

אונטולוגיות במדעי המידע וגישות לפיתוחן

מאת: מעיין ז'יטומירסקי-גפת

מילות מפתח: אונטולוגיה, ניהול ידע, נתונים מקושרים, רשת סמנטית, ווב 3.0, מיקור המונים, חוכמת ההמונים, עיבוד שפה טבעית, סוכנים חכמים, בינה מלאכותית.

תקציר

הגדרת אוצר מלים - לקסיקון פורמלי מוסכם ומשותף עבור תחומי ידע שונים עומד בבסיסם של יישומים ומערכות רבות. לקסיקון פורמלי כזה מאפשר עיבוד וניתוח אוטומטיים או אוטומטיים למחצה יעיל ומהיר של כמויות גדולות של מידע ממקורות הטרוגניים המציפות אותנו בעידן האינטרנט ומאגרי המידע האדירים. יצירת לקסיקון כזה חוסכת זמן ומאמץ אנושי רב ומאפשרת הצלבת נתונים מבוזרים והפעלת היסק לוגי ממוכן על ידי אלגוריתמים של למידת מכונה. לתהליכים אלה פוטנציאל לגילוי תופעות חדשות ולפריצות דרך בתחומי ידע שונים שלא היו מתאפשרים בניתוח אנושי בלבד. לקסיקון פורמלי כזה נקרא בעגה המקצועית "אונטולוגיה". מאמר זה סוקר את הספרות הקיימת על אונטולוגיות במדעי המידע ומפרט מודלים אונטולוגיים הקיימים בספרות, תוך ציון יתרונותיהם והגישות השונות לפיתוחם.

מבוא

אנו חיים בעידן המידע והתקשורת. בשני העשורים האחרונים גדלה רשת האינטרנט בקצב מעריכי (exponential growth). נכון להיום, מדענים מעריכים שרשת האינטרנט מכילה כ-4.74 מיליארד דפים (hypertext pages) ממופתחים (<http://www.worldwidewebsite.com>) על ידי מנועי החיפוש Google ו-Bing ואלה מהווים רק חלק קטן מסך מספר הדפים שיש ברשת (van den Bosch et al., 2016). בשנים האחרונות גם מכשירים ניידים חכמים מאפשרים לגשת לרשת, לגלוש בה ולחפש מידע בכל מקום וזמן וכמעט ללא הגבלה. אי לכך, מנועי חיפוש ובפרט מנוע Google הם האתרים השימושיים ביותר כיום, ומנוע גוגל לבדו, מקבל 3.5 מיליארד חיפושים בכל יום (<http://www.internetlivestats.com/google-search-statistics>). כחלק מתהליך חיפוש המידע ואחזורו משתמשים מקלידים במילות מפתח, שעל פיהן מעבדים מנועי החיפוש את המידע ברשת במהירות באמצעות אלגוריתמים אוטומטיים מתוחכמים ומחזירים מידע תואם ורלוונטי לשאלות של המשתמשים. בחינת רלוונטיות מבוססת על דמיון בין מילות השאלות למילות המסמך המאוחזר. ואף על פי כן, במקרים רבים, מנועי החיפוש אינם מסוגלים לספק תשובות מדויקות לשאלות מורכבות, כמו לדוגמה, "מהם החלבונים המעורבים בהולכת אותות בעצב פירמידלי?" (Berners-Lee, 2009), או "מהן המדינות האירופיות שאוכלוסייתן הצטמצמה בעשור

האחרון?" ו"מי הבעלים של החברות היפניות שהרוויחו יותר מעשרה מיליון דולר בשנה האחרונה?". כל המידע הנחוץ בכדי לספק תשובות לשאלות הללו כבר נמצא ברשת האינטרנט בצורת דפים או מסמכים הכתובים בשפה הטבעית (שפות כמו עברית, אנגלית וכיו"ב). הדפים הללו בדרך כלל מובנים ומעוצבים גרפית בשפות תיוג ועיצוב כמו CSS/HTML. ובכל זאת, מנועי החיפוש אינם מבינים את המסמכים הללו וגם לא את צורכי המשתמש. מהו הקושי שניצב בפניהם?

הקושי נעוץ בשני המאפיינים העיקריים של שפה טבעית: הראשון הוא הגיוון הלשוני והשני הוא ריבוי משמעויות. משמעותו של הגיוון הלשוני הוא העובדה שבכל שפה טבעית יש מילים שונות אך דומות במשמעותן, ולפעמים קטעי טקסט שונים מבטאים את אותו הדבר (ז'יטומירסקי-גפת, 2009).

לדוגמה: חברות – רעות, שמחה – חדווה. רב משמעות הוא צירוף המציין תופעה בלשנית של מלה שיש לה יותר ממשמעות אחת, והמשמעות היא תלוית הקשר. כך לדוגמה, משמעותה של המלה "חברה" יכולה להיות ארגון מסחרי, קבוצת אנשים (society) או סביבה אנושית.

מאפיינים אלו הופכים את האתגר לקשה במיוחד עבור מנועי החיפוש ואפליקציות מידע אוטומטיות אחרות. זאת כיוון שאם בדף אינטרנט מסוים השתמשו במונח "חברה" (אפילו אם הודגשה ככותר בגופן וצבע בולט בדף), על מנוע החיפוש לגלות לאיזו מן המשמעויות הרבות של מילה זו התכוונו. המבנה הוויזואלי-עיצובי של הדף אינו מסייע למנועי החיפוש לפענח את משמעותן של המלים ואת ההקשרים שבהם הן מופיעות. אם המשתמש כתב שאילתה "מהן היצירות של ליאונרדו דה וינצ'י" ובמסמכים במקום המונח "יצירה" מופיעים מונחים כגון "ציור", "חיבור" ו"פסל", על מנוע החיפוש "להבין" שיש קשר סמנטי בין מילת החיפוש (הכללית יותר) למלים השונות הדומות לה במסמך בכדי להחזיר תשובה מתאימה למשתמש. גם הקישורים (hyper-links) בין הדפים אינם תורמים להבנת המשמעות, שכן יש צורך בקישור בין מושגים בדף אחד למושגים הקשורים אליהם במשמעות בדף אחר. במלים אחרות, אתרים שונים מחזיקים מידע רלוונטי, אבל מכיוון שהם אינם משתמשים באותן המלים, כלומר אינם "מדברים" באותה השפה והנתונים שלהם לא מקושרים ביניהם, בדרך כלל לא ניתן לקשר ולהצליב את המידע בהם בצורה אוטומטית בקלות וברמת דיוק מספקת. על כן, המשתמשים נאלצים להשקיע זמן ומאמץ רב כדי לברור את המידע במסמכים המאוחזרים ולמצוא בהם תשובות לשאלותיהם באופן ידני. כתוצאה מכך, חוסר קישוריות סמנטית בנתונים המפוזרים ברשת הופך את אחזור המידע האוטומטי המדויק והממוקד לנחלתם של מסדי נתונים מובנים, סגורים ומבודדים בלבד.

אחד הפתרונות שמוצע כיום בספרות שעוסקת במדעי המידע ובמדעי המחשב הוא יצירת לקסיקון פורמלי של תחום הידע או בעגה המקצועית - בניית אונטולוגיות לתחומי ידע שונים. אונטולוגיות פותרות את שתי הבעיות שמציגה השפה הטבעית ומאפשרות הוספת משמעות לדפים ברשת (סמנטיזציה) ויצירת קישוריות בין הנתונים באינטרנט. בפרקים הבאים נתאר מהן אונטולוגיות וכיצד הן מסייעות לסמנטיזציה של הרשת.

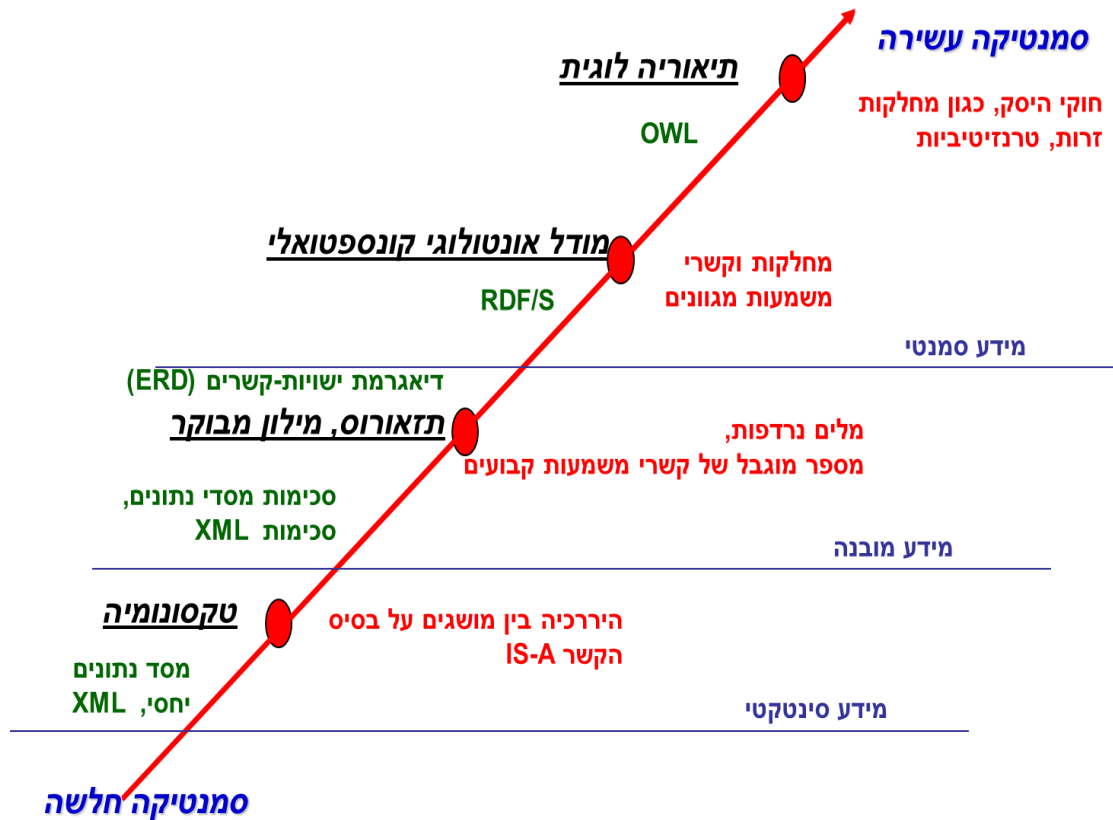
מהי אונטולוגיה?

האטימולוגיה של המלה "אונטולוגיה" היא מיוונית ($\delta\upsilon\tau\omicron\varsigma$, *ontos*) "מה שקיים" ו- ($\lambda\omicron\gamma\iota\alpha$ –*logia*) "תיאוריה, מדע". זהו מונח פילוסופי שמקורו בתורת ההוויה, תיאוריה שעוסקת בישויות הקיימות בעולם, במאפייניהן ובסיווגן שהופיעה כבר בקטגוריות של אריסטו (Stough, 1972). בסוף שנות ה-90 אימצו אנשי מדעי המידע ומדעי המחשב את תמצית הרעיון הפילוסופי הטמון במונח "אונטולוגיה". אנסה להסביר כיצד משתמשים בו אנשי מדעי המידע: האונטולוגיה, כפי שהיא משמשת אותם, היא מפרט (ספסיפיקציה) מפורש ופורמלי של מושגים משותפים (Gruber, 1993). המפרט מתייחס למודל מופשט של תופעה, שמוגדר באמצעות זיהוי המושגים הרלוונטיים שלה. מפורש משמעותו, שהמושג ומגבלות השימוש בו מוגדרים באופן ברור. פורמלי פירושו, שהוא מוגדר היטב בשפה לוגית (לא בשפה טבעית) ובאופן חד משמעי, כך שמכונה (אלגוריתם מחשב או סוכן אוטומטי) תוכל לקרוא ולפענח את תוכנה של האונטולוגיה. שיתוף - משקף את העובדה שאונטולוגיה לוכדת ידע מוסכם שאינו פרטי לאדם מסוים, אלא מקובל על קבוצה.

הגדרה משלימה וטכנית יותר לאונטולוגיה גורסת, שהיא אוצר מלים פורמלי, מודל עשיר של ידע משותף לחוקרים ולסוכנים (תוכנות מחשב, אלגוריתמים) אוטומטיים, שמכיל אוסף מושגים, הגדרותיהם, מאפייניהם וקשרים סמנטיים ביניהם (Uschold & Gruninger, 1996). היא כוללת הגדרות מובנות מכוונות-מכונה של מושגים מופשטים בסיסיים בתחום (המכונים בעגה המקצועית "מחלקות"). מאפייניה של כל מחלקה הם תכונותיה והקשרים הסמנטיים שקיימים בינה לבין מחלקות אחרות באונטולוגיה. בנוסף למחלקות ולמאפיינים, האונטולוגיה יכולה להכיל גם מופעים ספציפיים, שהם אובייקטים בעולם, מקרים פרטיים, של המחלקות שהוגדרו בה. למשל, המושג "עיר" היא מחלקה באונטולוגיה עם מאפיינים כמו "מיקום גיאוגרפי", "גודל האוכלוסייה", "שנת הקמתה" ועוד, ו"תל אביב" היא מופע ספציפי של המחלקה "עיר" עם מיקום מסוים (קו רחב: $32^{\circ} 5'$ וקו אורך: $34^{\circ} 46'$) ואוכלוסייה של כ-432,000 תושבים שהוקמה ב-1909.

אונטולוגיה היא גם השלב האחרון באבולוציה של לקסיקון פורמלי כפי שניתן לראות בתרשים מספר 1 והיא מהווה מילון מבוקר ועשיר שבו כל מלה נושאת משמעות אחת ויחידה. בדומה לקודמיה באבולוציה של הלקסיקונים, כגון טקסונומיות ותזאורוס, גם בנייתה של אונטולוגיה מבוססת תמיד על קשר היררכי בין מושגים: קשר IS-A או "סוג של" בעברית. טקסונומיה (מיוונית – $\tau\acute{\alpha}\xi\varsigma$: סיווג, סידור; $\nu\acute{o}\mu\omicron\varsigma$: שיטה, חוק) היא סיווג שיטתי, היררכי של עצמים מאותו תחום (Lamberts & Shanks, 1997). תזאורוס, על פי הגדרתו על ידי הסטנדרט הבינלאומי ISO 25964, הוא מילון מובנה ומבוקר שבו מושגים מיוצגים על ידי מלים שמאורגנות כך, שהקשרים בין המושגים הם מפורשים ומלווים במלים נרדפות ובמלים בעלות משמעות דומה (<http://www.niso.org/schemas/iso25964>). גם בתזאורוס המושגים מסודרים בצורה היררכית

כאשר כל מושג מקושר למושג רחב יותר מצד אחד, וצר יותר מצד שני. הקשרים בתזאורוס הם קבועים, כמו לדוגמה קשר היררכי, קשר של שוויון או קשר אסוציאטיבי. באונטולוגיה לעומת זאת, ניתן להגדיר גם קשרים שונים שאינם היררכיים אלא רחביים בהתאם לצורכי תחום הידע וללא הגבלה לרשימת קשרים סגורה.



תרשים מספר 1: התפתחות המודלים והטכנולוגיות הסמנטיות (מלמטה – מודלים בסיסיים, עד למעלה – מודלים עם סמנטיקה עשירה). מבוסס על הדיאגרמה של MITRE Corporation 2006 (<http://slideplayer.com/slide/697642/>, slide 9).

שלבי בניית אונטולוגיה

בניית האונטולוגיה לפי (Noy & McGuinness, 2000) מורכבת מהשלבים הבאים:

1. מכיוון שהגדרת השימוש באונטולוגיה היא מהקווים המנחים בבנייתה, ראשית יוגדר תחום האונטולוגיה והיקפה, בניסיון לעמוד על התחומים שאותם היא מכסה ועל אוכלוסיית היעד שלה. אחר כך מתגבש תוכנה של רשימת שאלות שעליהן אמורה האונטולוגיה לתת מענה. בנוסף לדברים הללו, יש להביא בחשבון שימוש חוזר באונטולוגיה, ויש לבדוק האם קיימת אונטולוגיה בתחום הידע הנתון אשר ניתן להשתמש בה, לשפרה או להרחיבה.

2. יש להגדיר את המחלקות (המושגים) החשובות לתחום הידע שייכנסו לאונטולוגיה. מחלקה היא מושג באונטולוגיה, אוסף של ישויות בעלות תכונות דומות (למשל, מחלקת ערים, בעלי חיים, יינות).
3. הגדרת ההיררכיה בין המחלקות על בסיס הקשר IS-A. ניתן לבנות היררכיה בשלושה אופנים שונים:
- Top-down: תהליך המתחיל מלמעלה בהגדרת המחלקה הכללית ביותר וממשיך למחלקות ספציפיות יותר.
 - Bottom-up: התהליך מתחיל מהגדרת המחלקה הספציפית ביותר ומסתיים בקיבוץ המחלקות למחלקה כללית.
 - Middle-out: זהו שילוב של שני התהליכים הקודמים, ובמסגרתו מגדירים מושג טיפוס מייצג למחלקה, כמו למשל "כלב", ומשם הולכים ומרחיבים את הקטגוריה שאליה הוא מתייחס, כמו במקרה שלפנינו, מה"כלב" עוברים ל"יונק" במעלה ההיררכיה, או שמצמצמים את הקטגוריה מ"כלב" ל"כלב זאב" במורד ההיררכיה וכן הלאה.
4. הגדרת מאפייני המחלקות, על פי שתי חלוקות: א) תכונות של המחלקה כמו שם, צבע וטעם. ב) מאפיינים שקושרים את המחלקה למחלקות אחרות, כמו למשל פריט שמיוצר במחלקה אחרת (לדוגמה, הקשר בין יין ליקב). למאפיינים יש תחום המגדיר את המחלקה אליה שייך המאפיין וטווח שמגדיר את המחלקות שמהן מגיעים הערכים של המאפיין. המאפיין צריך להוסיף מידע חדש למחלקה.
5. הגדרת הגבלות על פי המאפיינים: למאפיינים יש הגבלות המתארות את סוג הערך החוקי אשר יכול לקבל המאפיין. כך לדוגמה, הערך של המאפיין "שם" יהיה מסוג מחרוזת (ערך אלפביתי), הערך של המאפיין שעוסק במספר הצאצאים יהיה "נומרי" והערך של מאפיין תאריך הלידה יהיה מסוג "תאריך".
6. אכלוס האונטולוגיה על ידי הגדרת המופעים: יש לבחור את המחלקה, וליצור מופע ספציפי שלה על ידי מילוי ערכי המאפיינים של המחלקה עם ערכים ספציפיים של המופע. למשל, "ירדן 2013" הוא מופע ספציפי של המחלקה "יין" ומאפייניו הם צבע - אדום, טעם - יבש, שנת ייצור - 2013.
7. הגדרת חוקי ההיסק: הגדרת מחלקות וקשרים סמנטיים בצורה פורמלית וחד משמעית מאפשרת בניית סוכן אוטומטי שמסוגל להבין את משמעותן של הישויות באונטולוגיה ולפרשן, ובכך להגביר את השימוש החוזר בה. יתרה מזאת, ניתן להגדיר חוקי היסק לוגיים שבאמצעותם ניתן להסיק קישורים חדשים בין מחלקות שלא קודדו באונטולוגיה הבסיסית

מלכתחילה. חוק היסק הנובע מהגדרת ההיררכיה של מחלקות הוא תורשה. לפי חוק זה, אם מחלקה א' הינה מחלקת על (כללית יותר) ומחלקה ב' הינה תת מחלקה (ספציפית יותר) של המחלקה א', הרי שמחלקה ב' יורשת את כל המאפיינים של המחלקה א' וכן כל מופע של מחלקה ב' הינו בהכרח מופע של מחלקה א'. זאת ועוד, ניתן להגדיר חוקי היסק נוספים בהתאם לתחום הידע. לדוגמה, אם האונטולוגיה מכילה את הקשרים (עובדות) הבאים: "האולימפיאדה השלישית נערכה בסנט לואיס", "סנט לואיס נמצאת בארצות הברית", נוכל, באמצעות חוק היסק מתאים, להסיק ש"האולימפיאדה השלישית נערכה בארצות הברית". יש לסייג ולומר, שאין אונטולוגיה אחת ויחידה שהיא נכונה לכל תחום. מכיון שנקודת המבט והבנתו את התחום של יוצר האונטולוגיה ישפיעו על מבנה האונטולוגיה, בניית האונטולוגיה היא תהליך יצירתי ואיטרטיבי, ועל כן, אין לו כללים חד משמעיים. כתוצאה מכך, שתי אונטולוגיות שנוצרו על ידי שני מומחים שונים, יהיו שונות זו מזו. לכן, בשנים האחרונות הוצעו מודלים אונטולוגיים המשלבים היבטים, הקשרים, ממדים ודעות שונות, אשר משרתים פעולות ומטרות שונות של המשתמש בתחום הידע (Carmagnola et al., 2005; Djakhdjakha et al., 2012). כך למשל, (Zhitomirsky-Geffet et al., 2010) הציעו אונטולוגיה רבת פרספקטיבות עבור תמונות; האונטולוגיה שנבנתה משקפת היבטים שונים של תמונות, כמו היבט תיאורי, גיאוגרפי, היסטורי, אמנותי, פוליטי, תרבותי ונושאי. במחקר אחר, בתחום התזונה, נבנתה אונטולוגיה מרובת דעות (Zhitomirsky-Geffet and Erez, 2014) המשקפת את המחלוקות בין החוקרים בדבר השפעתם של מוצרים שונים כמו מוצרי חלב, בשר וסויה, על בריאות האדם. בעבודה נוספת נבנתה אונטולוגיה רב ממדית לפתגמים, שכידוע מורכבים מממד מילולי פשוט, וממד מטאפורי המייצג את המשמעות והפירוש של הפתגם (Zhitomirsky-Geffet et al., 2016).

לבסוף, ניתן לבחון את איכותה של האונטולוגיה שנבנתה באמצעות השימוש בה, דהיינו, באמצעות האפשרויות השונות שהיא פורשת ליישומים ייעודיים כגון אחזור מידע, מענה על שאלות, סיכום מידע אוטומטי, שירותים אוטומטיים חכמים, תיוג סמנטי של מידע ועוד. מחקרים מראים שמחפשי מידע מעוניינים בדעות ובפרספקטיבות שונות על הנושא שמעסיק אותם, ולכן שימוש באונטולוגיה המשקפת פרספקטיבות ונקודות מבט מגוונות תתרום לאפליקציות ולשירותים הללו (Paul et al., 2010; Liu et al., 2012; Ouzrout et al., 2009; Powell and French, 1998; Geryville, 2007).

רשת סמנטית (semantic web) ונתונים מקושרים (Linked data)

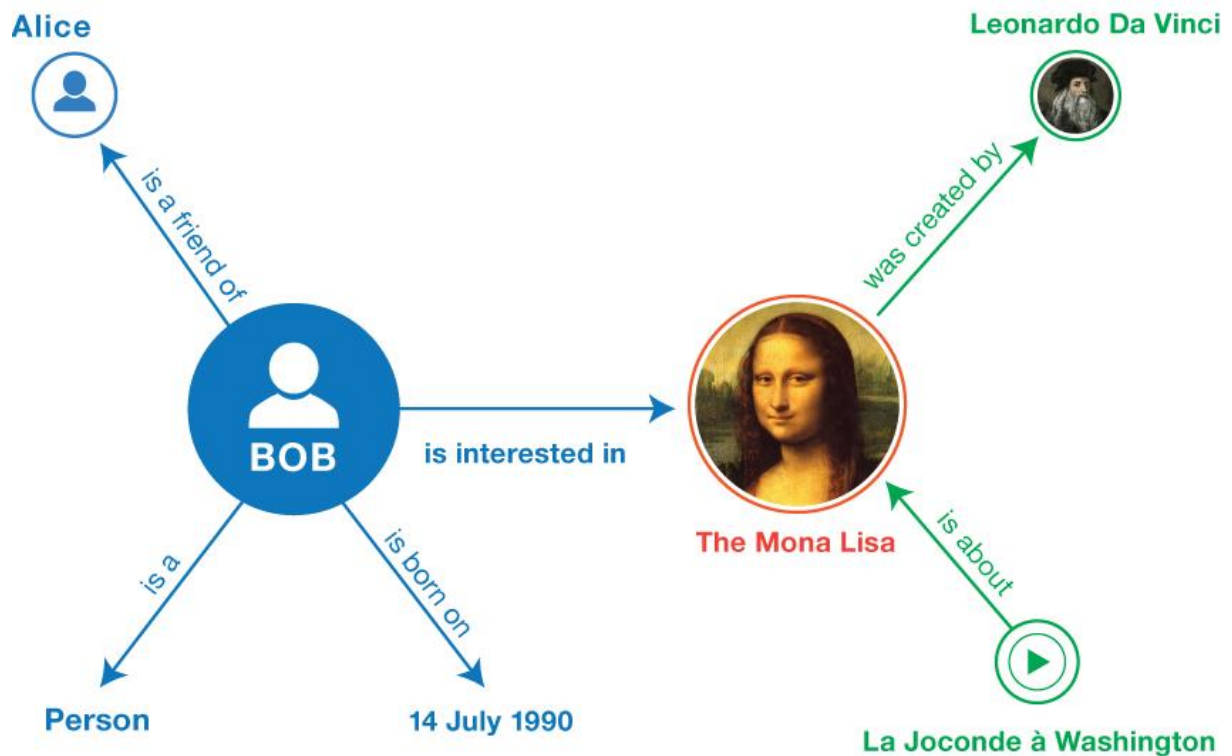
דמיינו שכל המידע העצום ברשת האינטרנט היה מקושר ויוצר מעין מסד נתונים אחד גלובלי ומוגדר היטב, כזה שכל מלה בו מגדירה מושג אחד שיש לו משמעות אחת וכל מושג (משמעות) מוגדר על ידי מחלקה (או מופע) סטנדרטיים מסוימים באונטולוגיה של תחום הידע שמפורסמת ברשת.

המחלקות והמופעים הללו קשורים זה בזה בקשרי משמעות שגם הם מוגדרים היטב באופן חד משמעי כחלק מאחת האונטולוגיות המפורסמת ברשת. כתוצאה מכך, הסוכנים האוטומטיים (תוכנות מחשב מיוחדות) הופכים לסוכנים חכמים. סוכנים אוטומטיים חכמים היו יכולים לדבר זה עם זה, להעביר ולשתף ביניהם מידע זה באמצעות שפה משותפת ומוסכמת, ולעבד במהירות את כל המידע המגיע ממקורות שונים ברשת, לבצע היסק לוגי מורכב ולהשיב תשובות מדויקות לשאלות מורכבות, ולעשות את כל הדברים הללו ללא התערבות אנושית. מכיוון שרוב עבודת ניפוי המידע, עיבודו, וסינונו ייעשה באופן אוטומטי, ניתן יהיה לחסוך מהמשתמשים בשירותים החכמים הללו מאמץ רב וזמן יקר. זהו, בתמצית, החזון של הרשת הסמנטית המכונה גם web3.0 שהוצע על ידי ממציא הווב טים, ברנרס-לי. הממציא, ברנרס-לי, חפץ ליצור רשת אינטרנט שמכילה מידע מובנה ומתויג סמנטית, באופן כזה, שגם מכונות ולא רק בני האדם יוכלו להבין את המידע ולנתחו (Berners-Lee et al., 2001). כדי לאפשר את הגשמת החזון הזה יש להגדיר את האונטולוגיות על כל רכיביהן באופן פורמלי ולפרסם הגדרות אלו ברשת. אם ברשת הסינטקטית שכולנו מכירים (web 1.0) לכל דף יש מזהה URI ייחודי והקשרים נוצרים בין דפים (שרובם אינם מובנים והמידע בהם אינו מובן לסוכנים אוטומטיים), אזי ברשת הסמנטית לכל אונטולוגיה יש מזהה URI ייחודי. בדומה לזאת, גם לכל נתון (מחלקה או מופע באונטולוגיה), מאפיין או קשר סמנטי ששייכים לאונטולוגיה כלשהי יש מזהה URI משלו. ההנחה היא, שהכתובת ה-URL התואמת למזהה הזה כוללת גם הגדרה פורמלית של אותה המחלקה, המופע או הקשר האונטולוגי שיכול לשמש את הסוכנים האוטומטיים. יתרה מזאת, ברשת הסמנטית מקושרים זה עם זה לא רק הדפים, אל גם הנתונים המובנים הללו. נתונים בעלי זיהוי ייחודי שמקושרים זה עם זה בקשרים סמנטיים כאלה נקראים "נתונים מקושרים" (Linked data) והם תשתיתה של הרשת הסמנטית.

כדי להפוך גם דפי אינטרנט לא מובנים שכתובים בשפה טבעית למובנים לסוכנים אוטומטיים יש להצמיד למושגים המופיעים בהם קישור או תיוג למחלקות, או למופעים מקבילים להם באחת האונטולוגיות המפורסמת ברשת. כך לדוגמה, ליד המלה "תל אביב" ניתן להוסיף את הקישור למזהה של המחלקה "city" מאונטולוגיה של ישויות גיאוגרפיות, וכן את הקישור למזהה של המופע "Tel Aviv", אם קיים באונטולוגיה כלשהי, ובאמצעות הקישורים הללו להבהיר לסוכנים האוטומטיים את משמעות המלה על כל מאפייניה וקשריה.

כדי לאפשר הגדרה פורמלית של אונטולוגיות פיתח ארגון התקנים הבינלאומי W3C כמה סטנדרטים וטכנולוגיות להוספת סמנטיקה לרשת האינטרנט. העיקריים שבהם הם (Resource RDF (Resource Description Framework Schema) RDF/S i Description Framework (Brickley & Guha, 1999) המבוססות על שפת XML (Extended Markup Language) להגדרת נתונים מובנים. ברשת הסמנטית מתייחסים למושגים ואובייקטים בעולם כמשאבים (Resources), מכאן, שהשפות הן "מסגרת לתיאור משאבים" ו"סכמה של המסגרת לתיאור משאבים". משאב יכול להיות כל מושג או אובייקט שמישהו ירצה לדבר עליו. אבני הבניין של שפת RDF הן שלשות

מהצורה הבאה: נושא (subject) – קשר (predicate) - מושא (object) המקשרות שני מושגים או אובייקטים בעולם (הנושא והמושא) על ידי קשר סמנטי מסוים. שלשות אלו מהוות עובדות על תחום הידע ואוסף של שלשות כאלה מגדיר את בסיס הידע בתחום באופן פורמלי ומסודר. תרשים מספר 2 מדגים את גרף השלשות של RDF.



תרשים מספר 2: דוגמה לקטע מגרף השלשות של RDF. מקור: RDF Primer,

<https://www.w3.org/TR/rdf11-primer/#fig1>

השלשות שמופיעות בגרף הן: בוב - מופע של - אדם, בוב - חבר של - אליס, בוב - נולד ב- 14.7.1990, בוב - מעוניין ב- מונה ליזה, מונה ליזה - נוצרה על ידי - ליאונרדו דה וינצ'י, ג'וקונדה מושינגטון - מספר על - מונה ליזה.

RDF היא אבן היסוד של הווב הסמנטי. בעזרת RDF ניתן לגשת לנתונים ללא צורך בבסיס נתונים מרכזי בעל סכמה מוגדרת. יישומי RDF שומרים נתונים כשלשות RDF ומקשרים נתונים ממקורות שונים על בסיס מושגים משותפים. היתרון של RDF על פני בסיסי נתונים אחרים הוא גמישות שמתבטאת ביכולת של RDF להוסיף מושגים חדשים באופן דינמי וכן היכולת לקשר בין מושגים מאונטולוגיות ומלקסיקונים שונים המאוחסנים במקומות שונים ברשת. ב-RDF גם ניתן לקשר בין כינויים שונים של מושג המכונה באופן שונה באונטולוגיות ולקסיקונים שונים, מבלי להעדיף את אחד הכינויים על פני אחרים.

בשפת RDF/S המשלימה את RDF ניתן להגדיר את הסכמה הסמנטית הכללית של השלשות (העובדות) המתוארות באמצעות RDF. כלומר, ניתן להגדיר מחלקות מופשטות ולא רק מושגים (מופעים) ספציפיים ואת תכונותיהם ברמה הגנרית בהתאם למידע המבוטא בשלשות RDF. לדוגמה, אם השלשות מכילות מידע על תל אביב ועל מאפייניה כגון, תל אביב - ממוקמת בישראל, תל אביב - בעלת אוכלוסייה של 432,000 נפש, רחוב דיזנגוף - נמצא בתל אביב, אזי בשפת RDF/S ניתן להגדיר את המחלקות "רחוב", "עיר" ו"מדינה" כתת מחלקות (הקשורות בקשר IS-A) של המחלקה הכללית "אובייקט גיאוגרפי" ולהגדיר את הקשר "נמצא" ב- כמקשר בין מחלקות הללו. שפה נוספת מאותה המשפחה המאפשרת להחיל מגבלות על מאפיינים באונטולוגיה וחוקי היסק לוגיים על ההגדרות של RDF/S היא שפת (Web Ontology Language) OWL (McGuinness and Van Harmelen, 2004).

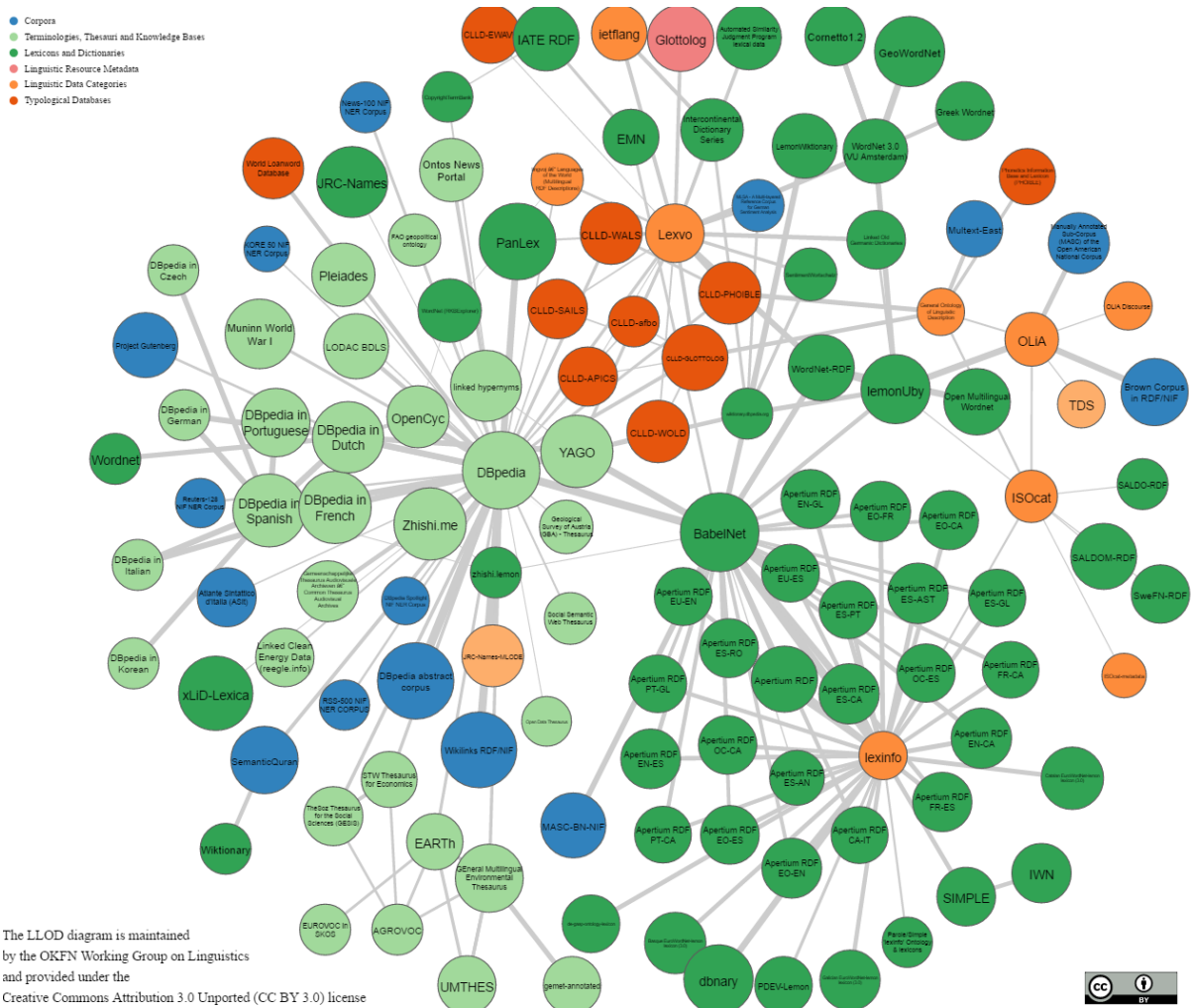
ככל שנדע לחבר בין נתונים ממאגרי מידע שונים שייתכן שהם נמצאים בשרתים שונים, כך יהיה הגרף שנצייר שלם וגמיש יותר. לשם כך, כל שרת צריך להיות מסוגל לחשוף כל ישות על כל מאפייניה.

יתרה מזאת, פרסום אונטולוגיות ברשת האינטרנט מאפשר קיום קשרים בין אונטולוגיות שונות והקמת מסד נתונים גלובלי מבוזר מבוסס RDF. מנועי חיפוש סמנטיים כמו Virtuoso Universal Server - <http://virtuoso.openlinksw.com/> מאפשרים לשלוח מידע ממסד זה באמצעות שפת שאילתות ייעודית לשליפת מידע משלשות RDF המכונה-[http://www.w3.org/TR/rdf-sparql-](http://www.w3.org/TR/rdf-sparql-query/) (SPARQL) query/ כתחליף למסדי נתונים לוקליים עם שפת שאילתה SQL. יתרונה של שפת SPARQL ועוצמתה טמונים ביכולתה לשלוח מידע ממספר רב של מאגרי שלשות RDF ואונטולוגיות בעת ובעונה אחת באמצעות שאילתה אחת מפורטת וממוקדת. כך הופכות הטכנולוגיות הסמנטיות שתוארו לעיל את רשת האינטרנט למערך נתונים מובנים ומקושרים במקום דפי טקסט לא מובנים ומקושרים.

אונטולוגיות בתחומי ידע שונים

בשנים האחרונות פותחו אונטולוגיות רבות בתחומי ידע שונים על ידי גופים שונים ופורסמו ברשת. עם הגופים שעוסקים בנושא ניתן למנות אוניברסיטאות (כגון, protégé ontology library – ספרייה כללית שניתן להעלות אליה אונטולוגיות בתחומים מגוונים שנוסדה בסטנפורד http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library), מיזמים חברתיים של W3C (כגון, - SIOC, <http://xmlns.com/foaf/0.1/>, FOAF - <http://www.w3.org/Submission/sioc-spec/>), פרויקטים ציבוריים (כגון, DBpedia – <http://dbpedia.org/About>, אונטולוגיה כללית ומקיפה ביותר שנבנתה מתוך המידע של ויקיפדיה), (Auer et al., 2007) ומשרדי ממשלה שמפרסמים את הנתונים שלהם בצורת נתונים מקושרים מובנים וסמנטיים (SemanticGov, www.data.gov/semantic). אונטולוגיות מפורסמות ברשת

ומקושרות ביניהן באמצעות קישורים בין מושגים שקולים זהים, ומהקישורים הללו נוצר ענן של אונטולוגיות שבמרכזו ה-DBPedia, שהיא האונטולוגיה הגדולה מביניהן והמקושרת לכולן, כפי שניתן לראות בתרשים מספר 3.



תרשים מספר 3: ענן הנתונים המקושרים ברשת האינטרנט, מקור:

<https://upload.wikimedia.org/wikipedia/en/f/f1/LLOD-Cloud-2016-05-24.png>

לאחרונה, ספריות הפכו ליצרן משמעותי של נתונים מקושרים ומידע סמנטי. כך למשל, פרסום הספרייה הלאומית של בריטניה (<http://www.bl.uk/bibliographic/datafree.html>) וכן ספריית הקונגרס האמריקאי (<http://id.loc.gov/ontologies/bibframe.html>) את המידע מהקטלוגים שלהן בצורת נתונים מקושרים כשלושות RDF. אונטולוגיות וסטנדרטים הנפוצים ביותר בתחום הביבליוגרפיה הם:

Dublin core (ISO 15836:2009, <http://dublincore.org/>), BIBO (D'Arcus & Giasson, 2009), CITO (Peroni & Shoffon, 2012), FRBR (Tillet, 2004), FRBRoo (http://www.cidoc-crm.org/frbr_inro.html)).

בתחום של מורשת תרבותית לאומית ומדעי הרוח הדיגיטליים פותחו בעשור האחרון שתי אונטולוגיות מרכזיות כלליות: CIDOC-CRM (Doerr, 2003, <http://www.cidoc-crm.org/>) ו-EuropeananaData Model (Winer, 2011, http://dm2e.eu/files/DM2E_Model_V1.2.pdf). מודל קונספטואלי מונחה עצמים CIDOC-CRM (CIDOC object-oriented Conceptual Reference Model) פותח על ידי קבוצת ICOM/CIDOC Documentation Standards Group והפך לסטנדרט בינלאומי של ISO 21127:2014. מטרת הפרויקט היתה ליצור אונטולוגיה משותפת שתחבר באופן סמנטי בין מקורות מידע לוקליים שונים של מורשת תרבותית לאומית שהיו מנותקים עד כה, ותהפוך אותם לנתונים מקושרים. אונטולוגיות אלו היו בשימוש והורחבו על ידי פרויקטים לאומיים רבים, כגון הפרויקט של המוזיאון הבריטי לארגון מורשת תרבותית ולדיגיטציה שלה, וחיבורו עם נתונים אחרים של מורשת תרבותית (<http://collection.britishmuseum.org>). בענף האונטולוגיות ברשת יש גם אונטולוגיה של השואה שמכילה מידע מובנה על עשרות אלפים מקורבנותיה של שואת יהודי אירופה (Mazzini and Brazzo, 2015). בנוסף לזאת, ספריות גם משתפות פעולה אלה עם אלה ביצירת קובצי זהויות המכילים רשימות של שמות פרטיים של אנשים ומקומות ובפרסומם. הקבצים הללו כוללים צורות כתיב שונות של כל שם, כתיב בשפות שונות ושמות היסטוריים של מקומות. המפורסמים שבקובצי הזהויות הם: Geonames (<http://www.geonames.org/ontology>), GND (http://en.wikipedia.org/wiki/Integrated_Authority_File), VIAF (<http://viaf.org/>), Library of Congress subject headings (<http://id.loc.gov/>), AAT (Getty thesaurus - <http://www.getty.edu/research/tools/vocabularies/aat/>).

בתחום הביו-רפואה יש ספריות של מאות אונטולוגיות (<https://bioportal.bioontology.org>) המכסות תת-תחומים רבים של המקצועות הקשורים לרפואה, כגון אונטולוגיית הגנים the Gene ontology (<http://amigo.geneontology.org/amigo>), אונטולוגיית התרופות DRON (<https://bioportal.bioontology.org/ontologies/DRON>) והקשרים ביניהם SNOMED CT (<https://www.nlm.nih.gov/healthit/snomedct/>). אחד התחומים הנתרמים ביותר מהסמנטיזציה של הרשת הוא תחום חיפוש מידע ואחזור. על כן, חברת גוגל בונה בשנים האחרונות את Google knowledge graph (<https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html>) שמזהה אובייקטים ברשת ואת הקשרים הסמנטיים ביניהם. גוגל משתמשת בגרף שנוצר כדי להחזיר תוצאות חיפוש מובנות סמנטית במקום רשימה ארוכה של דפים. כך למשל, תשובה לשאלתה על

אוניברסיטת בר אילן תניב "תיבת מידע" עם תיאור של האוניברסיטה או הגדרה קצרה שלה ובהמשך יופיעו נושאים קשורים אליה סמנטית, כמו אנשים שלמדו בה ואוניברסיטאות נוספות בישראל. יוזמה משותפת נדירה של שלוש ענקיות החיפוש, גוגל, מייקרוסופט ויאהו, שיצאה ב-2011 היא אתר סכמה (www.schema.org). באתר זה פורסמה אונטולוגיה גנרית כללית המכסה בעיקר (אך לא רק) מחלקות וקשרים מעולם המסחר האלקטרוני, וניתנת להרחבה דינמית על ידי קהל המשתמשים. כדי לעודד מפתחי אתרים ומשתמשי רשת להוסיף מבניות וסמנטיקה למידע שהם מעלים לרשת, הצהירו החברות הללו שיקדמו אתרים שיתייגו את המידע שלהם באמצעות המחלקות מאונטולוגיה ז.

אונטולוגיות כמודל לארגון מידע וידע

בנוסף לתפקידן המרכזי בהפיכת הרשת לסמנטית ומובנת לסוכנים אוטומטיים, אונטולוגיות גם מהוות בסיס למערכות מידע בתחומי ידע שונים. גרובר (Gruber, 2009) התייחס לאונטולוגיה כמגדירה סדרת רכיבים תיאוריים, שניתן באמצעותה לתאר תחום ידע. למעשה, אונטולוגיה שמכילה גם מופעים ספציפיים של מחלקות מהווה בסיס ידע. פיתוח אונטולוגיה דורש הבנה אנושית של התחום, לוגיקה, הנמקה, ובהירות לגבי השימוש המיועד לה. התהליך הארגוני של הגדרה והגעה להסכמה על טרמינולוגיה ומערכות יחסים באונטולוגיה משפר את בהירות המשמעות של מה שחשוב בארגון ובתחום הידע. הערך הנובע מאונטולוגיה יכול להתבטא בכמה וכמה דרכים:

1. בהירות משופרת של מודל המידע המסייעת להבנה אחידה.
2. הבנה משותפת של מבנה המידע של אנשים וסוכני תוכנה.
3. שימוש חוזר במידע בתחום מסוים.
4. הפשטה, שתוביל להפחתת קונפליקטים וכפילויות במידע.
5. ניתוח של מידע, אחזור וייצוג הוויזואלי בצורה אוטומטית וידידותית למשתמש.
6. יכולת פעולה הדדית ושילוב בין מערכות ונתונים בעלות ובסיכון נמוכים יותר.
7. עלויות תחזוקה מופחתות הודות לארכיטקטורה מבוססת מודל ידע המקילה על ההסתגלות לנסיבות משתנות ומאפשרת הוספת יכולות חדשות.

לסיכום, אונטולוגיה מאפשרת יישום אוטומטי של לוגיקה וחוקי היסק בדרכים המפחיתות מורכבות מיותרת או משפרות את יעילות הפתרונות של הארגון (Noy & McGuinness, 2001).

גישות שונות בבניית אונטולוגיה

כפי שפירטנו בפרקים הקודמים, בניית אונטולוגיות הכרחית לקידום הרשת הסמנטית וכן מסייעת לארגון מידע וידע, אלא שתהליך פיתוח האונטולוגיה ברמה הפרקטית הוא תהליך מורכב למדי. בפועל, יש כמה גישות מעשיות לבניית אונטולוגיות ולפיתוחן, כפי שנפרט בסעיפים הבאים.

בנייה ידנית של אונטולוגיות

הגישה הראשונה והמדויקת ביותר לבניית אונטולוגיה היא בנייה ידנית על ידי מומחים לתחום הידע שבו היא עוסקת (Sarasua et al., 2012). בדרך כלל, התהליך מצריך מעורבות של מומחים בתחום הידע שאותו תייצג האונטולוגיה וכן את מעורבותם של מומחים לבניית אונטולוגיות, Kotis & Vouros, (2006). דוגמה לאונטולוגיה גדולה שנבנתה ידנית על ידי קבוצת לקסיקוגרפים ופסיכולוגים בלשניים היא WordNet (Fellbaum, 1998). זהו לקסיקון לשפה האנגלית, שפיתוחו ארך כ-20 שנה, והוא מכיל יותר מ-100,000 מושגים, כשהמשמעויות השונות של כל מלה שמופיעה בו מפורטות כמושגים נפרדים. בנוסף לזאת, לכל מושג (משמעות) מצורפת רשימת המלים הנרדפות שלו וכן מלים שקשורות אליו בקשרי משמעות שונים, כגון, קשרים היררכיים של כלל ופרט ושלם מול חלקיו. אמנם, כדי להבטיח מבנה לוגי עקבי ותואם לחוקים, אונטולוגיות צריכות להיבנות על ידי מומחים להנדסת אונטולוגיות, אולם מחקר קודם מצביע על כך שיש לשתף גם מומחים לתחום הידע בתהליך בניית אונטולוגיה (Denaux et al., 2011; Kalbasi et al., 2014). הקושי בשילובם של מומחים אלה נובע מהעובדה שעל פי רוב אין להם ידע והבנה במידול אונטולוגיות. הם בדרך כלל נעדרים כל ניסיון בהנדסת ידע, ואינם מעוניינים ללמוד פרטים טכניים על אודות RDF, OWL או לוגיקה פורמלית. אוצר המלים המשמש אותם לתקשורת עם עמיתיהם אינו פורמלי, ועשוי להכיל מלים רבות משמעות וכינויים שונים לאותה תופעה. לעומת זאת, רוב הכלים לבניית אונטולוגיות נועדו לשימוש על ידי מומחים בעלי ידע הנדסי מתאים ומיומנויות לוגיות, כמו DRUPAL (Corlosquet et al., 2009) ו-PROTÉGÉ (Noy et al., 2001). כלים אלו אינם ידידותיים ואינטואיטיביים לשימוש מספיק עבור מומחים בתחום הידע שאין להם רקע בהנדסת אונטולוגיות ומידול ידע (Fluit, Sabou & Van, 2004). אולם מומחים למידול אונטולוגיות הם בדרך כלל חסרי יכולות בתחום הידע הרלוונטי ונעדרים מומחיות בו.

לכן הגישה האופטימלית היא שמומחי תחום הידע יתרמו בשלבים הבאים:

1. זיהוי טווח האונטולוגיה ומטרתה.
2. איסוף ידע ומקורות מידע, כולל אונטולוגיות קיימות לשימוש חוזר.
3. זיהוי צרכי המשתמש והגדרת שאלות שהאונטולוגיה תוכל לענות עליהן, מתוך הנחה כי קיימת חפיפה, לפחות חלקית, בין מומחי תחום הידע לבין משתמשי הקצה.
4. הגדרת מבנה של אונטולוגיית על המורכבת ממושגי ליבה, שהם מושגים כלליים ומופשטים ביותר בתחום וקשרי משמעות גנריים ביניהן. זה כולל גם בניית היררכיית מושגים בסיסית או טקסונומיה של תחום הידע.

מהנדסי אונטולוגיה מנחים את המומחים לתחום הידע, מבקרים מבנים אונטולוגיים ומאמתים אותם מבחינת עקביות לוגית של האונטולוגיה שנוצרת. כמו כן, הם ממירים מושגים שהוגדרו על ידי המומחים לתחום הידע למחלקות ומופעיהם ועושים זאת גם לקשרים למאפיינים אונטולוגיים (Properties) בשפה פורמלית כמו RDF/S (Denaux et al., 2011). כלומר, מומחי תחום הידע

אחראים לכך שהאונטולוגיה תשקף את התחום ותהיה נטולת טעויות מצד הידע בתחום. המומחים לבניית אונטולוגיות אחראים לעקביות האונטולוגיה וליעילותה.

היתרונות של בניית אונטולוגיה בצורה ידנית על ידי מומחים לאונטולוגיות ולתחום הידע הם דיוק ועקביות, לצד איכותו הגבוהה של המידע המתקבל ובקרה אנושית שמתבצעת בכל שלבי הפיתוח. לעומת יתרונותיה של הבנייה הידנית, ראוי לציין, שתהליך ידני כזה הוא יקר, תובע זמן רב ומאמץ אנושי והעובדות הללו הופכות אותו לבלתי נגיש לארגונים רבים. בנוסף לזאת, אונטולוגיה כזאת בדרך כלל תהיה מוגבלת מבחינת הנושאים שהיא יכולה לכסות, ולעתים גם תשקף דעה סובייקטיבית של הקבוצה המצומצמת של מומחים שפיתחו אותה, ולא תמיד תיקח בחשבון את דעתם של המשתמשים בה. לכן, כדי להתגבר על מגבלות אלו של הגישה הידנית, פותחו שיטות אוטומטיות וסמי-אוטומטיות לבניית אונטולוגיות כפי שנתאר בסעיפים הבאים.

בנייה אוטומטית של אונטולוגיות

הגישה השנייה לפיתוח אונטולוגיות היא בנייה אוטומטית של אונטולוגיות. לשם כך, מופעלות טכניקות של למידת מכונה לניתוח שפה טבעית. באמצעות טכניקות אלו מושגים אונטולוגיים וקשרים סמנטיים ביניהם נשלפים מתוך קורפוס של טקסט גדול המייצג תחום ידע נתון. בבסיס השיטות האלו עומד ניתוח סטטיסטי של הופעות משותפות של מלים בתוך הטקסט, לעתים תוך שילוב תבניות וחוקים בלשניים. שתי גישות מרכזיות מוצעות בספרות לבניית אונטולוגיה בצורה אוטומטית:

1. למידת מכונה מבוקרת עבור סוגים נתונים מראש של קשרי משמעות בין מלים בטקסט כמו פרט-כלל (hyponymy), חלק-שלם (meronymy) או קשרים תמטיים ספציפיים לתחום הידע כמו "מחבר של", "נכתב על ידי", "מיוצר ב" ועוד. כדי לאתרם מחפש אלגוריתם הלמידה בקורפוס של טקסט הופעות משותפות של מלים בתוך תבניות לקסיקליות-סינטקטיות (Hearst, 1992; Mirkin et al., 2006; Kozareva and Hovy, 2010; Panchenko et al., 2016). כך לדוגמה, התבנית "מושג א' כמו מושג ב', מושג ג' ומושג ד'" (כגון, "חברות היי טק כמו IBM, מייקרוסופט ואינטל") מייצגת קשר "סוג של" או "מופע של" בין מלה א' לבין מלה ב', ג' וד' במשפט. מכאן ניתן להסיק, ש-IBM, מייקרוסופט ואינטל הן מקרים פרטיים או מופעים של חברות היי טק.

2. למידת מכונה לא מבוקרת לצורך זיהוי מלים שדומות זו לזו במשמעותן מתוך קורפוס גדול של טקסט (Lin, 1998; Maedche & Staab, 2012; Geffet and Dagan, 2005; Gherasim et al., 2013). העיקרון המנחה של טכניקות אלה הוא עיקרון הדמיון ההתפלגותי שמבוסס על הנחת עבודה שמלים דומות במשמעותן מופיעות בהקשרים דומים או בסביבות מלים דומות בטקסט מכיוון שהן משמשות לתיאור דברים דומים זה לזה ("מכונית לבנה נסעה בכביש", "פג'ו לבנה נסעה בכביש", "רכב לבן נסע בכביש"). לשם כך, עבור כל מלה נבנה "וקטור תכונות", שאינו אלא אוסף של הקשרים (מלים וביטויים שמופיעים יחד עם המלה הנתונה בטקסט). לאחר מכן, האלגוריתם משווה את הווקטורים של

המלים השונות על ידי מטריקות דמיון הסתברותיות וכך מחשב את מידת הדמיון הסמנטי בין המלים. שיטה זו מוצאת מלים דומות שאינן מופיעות יחד בטקסט אך הקשריהן דומים, כמו לדוגמה מלים נרדפות או מלה כללית יותר או ספציפית יותר מאותו סוג, שיופיעו פעמים רבות בתוך משפטים או בסביבות מלים דומות זו לזו.

לעתים יש כמה אונטולוגיות לאותו תחום ידע. כדי להגיע לייצוג מוסכם ושלם של תחום הידע יוצרים אונטולוגיה מאוחדת אחת מתוך האונטולוגיות הרבות הללו, או שממפים רכיבים זהים באונטולוגיות שונות ומקשרים ביניהם. המתודולוגיה הגנרית למיזוג אונטולוגיות באופן אוטומטי כוללת את השלבים הבאים (Euzenat and Valtchev, 2003; Euzenat and Shvaiko, 2013):

1. השוואת שמותיהם של מחלקות, מופעים ומאפיינים מאונטולוגיות שונות כדי לזהות רכיבים זהים.
 2. השוואת שמותיהם של מחלקות, מופעים ומאפיינים מאונטולוגיות שונות כדי לזהות רכיבים עם שמות חופפים חלקית אך לא זהים לחלוטין.
 3. זיהוי מחלקות עם שמות שאינם זהים אך דומים במשמעותם (כאלה המקיימים ביניהם יחסי נרדפות, או יחסי כלל ופרט) מאונטולוגיות שונות.
 4. השוואת מיקומיהן של המחלקות שזוהו כחופפות בסעיפים 1-3 בהיררכיות של האונטולוגיות השונות או בטקסונומיות שלהן.
 5. השוואת הגדרותיהם של מאפייני המחלקות שזוהו כחופפות בסעיפים 1-4 (כגון מספר הערכים של כל מאפיין והסוג שאליו הוא שייך) מאונטולוגיות שונות.
 6. השוואת המופעים של המחלקות שזוהו כחופפות בסעיפים 1-5 מאונטולוגיות שונות.
- במקרה של מיזוג אונטולוגיות שונות יש גם לקחת בחשבון נקודות מבט שונות שעשויות להשתקף באונטולוגיות הממוזגות. יש שתי גישות בסיסיות שמתמודדות עם נקודות המבט השונות המשתקפות באונטולוגיות (de Bruijn et al., 2006): הראשונה היא איחוד אונטולוגיות – כאשר האונטולוגיה הסופית מכילה את כל הרכיבים שקיימים בכל אונטולוגיות המקור השונות שמוזגו, והשנייה היא חיתוך אונטולוגיות – כאשר האונטולוגיה הסופית מכילה רק את הרכיבים המשותפים שמופיעים באונטולוגיות המקור.
- היתרונות של הגישות האוטומטיות ליצירת אונטולוגיות הם מהירות הביצוע, כיסוי רחב וחיסכון בעבודה אנושית. ואולם יש לגישות האוטומטיות גם מחיר, והוא "רעש" (מושגים, מאפיינים וקשרים שאינם רלוונטיים) וחוסר דיוק באונטולוגיה. על כן, הגישה האופטימלית צריכה לשלב כלים טכנולוגיים ואלגוריתמים של למידת מכונה עם אינטלקט אנושי, ידע ובקרה אנושיים (Simperl and Luczak, 2014). כלומר, הגישה האופטימלית היא גישה סמי-אוטומטית שמשלבת בין ידע ועבודה אנושית לכלים או אלגוריתמים אוטומטיים. אחד המחקרים המוקדמים בנושא זה (Maedche and Staab, 2002) הציג מנשק גרפי לעריכה ידנית של אונטולוגיות בצירוף עם כלי עזר אוטומטי ששולף ישויות אונטולוגיות (מושגים ומאפיינים) מתוך הטקסט, באמצעות אלגוריתמים של עיבוד שפה

טבעית. בעבודה מאוחרת יותר, חוקרים (Denaux, Dolbear, Hart, Dimitrova, & Cohn, 2011) פיתחו מתודולוגיה המשלבת עבודה של מומחים לתחומי הידע עם כלי עזר אוטומטיים המקילים על עבודתם וממירים את המבנים האונטולוגיים שמגדירים המומחים בשפה טבעית מבוקרת לשפת אונטולוגיות פורמלית OWL.

גוב חברתי בשירות הווב הסמנטי

כפי שתיארנו לעיל, יש שלל גישות ויישומים לטכנולוגיה לאיסוף מידע וידע, לצבירתו, להפצתו ולניהולו. אלא שטכנולוגיה בלבד אינה יכולה להוות פתרון למערכות מידע וידע. ההתפתחות המשמעותית בתחום היא בהרחבת קהל צרכנים חדשים, הציבור הרחב, שרואה כמובן מאליו את זמינות המידע והאפשרויות הגלומות בו. הרחבה זו מתווספת לכך שכל צרכן מידע היום הוא גם יצרן מידע. יתרה מזאת, בעוד שהאונטולוגיות אמורות לענות על הצרכים המשתנים של המשתמשים, רוב המשתמשים בהן אינם שותפים לתהליך בנייתן (Vrandečić, Pinto, Tempich, & Sure, 2005). כתוצאה מכך, אונטולוגיות מקצועיות, למשל בתחום הביורפואי, כמו SNOMED CT, המשולבות באפליקציות מכוונות צרכנים לא מספקות תמיכה משביעת רצון, מכיוון שהן אינן מכילות את אוצר המלים המשמש הדיוטות בתחום הזה, והוא נחוץ באפליקציות אלו (Muresan and Klavans, 2013).

לכן, הגישה השלישית לבניית אונטולוגיות שנסקור כאן היא ניצול התרומה של הווב החברתי לוווב הסמנטי כדי להגיע לאיכות ולדיוק מרביים, תוך שימוש בכלי עזר אוטומטיים שיחסכו זמן ומאמץ אנושי. שתי השיטות המרכזיות בגישה זו הן שימוש בשיתוף "המונים" ובעבודת "המונים". ההבדל העיקרי בין שתי השיטות הוא, שבשיתוף "המונים" המשתתפים עובדים יחד ויכולים לעיין בשינויים ובעדכונים של כל אחד מהם באונטולוגיה ולדון בהם, ואילו בעבודת ההמונים אין קשר בין המשתתפים וגם אסור שיהיה קשר כזה, וכל אחד מהם מבצע את עבודתו לחוד באופן בלתי תלוי לחלוטין (Surowiecki, 2005).

יצירת אונטולוגיה שיתופית

בעשור האחרון הוצעו פלטפורמות שונות ליצירה שיתופית של אונטולוגיות (Maedche and Staab, 2012; Pereira, 2008; Shvaiko and Euzenat, 2013; Euzenat and Shvaiko, 2013; Simperl and Luczak-Rösch, 2014).

מכיוון שבניית אונטולוגיה אינה מהלך חד פעמי, אלא תהליך מבוקר וקבוע של שיפור האונטולוגיה בהתאם לצורכי המשתמשים, האתגר הגדול ביותר ביצירתן של פלטפורמות כאלה הוא השגת הסכמה בין המשתתפים ופתרון קונפליקטים שמתגלעים בתהליך העבודה. אפשרות אחת היא בקרה של מנהל האונטולוגיה שהוא מומחה להנדסת אונטולוגיות. במסגרת האפשרות הזאת, רק המנהל או קבוצה מצומצמת של מנהלים מומחים מחליטים אלו רכיבים שהוזנו על ידי המשתתפים שאינם

מומחים ("ההמונים") ייכללו באונטולוגיה הסופית. כך לדוגמה, (Vrandečić, Pinto, Tempich, & Sure (2005) הציעו מתודולוגיה בשם DILIGENT אשר מאפשרת למשתמשים לעדכן את האונטולוגיה שנבנית על ידי המומחים ולשנות אותה באופן דינמי אך מבוקר על ידי מנהל האונטולוגיה.

גישה אלטרנטיבית דוגלת במתן חופש למשתתפים ובנייה שיתופית ללא בקרה של מנהל האונטולוגיה. במקרה כזה, כדי להגיע להסכמה, המשתתפים יכולים להתכתב ביניהם בצ'ט ולנסות לשכנע זה את זה (Heflin, 2001; Gómez-Gauchía et al., 2002; Holsapple and Joshi, 2006; Karapiperis and Apostolou, 2008), לשנות רכיבים שונים ולעדכןם, בדומה לאופן שבו מתנהלים העניינים בוויקיפדיה, ולהצביע על רכיבים אונטולוגיים שצורתם הסופית תיקבע על פי דעת הרוב (Tudorache et al., 2008).

מקור חברתי נוסף למידע סמנטי הוא תיוגים חברתיים שיתופיים של אובייקטים ברשת, כמו אתרי שיתוף תמונות Flickr, ורשתות חברתיות כמו פייסבוק, טוויטר ואינסטגרם המשלבים גם תיוג מובנה באמצעות סולמיות (hashtags). על בסיס תגיות אלו הופעלו אלגוריתמים אוטומטיים לסינון "רעש" (תגיות מקריות ונדירות) וסיווג המידע ויצירת פולקסונומיות – אלו הן טקסונומיות או אונטולוגיות עממיות (folk – "של החברה") (Chiang et al., 2016; Bar-Ilan et al., 2010; Zhitomirsky et al., 2010; Trant, 2009; Schmitz, 2006; Mika, 2007).

יצירת אונטולוגיה באמצעות עבודת המונים

עבודת "המונים" או מיקור "המונים" (crowds work) משמשת לביצוע משימות, אשר באופן מסורתי מבוצעות על ידי מומחים, באמצעות הוצאתם החוצה לקבוצה גדולה של עובדים שאינם מומחים, תמורת תשלום סמלי (Surowiecki, 2005).

עבודה זו מבוססת על קבוצה אנונימית של עובדים שאינם מומחים בתחום, שמבצעים באופן בלתי תלוי סדרת פעולות פשוטות, כגון מענה על שאלות שיש להן כמה תשובות אפשריות שמתוכן יש לבחור אחת. תוצאות עבודתם נאגרות ומעובדות בצורה אוטומטית כדי להגיע לתוצאת ה"אמת הקולקטיבית" האופטימלית. מחקרים רבים מראים כי תוצאה זו יכולה להיות לא פחות מדויקת ואיכותית מתוצר עבודת המומחים, והיא זולה ממנה ומהירה ממנה לאין שיעור (Howe, 2006; Quinn & Bederson, 2011; Cooper et al., 2010).

גיוס ההמונים למשימות מתבצע דרך אתרי רשת המיועדים לכך, כמו Amazon Mechanical Turk (<https://www.mturk.com/mturk/welcome>) או CrowdFlower (<https://www.crowdfLOWER.com/>). כדי לשפר את איכות התוצאות, העובדים הפוטנציאליים ניגשים למבחן סינון ראשוני באתר שבוחן את בקיאותם הבסיסית בתחום ידע נתון, ורק אלה שעברו בציון מינימלי מסוים, נקראים לבצע את המשימה.

לאחרונה, מספר עבודות מחקר יישמו טכניקות של עבודת "המונים" ליצירת אונטולוגיות בתחום הרפואה והתזונה, להערכת איכותן ולמיזוגן (Sarasua, Simperl and Noy, 2012; Noy, Mortensen, Alexander and Musen, 2013; Mortensen et al., 2013; Mortensen et al., 2015; Mortensen et al., 2016; Zhitomirsky-Geffet et al., 2016). בעבודות אלו החוקרים הציגו לעובדים "ההמוניים" שאלות כגון "האם לב הוא סוג של איבר גוף", והשתמשו באלגוריתמים של למידת מכונה מבוקרת כדי לעשות אגריגציה של תשובות ולהשוות את התשובה הקולקטיבית של העובדים "ההמוניים" לדעתם של המומחים. העובדים "ההמוניים" (הלא מומחים) דייקו בשיעור של 90 אחוז או יותר ברוב המקרים. אלא שמידת הדיוק היתה תלויה באופן הצגת השאלות לעובדים ובמדד (שיטת האגריגציה) שהשתמשו בו בחישוב ההחלטה הקולקטיבית (Zhitomirsky-Geffet et al., 2013; Mortensen et al., 2016). כך למשל, רמות דיוק היו משמעותית גבוהות יותר כאשר העובדים התבקשו להעריך את דעת המומחים בנושא מאשר כשהתבקשו להביע את דעתם האישית.

דיון וסיכום

הווב הסמנטי הוא הצעד הבא לקידום ארגון המידע העצום הנמצא כיום ברשת האינטרנט, ניהולו ואחזורו. הרעיון הוא להוסיף רכיבי "משמעות" לדפי הווב ובכך להנהיר אותם הן לסוכנים אוטומטיים והן למשתמשים אנושיים. בבסיס המימוש של הרשת הסמנטית עומדות אונטולוגיות – אוספים של מונחי מפתח שמאפיינים תחום ידע מסוים שיש ביניהם קשרי משמעות מוגדרים. אונטולוגיה מהווה למעשה שפה, אוצר מלים סטנדרטי משותף לתחום ידע נתון. באמצעות אונטולוגיה אפליקציות מחשב הופכות להיות "חכמות" ויכולות להתקשר ביניהן, לחלוק מידע ולהעבירו זו לזו, ולבצע משימות מורכבות יחד ללא התערבות אנושית. אונטולוגיות גם מהוות כלי למידול נתונים וניהול ידע, ויצרות בסיס לטכנולוגיות ביג דאטה ובינה מלאכותית.

הפיכת המידע ברשת לנתונים מקושרים, בניית אונטולוגיות ופרסומן מאפשרים לשפר את אחזור המידע, לשבצו בהקשר ולהפוך את החיפוש לחיפוש סמנטי "חכם" ומותאם לצורכי המשתמש ואת הצגת תוצאותיו לידידותית ומובנית עבור המשתמש. בתחומים רבים כמו רפואה ומדעי הרוח הוספת סמנטיקה, הפיכתם של הנתונים לנתונים מקושרים ושימוש בטכנולוגיות סמנטיות אף מאפשרות לעשות מחקרים וניתוחים סטטיסטיים בקנה מידה גדול, חוצה מאגרים ומקורות מידע ובכך לגלות תופעות ותבניות חדשות שלא ניתן היה לגלותם על ידי מחקר לוקלי מסורתי המבוצע לרוב באופן ידני (Zhitomirsky-Geffet and Prebor, 2016).

למרות התפתחות הטכנולוגיות הסמנטיות, ועל אף פיתוח מואץ של אונטולוגיות בשנים האחרונות, יש עדיין תחומי ידע רבים שבהם חסרות אונטולוגיות. כך למשל, יש מחסור באונטולוגיות ובנתונים מקושרים לתחומי ידע רבים בשפה העברית ובשפות לאומיות אחרות, בייחוד במה שקשור בהיסטוריה ובמורשת תרבותית. בנוסף לזאת, דפים רבים ברשת עדיין מופיעים בצורת טקסט לא מתויג סמנטית. על כן, כחלק מעבודה עתידית, כדי לעבור מעידן המידע לעידן הידע, לפתח

אפליקציות אוטומטיות של העתיד ולהביא את רשת האינטרנט למיצוי מלוא הפוטנציאל הטמון בה, צריך להמשיך להשקיע מאמצים בסמנטיזציה של מידע ברשת בתחומים ובשפות נוספים.

ביבליוגרפיה

- ז'יטומירסקי-גפת מ., 2009, מהן מילים הדומות במשמעות וכיצד הן מסייעות לאחזור מידע? מידעת, 5, 59-69.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *Proceedings of the 8th International Semantic Web Conference (ISWC2007)*, Lecture Notes in Computer Science, 4825, pp. 722-735. Springer Berlin Heidelberg.
- Bar-Ilan, J., Zhitomirsky-Geffet, M., Miller, Y., & Shoham, S. (2010). The effects of background information and social interaction on image tagging. *Journal of the American Society for Information Science and Technology*, 61(5), 940-951.
- Berners-Lee T. (2009). *Presentation at the International Ted Conference*, https://www.ted.com/talks/tim_berniers_lee_on_the_next_web#t-740284, last accessed in September, 2016.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 28-37.
- Brickley, D., & Guha, R. (1999). Resource Description Framework (RDF) Schema Specification. *W3C Proposed Recommendation*. March 1999. Retrieved August 05, 2015 from: <http://www.w3.org/TR/1999/PR-rdf-schema-19990303/>.
- de Bruijn, J., Ehrig, M., Feier, C., Martin-Recuerda, F., Scharffe, F., & Weiten, M. (2006), Ontology mediation, merging and aligning. In: Davies, J., Studer, R., Warren, P. (Ed.), *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*. Chichester: John Wiley & Sons Ltd., pp. 95-114.
- van den Bosch, A., Bogers, T., & Kunder, M. (2016). Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics*, 107(2), 839-856.
- Carmagnola, F., Cena, F., Gena, C., & I. Torre. (2005). A multi-dimensional framework for the representation of ontologies in adaptive hypermedia systems. *AI*IA 2005: Advances in Artificial Intelligence*, pp. 370-380.
- Chiang, A. L., Vartabedian, B., & Spiegel, B. (2016). Harnessing the Hashtag: A Standard Approach to GI Dialogue on Social Media. *The American Journal of Gastroenterology*, 111(8), 1082-1084.
- Cooper, S., Khatib, F., Treuille, A., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, 466, 756-60.
- Corlosquet, S., Delbru, R., Clark, T., Polleres, A., & Decker, S. (2009). Produce and Consume Linked Data with Drupal! In *Proceedings of the 8th International Semantic Web Conference (ISWC2009)* volume 5823 of Lecture Notes in Computer Science (pp. 763-778). Springer Berlin Heidelberg.

- D'Arcus, & Giasson, F. (2009). *Bibliographic ontology specification*. Retrieved August 05, 2016 from: <http://bibliontology.com/>.
- Doerr, M. (2003). The CIDOC Conceptual Reference Module - an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3), 75-92.
- Doerr, M., Kritsotaki, A., & Boutsika, A. (2011). Factual argumentation - a core model for assertions making. *Journal on Computing and Cultural Heritage* 3(3), article no. 8.
- Denaux, R., Dolbear, C., Hart, G., Dimitrova, V., & Cohn, A. G. (2011). Supporting domain experts to construct conceptual ontologies: A holistic approach. *Journal of Web Semantics*, 9(2), 113-127.
- Djakhdjakha, L., Hemam M., & Boufaïda Z. (2012), Multi-viewpoints ontology alignment based on Description Logics. *Communications in Computer and Information Science*, 294, 109-122.
- Euzenat, J., Meilicke, C., Shvaiko, P., Stuckenschmidt, H. & C. Trojahn dos Santos. (2011), Ontology alignment evaluation initiative: Six years of experience. *Journal of Data Semantics*, 15, 158-192.
- Euzenat, J. & Valtchev, P. (2003), An integrative proximity measure for ontology alignment. *Proceedings of ISWC-2003 workshop on semantic information integration*, Sanibel Island (FL US), pp. 33-38.
- Euzenat J. & Shvaiko P. (2013). *Ontology matching*, Heidelberg, Germany:Springer.
- Fellbaum, C. 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Fluit, C., Sabou, M., & van Harmelen, F. (2004). Supporting User Tasks through Visualisation of Light-weight Ontologies, in S. Staab, & R. Studer, (eds.), *Handbook on Ontologies in Information Systems*, (pp. 1-20.)
- Geffet, M. & Dagan, I. (2005, 5-10 June). Distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics ACL 2005* (pp. 107-114), Ann Arbor, MI, USA. Association for Computational Linguistics.
- Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition*, 5, 199-220.
- Gruber T. R. (2009). Ontology. *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag.
- Gherasim, T., Harzallah, M., Berio, G., & Kuntz, P. (2013). Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application. In *Advances in Knowledge Discovery and Management* (pp. 177-201). Springer Berlin Heidelberg.
- Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, 2, 539-545.
- Heflin, J. (2001), *Towards the Semantic web: Knowledge representation in a dynamic, distributed environment*, Unpublished doctoral dissertation, University of Maryland, College Park.
- Holsapple, C. W. & Joshi, K. D. (2002). Ontology applications and design: A collaborative approach to ontology design. *Communications of the ACM*, 45(2), 42-47.
- Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine* 14, 1-4.

- Kalbasi, R., Janowicz, K., Reitsma, F., Boerboom, L., & Alesheikh, A., (2014). Collaborative ontology development for the geosciences. *Transactions in GIS* 18(6), 834–851.
- Karapiperis, S., & Apostolou, D. (2006). Consensus building in collaborative ontology engineering processes. *Journal of Universal Knowledge Management*, 1(3), 199-216.
- Kotis, K., & Vouros, G. A. (2006). Human-centered ontology engineering: The HCOME methodology. *Knowledge and Information Systems*, 10(1), 109-131.
- Kozareva, Z. & Hovy, E. (2010). Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1482–1491, Uppsala, Sweden.
- Lamberts, K. & Shanks, D.R. (Eds). (2013). *Knowledge, Concepts, and Categories*. Psychology Press.Hove, UK.
- Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING–ACL98*, Montreal, Canada.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.
- Liu, J., & Gruen, D.M. (2008). Between ontology and folksonomy: A study of collaborative and implicit ontology evolution. In *Proceedings of the 13th International Conference on Intelligent User Interfaces, ACM*, Gran Canaria, Canary Islands, Spain, pp. 361-364.
- Maedche, A. (2012). *Ontology Learning for the Semantic Web* (Vol. 665). Springer Science & Business Media.
- Maedche, A., & S. Staab. (2002). Measuring similarity between ontologies. Proceedings of the International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Sigüenza, Spain, pp. 15-21.
- Mazzini S., & Brazzo L. (2015). From the Holocaust Victims Names to the Description of the Persecution of the European Jews in Nazi Years: the Linked Data Approach and a New Domain Ontology. *The Italian Pilot Project. In Proceedings of the International Conference on Digital Humanities*, Sydney, Australia. 30.6-3.7.2015.
- McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, 10(10).
- Mirkin, S., Dagan, I., & Geffet, M. (2006, August). Integrating pattern-based and distributional similarity methods for lexical entailment acquisition. *In Proceedings of the 3rd Joint COLING/ACL-06 Conference (pp. 107-114)*. Association for Computational Linguistics, Sydney, Australia.
- Mortensen, J. M., Minty E. P., Januszuk, M., Sweeney, T. E., Rector, A. L, Noy, N. F., & Musen, M. A. (2015). Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. *Journal of the American Medical Information Association*, 22(3), 640-648.
- Mortensen, J. M., Musen, M. A., & Noy, N. F. (2013). Crowdsourcing the verification of relationships in biomedical ontologies. In: *Proceedings of the AMIA 2013 Annual Symposium*, pp. 1020-1029.

- Mortensen, J. M., Telis, N., Hughey, J. J., Fan-Minogue, H., Van Auken, K., Dumontier, M., & Musen, M. A. (2016). Is the crowd better as an assistant or a replacement in ontology engineering? An exploration through the lens of the Gene Ontology. *Journal of Biomedical Informatics*, 60, 199-209.
- Muresan, S. & Klavans, J. L. (2013). Inducing terminologies from text: A case study for the consumer health domain. *Journal of the American Society of Information Science and Technology*, 64, 727-744.
- Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R.W., & Musen, M.A. (2001). Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16(2), 60-71.
- Noy, N. F. & Musen, M. A. (2003). The PROMPT suite: interactive tools for ontology merging and mapping. *International Journal of Human Computer Studies*, 59(6), 983-1024.
- Noy, N. F., Mortensen, J. M., Alexander, P. R., & Musen, M. A. (2013). Mechanical Turk as an ontology engineer? Using microtasks as a component of an ontology-engineering workflow, In: *Proceedings of the 5th ACM Web Science 2013 Conference*, (pp. 262-271).
- Noy, N. F. & D. L. McGuinness, (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Retrieved from: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
- Op't Land M., Zwitser H., Ensink P. & Q. Lebel, (2009). Towards a fast enterprise ontology based method for post merger integration. In *Proceedings of the ACM Symposium on Applied Computing, Honolulu, HI, USA*, (pp. 245-252).
- Ouzrout, Y., Geryville, H., Bouras, A., & Sapidis, N. S. (2009). A product information and knowledge exchange framework: a multiple viewpoints approach. *International Journal of Product Lifecycle Management*, 4(1), 270-289.
- Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, C. & Biemann, C. (2016). TAXI at SemEval-2016 Task 13: a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In *Proceedings of SemEval Conference*, (pp. 1320-1327).
- Paul, M. J., Zhai C. & Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 66-76). MIT, Massachusetts, USA.
- Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17, 33-43.
- Pereira, C. S. (2008). *Collaborative ontology specification*. *Doctoral Symposium on Informatics Engineering*, Porto, Portugal. Retrieved from <http://paginas.fe.up.pt/~prodei/DSIE08/papers/38.pdf>
- Pirro, G. & D. Talia. (2008). LOM: a linguistic ontology matcher based on information retrieval. *Journal of Information Science*, 34(6), 845-860.

- Powell, A. L. & J. C. French. (1998). *The potential to improve retrieval effectiveness with multiples viewpoints* (Tech. Report N° CS-98-15). Charlottesville, VA: University of Virginia, Department of Computer Science.
- Quinn, A. J., & Bederson, B. B. (2011). Human computation: a survey and taxonomy of a growing field. In *Proceedings of the Annual Conference on Human factors in Computing Systems—CHI'11* (pp. 1403-1412). Vancouver, BC: ACM.
- Sarasua, C., Simperl, E. & Noy, N. F. (2012). Crowdmap: Crowdsourcing ontology alignment with microtasks. In *Proceedings of the International Semantic Web Conference* (pp. 525-541). Boston, USA.
- Schmitz, P. (2006, May). Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW2006 (50)*, Edinburgh, Scotland.
- Shvaiko, P., & Euzenat, J. (2013). Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 158-176.
- Simperl, E., & Luczak-Rösch, M. (2014). Collaborative ontology engineering: a survey. *The Knowledge Engineering Review* 29, 101-131.
- Stough, C. L. (1972). Language and Ontology in Aristotle's Categories. *Journal of the History of Philosophy*, 10, 261–272.
- Surowiecki, J. (2005), *The wisdom of crowds*. New York: Doubleday.
- Tillett, B. (2005). What is FRBR? A conceptual model for the bibliographic universe. *The Australian Library Journal*, 54(1), 24-30.
- Trant, J. (2009). Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information*, 10(1).
- Tudorache, T., Noy, N. F., Tu, S., & Musen, M. A. (2008). Supporting collaborative ontology development in Protégé. In: *Proceedings of the 7th International Semantic Web Conference (ISWC)* (pp. 17-32).
- Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, 11(02), 93-136.
- Vrandečić, D., Vr, D., Pinto, S., Sure, Y., & Tempich, C. (2005). A diligent. The DILIGENT knowledge process. *Journal of Knowledge Management* 9(5), 85-96.
- Winer, D. (2014). Judaica Europeana: An Infrastructure for Aggregating Jewish Content. *Judaica Librarianship*, 18(1), 88–115.
- Zhitomirsky-Geffet M. & Prebor G. (2016). Towards an ontopedia for historical Hebrew manuscripts. *Frontiers in Digital Humanities, section of Digital Paleography and Book History*, 3, 3. Retrieved from: <http://dx.doi.org/10.3389/fdigh.2016.00003>.
- Zhitomirsky-Geffet M. & Erez E. S. (2014). Maximizing agreement on diverse ontologies with "wisdom of crowds" relation classification. *Online Information Review*, 38(5), 616 - 633.
- Zhitomirsky-Geffet M., Bar-Ilan, J., Miller, Y., & Shoham, S. (2010). A generic framework for collaborative multi-perspective ontology acquisition. *Online Information Review*, 34(1), 145-159.
- Zhitomirsky-Geffet M., Erez E. S., & Bar-Ilan J. (2016). Towards multi-viewpoint ontology construction by collaboration of non-experts and crowdsourcing: the case of the effect

of diet on health. *Journal of the Association for Information Science and Technology*,
68(3), 681-694.

ISO 25964 – *The International Standard for Thesauri and Interoperability with other
Vocabularies*