# **נושאים במדעי הרוח הדיגיטליים – ארכיון המדינה, העשור הראשון**

## **מגישים: עדי אפל, גל שרגר ועומר סלע**

**נושא הפרויקט:**   
הנגשת תיקים מארכיון המדינה מהעשור הראשון של מדינת ישראל וניתוח הmetadata שלהם על מנת לקבל פרספקטיבה כוללת ומסקנות על התיקים שאספנו.  
בעבודתנו ניסינו למפות את סוגי התיקים אשר נמצאים בארכיון הנוגעים בעשור הזה, להפנות זרקור כלפי הנושאים והאישים המרכזיים המופיעים בתיקים ובאופן כללי להציג באופן מאוגד ונגיש את רבים מהתיקים מהעשור הראשון למדינה.

**פתיחות ונגישות:** האוסף מיועד לציבור הרחב ללא תשלום והרשמה על מנת להשתמש באוסף. כפי שיפורט בהמשך, המטרה המרכזית של האוסף הנ״ל הינה הנגשת מידע מעניין ומעשיר על הmetadata- בראי העשור הראשון למדינת ישראל.

האוסף הינו סגור בפני עריכה ע"י משתמשים חיצוניים אולם, אנו מציגים את האוסף ואת הקוד שבעזרתו יצרנו את האוסף באתר פתוח ונגיש לכל על מנת לעודד משתמשים נוספים להשתמש בקוד שלנו על מנת לייבא מידע בעצמם מארכיון המדינה ולהכין אוספים דומים על סמך שנים או נושאים שמעניינים אותם. כמות התיקים עליהם עוברים, השנים הרלוונטיות, מילות החיפוש ואישים המוזכרים בתיקים כולם מוגדרים כמשתנים גלובליים הניתנים לשינוי בקלות.

**סטנדרטים ואחידות:**

כל פריט אשר נוסף לאוסף הוא תיק מארכיון המדינה אשר מתוארך בעשורה הראשון למדינת ישראל (1948-1958). לכל תיק קיים מידע אשר נאסף משלושה endpoint-ים שונים מאתר ארכיון המדינה (כל אחד מהם מכיל מידע שונה על התיק), התיק נוסף לאוסף אך ורק קיים מידע עבורו בשלושת הendpoint-ים אותם אנו מאחדים לרשומה אחת באוסף.

**מטרת הפרויקט:**

כאשר התחלנו את הפרויקט בחנו מאגרי מידע וארכיון המדינה צד את עינינו, נמשכנו לכמות העצומה של המידע בנושאים שונים ומגוונים בהיסטוריה של מדינת ישראל. כשהתעמקנו בארכיון ובמידע שנמצא בו נוכחנו לדעת כי לא קיים שום מידע אגריגטיבי על נושאים שונים ואין כמעט שימוש ב-metadata על התיקים בארכיון.   
החלטנו להתמקד בעשור הראשון של מדינת ישראל משום שלראייתנו, ניתוח המידע שנאסף בעשור זה, בו התעצבה הלכה למעשה החברה הישראלית, יכול להיות מעניין עבור אנשים רבים ויש לניתוח זה ערך רב. כלומר, הדגשים העיקריים בעבודה שלנו היו: יצירת והנגשת אוסף המתמקד כולו בעשור הראשון למדינת ישראל ומיצוי המידע מהתיקים שנאספו ויצירת נדבך נוסף של מידע מתוך ה-metadata.

**מבנה ה-Data:**

על מנת לאסוף את ה-metadata על התיקים בארכיון פנינו לשלושה endpointים שונים. כל endpoint החזיר מידע גולמי אחר על כל תיק ותיק, פרסרנו את המידע הזה ל-json וטענו את כל קבצי ה-json ל-dataframes. לבסוף, היו בידינו שלושה dataframes שונים כאשר כל אחד מהם מכיל מידע אחר על התיקים (כל שורה מייצגת תיק). איחדנו את שלושת ה-dataframes לאחד אותו הצגנו באתר וכן כקובץ csv ב-GitHub. הdataframe- הסופי לאחר האיחוד מונה 12,981 רשומות (תיקים) ו-261 שדות. בנוסף, ייבאנו את התיקים עצמם (תיקי מסמכים כpdf, תמונות וכו') מendpoint- נוסף אשר הצגנו באתר תחת הלשונית The Files.

**במה שונה המידע מזה שהיה לפני העבודה?**

קיבצנו תיקים מארכיון המדינה שלכאורה אין ביניהם קשר אולם כולם מתוארכים לעשור הראשון למדינת ישראל. לא קיים (נכון לרגע זה) אוסף כזה באתר ארכיון המדינה ולכן האוסף שלנו נותן מבט כולל על התיקים מתקופה זו שלא קיים באוסף המקורי (אתר ארכיון המדינה). ע"י איסוף הmetadata- מתיקים אלו יכולנו לנתח את המידע ועל כן ליצור תמונת מצב כוללת על התיקים מתקופה זו בנושאים שונים כמו: כמה אחוזים מכלל התיקים שנאספו גלויים לציבור או חסויים ממנו? אילו אישים חשובים הוזכרו בהכי הרבה תיקים? איזה ארגון יצר הכי הרבה תיקים בעשור זה? וכו'.

**תהליך העבודה:**

הקוד פייתון המצורף עובר על 200,000 תיקים מארכיון המדינה וטוען רק את התיקים שמתוארכים בשנים שנתנו בקוד כפרמטר (במקרה שלנו - העשור הראשון למדינה). עבור כל תיק (הפעולה נעשית בלולאה), אנו מתחילים מendpoint ראשוני שמכיל metadata על כל תיק, ובעזרת המידע שיש בו, אנו פונים לשני endpointים נוספים המכילים metadata נוסף על התיק וכן לendpoint שמחזיר לנו את הקובץ המצורף לתיק אם קיים כזה (התיק עצמו, כקובץ pdf או תמונה). עבור כל תיק יש לנו שלושה קבצי metadata בפורמט json אותם אנחנו מפרסרים וטוענים לdataframes שונים (אחד עבור כל קובץ metadata) באמצעות ספריית PANDA של פייתון. בסיום הריצה של הלולאה נוצרים שלושה dataframes מלאים במידע על התיקים הרלוונטים אותם אנו מאחדים ע"י join על השדות המתאימים לdataframe אחד. אנו משתמשים בו כדי להוציא סטטיסטיקות וגרפים על הmetadata שאספנו באמצעות הספריות PANDA ו- MATHPLOTLIB של פייתון. (הערות מפורטות ודוקומנטציה ניתן למצוא בקוד המצורף)