

השימוש במאגרי מידע לעריכת ספרים*

בית הועד - תשס"ג

- מרשימות פרטיות ועד למאגרי מידע אלקטרוניים
- מה הן הדרישות ממאגר מידע מודרני?
- דרישות התוכנה של מאגר המידע האידיאלי
- מאגרי מידע פרטיים לשימוש הציבור

מרשימות פרטיות ועד למאגרי מידע אלקטרוניים

הביטוי 'מאגר' כמציין אוסף של נתונים הוא ביטוי חדש, שנוצר לאחרונה. אכן איסוף נתונים, רישום, והשימוש בהם אינו חדש. כל אדם המעיין בספר זה או אחר יודע כי אין הוא יכול לשמור את כל הפרטים בזכרונו, ולפיכך הוא רושם לעצמו נתונים מסויימים, שלדעתו יהיה להם חשיבות ביום מן הימים. אין אנו רוצים לתת כאן סקירה על תופעה זו. נציין אפוא לדוגמה כי בין חכמי הספרדים מצאנו את הביטוי "זכרונותי", הבא לציין רישום של נתונים שרשמו חכמים לעצמם. בין חכמי האשכנזים יש שקראו לכך "ליקוטים", ולעתים גם "ילקוטים". גם אנציקלופדיות למיניהן הן למעשה אוסף של נתונים על פי שיטת רישום מסויימת, וכאמור לא נאריך בכך.

הואיל ומטרת רישום הנתונים היא לשם שימוש ביום מן הימים, ברור הוא שיש חשיבות לצורת הרישום, שהרי היא תאפשר לנו בבוא הזמן למצוא נתון מסויימ במהירות וביעילות. היו שרשמו את זכרונותיהם לפי סדר א"ב, היו שרשמו אותם לפי נושאים, והיו עוד שיטות שונות. ברם, ברור לנו כי כאשר כמות הנתונים תגדל ותתפח, גם שיטת הרישום המשוכללת ביותר לא תצליח לתת ביד רושמה מענה מידי לשאלותיו ולחיפושיו. ואכן, לא אחת ולא שתיים אנו שומעים על חכם המתלונן כי כעת לא מצא דבר מסויימ ברשימותיו.

צאו וראו כי חכמים נתנו דעתם לחשיבות ההתמצאות, לכן הם לא הסתפקו בתוכן העניינים. בחלק מן הספרים, ובמיוחד הדבר בולט אצל חכמי הספרדים, מצאנו בסוף הספר מדורים נוספים הנקראים מפתחות. לאמור, לפני המעיין בספר מוצגת רשימת מקורות הנזכרים בספר, וכן רשימת נושאים ועניינים שנידונו בספר. סיבת הדבר מובנת מאליה: אין אפשרות להגיע לנתונים אלו על פי תוכנו של הספר, או על פי עיון בספר לפי נושא מסויימ, כי אלו הם עניינים שנידונו בספר והוזכרו בו אגב אורחא. הדוגמה הבולטת לכך בימינו היא האנציקלופדיה התלמודית: העורכים הגיעו למסקנה שהן סידור הערכים לפי סדר הא"ב, והן תוכן העניינים בראש כל כרך, אינם מספיקים, והמעייין עדיין אינו יכול למצוא דברים רבים שנידונו בספר. לכן, לאחר שיצאו כעשרים כרכים, יצא לאור כרך מיוחד של האנציקלופדיה שכל כולו הוא מפתחות. ואכן, כיום אין כמעט ספר תורני ראוי לשמו ללא מפתחות מפורטים בסופו. מפתחות אלו נקראים בלועזית בשם 'אינדקס'.

אין מן הצורך לומר כי גם המפתחות המשוכללים ביותר אינם מאפשרים למצוא כל דבר במהירות וביעילות, וכל לומד יודע זאת מנסיונו האישי. יתר על כן, בחלק ניכר מהספרות התורנית, ובמיוחד בספרי חז"ל וראשונים, אין כלל תוכן עניינים, שלא לדבר על מפתח.

דומה שהמצאת המחשב הביאה אותנו לפתרון כמעט מוחלט של הבעיה שהוצגה לעיל. המצאה זו הביאה לעולם את התקליטור המשמש כמאגר מידע, שבאמצעותו ניתן להגיע לנתונים מבוקשים במהירות רבה וביעילות. אני מתאר לעצמי שהנוכחים כאן משתמשים בתקליטור השו"ת של אוניברסיטת בר אילן, או בתקליטורים הדומים לו, ויכולים לאשר את דבריי. ניתן לומר, ללא הגזמה, כי כל מהדיר ספר תורני כיום חייב להשתמש

* מתוך דברים שנאמרו ביום העיון. הדברים נכתבו בדרך של הרצאה, ולפיכך לא הובאו כאן מראי מקורות והפניות.

בתקליטורים של מאגרי מידע; אף כאן הדברים ידועים, והאריכות למותר. נציין כאן לדוגמה כי רי"ח סופר בספרו מנוחת שלום ח"ד וגם ח"ה (שניהם מוקדשים לספרים וסופרים) מודיע בהקדמתו כי כל האינפורמציה שהוא אסף בספרו אינה בעזרת תקליטור השו"ת, ומדבריו אתה למד על דרכם של אחרים.

אין לי אפוא הרבה מה לחדש בנושא השימוש בתקליטור, ולכן אתרכז כאן יותר במאגרי מידע לכשעצמם.

מה הן הדרישות ממאגר מידע מודרני?

מדובר למעשה על שתי דרישות. האחת, כיצד לבנות את המאגר עצמו. השניה, כיצד לבנות את החיפוש במאגר.

על מנת שהדברים יהיו פשוטים וברורים נלך בשיטת ר' יהודה, שהיה נותן בהם סימנים. לאמור, המילה 'מאגר' מציינת בראשי תיבותיה את הדרך שבה נבנה מאגר נתונים שיוכל להיות לנו לעזר.

מ - מהדורה

כאשר אנו קולטים אל תוך המאגר ספר מסויים, עלינו לבדוק היטב מהי המהדורה שאותה נקלוט. כידוע, חלק גדול ממקורות חז"ל וגם מספרי הראשונים יצאו לאור בכמה וכמה מהדורות, השונות זו מזו בכמה פרטים. עלינו להביא בחשבון מצד אחד מהי המהדורה הטובה והמדויקת ביותר, שהרי טובת המשתמשים לנגד עינינו. מצד שני, עלינו להביא בחשבון שמהדורות מסויימות, על אף חשיבותן הרבה, לא תמיד מצאו את הדרך לשימוש רוב המעיינים. לדוגמה, המהדורה הנפוצה של מדרשי חז"ל כיום היא מהדורת וילנא, וראוי שהנוסח שבה יתפוס את מקומו במאגר המידע; מצד שני, יש כיום גם הוצאות מדעיות של המדרשים הללו, המיוסדות על כתבי יד, וחשיבותן רבה ביותר. במקרה זה נקטנו ב'פרוייקט השו"ת' של אוניברסיטת בר אילן בדרך של 'טובים השניים', וקלטנו את שתי המהדורות. אבל לא תמיד כך הם פני הדברים, ולעיתים קרובות דרושה הכרעה, שאינה קלה, לגבי המהדורה שאותה נקלוט.

א - אמין

המאגר צריך להיות אמין. לא די בהחלטה שעלינו לקלוט ספר מסויים, אלא עלינו לוודא כי הקליטה נעשתה כהלכה. לשם כך עלינו לבדוק את עבודת ההקלדה. לדוגמא, אנשים שואלים מדוע אין אנו משתמשים בקורא אופטי (scanner) כדי להכניס טקסטים למאגר פרוייקט השו"ת. התשובה פשוטה: מתברר כי כמות השגיאות הנעשית בקריאה בדרך זו היא רבה ביותר, ותקנתה - קלקלתה. יתר על כן, לעיתים אנו מקבלים ממוציא לאור מסויים דיסקט שבו נמצא ספר שיצא זה עתה לאור. והנה, כשאנו בודקים את החומר הנמצא בדיסקט מתברר כי לא קרב זה אל זה. המדפיס הכניס שינויים לפני הדפסת הספר, ושינויים אלו אינם נמצאים בדיסקט. אילו היינו קולטים את החומר הנמצא בדיסקט כמו שהוא, היינו קולטים למעשה ספר שאינו זהה לספר הנדפס, דהיינו ספר פגום.

ג - גמיש

כבר קבעו הקדמונים כי כשם שאי אפשר לבר בלא תבן, כך אי אפשר לספר בלא שגיאה, והמליצו על כך את הכתוב 'שגיאות מי יבין'. גם אנו מוצאים מידי פעם שגיאות במאגר שלנו. יבורכו כאן המשתמשים שמעירים לנו על שגיאות שמצאו כדי שנוכל לתקנם. על המאגר להיות בנוי בצורה כזו שנוכל להכניס בו תיקונים ללא קושי. יתר על כן, עליו להיות בנוי בצורה כזו שהוא יכול להיות בסיס נוח לשינויים עתידיים. ידוע כי עולם המחשבים מתפתח בצורה מהירה ביותר, ותוכנות שונות מפותחות ומופצות חדשות לבקרים. לכן יש לבנות את המאגר בגמישות מירבית, כדי שנוכל להתאימו לשינויים אפשריים מכל סוג שהוא.

ר - רחב

על המאגר להיות רחב ככל האפשר. זו דרישה מובנת, וכמעט שאין צורך להסבירה. אולם זו הדרישה שגורמת לקשיים הגדולים ביותר. הרי ברור שאי אפשר בן-לילה לקלוט את כל הספרים, ולכן השאלה היא מה המדיניות שעלינו ללכת בה: האם להתרכז רק בנושא אחד ולהרחיב בו ככל האפשר? לדוגמה, בתחילה קלטנו רק ספרי שו"ת, וגם כאן כבר קיימת אותה שאלה. הרי נמצאים כיום למעלה משלושת אלפי ספרי שו"ת. מה עלינו לעשות? לקלוט את ספרי הראשונים בלבד, ולוותר על השאר? לקלוט נציגים מכל תקופה ותקופה? מי יהיו הנציגים של כל תקופה? אני יכול לומר שבנושא זה לא כולם משלימים עם הנעשה בפרוייקט השו"ת. אנו מקבלים הצעות שונות מה לקלוט ואת מי לקלוט, וכמובן, כדרכם של בני ישראל, אנו מקבלים הצעות ודרישות גם את מי לא לקלוט. לא נאריך בזה מפני הכבוד.

עד כאן החלק הראשון. ועתה נעבור לחלק השני.



דרישות התוכנה של מאגר המידע האידיאלי

גם אם יהיה בידינו מאגר בהתאם לאמור לעיל, הרי שאין בו תועלת אם לא נפעל להפיק ממנו את המידע באמצעות תוכנה מתאימה. מה נדרש מתוכנה זו? אף כאן יסייעו בידינו ראשי התיבות של המילה 'מידע'.

מ – מהירות

תוכנה טובה ככל שתהיה אין לה ערך רב אם אין היא מהירה, ללא קשר עם גודלו של המאגר. להמחשה בלבד: מאגר המלים בתקליטור הנוכחי של השו"ת הוא כמאה מיליון מילה. אתם יכולים לתאר לעצמכם עד כמה צריכה להיות מהירה התוכנה כדי להספיק לסרוק את כל המאגר הזה ביעילות בעקבות כל שאילתא.

י – ידידותיות

על התוכנה להיות פשוטה להפעלה ככל האפשר. נכון הוא שלכל תוכנה מצורף ספר המסביר את אופן תפקודה של התוכנה, אבל אין כוונתנו רק לספר. כוונתנו שהמימשק של התוכנה צריך להיות בנוי באופן יפה ונוח למשתמש; רשימת הפקודות צריכה להיות ברורה; הצורה החזותית שלה צריכה להיות נאה; יש לשאוף כי המשתמש, גם זה שאינו מצוי במחשבים, יוכל להפעיל את התוכנית עם סיוע מינימלי מן הספר.

ד – דינמיות

תוכנת החיפוש "יושבת" על תוכנות הפעלה אחרות. בעוד שתוכנת החיפוש היא פיתוח מיוחד של אוניברסיטת בר אילן, הרי שהתוכנות האחרות הן תוכנות של חברות המחשבים העולמיות, דוגמת מייקרוסופט ואחרות. חברות אלו משנות את התוכנות שלהן במהירות רבה ביותר. לדוגמה, לא מכבר הודיעה חברות מייקרוסופט כי היא מפסיקה לתמוך ולתת שירות לתוכנית ההפעלה חלונות 95, זאת משום שיש בשוק כבר כמה תוכנות חדשות יותר. נמצא כי תוכנת החיפוש שלנו צריכה להיות דינמית, ולהתאים את עצמה במהירות לתוכנות ההפעלה החדשות ביותר של שוק המחשבים. נקודה זו היא חשובה ביותר לצורך השימוש בתוכנת החיפוש. לדוגמה: אדם רגיל, שאינו בקי יתר על המידה בענייני מחשבים, קונה תוכנת חיפוש נהדרת למאגר מסויים ומשלם עבורה ממון רב, אבל הוא אינו יודע כי אפשר שהוא שם את כספו על קרן הצבי. הרי ברור הוא כי תוך שנה או שנתיים תהייה תוכנות הפעלה חדשות, וכלל לא בטוח כי תוכנת החיפוש הנהדרת שלו תמשיך לפעול. תוכנת חיפוש שהתאימה דרך משל למערכת הפעלה של חלונות 95, אפשר שלא תפעל בחלונות 2000. כדי שנבין את חומרת הבעיה, נציין כי תוכנת החיפוש של פרוייקט השו"ת נמצאת כל הזמן בדינמיקה. היא מצד אחד מתאימה למערכות הפעלה קודמות, כי אנו יודעים שציבור המשתמשים שלנו אינו ממחר להחליף את התוכנות הישנות שלו; מצד שני יש לנו ציבור רחב שפועל עם מערכות הפעלה מתקדמות, ועלינו להתאים את תוכנת החיפוש גם עבורו. דומה שמצב מעין זה כמעט ולא קיים בחברות המחשבים, לאמור,

הצורך לתת שירות בטווח כה רחב של תוכנות הפעלה, וכבר ציינתי לעיל מה מנהגה של חברת מייקרוסופט ביחס לנתינת שירות עבור תוכנות "ישנות".

ע – עממי

בכל האמור עד עתה יש עוד מרכיב שיש להביא אותו בחשבון, והוא מחירו של המוצר. יש לדאוג כי המחיר לא יהיה גבוה יתר על המידה, ובכך המוצר יהיה עממי ושווה לכל נפש, דהיינו שהמוצר יהיה בשימוש נרחב ביותר, לתועלת לומדי התורה בכל בית ובכל מקום.

אפשר להוסיף גם 'ע – עומק', כלומר שאנו מעוניינים כי התוכנה תוכל לחפש בצורה מעמיקה בתוך המאגר הקיים. ככל שהתוכנה תהיה מעמיקה יותר, כן נוכל לדלות חומר מתוך המאגר בצורה יעילה. לדוגמה: אנו מעוניינים לחפש מילה מסויימת, והתוכנה צריכה לחפש לא רק את המילה כפי שהיא - אלא גם את המילה עם הקידומות שלה, עם הסיומות שלה, עם הקיצורים שלה, ואולי גם מילה נרדפת שלה. התוכנה צריכה לתת אפשרות לחפש צירופי מלים, ועוד ועוד. מיגוון אפשרויות החיפוש האפשריות בתוכנה מעמיקה צריך להיות גדול ביותר, לתועלת המחפש.

מאגרי מידע פרטיים לשימוש הציבור

עד כאן עסקנו ביוזמתנו בהכנת תקליטור פרויקט השו"ת, המשמש כמאגר מידע, כעזרה לציבור הלומדים. אבל גם ציבור הלומדים יכול להיות לנו לעזר. אנו פונים בזה אל הציבור בבקשה פשוטה ביותר: הואיל ורובם ככולם של המהדירים כיום עובדים באמצעות מחשב, הרי שברשותם נמצאים מאגרי מידע. כל ספר שהוקלד לצורך ההדרה הוא בבחינת מאגר מידע. אלא שמאגר מידע זה לוקה בשניים. האחד, הוא מאגר פרטי, ואינו עומד לרשות הציבור. השני, אי אפשר להפעיל עליו תוכנת חיפוש טובה. בקשתנו היא אפוא כי תאפשרו לפרויקט השו"ת להשתמש במאגרים הפרטיים הללו. אם המאגר ייקלט בפרויקט השו"ת הוא יעמוד לרשות ציבור הלומדים כשפועלת עליו תוכנת חיפוש טובה, לתועלת כלל הלומדים.

סיימתי את הרצאתי בעניין הקשור בנושא ההרצאה, לאמור, הצעת אפשרות של שימוש בספרים שהוחדרו להרחבת מאגרי המידע. שיתוף הפעולה וההיזון ההדדי בין המהדירים למפיצי מאגרי המידע עשוי בעז"ה להגביר באופן משמעותי את איכות לימוד התורה והפצת התורה, וזוהי הרי מטרת כולנו.