

גירסה 1.00 - 16.9.2002

הגבלת תנועת רובוטים באינטרנט

מסמך זה הורד מהאתר <http://underwar.livedns.co.il>.
אין להפיץ מסמך זה במדיה כלשהי, ללא אישור מפורש מאת המחבר.
מחבר המסמך איננו אחראי לכל נזק, ישיר או עקיף, שיגרם עקב השימוש במידע
המופיע במסמך, וכן לנכונות התוכן של הנושאים המופיעים במסמך. עם זאת,
המחבר עשה את מירב המאמצים כדי לספק את המידע המדויק והמלא ביותר.

כל הזכויות שמורות לניר אדר

Nir Adar

Email: underwar@hotmail.com

Home Page: <http://underwar.livedns.co.il>

אנא שלחו תיקונים והערות אל המחבר.

הגבלת תנועת רובוטים באינטרנט

רובוטים באינטרנט (המכונים לעתים spiders או wanderers) הם תוכנות המשוטטות באתרים רבים באינטרנט בצורה רקורסיבית, על ידי קפיצה מקישור לקישור. החל מסביבות 1993 התעורר הצורך להגביל את תנועתם של הרובוטים, לאחר שנכנסו למקומות שלא היו צריכים להגיע אליהם, כגון קבצים זמניים בשרת, מידע כפול, קבצי CGI שמשפיעים על השרת (דוגמת טפסי הצבעות) וכו'. עקב צורך זה הוחלט ליצור מנוע שבעזרתו יהיה ניתן להגביל את תנועת הרובוטים.

השיטה שנבחרה היא הוספת קובץ לשרת, שיגדיר עבור רובוטים לאן מותר להם להיכנס ולאן אסור.

לפי שיטה זו, הקובץ צריך להיות נגיש דרך HTTP בכתובת בשרת robots.txt/ היתרון הגדול של השיטה הוא שניתן לממש אותה על כל שרת HTTP קיים, ללא צורך בשינויים בתוכנה, וכן הרובוט מקבל את המידע על זכויותיו בבקשת HTTP בודדת, ואינו מעמיס את השרת.

מבנה הקובץ robots.txt

הקובץ מורכב משורות שמופרדות ביניהם על ידי סימן מעבר שורה. כל שורה בקובץ היא מהמבנה:

```
<field>:<optionalspace><value><optionalspace>
```

ניתן להוסיף הערות לקובץ על ידי הסימן #. כל תו שיופיע אחרי הסולמית יחשב כהערה.

כל רשומה בקובץ מתחילה ב User-agent, ולאחר פקודה/פקודות Disallow. לפי הפרוטוקול, הרובוטים מתעלמים משורות שאינם במבנה זה.

User-agent זהו משתנה שקובע את שם הרובוט עבור ההרשאות הבאות מיועדות. לאחר שנכתוב User-agent נציין את שם הרובוט הרצוי, מופרד על ידי נקודותיים. ניתן לתת הוראות עבור כל הרובוטים על ידי שימוש בתו *.

Disallow זהו משתנה שלאחריו תופיע מחרוזת המציינת כתובת, אליה לרובוט אסור להיכנס. הכתובת יכולה להיות כתובת מלאה או כתובת חלקית. הרובוט לא יכנס לשום דף שמתחיל בכתובת זו.

למשל, ערכים אפשריים הם /files/ או /files/vbrun.exe אם לאחר Disallow לא מוכנס ערך, המשמעות היא שהרובוט רשאי לעשות כרצונו.

אם הקובץ robots.txt קיים וריק, ללא כל תוכן, כל הרובוטים יראו עצמם כמוזמנים.

מספר דוגמאות לקבצי robots.txt

```
# robots.txt for http://www.example.com/

User-agent: *
Disallow: /cyberworld/map/
Disallow: /tmp/ # these will soon disappear
Disallow: /foo.html
```

בדוגמא זו, אף רובוט לא יכנס אל /cyberworld/map/ וכן לא אל הספרייה /tmp/ ואל הקובץ foo.html.

```
# robots.txt for http://www.example.com/

User-agent: *
Disallow: /cyberworld/map/

# Cybermapper knows where to go.
User-agent: cybermapper
Disallow:
```

בדוגמא זו, אף רובוט לא יכנס אל /cyberworld/map/, מלבד רובוט בשם cybermapper, שלו ניתנת גישה חופשית לאתר.

```
# go away
User-agent: *
Disallow: /
```

קובץ robots.txt שיכיל טקסט זה, ימנע מרובוטים להיכנס אל כל קבצי האתר.

EOF