קריטריונים לבחירת מנוע אחזור טקסט

גרסה 5 – מאי 2009

עפר דרורי

offerd@gmail.com

מבוא

מטרת מסמך זה להגדיר טבלת קריטריונים לצורך השוואה בין מנועי חיפוש לאחזור טקסט. עם הגידול במאגרי המידע ובכמויות המידע המילולי בהן (בעיקר טקסטים שאינם מפורמטים) עולה הצורך בכלים / מוצרים אשר ינהלו את המידע הרב הזה. ניהול טקסטים חופשיים בצורה איכותית, בעיקר בכל הקשור לאחזור המידע מהם, הוא משימה מחשבית שאינה קלה כלל ועיקר. בגלל מורכבות המשימה מחד והצורך בפתרון בעל ביצועים סבירים מאידך נוצרה בתחום זה התמחות אשר מסופקת על ידי מספר רב של חברות בעולם. נפיצותם של מערכות לאחזור טקסט ומנועי חיפוש גדלה באופן משמעותי יחד עם התפתחות רשת האינטרנט והצורך לאיתור מידע מהרשת. ככל שהמאגרים גדולים יותר וככל שהמידע המוזרם לתוכם לא עובר תהליכי סינון ובקרה, כך גדלה חשיבותם של המנועים ובמיוחד תכונותיהם. במאגרים מקצועיים בהם הבקרה על המסמכים רבה, הצורך בתכונות משוכללות של המנוע קטנה במידה מסוימת אף שגם במאגרים מסוג זה קטלוג החומר ומפתוחו לא תמיד עונים על כל הצרכים. בנוסף אי אפשר להתעלם שגם במאגרים המקצועיים נעשים ניסיונות "לעיגול פינות" בגלל העלויות הגדולות הכרוכות בבקרה זו.

המסמך כאמור, מרכז קריטריונים רבים לצורך בדיקה והשוואה בין מוצרים שונים והוא מהווה כלי עזר לבחירה של מוצר לארגון על פי הצרכים הארגוניים שנקבעו מראש. הטבלה כוללת התייחסות למספר נושאים:

נתוני זיהוי של המוצר, החברה המפתחת והנציגות בארץ אם קיימת
תכונות המוצר, בקריטריונים השונים
דרישות פיתוח, ההנחה היא שמנוע החיפוש ישתלב במערכת גדולה יותר וחלק מהיכולות שלו יבואו לידי ביטוי במערכת זו באמצעות כלי פיתוח מתאימים
טכנולוגיה של המוצר, יכולת טכנולוגית, תמיכה וקישוריות למוצרים אחרים
ביצועים של המוצר
מידע על הספק וניסיונו, נושא חשוב לעיתים אף עולה בחשיבותו על תכונות מסוימות במוצר
מחיר המוצר, בהתייחס לתצורות השונות

מומלץ להעביר את הטבלה המצורפת להתייחסות הספקים השונים. יש להחליט כמובן על השקלול המתאים של התכונות השונות עפ"י מודל דלפי לברירת חלופות.

פורמט Word של המסמך נמצא בכתובת
http://www.sigtrs.org/a.php?c=sigtrs&a=287&rc=sigtrs

פורמט PDF של המסמך נמצא בכתובת
http://www.sigtrs.org/a.php?c=sigtrs&a=287&rc=sigtrs

מסמך המפרט את ספקי מנועי האחזור התומכים בעברית נמצא בכתובת
http://www.global-report.net/a.php?c=drori&a=57&rc=drori

מידע נוסף על תכונות המנועים, אחזור טקסט ועוד נמצא באתר קבוצת העניין אחזור טקסט – SIGTRS
http://www.sigtrs.org

פרטים מזהים

שם המוצר	
שם קודם של המוצר	
מספר גרסה נוכחית	
שם החברה המפתחת	
כתובת החברה	
שם הנציגות בארץ	
כתובת הנציגות	
שם איש הקשר בארץ	
טלפון איש קשר	
פקס איש קשר	
דואר אלקטרוני	
אתר אינטרנט של המוצר	
רקע כללי על המוצר והחברה	

תכונות

קריטריון	הסבר	התייחסות הספק
אחזור על מסמכי טקסט	תכונה בסיסית, בהתייחס לטקסט חופשי	
אחזור על מידע מפורמט	היכולת לבצע אחזור על שדות מידע רגילים מפורמטים	
אחזור לשדות מפורמטים בתוך מסמך טקסט	כמו תאריך, מחבר וכו'	
אופרטורים בולאנים	AND, NOT, OR, <, >, =<, >=, שימוש ב- ( ) לביטויים מורכבים	
אופרטורים מטריים (מטריקה)	"מילה" AND "מילה שנייה" במרחק X מילים, באותו משפט, באותה פסקה וכו'	
אחזור לשפה עברית	הכוונה לטיפול מיוחד בשפה ולא לאחזור על בסיס ייצוג האותיות העבריות במאגר	
אחזור לשפה אנגלית		
אחזור ל- 2 השפות במעורב	באותו מסמך	
אחזור לשפות נוספות	נא לפרט	
מילון מורפולוגי כחלק מהמוצר	האם קיים כזה ואם כן מהן תכונותיו	
שילוב מילון מורפולוגי חיצוני	האם השילוב אפשרי, אם כן ציין איזה מילון (כמו מורפיקס של מלינגו) ואיזה גרסה	
תמיכה בטבלאות תזאורוס במוצר	האם קיימת תשתית לשימוש בטבלאות תזאורוס חיצוניות, האם המוצר כולל תזאורוס משלו, אם כן לאיזה תחום ובאיזה שפה	
ניהול תזאורוס במוצר	במידה ואנו רוצים לייצר את המילון לבד, האם יש תמיכה לשימוש בטבלה ריקה שתוזן ע"י המשתמש	
ניהול תזאורוס היררכי	האם קיימת תשתית במוצר לתזאורוס היררכי	
הצגת תוצאת החיפוש ע"י המוצר	האם המוצר כולל ממשק משתמש להצגת תוצאות החיפוש, אם כן באיזו שיטה: רק כותרות, תחילת מסמך, משפטיים רלוונטיים לחיפוש וכו' – יש לפרט	
הצגת שאילתת החיפוש ע"י המוצר	האם המוצר כולל ממשק משתמש לביצוע שאילתת החיפוש	
הדגשת מילים המקיימים את תנאי החיפוש	האם המוצר כולל תמיכה בהדגשת מילים המקיימות את תנאי החיפוש במסמכים שאותרו	
ביצוע שאילתות על שאילתות	יצירת סטים של תשובות וביצוע אחזור נוסף עליהם	
אחזור על מספר מאגרים במקביל	האם אפשרי ואם כן האם גם בפורמטים שונים של המאגרים	
אחזור על מסמכים מפוצלים	טיפול במצב מורכב בו כותרות המסמכים נמצאות בבסיס נתונים והמסמכים עצמם במערכת לניהול קבצים	
אחזור מדויק בלבד	האם קיים בלי הרחבות אוטומטיות	
אחזור פונטי	האם קיים	

הרחבות אחזור לחלקי מילים (Wildcard)	ראשיות, סופיות ואמצעיות אפשרות להחלפת תווים בסימן "*"	
ניהול אינדקסים במוצר	האם נעשה במוצר עצמו, האם מתבסס על בסיסי נתונים חיצוניים, תיאור שיטת עבודה	
ניהול הרשאות גישה לקטעים במאגר	האם יש תמיכה להרשאות שונות על קטעים במאגר	
גיבוי ושחזור במוצר	האם כולל מנגנון גיבוי ושחזור עצמי או שמתבסס על גיבוי ושחזור חיצוני של כל סביבת העבודה	
תמיכה באחזור מסמכי טקסט ASCII	האם קיים, מגבלות אם יש נא לציין	
תמיכה במסמכי UNICODE Text		
תמיכה במסמכי XML כולל מסמכי אופיס שנשמרו כ-XML		
תמיכה באחזור מסמכי טקסט מסוג WORD	להתייחס לאיזה גרסאות WORD, הכוונה לאחזור ישיר ממסמכי WORD ללא הסבתם	
אחזור מתוך מסמכי RTF		
תמיכה באחזור מסמכי טקסט מסוג HTML	האם קיים	
תמיכה באחזור מסמכי טקסט HTML בעברית לוגית וחזותית	האם קיים, ציין מגבלות לגבי לוגית וחזותית	
אחזור מתוך מסמכי Excel	האם קיים, הכוונה לאחזור ישיר ממסמכי Excel	
אחזור מתוך מסמכי Powerpoint		
אחזור מתוך מסמכי PDF	האם קיים, ישירות מול קבצי אקרובט PDF	
אחזור מתוך מסמכי PS	האם קיים, ישירות מול מסמכי Post Script	
תרגום פורמטים	האם המוצר כולל תרגום פורמטים שונים לטקסט נקי, אם כן פרט אילו	
אחזור לשורשים בשפה העברית והאנגלית	האם קיים	
אפשרות להוספת מסמכים בצורה מקוונת	מתוך מערכת O.L., הכוונה למנגנון המאפשר הוספה מיידית של מסמך וביצוע אחזור מיידי למסמך במסגרת המוצר ולא להפעלת אצווה כל מספר שניות או דקות	
תמיכה בטיפול בטבלאות	אחזור וניהול המידע כאשר הוא בטבלאות	
דרוג (Ranking) תשובות	האם המוצר כולל מנגנון לדירוג רשימת התוצאות, פרט אלו אלגוריתמים קיימים לדירוג	
מנגנונים לסיווג	האם המוצר כולל מנגנוני סיווג כמו קטגוריזציה, אשכולות דינמיים ועוד. נא לפרט גם בהתייחס לשפה	
אבטחת מידע	יכולת המוצר לתמוך במנגנוני אבטחה (קבצים או בסיסי נתונים) כך שהמשתמש יקבל את רשימת תוצאות	

	החיפוש בהתבסס על ההרשאות שלו	
חילוץ ישויות	יכולת המוצר לחלץ מידע על ישויות מתוך טקסט	.

דרישות פיתוח

קריטריון	הסבר	התייחסות הספק
תמיכה בסביבות פיתוח	יכולת המוצר בתמיכה בסביבות פיתוח שונות SDK ל- Java .Net וכו׳	
קיום SDK	קיום SDK למגוון צרכים כמו: חיפוש, קלסיפיקציה, תחזוקת אינדקסים, תחזוקה כללית של מנהל המערכת, יכולות חיפוש מורחבות וכו׳	
תפיסת פיתוח תהליכי אינדוקס ואחזור	קיום רכיבים לביצוע עיבודים שונים בתהליכי האינדוקס לדוגמא: ניתוח לשוני, קטגוריזציה, קלסיפיקציה, יצירת אשכולות וכו׳.	

טכנולוגיה

קריטריון	הסבר	התייחסות הספק
תמיכה בשרת/לקוח	האם קיימת	
תמיכה בלקוח תחת חלונות XP ,Vista וכו' דפדפנים לסוגיהם	התייחס לגרסאות התומכות בסביבות העבודה השונות, לגבי דפדפנים התייחס לגרסאות השונות, אם ידועות בעיות בנושא - נא ציין	
תמיכה בשרתי NT, UNIX	האם קיימת	
קישור לבסיסי נתונים מסוג SQLServer , אורקל	האם קיימת תמיכה למידע המאוחסן בבסיסי הנתונים הנ"ל, אם קיימת תמיכה בבסיסי נתונים מקומיים אחרים נא ציין	
קישור ל- Exchange	האם קיים בכל הקשור לאחזור מידע המוטמע בשרתי Exchange	
קישור ל- Notes	האם קיים בכל הקשור לאחזור מידע המוטמע בשרתי Domino	
ממשקים מסוג RPC, OLE, Java, ACTIVEX	ציין אילו קיימים ואם קיימים נוספים נא ציין	
ממשק API	האם קיים ממשק API המאפשר ביצוע של כל התכונות מתוך תוכנה אחרת חיצונית	
קישור לשרת אינטרנט IIS	האם קיים	
קישור לשרת אינטרנט של אורקל	האם קיים	
קישור לבסיס נתונים ב- MF כאשר אינדקסים נמצאים ב- MF או בשרת	הכוונה למצב בו המידע המיועד לאחזור מאוחסן בבסיס נתונים מרכזי ב- M.F. כדוגמת ADABAS והאינדקסים לאחזור נמצאים בשרת או ב- M.F. (נא ציין במה המוצר תומך)	
תמיכה בריבוי אינדקסים	מה המספר המקסימאלי של אינדקסים שניתן להקים ולנהל?	
אינדוקס ישירות מבסיס הנתונים	האם מתבצע אינדוקס אוטומטי עם עדכון בסיס הנתונים?	
שמירת האינדקס בבסיס הנתונים	האם יש אפשרות לכלול את האינדקסים בבסיס הנתונים עצמו?	
תמיכה בבסיס נתונים מעל 10 מיליון מסמכים	האם קיימת	
יכולת Scale Out	האם ניתן לתמוך בנפחי מסמכים / שאילתות גדולים ע"י הפעלת מספר שרתים במקביל	
ניצול מעבדים מרובי ליבות (core)	האם תהליכי המערכת בנויים לריבוי ליבות עיבוד ולריבוי נימים (Multithreading)	

ביצועים

קריטריון	הסבר	התייחסות הספק
זמן טעינה של 100,000 מסמכים	אינדוקס של 100,000 רשומות חדשות, נא ציין ביחס לשרת סטנדרטי (ציין חוזק השרת, מספר CPU וכו')	
זמן טעינה של 10,000 מסמכים	כנ"ל לגבי 10,000	
זמן תגובה לחיפוש ביחס לגודל המאגר	עפ"י מבחני ביצועים של המוצר (נא ציין המבחן)	
נפח אינדקס ביחס לגודל המאגר	בזמן נתון ועם גידול המאגר, נא לצרף נתונים	
דוגמת ביצועי אינדוקס	נא צרף דוגמא ספציפית לקליטת X מסמכים בנפח Y בדקות למעבד בגודל נתון	

מידע על הספק וניסיונו

קריטריון	הסבר	התייחסות הספק
ניסיון בשנים בפיתוח מנוע אחזור טקסט		
מספר התקנות בארץ		
מספר התקנות בחו"ל		
כמות התקנות למאגרים מעל 10 מיליון מסמכים		
כמות התקנות מעל מיליון מסמכים		
כמות התקנות מעל 100,000 מסמכים		
כמות התקנות בסביבת אינטרנט/אינטרה-נט		
כמות התקנות המשלבת בסיס נתונים ארגוני על PC	נא לציין את סוג בסיס הנתונים: אורקל, SQL-Server וכו'	
כמות התקנות המשלבת בסיס נתונים ארגוני על MF	נא לציין את סוג בסיס הנתונים: אדבס, DB2 וכו'	
כמות התקנות עם Exchange		
כמות התקנות עם Notes		
סוגי תמיכה	נא לפרט: תמיכה טלפונית או אחרת	
פרק זמן בין גרסה לגרסה		
מחויבות החברה למוצר	האם מוגבלת בזמן, האם מרכז עיסוק החברה בתחום, תוכניות לעתיד בהקשר למוצר	
רשימת לקוחות	עפ"י הקריטריונים השונים: לקוחות עם בסיס נתונים אדבס, DB2 אחר לקוחות עם בסיס נתונים אורקל, SQL-SERVER וכו'.	

מספר שנים שהמוצר הנוכחי מותקן אצל לקוחות		
השתתפות במכרזים או בחירות קודמות	אם המוצר השתתף במכרזים או בחירה אחרת ונבחר, נא ציין את הגורם הבוחר, שנת הבחירה ואת המוצרים שמולם עמד לתחרות	

מחיר

קריטריון	הסבר	התייחסות הספק
מחיר המוצר	נא להתייחס לגרסאות שונות של המוצר (אם קיימות) בהתייחס למספר שרתים, לאתר, שיטות רישוי שונות	
אחריות	כולל משך זמן האחריות, ניתן לציין מחיר אפס	
מחיר אחזקה ותמיכה לשנה לאחר תקופת האחריות	נא לציין את העלות	