ארכיון אישי בעידן ה Born Digital
היש חיה כזו?

כנס טלדן 13 מאי 2014,

עמיחי פיגנבוים
המחלקה ללימודי מידע, אוניברסיטת בר - אילן
fgnbmcs@computing-services.com

ארכיון אישי- הגדרה ומאפיינים עיקריים

- ארכיון אישי הינו מאגר מידע אשר מקור המידע העיקרי ולרוב הבלעדי שבו הוא הגורם היוצר או האוסף את המידע.
- ארכיון אישי מנוהל באופן עצמאי ע"י יוצר/ אוסף המידע ללא גוף מתווך בהקשר של צורת ניהול ארגונו הפנימית ומבחינת היקף תוכנו.
- המידע אשר בארכיון האישי הוא בחזקת אמין ולפחות ברמת "נאמן למקור" אם לא המקור עצמו.
- חלק מתוכן הארכיון האישי פרטי - קרי: לא את כול תוכן הארכיון אישי אנו מעוניינים שיהיה ניתן לגישה שלא על ידינו.

ארכיון אישי שתוכנו דיגיטלי

תוכן ארכיון כזה יכול להגיע מ 2 סוגי מקורות עיקריים:

1. מידע אשר מקורו בפורמט פיסי אחר והומר לדיגיטלי

   לדוגמא: סריקת מסמכים, דגימת הקלטות

2. מידע אשר נוצר מלכתחילה בצורה דיגיטלית Born Digital,

   לדוגמא: קבצי מעבד תמלילים, גיליונות חישוב אלקטרוניים, דוא"ל, תמונות ממצלמה דיגיטלית

ארכיון אישי שנולד דיגיטלי- 1

ככול שאנו עובדים עם מחשב ואמצעים אחרים לניהול ויצירת מידע אנו צוברים מידע באופנים שונים:

1. *באופן מודע מלא*

   (כגון על ידי ביצוע שמירה- save למסמך)

2. באופן *שאיננו מודעים אליו באופן מלא*

לדוגמא: כשאנו גולשים לאתר אנו יוצרים רשומות ביקור/ היסטוריה עם פרטינו לפחות ב3 מקומות [בדפדפן

הגולש (לזה אנו מודעים) אך גם בשרת בו מאוחסן המידע ואצל ספק האינטרנט שלנו]

ארכיון אישי שנולד דיגיטלי- 2

כיום עקב ריבוי המכשירים הפיסיים לנגישות וליצירת מידע שיש לנו מעבר למחשב האישי, כגון: טאבלט, טלפון חכם, טלוויזיה חכמה, מצלמה דיגיטלית ועוד וכן גם עקב מדיניות B.Y.O.D בארגונים, נוצרים עבורנו לעיתים ללא ידיעתנו גם הרבה מאגרים מידע אישיים *שאנו מודעים רק חלקית לעצם קיומם* אך לא למלוא משמעותם.

לדוגמא:

- מאגר התצלומים בטלפון או במצלמה אשר עשויים להיות מסונכרנים אוטומטית עם facebook,
- רשימות המועדפים עשויה להיות מסונכרנת בין דפדפנים (לדוגמא למשתמשים בכרום ביחידות גישה שונות),
- מיילים שקבלנו / שלחנו דרך אתר רשת,
- היסטוריית הגלישה באינטרנט גוררת המלצות גלישה
- היסטוריית הגלישה באינטרנט היסטוריית הצפייה בסרטים ב youtube גוררת המלצות צפייה
- המלצות רכישה מאמזון שנובעות מרכישות קודמות שלנו או של אחרים עם מאפיינים דומים.

מאפיינים של מידע שנולד דיגיטלי

- מידע אשר מקורו דיגיטלי שונה ממידע שאינו כזה.
- מידע כזה באופן אינהרנטי "סובל" ו"מרוויח" מהיותו כזה:

א- בפריט המידע הדיגיטלי **עצמו** קשה להצביע על מקור מול העתק באופן קטגורי רק באמצעות כלי והגדרות מטאדטא שונים ניתן לדעת זאת ברמת אמינות סבירה (בניגוד למידע דיגיטלי שמקורו מהסבה שניתן להשוות למקור).

ב- תוכן פריט המידע הדיגיטלי עצמו קל לשכפול- כלומר ליצירת העתק אשר הינו זהה מבחינה התוכן למקור (כפוף לנגישות אליו כפריט עצמאי).

ג- במרבית המקרים בכדי להשתמש/ לצפות בפריט שצברנו לא מספיק שיש לנו את המידע/ פריט עצמו, יש צורך גם בסביבה תומכת שיודעת להציג אותו או להשתמש בו.

מקומות צבירת המידע הדיגיטלי בעידן הנוכחי

- בעוד שבתחילת עידן המחשוב האישי, לרוב כל המידע אשר היה לנו היה נצבר לרוב על מכונה פיסית יחידה, כאשר גם בה לרוב בתוך "תיקייה" אחת, לדוגמא תיקיית הMy documents על תתי תיקיותיה וכן בתוך קובץ Post אשר הכיל את כלל דברי הדואר שלנו אשר החלטנו לשמור.
- היום בעידן הרשת- המידע שלנו, פיזורו הטרוגני ביותר ונמצא מלכתחילה בכמה מקומות זאת בנוסף למה שאחסנו במפורש במחשב האישי שלנו. לדוגמא:
  - עקב שיקולי נגישות, הרבה אנשים מעדיפים היות לאחסן את הדואר שלהם אצל ספק שירות הדואר כ gmail,
  - לאחסן תמונות סמוך למקום השימוש/ צפייה בהם כגון ב facebook.
  - המסמכים ב dropbox, קבצים גדולים ב filemonster ועוד .

משמעויות ההטרוגניות במקומות צבירת המידע

1. המידע לא נמצא בשליטתנו המלאה ובמקום אחסון יחיד ולכן לא ניתן למעשה לבצע לו "גיבוי כולל".
2. פרטיות המידע אינה מובטחת.
3. לפעמים רק כדי שהוא ימשיך להתקיים אנו צריכים ל"תמוך" בו.
4. אין לנו דרך פשוטה לשמור ולוודא את שלמות התוכן.
5. הסדר שמאורגן המידע לרוב איננו בשליטתנו המלאה.
6. יכולת הגישה למידע מחייבת אותנו לא רק באספקת פרטי זכויות הגישה (Credentials) למקום אחסונם אלא גם בתלות בצדדים שלישיים כגון ספקי תווך האינטרנט (קווי ו/או סלולרי) וכן באי חסימה של הגישה אליו בארגון הספציפי בו אנו כרגע נמצאים.
7. יש במצב זה גם יתרון מסוים, שאם משהו הולך לאיבוד אז רק החלק היחסי.

סכנות עיקריות בקיום ארכיון אישי שלא בשליטתנו המלאה = ברשת

- הסכנה העיקרית היא שהמידע מאבד את תקיפותו ואמינותו - כלומר את ודאות נאמנותו למקור, לדוגמא: כאשר אנו מעלים ל youtube סרטון, הוא משנה אותו מהותית למעשה ממיר אותו מפורמט אחד לאחר עד כדי איבוד ושינוי פרטים בתוצר.
- אין לנו פרטיות אמיתית! אפילו אם אנו לכאורה מגדירים משהו ככזה, לדוגמא ב facebook לתמונה פרטית, אנשי facebook עצמם יש להם יכולת גישה מלאה אליה!
- אנו מאבדים את יכולת המעקב גם על המטאדטא ועל כול רישום הפעילות המדוקדקת עימו, דוגמא: הוא עשוי להיות חשוף לעין לא צפויה ולא אנושית כמו שבג'ימייל מהמפורסמות היא שהם מציגים פרסומות הנוגעות לדעתם לתוכן המייל שאנחנו קוראים או בדוגמא מיוטיוב מקודם, אפילו גודל הקובץ יהיה שונה וגם לא נגיש!
- ישנם גם מצבים של איבוד זכות הגישה למידע עקב איבוד הCredentials ואז עלינו לחשוב על איך ניגשים למידע.
- בעיית השלמות: כאשר מעלים כמות גדולה של מסמכים נניח לdropbox הסיכוי שנרגיש כי מסמך יחיד חסר לחלוטין או שונה בתוכנו, נמוך ביותר במיוחד כשהוא בתוך מבנה היררכי.

פתרונות עקרוניים אפשריים לבעיות

1. ריכוז כל המידע שלנו במדיות אישיות פרטיות גדולות נפח בעלות רמת אמינות חומרה גבוהה אשר כוללות בתוכם רמת יתירות Redundancy מובנית ומוגברת הן ברמת החומרה והתוכנה.
2. שימוש במאגרי הרשת תוך הקפדה שלא להעלות חומרים בעלי חשיבות עתידית ושאנו מוכנים שיהיו חשופים לעיניים זרות.
3. ניהול רשימת הCredentials במנותק מהרשת.

פתרונות עקרוניים לבעיית האחסון ברשת

- הכרח בקיום של מנגנון סטנדרטי, אשר לכל פריט מידע שאנו נעביר למקום אחסון שלא בשליטתנו המלאה, קרי ברשת, ינוהל מעקב קפדני עליו הן מבחינת עצם הגישה והן מבחינת מעקב על השינויים שחלים בו, כמובן עם עדיפות לאי שינוי כלל. כלומר לספק למעשה את ה Provenance שלו, כלומר למעשה הכרח במערכת ניהול מסמכים ומידע אשר באמצעותה נשמור המידע.
- סטנדרט Prov נמצא כעת בפיתוח ע"י ארגון הW3C. מטרתו הסופית היא לאפשר לספק לפריט מידע בעל URI, כלומר למעשה לכל דבר בעל כתובת באינטרנט את המנגנון הזה. זאת באמצעות הכמסת פריט הידע בתוך מעין מעטפת, אשר תספק באופן אינהרנטי את כל מה שאנו צריכים לשמור לצורך קיום ה Provenance של פריט המידע על כל פניו, הסטנדרט כולל גם יכולת התאמה והרחבה לצרכים נוספים.

סטנדרט Prov

- בסטנדרט הProvenance מוגדר כך:

"***Provenance*** is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. "

- כלומר מעבר לפריט המידע עצמו אנחנו כוללים גם את כל מה שהביא אותו למצבו הנוכחי. ואלו בדיוק מה שהיינו כוללים במטאדטא- התיאורים ברמות שמעל רמת המסמך הבודד בארכיון.
- הסטנדרט מוצע למימוש (בין השאר) באמצעות אונטולוגיה סמנטית מתאימה, אשר בעצם היותה כזאת היא ניתנת להתאמה לצרכים של כל תכולה שהיא של ארכיון אישי ואפילו ארגוני.

יש פתרון רשתי *אחר* שכבר עובד

- באתר Wikipedia, כל שינוי בתוכן ערך נרשם וכולל את מזהה מבצעו ועיתויו דבר אשר מאפשר לעקוב על התוכן והתפתחותו,
  אבל יש לשים לב כי הוא מיושם רק למידע שהוחלט שיהיה גלוי לציבור כלומר אנחנו לא יכולים לראות ערכים אשר לא אושר לפרסם גרסה כלשהיא שלהם או שהינם בשלבי הכנה או שכבר הוסרו מתצוגה או שלא נמצאים בשפת האנציקלופדיה שבה אנו כרגע צופים.
- נא שימו לב כי הויקיפדיה לא משתמשת מבחינה טכנית בתקן שעכשיו עובדים עליו, אלא בנתה מערכת משלה אשר עונה לדרישות של מפתחיה/ משתמשיה, כאשר לפחות לגבי ניהול הטקסטים שלה יתכן והינה מספיקה גם לרמת ארכיון אישי.

לסיכום

- באם נרצה שלמידע שלנו בארכיון האישי תהיה אותה רמת אמינות ונאמנות למקור כמו בארכיון מוסדי, עלינו לנהל לא רק את המידע וסידורו אלא גם את המטאדטא שלו.
- לדעתי, נכון לעכשיו עדיפה הדרך של ניהול המידע בארכיון אישי באופן עצמאי על מדיה אישית פרטית אמינה בעלת אמינות מובנית גבוהה.
- שימוש באתרי אחסון חסרי היבט ויזואלי כגון: dropbox ו filemonster אפשרי אם הקבצים שנאחסן יהיו מוצפנים בצורה סבירה, אז לפחות על בעיית הפרטיות אנחנו ברמה מסוימת מתגברים.

שאלות:

ניתן לפנות אלי גם במייל:

fgnbmcs@computing-sevices.co.il

מקורות

1. Personal ARCHIVING, Preserving our Digital Heritage, Donald T. Hawkins (editor) 2013, Information Today Inc., ISBN: 978-1-57387-480-9.
2. The future of Personal Information Management Part I: Our Information, Always and Forever, William Jones, 2012, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, ISBN: 9781598299366.
3. Transforming Technologies to Manage Our Information, The future of Personal Information Management Part 2, William Jones, 2012, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers , ISBN: 9781627050173.
4. Provenance an Introduction to PROV, Luc Moreau and Paul Groth, 2013, Synthesis Lectures on Semantic Web: Theory and Technology, Morgan & Claypool Publishers, ISBN: 9781627052221.

אתרים בנושא

- http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
- http://digitalpreservation.gov/personalarchiving/
- http://library.columbia.edu/locations/dhc/personal-digital-archiving/online-resources.html
- http://www.nationalarchives.gov.uk/information-management/projects-and-work/digital-preservation-faqs.htm
- http://en.wikipedia.org/wiki/Personal_wiki
- http://www.wdc.com/en/products/products.aspx?id=870
- http://plonter.co.il/stores/main.tmpl?store=TicTac&cart=13998964772926479 2&lang=heb
- http://www.pcmag.com/article2/0,2817,2399582,00.asp