

ההתחלה: פרויקט *Palestine Post*

רונלד ו' צוויג

כחוקר היסטוריה עכשווית, מחקרי התבססו תמיד על עיון בארכיונים ציבוריים ופרטיים שבהם כמות עצומה של מסמכים העלולים להכביד על המלאכה. ניסיוני במחקר בארכיונים החל בשנת 1975, בראשית עידן המחשוב האישי. שנים רבות הרהרתי בדרכים לנצל את הטכנולוגיות החדשות לייעול תהליך החיפוש באוספים טקסטואליים רחבי היקף, אך הכלים לכך טרם היו בנמצא. הצעד הראשון שהתבקש היה שיטה לזיהוי תווים אופטי (OCR – Optical Character Recognition) לשימוש בכלי חיפוש ממוחשבים במסמכים טקסטואליים. עקבתי בעניין רב אחר ההתפתחות בתחום זה. בשנות התשעים השתפרה יכולתן של תוכנות ה-OCR להמיר תמונת טקסט מודפס לגרסה דיגיטלית המאפשרת חיפוש ממוחשב וניתן היה לשקול דיגיטציה של אוספים מודפסים גדולים.

בתקופה זו התייעצה אתי הספרנית/ארכיונאית של העיתון *ג'רוזלם פוסט* בשאלה אם יש אפשרות לדיגיטציה של ארכיון גזירי העיתון שמוינו לפי נושאים. המטרה הייתה להמיר את האוסף הענק לפורמט דיגיטלי כדי לאפשר איתור מידע מהיר, ולא פחות חשוב – להשליך את כמיות הנניר הגדולות ולפנות שטח משרד יקר. האתגר היה מפתה, אך מבט חטוף באוסף הגזירים הבהיר מייד שהמשימה בלתי־אפשרית. הקטעים הודבקו לנייר בדבק זול והיו מקומטים מאוד. גרוע מכך, כל קטע נשא חותמת תאריך שהודפס על הטקסט. הגזירים היו קריאים לעין האנושית, אך לא הייתה מערכת OCR שיכולה הייתה להפיק מהם טקסט דיגיטלי בר שימוש באותה תקופה. עם זאת, מעל למדפי הקופסאות שבהן אוחסנו הגזירים הבחנתי בשמונים כרכים עבים של העיתון *Palestine Post* החל משנת 1932, עת נוסד העיתון, ועד 1950, אז שונה שמו ל-*Jerusalem Post*. כרכים אלה היו מפתים הרבה יותר מהגזירים למיזם דיגיטציה.

כהיסטוריון של תקופת המנדט ידעתי שהעיתון *Palestine Post* הוא מקור רב ערך, אך כמו כל עיתון היסטורי השימוש בו לצורכי מחקר היה תהליך מייגע ומורכב. ה-*Post* היה עיתון באנגלית של היישוב היהודי ונוסד במטרה להציג לפני פקידים בריטים, חיילים ואחרים ברחבי המזרח התיכון את נקודת המבט של היישוב על אירועים בארץ ישראל. היותו עיתון יומי הפך אותו למקור מידע עשיר ומקיף על חיי היומיום בתקופת המנדט. בהיותו עיתון בבעלות יהודית, הוא תיעד גם את ההידרדרות במצבן של הקהילות היהודיות באירופה בשנים לפני מלחמת העולם השנייה. פרויקט הגזירים לא צלח, אך העיתון השלם והמקורי היה אתגר שלא יכולתי להתעלם ממנו ובעלי ה-*Post* הסכימו לאפשר לי לנסות זאת.

הרעיון שלי היה לסרוק את הטקסט (של כל 27,000 הדפים) בדפים בגודל של גיליון רחב ולאחר מכן להעביר את התמונה הסרוקה לזיהוי תווים אופטי (OCR) כדי שתהפוך לפורמט דיגיטלי שאפשר לערוך בו חיפוש ממוחשב. אך בכך לא היה די. עימוד העיתון ועיצובו מעבירים מידע רב מעבר לטקסט עצמו. מיקום המאמר בעיתון ובעמוד, גודל הכותרת, איורים או צילומים – כל אלה נושאים מידע חשוב. לאחר שתמונה של עיתון נסרקה והומרה לטקסט דיגיטלי באמצעות OCR צריך היה לקשור את הטקסט לתמונה המקורית של המאמר. דיגיטציה בעלת ערך של עיתונים היסטוריים יכולה הייתה לאפשר לחוקר לחפש טקסט ולהציג את העמוד המקורי שבו הופיע בעיתון. הבנתי מה נדרש ומה יהפוך את המאמץ לסרוק את כל שמונה-עשרה השנים של העיתון *Palestine Post* לכדאי, אך לא ידעתי אם הדבר אפשרי מבחינה טכנית. כיום הטכנולוגיה זמינה בכל מחשב נייד הודות ל-Adobe ותוכנות דומות; אך לפני שלושים שנה היה זה בגדר חזון דמיוני ולכן התחלתי לנסות להפוך אותו למציאות. כל שנותר לעשות היה למצוא את התוכנה שתאפשר זאת.

במקביל לחשיפתי ל-*Palestine Post* מוניתי לנציג הפקולטה למדעי הרוח באוניברסיטת תל אביב בוועדה האחראית להקצאת משאבי מחשוב בין הפקולטות (מדעי המחשב, מתמטיקה, הנדסה וכו'). היה ברור כי לא נמנית עם המומחים בתחום זה, אך עד מהרה התברר מדוע הזמינו איש הפקולטה למדעי הרוח לוועדה. מדי שנה ערך הגוף המנהל את ההשכלה הגבוהה בישראל, ות"ת (הוועדה לתכנון ולתקצוב), תחרות בין האוניברסיטאות לתקציבים משמעותיים עבור פרויקטים מקוריים ויצירתיים בתחום המחשוב. אחד התנאים היה שהפרויקטים המוצעים יהיו גדולים ומקוריים ובתקציב העולה על 100,000 דולר – סכום גדול במיוחד באמצע שנות התשעים. כל פקולטה יכולה הייתה להשתתף ללא הגבלה במספר הפרויקטים. ככל הנראה סברו הגורמים המוסמכים כי הגשת פרויקט בנושא החדש של מחשוב מדעי הרוח תיטיב עם המוניטין של האוניברסיטה וביקשו ממני להגיש הצעת מחקר, אף שלא היה סיכוי אמיתי שתתקבל בהגשה הראשונה.

הכנתי תקציב וטקסט מנוסח בז'רגון מתאים, שהציג את אפשרויות המחשוב במדעי הרוח. התקציב כלל בקשה לעמדת עבודה יוניקס, עוזר מחקר וסורק בגודל תעשייתי; הסכום הכולל היה 70,000 דולר – סכום נמוך מהדרישה המינימלית של ות"ת. מאחר שהוועדה לא האמינה שההצעה תתקבל, הורו לי פשוט להכפיל את הבקשה כדי להגיע לסכום של מעל 100,000 דולר, אחרת לא תעמוד בכללים. לתדהמתנו ות"ת קיבלה את ההצעה. לפתע, ובאופן בלתי צפוי, קיבלתי את כל הדרוש לשלב הראשון של דיגיטציה של *Palestine Post*: סורק מקצועי, תקציב להעסקת סטודנטים שיפעילו אותו ושתי עמדות עבודה יוניקס, שלא ידעתי כיצד להשתמש בהן.

האוניברסיטה הקצתה שטח במרתף לא מנוצל של בניין קרטל בקמפוס, ובו הותקן סורק Scangraphics 36/400 (ברוחב סריקה 90 ס"מ וברזולוציה 400DPI, בשחור-לבן). גיליונות ה-*Palestine Post* רוכזו בשמונים כרכים בפורמט רחב ונמסרו לנו על

ידי משרדי ה-*Jerusalem Post*. נייר העיתון היה דק ובאיכות נימוחה ולא שמר על איכותו לאורך זמן; הכרכים המוקדמים ביותר, בני שישים שנה, התפוררו ללא טיפול זהיר. היה צורך לפרק את הכרכים (להסיר את הכריכה ואת הדבק או החוטים שאיחדו את הדפים) ולהעביר את הדפים אחד-אחד דרך הסורק. כדי לשמר את הנייר ולהקל על הטיפול יצרנו מעין מעטפת פלסטיק; שתי יריעות פוליקרבונט שקופות מחוברות בצידן כמעין "כריך". כל דף מהכרכים המפורקים הונח בין יריעות הפוליקרבונט והועבר דרך הסורק – תהליך פשוט שחזר על עצמו 27,000 פעמים.

צוותי סטודנטים עבדו על הפרויקט חודשים רבים. חלק מהעמודים היו חסרים, קרועים או הודבקו בכריכה המקורית באופן גרוע כל כך שאי-אפשר היה לסרוק אותם. נאלצנו לפנות לספריות אוניברסיטאיות שונות כדי למצוא עמודים חלופיים. בגיליונות מסוימים של העיתון היו חסרים עמודים 5 ו-6 הן בכרכים הכרוכים הן בכל הספריות ועבר זמן עד שהבנתי שעמודים אלה לא נדפסו כלל. רוב שנות קיומו מנה העיתון שמונה עמודים. עמודים 7 ו-8 הכילו מודעות ומידע קבוע נוסף, אשר הוקלד מראש בזמן איסוף החדשות וכתיבתן. אם היו חדשות לארבעת העמודים הראשונים בלבד, עמודים 5 ו-6 פשוט לא הודפסו. אולם מאחר שעמודים 7 ו-8 כבר הוקלדו והוכנו לדפוס, אי-אפשר היה לשנות את מספורם. לפיכך העיתון יצא לאור עם מספור עמודים 1 עד 4 ו-7 עד 8 בלבד. מגבלות אלה שיקפו את טכנולוגיית הדפוס של עיתונים באותה תקופה.

העיתון *Palestine Post* צולם פעמים מספר למיקרופילם, אולם אוספי המיקרופילם לא היו שלמים; דפים רבים היו חסרים או לא היו זמינים לפרויקט בשל זכויות יוצרים של המו"ל. נוסף על כך הטכנולוגיה להמרת מיקרופילם לעותקים דיגיטליים ברזולוציה גבוהה לא הייתה זמינה, לפני שלושים שנה. אוסף הכרכים הכרוכים של *Jerusalem Post* היה האוסף השלם ביותר של העיתון שעמד לרשותנו. לא הייתה אפשרות להימנע מהמשימה שעמדה לפנינו: סריקה ידנית מייגעת. זו הייתה הדרך היחידה שבה יכולנו להפיק את הגרסה הדיגיטלית המלאה והאיכותית שנדרשה לפרויקט.

לאחר הסרת הכריכה והסריקה של כל כרך, הם נשלחו כרך אחר כרך לבית מלאכה מקצועי לצורך כריכה מחדשת, על מנת להחזירם למשרדי *Post* בירושלים. משימה זו התגלתה כקשה מהצפוי. מקצוע כריכת הספרים הולך ונעלם ובישראל מצאנו רק בעל מלאכה אחד שיכול היה להתמודד עם כרכים גדולים כמו אלה של *Palestine Post*. בהיותו גמלאי הוא עבד במשרה חלקית בלבד. תהליך הפירוק והכריכה מחדש היה משימה טכנולוגית שכמעט שיבשה את כל הפרויקט.

תהליך הסריקה ארך חודשים רבים. במהלך תקופה זו התחלתי לחפש תוכנה שתוכל להגשים את חזוני – להמיר את התמונה הסרוקה בטקסט קריא שאפשר יהיה לחפש בו באופן אוטומטי וליצור חיבור לאימג' המקורי של הכתבה בעיתון. פניתי לכל חברת תוכנה בישראל שהתמחתה בזיהוי תווים אופטי (OCR), אך אף אחת מהן לא הייתה מוכנה להתמודד עם האתגר שהציב הפרויקט שלנו. עם זאת קיבלתי טיפ יקר ערך

על קיומה של חברה שמפתחת תוכנה לדיגיטציה של מדריכים טכניים בהיקף נרחב, המבוססת על רעיון דומה לזה שהגיתי עבור ה-Post: מנגנון חיפוש ושליפה המקושר לדף המודפס המקורי; הייתה זו חברת IOTA, סטארט-אפ קטן שפעל מעל מוסך בגבעתיים. שם הסבירו לי כי המדריך הטכני שמצורף למטוס קרב F-16 שוקל יותר מהמטוס עצמו, ואף גרוע מכך – הוא מתעדכן בקצב של 1,000 עמודים בחודש (או שמא בשנה?). תהא תדירות העדכון אשר תהא, היקף המידע היה עצום ומשימתה של IOTA הייתה להנגיש אותו בדיגיטציה, כלומר להפוך את המידע לחיפוש ממוחשב ולקשור את המידע שנשלף לתרשימים הטכניים המקוריים אשר לא הייתה דרך להמירם לדיגיטל.

הבאתי עמי גיליון של Post כדי להמחיש את האתגר של הפרויקט שלנו. התגובה הייתה מיידית: כולם גילו עניין רב בעיתון ההיסטורי. החדשות משנות השלושים, וביתר שאת המודעות הקטנות מאותה תקופה – כמו למשל מגרשים נבחרים בלב תל אביב במחיר של 100 לירות פלשתיניות – משכו תשומת לב. נייר העיתון היה אתגר גדול בהרבה מספרי המדריכים הטכניים שעליהם עבדו קודם לכן. דפי העיתון (60 על 70 ס"מ) היו גדולים משמעותית מהמדריכים, הנייר עצמו התיישן ודהה עם השנים, והטיפוגרפיה הייתה רחוקה מהסטנדרטים של ההדפסה לאחרונה. בסופו של דבר דובר בעיתון יומי אשר נועד להיזרק לאחר הקריאה. מדריכים מודרניים מודפסים בטקסט שחור על נייר לבן ונקי וטיפוגרפיה מושלמת; הניגוד (contrast) בין הטקסט לנייר הופך את משימת זיהוי טכנולוגיית פענוח הטקסטים (OCR) לפשוטה יחסית. אלא שהעיתונים ההיסטוריים מציבים אתגר שונה בתכלית, שכן רוב הנייר התיישן וקיבל גוון צהבהב; אך בשנים 1942-1943, בזמן מלחמת העולם השנייה, אספקת נייר העיתון ל-Post הגיעה מהודו ונייר זה קיבל גוון ורדרד עם הזמן. אם ורוד ואם צהוב, הניגוד היה רחוק מלהיות אידיאלי עבור OCR. יתר על כן, חלק מהאותיות היו שחוקות ולא שלמות, וחלק מהשורות לא היו מקבילות זו לזו. טכנולוגיית סדר הלינוטיפ (Linotype), אשר שימשה לדפוס של עיתונים במאה התשע-עשרה ועד שנות השבעים של המאה העשרים, הייתה תהליך לא מושלם אך די היה בו לצורכי העיתון. עם זאת היא הפכה את משימת הדיגיטציה במאה העשרים ואחת למורכבת במיוחד.

האתגר בפני מהנדסי התוכנה בחברת IOTA היה משמעותי: דיגיטציה של עיתון היסטורי. היה זה שנים לפני הופעתן של תוכנות אחרות שיכלו לספק את הפתרון הנדרש, אך חברת IOTA נענתה לקריאה. הכוח המניע לשיתוף הפעולה עם IOTA היה מר יוני שטרן, ששימש אז כראש מערך הטכנולוגיה בחברה. במסגרת הפרויקט ייצרנו כמות עצומה של דפים סרוקים והגענו להסכם בלתי פורמלי: הענקנו לחברה גישה חופשית למסד הנתונים של תמונות הסריקה והם השתמשו בו לפיתוח התוכנה הנדרשת. לאחר חודשים מספר של עבודה מאומצת, הציגו יוני שטרן וצוותו דגם ראשוני של תוכנה אשר הניבה תוצאות שימושיות. אמנם מערכת ה-OCR לא הייתה מדויקת במיוחד, אך מאחר שהיה זה חומר עיתונאי, שביטוי משמעותי צפוי להופיע בו פעמים

מספר בכל כתבה, שיעור האחזור של מאמרים רלוונטיים היה רב. מטרת הפרויקט הייתה לאתר כתבות שלמות ולא רק מילים בודדות, ולכן התוצאה הייתה מספקת. עם השנים השתפר דיוק התוכנה בשיעור משמעותי. ההתקשרות עם IOTA הייתה לא-פורמלית: הם פיתחו כלי תוך שימוש במשאבים שלנו ואפשרו לנו להשתמש בו ללא תשלום; בתמורה נשמרה להם האפשרות לשווק את התוכנה שפיתחו.

השאלה שנותרה ללא מענה הייתה אם קיימת דרישה לעיתונים דיגיטליים. שטרן ביקש לבחון את הפוטנציאל למסחור, ובשנת 1997 לקח את הפרויקט של ה-*Palestine Post* ואת הגרסה הראשונית והגולמית של התוכנה לתערוכה המקצועית המרכזית בארצות הברית (Seybold Publishing Report Seminar). איש מאתנו לא ידע אם יתעורר עניין כלשהו במוצג, אולם כבר ביום הראשון התקשר אלי שטרן מבוססטון וסיפר כי ההדגמה של IOTA עוררה התלהבות רבה. מסתבר שהיה צורך ממשי בכלי התוכנה שפותחו במסגרת הפרויקט. התגובה הייתה כה מעוררת עד כי IOTA החליטה להמשיך לפתח את התוכנה, לשפר את יכולות הזיהוי האופטי (OCR), את הקישור לתמונה המקורית ואת מנגנון החיפוש והשליפה. פרויקט ה-*Palestine Post* המשיך ליהנות משימוש חופשי בתוכנה המתפתחת, ו-IOTA קיבלה גישה מלאה לעמודים הסרוקים של העיתון. הייתה זו שותפות אידיאלית. שטרן וצוותו עבדו על העמודים תוך כדי תהליך הסריקה. מדי 24 שעות נסרקו 50-100 מגה-בייט של קובצי TIFF (100-200 עמודי עיתון). כדי לחסוך את הצורך בהעברה מתמדת של קבצים גדולים – אתגר קטן היום, אך משמעותי וממושך לפני שלושים שנה (טרם המצאת Dropbox והענן) – עבד צוות IOTA באוניברסיטת תל אביב בגישה ישירה לכוננים הקשיחים. הענקתי להם מפתח למשרדי ובכל ערב, לאחר שסיימתי את ההוראה, השתמשו בתחנות העבודה של UNIX ובקישור הישיר לכוננים שבהם נשמרו העמודים הסרוקים. סוף-סוף נמצאה הצדקה לא לתחנה אחת בלבד אלא לשניים, שכן היו אנשים שהיטיבו לנצל. גיבויים לעמודים הסרוקים הועברו כמעט מדי יום לדיסקי CD-ROM.

תוכנת IOTA, הידועה בשם "Newshare", עברה תהליכי פיתוח ושדרוג מתמידים. כל שדרוג תרם לשיפור איכות זיהוי התווים, כלי האחזור ותהליכי המפתוח. עם זאת כל שיפור בתוכנה חייב מפתוח מחודש של בסיס הנתונים ויצירת סדרה נוספת של דיסקים מסוג CD-ROM. בעת קבלת ההחלטה להציג את הפרויקט לשימוש הכלל, הוא כלל מערך של 25 דיסקים. תהליך האחזור דרש החלפה רבה של דיסקים, מצב שאינו אידיאלי בעליל.

עם זאת יצרנו את הגרסה הראשונה של כלי שעשוי היה לחולל מהפכה בשימוש בעיתונים היסטוריים למחקר. עותקי המיקרופילם של העיתונים הללו פתרו את בעיית המחסור – הם אפשרו גישה לעותקים של עיתונים מכל מקום, ולא רק בכמה ספריות מרכזיות שבהן נשמרו אוספים מלאים של המקור. אך המיקרופילם אינו נוח לחיפוש ומסורבל לשימוש. פרויקט *Palestine Post* הדגים את היתכנותו של רעיון רדיקלי

בהרבה – חיפוש ממוחשב ושליפת מידע מתוך עיתונים המוצגים לחוקר בפורמט המקורי שלהם. הגרסאות הראשונות של ה-OCR היו רחוקות מלהיות מושלמות והתוכנה הצליחה לאתר רק בין 60 ל-85 אחוז מכל מופע של מילה או מונח שחיפשו. אולם מכיוון שהתוכנה שלפה את המאמר כולו שבו הופיעו המילה או המונח ומונחי חיפוש סבירים בדרך כלל יותר מפעם אחת במאמר הרלוונטי, שיעור ההצלחה בשליפת מאמרי עיתון היה גבוה בהרבה משיעור ההצלחה באיתור כל מופע מודפס של מונח החיפוש. חברת IOTA (שכיום נקראת Olive Software) המשיכה לשפר את התוכנה וההכרה האופטית OCR השתפרה גם היא. כמות המידע הרלוונטי שנשלפה אף בגרסאות הראשונות של התוכנה הייתה בדרך כלל עצומה, גם אם השליפה לא הייתה ממצה לחלוטין.

הפרויקט התפתח באמצע שנות התשעים במקביל להתחלת השימוש הנרחב באינטרנט. עד מהרה הבנו שהאינטרנט, ולא תקליטורים (CDROM), הוא הדרך הנכונה להנגיש את הפרויקט. עם העלאתו לרשת, היה העיתון *Palestine Post* לא רק נגיש אלא אפשר היה לבצע חיפוש מקיף בכל העיתון בבת אחת. גרסת התקליטור נשכחה במהירות והפרויקט של *Palestine Post* היה נגיש מכל מקום.

היה זה השלב האחרון בפיתוחו של משאב מרשים והגשמת החלום שחלמתי כאשר נתקלתי לראשונה בשמונים הכרכים הנשכחים של העיתון. התמיכה של הפקולטה למדעי הרוח באוניברסיטת תל אביב והמימון של ות"ת, לצד ההתלהבות והיצירתיות של צוות IOTA/Olive Software, אפשרו את קיומו של הפרויקט.

לאחר שעברתי מאוניברסיטת תל אביב, אחרי עשרים ושתיים שנה, לתפקיד חדש ב-NYU, קיבל על עצמו חברי לעבודה וידידי הקרוב, פרופ' ירון צור, להמשיך את הפרויקט. בניהולו ולאור חזונו הגיע הפרויקט להישגים גדולים בהרבה. השלב הראשון הוכיח שהחזון של מחשוב עיתון היסטורי היה אפשרי. היוזמה של צור להרחיב את הפרויקט ולכלול בו את כל העיתונות היהודית מכל התקופות באשר הן, הביאה לעולם את הפרויקט רחב הממדים של העיתונות היהודית ההיסטורית (Jewish Historical Press) כפי שהוא היום.