzen. Diese stellt einen Meilenstein in der Entwicklung des Programms dar, bietet sie doch die Möglichkeit, international einsetzbar zu sein. Der Grund dafür ist die volle Unterstützung des Unicode-Zeichensatzes nach UTF-16.

Dabei sind einige Aspekte zu beachten. Hier nun die Neuerungen:

### Unicode, was ist das?

Für Laien ausgedrückt ein Zeichensatz, der die meisten Schriftzeichen der Welt beinhaltet. Egal ob Russisch, Arabisch, Chinesisch oder Englisch – die meisten finden sich im Unicode-Zeichensatz wieder.

So sehen Datei- und Pfadnamen in der Version 9 aus:

<input type="checkbox"/> ψαρστεν.тχт.txt	108	c:\ch-soft\dupehunter unicode
<input type="checkbox"/> ψαρστεν.тχт.txt	108	e:\tests\ψαρστεν\
<input type="checkbox"/> ψαρστεν.тχт.txt2	108	e:\tests\ψαρστεν\
[-] 2F78DA9D55A28B06B410D...		
<input type="checkbox"/> تيايتزفاء.txt	78	e:\tests\ياتسيتي\
<input type="checkbox"/> تيايتزفاء.txt2	78	e:\tests\ياتسيتي\
[-] 26F63A6A5D16D778D623B...		
<input type="checkbox"/> שלום רב שוברבםפז.txt	56	e:\tests\דחלחג'טטר\
<input type="checkbox"/> שלום רב שוברך.txt	56	e:\tests\דחלחג'טטר\
[-] 23D813F6C457A41BDC625...		
<input type="checkbox"/> গৌড়ীভূজান.ফাঙ্ককা.txt	46	e:\tests\গৌড়ীভূজান ফাঙ্ককা\
<input type="checkbox"/> গৌড়ীভূজান.ফাঙ্ককা.txt2	46	e:\tests\গৌড়ীভূজান ফাঙ্ককা\
[-] 8E85B815D39C332D49E64B...		
<input type="checkbox"/> новый файл2.txt	6	e:\tests\тестовая папка\
<input type="checkbox"/> новый файл.txt	6	e:\tests\тестовая папка\

### Warum ist das für Dupehunter Professional so wichtig?

Ältere Versionen bis Version 8 konnten Dateinamen und Ordnernamen, welche Sonderzeichen aus diesem Zeichensatz enthielten nicht lesen oder auf sie zugreifen. Dies geschah unbemerkt ohne Fehlermeldungen, aber leider wurden unter Umständen auch Duplikate unter diesen Dateien nicht gefunden. Ab jetzt werden auch solche Dateien und Ordner erkannt und ausgewertet.

### Für wen ist das wichtig?

Für alle, die Kontakte in andere Sprachräume pflegen. In erster Linie natürlich für geschäftliche Verbindungen, aber auch bei privaten Kontakten und Austausch von Bildern oder Dokumenten kann die Gefahr lauern, dass diese Sonderzeichen in ihren Datei- oder Ordnernamen enthalten.

## Was bedeutete dies für die neue Version 9?

Da die Unicode-Fähigkeit über alles stand, mussten teilweise neue Wege beschritten werden, alte Bestandteile durch Alternativen ersetzt werden, als auch vieles modifiziert werden.

## Exporte aus Dupehunter Professional heraus:

Dupehunter Professional Version 9 schreibt alle Dateien im **UTF-16 Format**. Dies gilt vor allem für die Exportlisten von gefundenen Dateien, die sie an andere Abteilungen oder Personen weiterreichen können. Denkbar, um zum Beispiel eine Abteilung in Ihrem Unternehmen anzuweisen, die betreffenden Duplikate nach Durchsicht zu entfernen, um den Server zu entlasten. Dort sollten Leseprogramme installiert sein, die diese UTF-16 Unicodes unterstützen.

**Dupehunter Professional speichert neben dem empfohlenen XML-Format außerdem das CSV-Format sowie reinen Text ab.** Diese ebenfalls textbasierten Formate waren früher meistens reiner ASCII oder ANSI-Text. Dupehunter Professional schreibt hier generell immer den Text als UTF-16 ab, auch wenn die gefundenen Dateien innerhalb des ANSI-Zeichensatzes liegen sollten. Dadurch können selbst kyrillische oder griechische Zeichen in der CSV-Datei ausgegeben werden.

Weitere Formate für die Listen der gefundenen Duplikate gibt es aktuell nicht. Das langsame, über eine OLE-Schnittstelle verbundene, DOC-Format von Microsoft Word war zu träge und nicht auf tausende Zeilen ausgelegt. Ebenso entfiel das native XLS-Format von Microsoft Excel, da es zwar native war, aber der Import in Excel mittels CSV-Dateien ebenfalls einwandfrei funktioniert und dem Anwender eine Aufbereitung für eine Präsentation ermöglicht.

Der HTML-Export entfiel ebenfalls, da über geeignete Parser für XML-Dateien ebenfalls HTML-Dateien erstellt werden können. Werkzeuge wie Altova XMLSpy oder andere helfen dort sicherlich weiter und bieten mehr Möglichkeiten, als dies Dupehunter Professional bereitstellen könnte.

## Job-Dateien von Dupehunter Professional:

Diese so genannten Job-Dateien von Dupehunter Professional sind Einstellungsdateien für jede einzelne Suche. Jede einzelne Datei mit der **Endung .dhjb** enthält damit alle Parameter, die für eine reproduzierbare Suche nötig sind. Natürlich werden diese Dateien von Dupehunter Professional ebenfalls Unicode-fähig geschrieben und gelesen.

Auf ein Editieren von Hand sollte man indes **ganz verzichten**, da es über das Programm erstens übersichtlicher und vor allem sicherer ist. Nur so ist sichergestellt, dass die Suche gültige Parameter enthält, mit denen das Programm arbeiten kann.

## Betrachter in Dupehunter Professional:

Das Programm nutzt einen externen Betrachter, der Unicode-fähig ist und dementsprechend sowohl Bilder als auch Texte anzeigen kann. Dieser kann für das Lesen und Betrachten der gefundenen Duplikate genutzt werden, sofern die jeweiligen Formate unterstützt werden.

## Speicherbedarf des Programms:

Bedingt durch den größeren Zeichensatz benötigen viele gesammelte Informationen wie etwa Dateinamen oder Ordernamen doppelt so viel Platz im Arbeitsspeicher des Computers. Dies lässt sich nicht vermeiden. Gleichwohl waren wir auch schon in der Vergangenheit immer bemüht, den Speicherhunger des Programms so klein wie möglich zu halten. **Ein Vergleich von ca. 7 Millionen Dateien in einer einzigen Suche ist daher trotzdem noch möglich.** Im eigenen Interesse sollten Sie aber die Suchen deutlich beschränken, um auch übersichtliche Ergebnisse zu bekommen und um auch andere Programme nicht über Gebühr zu behindern.

## ZIP-Komprimierung als Ersatz für die SQX-Komprimierung:

Mit etwas Wehmut mussten wir das SQX-Format durch ein ZIP-Format für die Datenkompression von doppelten Dateien ersetzen. Der Grund hierfür war leider nicht die Unicode-Unterstützung von SQX. Denn es war bereits Unicode-kompatibel. Nein, vielmehr hat der Hersteller die Weiterentwicklung eingestellt, so dass es früher oder später zu einer Situation gekommen wäre, bei der Änderungen notwendig, aber nicht mehr realisierbar gewesen wären. Dies wollten wir sofort bei dem kompletten Umbau des Programms ausschließen. Ein weiterer Grund war zudem die schlechte Akzeptanz von SQX bei anderen Packer und Dateimanagern, denn uns waren nur die ebenfalls hauseigenen Tools von SpeedProject, Speed Commander und ZipStar, fähig, SQX-Archive zu erstellen und zu entpacken.

Der Wechsel zu ZIP ist aber nicht von Nachteil, bietet es doch ebenfalls eine hochsichere Verschlüsselung nach AES mit einer Stärke von 256 Bit und ist ebenfalls 64 Bit-kompatibel. Was dieser Packer nicht beherrscht, ist ein Passwort auf Header-Ebene, um auch schon das Inhaltsverzeichnis zu verschleiern und nicht nur den Inhalt. Neu ist in diesem Zusammenhang ebenfalls die Option, dass ein Administrator im geschützten Registrierungszweig **HKEY\_LOCAL\_MACHINE** einen Schlüssel namens „**GlobalZIPPasswordRequired=1**“ erstellen kann, um ein Passwort für jedes erstellte Archiv von Dupehunter Professional zu erzwingen.

## Wo gibt es keine Unicode-Unterstützung in Dupehunter Professional?

Auch wenn es ein Widerspruch zu „Volle Unterstützung für Unicode“ zu sein scheint, so gibt es doch sekundäre Einschränkungen, bei denen der ANSI-Zeichensatz notwendig ist. Begründet ist das durch Standards von Schnittstellen, an die sich Dupehunter Professional halten muss. Folglich ist zurzeit eine Beschränkung auf den ANSI-Zeichen an folgenden Stellen gegeben:

- IPTC/EXIF-Daten von Bilddateien, die Auflösungen etc. anzeigen (Nur lesen)
- MP3-TAG-Informationen von Musikdateien, die Interpret etc. anzeigen (Nur lesen)
- **ZIP-Passwort für erstellte Archive!**

---

Sofern Sie als Administrator Mitarbeiter betreuen, die auch ZIP-Dateien mit Passwortschutz erstellen sollen, weisen Sie diese bitte auf diesen Sachverhalt hin.

## Ihre Mitarbeit hilft uns sehr:

Sollten Sie in diesem Dokument oder im Programm noch Unzulänglichkeiten betreffend der Unicode-Unterstützung finden, scheuen Sie sich bitte nicht, uns dies mitzuteilen. Wir nehmen jede Anregung gerne auf, um weitere Verbesserungen zu erzielen.

Sie können uns jederzeit eine Nachricht an [support@dupehunter.com](mailto:support@dupehunter.com) schicken.

Vielen Dank für Ihre Geduld und Mitarbeit. Wir hoffen, dass Ihnen dieses Dokument geholfen hat.