10.17625/NKE.2018.024

PhD értekezés

Vadász János Pál

- 2018 -

NEMZETI KÖZSZOLGÁLATI EGYETEM

KATONAI MŰSZAKI DOKTORI ISKOLA

Vadász János Pál

A szemantikus keresés módszerei és alkalmazási lehetőségei a védelmi szférában, a közigazgatásban, illetve a gazdasági életben

Doktori (PhD) értekezés

Témavezető: Prof. Dr. Munk Sándor ny. ezds. (DSc)

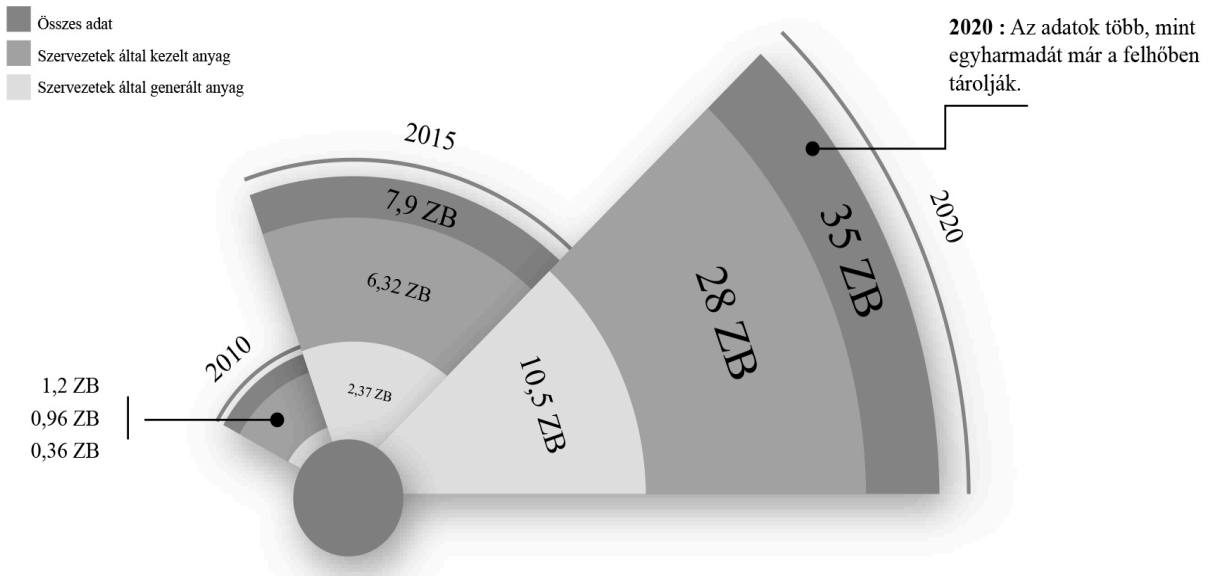
egyetemi tanár

BUDAPEST, 2018

TARTALOMJEGYZÉK

BEVEZETÉS

A XXI. század második évtizedének végén látható trendek egyértelműen mutatják, hogy napjainkban az információkezelés területén **paradigmaváltás zajlik**. Az elérhető adatok zöme már elektronikusan elérhető zárt vagy nyílt (pl. felhőben) rendszerekben, így az információkeresés jórészt elektronikus információkeresést jelent. Az adatok tárolásának költsége folyamatosan csökken. Az adatfeldolgozás bizonyos esetekben olcsóbb felhőben, mint saját rendszeren belül. A létrejött szöveges állományokat egyre inkább nem beszkennelt képi formátumban tárolják és teszik hozzáférhetővé, hanem az eredeti géppel olvasható változatban. A számítógépes feldolgozási kapacitások eleddig elképzelhetetlen feladatokat tudnak megoldani, így például komplex mesterséges intelligencia alkalmazásokat képesek futtatni, mint például a mélytanulásra épülő gépi fordítás, kérdező-válaszoló alkalmazások, kivonatolók stb. Az olcsó és gyors, vonalas vagy mobil adatátviteli csatornák mindent elérhetővé tesznek a felhasználó számára. Ezért az adatok óceánjában az igazi mai **kihívás megtalálni a lényeget, legyen az strukturált, strukturálatlan vagy félig strukturált formában**.



1. ábra: az elektronikusan elérhető adatok mennyisége.[1]

A világszerte elektronikusan tárolt szövegmennyiség exponenciális növekedése, a Moore-törvénynél már lassabban, de még mindig ugyancsak exponenciálisan növekvő processzorsebesség, a szinte korlátlan tárolókapacitások, az online elérhetőség és nem utolsósorban a mesterséges intelligencia gomba módra szaporodó alkalmazásai alapvető, a

[1] Forrás: [1] alapján saját szerkesztés.

szolgáltatások és a felhasználás jelentős mértékű változásait idézték elő az információkeresés területén is. A keresési technológiák egyik rendkívül hatékony és kifinomult válfaja a jelentésalapú vagy szemantikus keresés. Ennek a jelentősége annál is nagyobb, mert az elérhető adatok 85%-a strukturálatlan vagy félig strukturált (lásd lejjebb) [2],[2].

A tudományos probléma megfogalmazása

Bár az információkeresés jelentős szerepet játszik az értekezésem címében foglalt alkalmazási területeken (védelmi szféra, a közigazgatás és a gazdaság), a tudományos kutatás célorientáltan keveset foglalkozott az információkeresés általam vizsgált határterületeivel, és szinte egyáltalán nem a hazai körülményeket feldolgozó módon. Számos cikk jelenik meg a téma matematikai és informatikai vonatkozásairól, míg a felhasználhatóságról, gazdaságosságról, jogi keretekről, az alkalmazás körülményeiről keveset olvashatunk. Úgy tűnik, az elméleti tudomány művelői számára ez a terület túl gyakorlati, míg a gyakorló szakemberek számára túl elméleti.

A témakört a konkrét gyakorlat oldaláról nézve, információkereső rendszereket készítő, szolgáltatásokat nyújtó vállalkozás vezetőjeként régóta foglalkoztattak az információkeresés – és azon belül a szemantikus keresés – tudományos megalapozást igénylő kérdései. Tapasztalataim azt mutatták, hogy számos olyan kérdés van az információkeresés mikro- és makrotársadalmi környezetében, amelyet a magyar szakirodalom szinte egyáltalán nem, és a nemzetközi is csak igen gyéren tárgyal. Ezek közé tartoznak többek között az alábbiak.

Mivel az információ és annak hatékony keresése a gazdaság és a közigazgatás területén is önálló erőforrássá vált, **az információkeresés is önálló beruházási igénnyel jelent meg**. Ugyanakkor az információkeresés mint beruházási terület kezelése a magyar ipari és közigazgatási tapasztalatok alapján gyakorlatilag ismeretlen fogalom. A felsővezetés sokszor ok nélkül ódzkodik a beruházási döntések meghozatalától a szükséges tudás hiánya miatt, és ez a technológiai megújulást, valamint a hatékonyság növelését visszafogja. Még a legegyszerűbb, megtakarított munkaidő-alapú számítási modell is megdöbbentő eredményeket mutat fel. Ha ehhez hozzávesszük a csak indirekt mérésekkel megállapítható, nem kézzelfogható jellemzőket is, akkor a döntéshozók számára komplett arzenál áll rendelkezésre a beruházási döntéseik meghozatalához. Ezek tudományos vizsgálata azonban hiányos.

[2] [3] 269. oldal

A mikrokörnyezet másik feltáratlan, de égetően fontos aspektusa az információkeresés humán és biztonsági oldala, azon gátló tényezők feltárása, amelyek akadályozzák egy kereső alkalmazás befogadását a működő szervezetek részéről. Ezen szempontoknak a feltárása és rendszerezése a hazai ipari tapasztalatok és a nemzetközi tudományos szakirodalom felhasználásával szintén tudományos problémaként azonosítható.

Végül fontos tudományos problémának tartom a magyar jogrendszer kereteinek rendszerezését az információkeresés szempontjából, azon hiányosságok és elmaradások feltárását, amelyek a technológiai fejlődéssel való lépéstartást akadályozzák. A hazai megoldások hátttereként kutatási feladatnak gondolom a nemzetközi kitekintést, hogy látható legyen a jogi szabályozás fejlődésének iránya az EU-ban és az angolszász országokban.

Közel 30 éves magyarországi szakmai múltam alatt mindvégig igyekeztem hidat építeni a tudományos elmélet és az ipari gyakorlat közé. A kutatásaim eredményét ötvözve több évtizedes tapasztalattal indultam el az értekezés megalkotásának.

Kutatási célok

Értekezésem alapvető kutatási célja az információkeresés eredményes és hatékony alkalmazási feltételeinek elemzése és meghatározása a védelmi szférában, a közigazgatásban és a gazdasági életben. Ennek részeként:

- az információkeresés kiemelt jelentőségű alkalmazási területeinek feltárása a védelmi szférában, a közigazgatásban és a gazdasági életben;
- az információkeresés magyar jogi környezetének feltárása és értékelése, az információkeresést akadályozó jogi normák és ezek hatásának meghatározása;
- a rendvédelmi és nemzetbiztonsági szervezetek számonkérhetőségét biztosító azon módszerek meghatározása, rendszerezése, amelyek növelik egyrészt az állampolgári bizalmat, másrészt a szervezetek munkájának hatékonyságát;
- az információkeresés hatékonysága és gazdaságossága mérési kereteinek, modelljének kidolgozása;
- az információkeresés mint tudásmenedzsment-eszköz szervezeti bevezetése akadályainak meghatározása, és javaslattétel azok elhárítására.

Kutatási hipotézisek

Az értekezés témájához, valamint a tudományos probléma fenti feltárásához kapcsolódó kutatási hipotéziseim a következőek voltak:

- a védelmi szféra nyílt forrású keresésének, valamint a gazdasági szféra hírszerzésének fogalma, értelmezése pontatlan, félreérthető (1. és 3. új tudományos eredmény);
- a szemantikus keresőrendszerek használata jelentősen növelheti a munka hatékonyságát (7. új tudományos eredmény);
- a nemzetközi trendek azt mutatják, hogy az információkeresés jogi környezete komoly átalakuláson megy keresztül (5. új tudományos eredmény);
- a magyar jogrendszer az információkeresést töredezetten, hiányosan és anakronisztikusan kezeli, ez a rendvédelmi, nemzetbiztonsági munkát egyrészt feleslegesen akadályozza, másrészt veszélyt jelent a terrorizmus és a szervezett bűnözés elleni fellépésben (2. és 4. új tudományos eredmény);
- a munkaszervezés és a kommunikáció legalább annyira fontos egy információkereső rendszer hatékony működésében, mint maga a technológia (6. új tudományos eredmény).

Az értekezés szerkezete

A doktori értekezés négy tartalmi fejezetből áll.

Az első fejezetben összegeztem az információkeresés információelméleti, tudáselméleti alapjait, a jelentésközpontú szemantikus keresés alapfogalmait, az információkeresés matematikai és informatikai alapjait, alkalmazott módszereit, végül az információkeresés nyelvészeti alapjait, a természetes nyelvfeldolgozásnak a téma szempontjából legfontosabb kérdéseit.

A második fejezetben bemutattam és értékeltem az információkereséshez, annak technológiai megoldásaihoz szorosan kapcsolódó három alkalmazási területet, megoldást: a szentimentelemzést, a keresési eredmények további feldolgozására épülő metakeresést, valamint a különböző forrásokból származó információk szintetizált feldolgozását is magában foglaló fúziós központokat.

A harmadik fejezetben megvizsgáltam az információkeresés egy-egy kiemelt szerepű, elsődleges alkalmazását az értekezés címében szereplő három alapvető alkalmazási területen (védelmi szféra, közigazgatás, gazdasági élet). Ezek esetében az egyes alkalmazási területeken meghatároztam az információkeresés legfontosabb sajátosságait, igazoltam alkalmazásának szükségességét, alapvető felhasználási területeit, javaslatot tettem hatékonyabb alkalmazásuk lehetőségeire, összegeztem várható jövőbeni fejlődési irányait.

A negyedik fejezetben az információkeresés eredményes és hatékony szervezeti alkalmazásának három alapvető külső és belső környezeti feltételét – a jogi környezetet, a szervezeten belüli humán és biztonsági körülményeket, valamint az információkeresés értékelhetőségét – vizsgáltam, értékeltem, és dolgoztam ki kapcsolódó javaslatokat.

Alkalmazott kutatási módszerek

Széles körű irodalomkutatás keretében feldolgoztam a vonatkozó releváns nemzetközi és hazai szakirodalmat, jogszabályokat és egyéb szakmai dokumentumokat.

Másodelemzéssel kutatási témám szempontjából elemeztem, feldolgoztam a témában készült korábbi kutatási eredményeket.

Személyes empirikus szakmai tapasztalataimat feldolgoztam, rendszereztem, empirikus kutatásokat végeztem.

Konzultációkat folytattam a kutatási témám szempontjából fontos alkalmazó szakmai szervezetek vezetőivel, szakembereivel, valamint a jogi, és információtechnológiai szakterület szakértőivel.

A megszerzett, rendszerezett információkat összehasonlító elemzés alá vetettem, általánosítottam, értékeltem, és ezek alapján következtetéseket, ajánlásokat, javaslatokat fogalmaztam meg.

A kidolgozás során érvényesült korlátozások

A kidolgozás során számos korlátozást kellett érvényesítenem, és számos korlátozás érvényesült szándékomtól függetlenül is. Az első és legfontosabb korlátozás a téma széleskörűségéből fakadt. Az értekezés terjedelme nem adott lehetőséget, és kutatási céljai nem igényelték például az információkeresés tudáselméleti, matematikai és nyelvészeti alapjainak mély vizsgálatát. Nem volt mód a három alapvető alkalmazási terület valamennyi információkeresési tapasztalatának vizsgálatára, csak egy-egy kiemelt jelentőségű alkalmazásra. Végül nyilvánvalóan önálló értekezések tárgyát képezhetik az információkeresés jogi keretei vagy az információkeresés értékelése. Mindezekre az egyes fejezetekben konkrét utalások is találhatóak.

A második korlátozás az információkeresés védelmi szférabeli alkalmazásához, a nyílt forrású hírszerzéshez kapcsolódik. Mint a nemzetbiztonság, honvédelem számos területén, az információk, megoldások, tapasztalatok jelentős része civil kutatók számára nem hozzáférhető, részvételük szakmai rendezvényeken, képzéseken, konferenciákon vagy azok egyes előadásain

korlátozott. Mint kapcsolódó területen dolgozó ipari szereplő nekem is látnom kellett, hogy ez igaz az információkeresésre is.

1. FEJEZET: AZ INFORMÁCIÓKERESÉS ALAPJAI

1.1. Bevezető gondolatok, a fejezet tartalma, célja

Az első fejezetben áttekintem az információkeresés, a számítógépes nyelvészet és a használt matematikai-informatikai apparátus azon alapfogalmait, amelyekre a későbbi fejezetekben szükség lesz. Külön figyelmet fordítok azokra a tudásreprezentációs eszközökre, amelyeket a magyar ipari gyakorlat során legsűrűbben alkalmaznak.

A hétköznapi nyelvet nem lehet tökéletesen értelmezni a mai tudásunk szintjén, de lehetséges a célt újabb és újabb technikákkal egyre jobban közelíteni. Nincs sem az adat-, sem a szövegbányászatnak egyedül üdvözítő módszere, nem létezik „a" megoldás. Sokkal inkább a feladattól függő kombinációi a többé-kevésbé felhasználóbarát, de olykor roppant bonyolult matematikai hátteret felvonultató eszköztáraknak, amelyek alkalmazása a tudomány és a művészet határán fekszik. Abszolút biztos, tökéletes eredmény sincs, csak valószínűségek, rangsorok, amelyeket az elemzőnek értékelnie kell a kívánt tudás kinyerése céljából. A módszerek és eszközök bizonyos határok közötti szabad kombinálása és kreatív felhasználása emlékeztet a hadtudomány versus hadművészet analógiára. Magának az információkeresésnek a tárgyalása előtt itt következzék az információkeresés definíciója.

„Az információkeresés fogalma alatt átfogó értelemben olyan tevékenységet javaslunk érteni, amely információreprezentációk (adatok) meghatározott köréből meghatározott információigény kielégítését segítő információreprezentáció(k), adat(ok) megtalálására, kiválasztására irányul." [4]

Felmerülhet a kérdés, hogy az internetes óriáskeresők elterjedése mellett van-e még élettere az egyéni, kidolgozottabb számítógépes nyelvészeti apparátust felhasználó keresőknek, miközben a Google 2017-ben 30 trillió oldalt vizsgál [5]. A Google gyenge jelentésalapú nyelvi technológiájával szemben az egyéni nyelvi technológiát alkalmazó keresők a számítógépes nyelvészet segítségével sokkal mélyebben tudják a szövegtestet a jelentése alapján értelmezni. Ettől sokkal teljesebb a felhasználói élmény. Több pontos, kevesebb téves találat, (közel) szabadszövegű lekérdezés, vagyis nem kell pontosan feltenni a kérdést, például szinonimát is elfogad a kereső. Az 1. táblázatban foglaltam össze **a két kereső típus előnyeit és hátrányait**.

1. táblázat: egyéni nyelvi technológiájú kereső főbb tulajdonságainak összehasonlítása az óriás internetes keresőkéivel.[3]

Egyéni nyelvi technológiájú kereső	**Google, Bing, Yahoo stb.**
• csak előre definiált helyen keres • automatikus nyelvfelismerés • komplex nyelvészeti technikák • tranzakciók „házilag” naplózva • hozzáférés korlátozása kiemelt fontosságú • rugalmas megjelenítési formák, elemzési eszközök	• nagyszámú nyílt helyet indexel • nem értelmez nyelvet • karakter- és statisztikai alapú tranzakciókat a felhő tárolja • nyílt, hozzáférés korlátozása nincs • merev formátumú megjelenítés, csak találati lista

1.2. Információkeresés, szemantika, szemantikus keresés

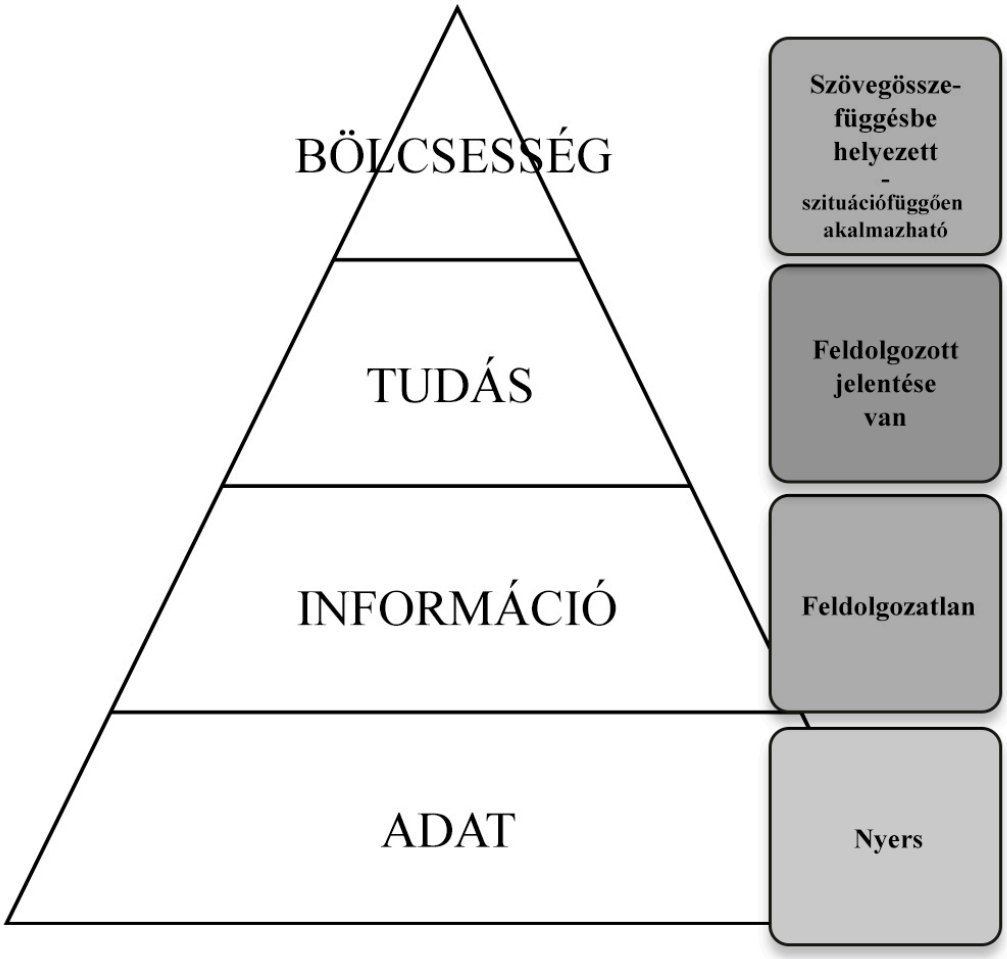
1.2.1. A tudás szintjei

Az ismeret elemeit a modern felosztásban négy kategóriába sorolják Rowley szerint [6]. Weinberger szerint [7] a gondolat nem is Russell Ackofftól származik, hanem egyenesen T.S. Elliot A szikla című művéig vezethető vissza, és az információ fogalmát üres fogalomnak (*hollow term*) tartja. Én nem tartom feladatomnak ebben a filozofikus vitában elmélyedni. **Egyet nem értve** Weinberger-féle megközelítéssel, informatikai szempontból **látok lényeges különbséget az adat és az információ között**, hiszen az egyik rendezetlen, a másik már valamilyen szempont szerint rendezett. A felbontás a 2. ábrán látható.

- Adat: minőségi és mennyiségi változó értékek halmaza. Ez a nyers, rendezetlen sokaság, amelyből feldolgozással információ keletkezik.
- Információ: az adatok valamilyen rendszer szerint feldolgozott, rendezett halmaza.
- Tudás: a megfelelő helyen, időben és formában megjelenő információ.
- Bölcsesség[4]: az a képesség, amellyel az információt egy adott helyzetben adekvátan, helyesen tudjuk felhasználni.

[3] Forrás: a szerző saját szerkesztése.

[4] Az angol *wisdom* nem különösebben szerencsés fordítása.



2. ábra: a tudás elemei.[5]

Azt gondolom, hogy érdemes elidőzni a 2. ábra felépítésén. Az emberi agyba az érzékszerveken át jutnak adatok. Ezeket az agy rendezi, feldolgozza, így információ keletkezik, amelyet az agy tárol. Adott helyzetben az információt az agy előhozza, ennek a segítségével az ember reagál az impulzusokra. A mély tudás vagy a tapasztalat az, amit „öreg róka" metaforával is jellemeznek. A sok korábbi élmény alapján a döntés nem is mindig tudatosan, hanem ösztönösen történik. Talán nem túl erőltetett a humán modellt az informatikaival összevetni. A bevitt nyers – akár strukturálatlan – adatokat a rendszer feldolgozza – pl. indexeli – és strukturálja, adatbázist épít. Az így keletkezett információt kellő időben és helyen előhívva tudás keletkezik. A tudást reprezentáló tanítóadatok bevitelével a rendszerben egyfajta

[5] Forrás: a szerző saját szerkesztése.

mesterséges bölcsesség keletkezik, amelyet az emberi agy mintájára alkotott neurális hálók tárolnak és szolgáltatnak.

1.2.2. Szemantika, jelentés, szemantikus keresés

A természetesnyelv-feldolgozás (*natural language processing*, NLP)[6] az informatika, a mesterséges intelligencia és a számítógépes nyelvészet közös területe [8]**. Feladata a természetes nyelvű szövegek számítógépes feldolgozása. Főbb területei a természetes nyelv gépi értelmezése, fordítás, nyelvek generálása, hangfeldolgozás, szöveges és hangalapú párbeszédre képes interaktív alkalmazások előállítása stb.**

Általában véve a nyelvészet főbb kutatási területei a következők.

- A szintaktika a nyelvtudománynak a mondatokat és a mondat belső viszonyait, szerkezeti részeit tanulmányozó ága.
- A szemantika a nyelv elemeinek (szöveg, mondat, szó, szimbólumok) tartalmi elemzése, jelentésének vizsgálata, értelmezése. A szemantika szoros kapcsolatban áll az ismeretelmélettel, a kognitív pszichológiával és a logikával. A szemantika eredetileg csak szavak jelentésével foglalkozott, de a múlt század második felében kiterjedt szószerkezetek, mondatok és nagyobb szövegtestek értelmezésére is.
- A szemiotika jeltudomány, nem kizárólag a nyelvi, hanem más, pl. grafikus jelekkel is foglalkozik.
- A pragmatika a jelentést adott kommunikációs kontextusban (háttértudás, a beszélő szándéka, rejtett tartalom stb.) vizsgálja.

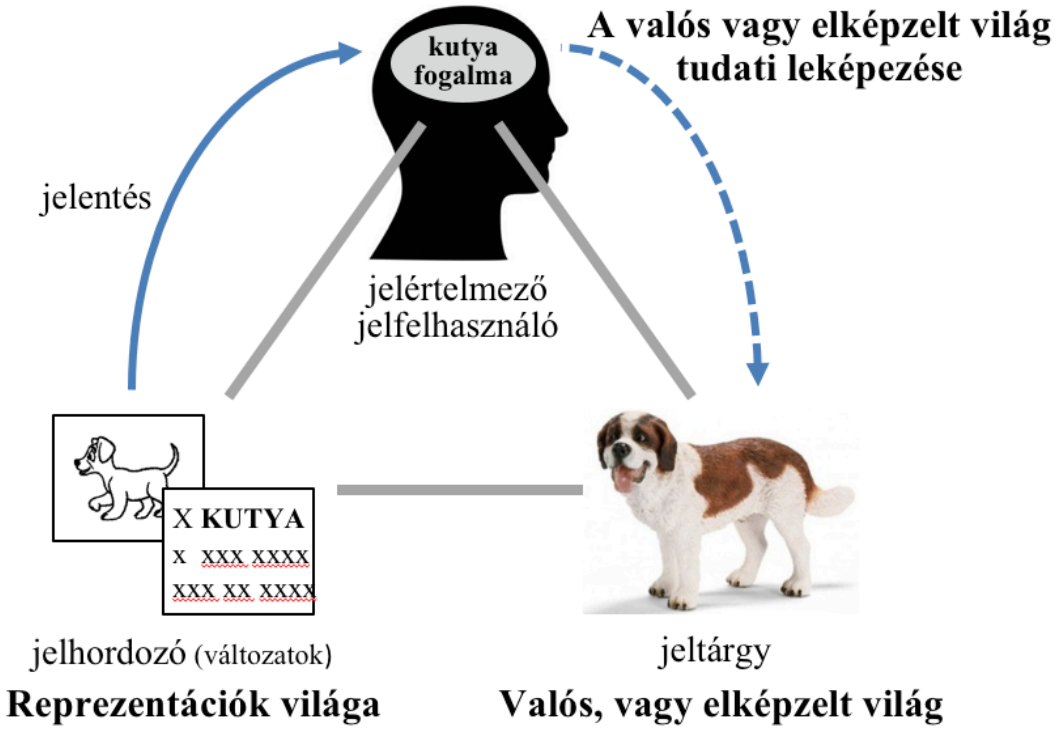
1.2.3. A jelentés jelentése

A szemantika szerint a jelek önmagukban értelmetlenek, a környezethez viszonyítva nyernek értelmet. A 3. ábra az eredeti saussure-i[7] [9] gondolatot feldolgozva mutatja be a három komponens viszonyát. A jeltárgy maga az objektum, amelyet a jel ábrázol. A jelhordozó az a kép, hang, betűsor stb., ami a jelfelhasználó számára a jeltárgyat szimbolizálja, reprezentálja. Nyilvánvaló, hogy a kutya, *dog, Hund, chien* betűsoroknak nincs önmagukban semmilyen *Canis lupus familiaris* jellege, a jel a jelhordozó emberi agyban képződő absztrakciója során

[6] A magyar szaknyelvben is az NLP a meghonosodott. Nem keverendő össze a Neuro-linguistic Programming nevű, ugyancsak NLP rövidítésű pszichoterápiai irányzattal.

[7] Ferdinand de Saussure svájci nyelvész, 1857–1913.

kap értelmet a korábbi beidegződések alapján. A tudás informatikai szempontból az adat és az információ, a gépi tárolás és megjelenítés tárgya, míg a bölcsesség eredeti megközelítésben már az emberi tudatban képződik meg. A mesterséges intelligencia világában a klasszikus felosztás eltolódik, mert a gép már képes az információkat megfelelő helyen és időben automatikusan, azaz megfelelő külső inger hatására előhozni, és azokat bonyolult következtetések, reakciók folyamán feldolgozni, egyes területeken egyre inkább az embernél magasabb színvonalon is (sakk, go stb.).



3. ábra: jeltárgy, jelhordozó és jelértelmező saussure-i hármasa.[8]

A jelalkalmazás legfontosabb területe a kommunikáció, vagyis a jelek egységes értelmezésén alapuló információcsere. Bár a világ globalizálódásával a jelértelmezés egyre inkább egységessé válik, távolról sem állítható, hogy különböző kultúrákban ugyanaz a jel ugyanazt a jelentést hordozza.

A 4. ábra mutatja az OK = „minden rendben” két ismert jelét. Az első jel az angolszász és dél-európai országokban honos, míg a második inkább közép- és észak-európai gyökerű, de például

[8] Forrás: [10]

a búvároknál is használatos. Ugyanakkor a második jel a mediterrán régió jelrendszerében az OK helyett igen durva szexuális tartalmat szimbolizál.

4. ábra: OK jelek.

A szavak jelentését definíciókkal és szövegösszefüggésbeli példákkal határozzuk meg. A definíciók szótárakban, lexikonokban találhatók. A szövegösszefüggésre példákat értelmező szótárak és mintaszövegek adnak. Mindkettőnek nagy jelentősége van a számítógépes nyelvészetben.

1.2.4. A szemantikus keresés fogalma, értelmezése

A szemantikus keresés a keresési pontosságot olyan technológiákkal javítja, amelyek a keresést annak jelentése alapján értelmezik a keresett térben, legyen az egy zárt rendszer, a nyílt web vagy korlátozott hozzáférésű adatállomány a nyílt weben [11].

1.2.5. Adattípusok

Az adatok lehetnek strukturáltak, félig strukturáltak és strukturálatlanok. **Strukturált adatok adatbázisokban és az azt leíró sémákban fordulnak elő, rekordonként egyező formátumú mezőkkel.** Ilyen például a telefonhívások vagy hitelkártyaszámok listája. **A félig strukturált adatok egyező elválasztóval, például pontosvesszővel elválasztott, de nem feltétlenül egyforma hosszúságú mezőket tartalmazó rekordok.** A strukturálatlan adatok szabadszövegű állományok, mint például a Word vagy a PDF (hordozható dokumentumformátum, *Portable Document Format*) fájlok, e-mailek stb. **Nem strukturált, szabad szövegű formán azt értjük, hogy az adat szemantikájára nem utal a tároló adatstruktúra** [9]. Hétköznapi nyelven úgy mondhatjuk, hogy az adatok nem rendezett formában jelennek meg, ahogy a telefonszámok egymás alatt, hanem sorban, de alakilag rendezelenül, mint például szavak, mondatok egy újságcikkben vagy egy honlapon.

[9] [12] 21. oldal

Az adatállományokat leíró adatokat metaadatoknak nevezzük. Ilyen egy Word-fájl keletkezésének vagy utolsó módosításának dátuma, a szerző neve stb. A metaadatok az ISO 11179 metaadattárház-szabvány [13] és a Dublin Core Metadata Initiative [14] szerint más adatokat meghatározó és leíró adatok. A szabvány meghatározza az adatelemek szemantikáját, reprezentációjának leírását és ezek regisztrációs rendjét. A metaadatok az információkeresésben fontos szerepet játszanak. Ilyen például olyan dokumentumok osztályozása, amelyek tartalma titkosított, de a keletkezésükről és forgalmazásukról szóló jellemzők megismerhetőek, például műholdról befogott rejtjelezett üzenetek vagy titkosított e-mailek.

1.2.6. A szemantikus web

Az interneten a kezdeti, statikus megjelenéseket (web1) mint például a statikus honlapokat az interaktív felületek követték (web2) mint a fórumok, wikik és más *folksonomy* (ld. lejjebb) alkalmazások, mint például a Facebook. Minőségileg más a szemantikus web (web3) [15]. A vízió lényege, hogy az egész kíberteret egy egységes, szemantikus alapú háló szövi át, ahol a találatok a fogalmak jelentései, és nem pusztán karaktersorozatuk után születnek, hanem egy egységes ontológia alapján. Sajnos a korábbi **remények**, miszerint az ilyen **ontológia hamar elterjed, túlzottan optimistának bizonyultak**. A Berners-Lee alapította W3C (Világszerte Működő Konzorcium, *World Wide Web Consortium*) saját ontológiakészítő nyelvet alkotott, az OWL-t (hálózati ontológiai nyelv, *Web Ontology Language*), amely az elterjedt RDF (forrásleíró keretrendszer, *Resource Description Framework*) nyelvre épül. A megjelenítést egy példával illusztrálom [16]: „A Dekameron szerzője Boccaccio” állítás a szemantikus weben.

> http://nektar.oszk.hu/resource/manifestation/2804140 (Dekameron [egy bizonyos kiadása])
>
> http://purl.org/dc/terms/creator (szerzője)
>
> http://viaf.org/viaf/64002165 (Boccaccio).

1.2.7. Címkézés, annotáció

A könyvtári gyakorlatból vette át a szemantikus technológia az annotációt. Ennek leggyakoribb formája a címke (*tag*). **A címkézés** (*tagging*) **lényege, hogy a fogalmak mellé** (könyvtár esetében ezek rendesen könyvcímek, szerzők stb.) **további kulcsszavakat, tárgyszavakat mellékelnek a hatékonyabb kereshetőség végett.** Ez a gyakorlat szinte kikerülhetetlen a régi papír alapú irattárak beszkennelésekor, ha az nem elfogadható minőségű OCR (optikai karakterfelismerés, *Optical Character Recognition*) technikával történik. Ugyanez igaz a

hanganyag írott szöveggé történő átalakításának (beszédből – írott, géppel olvasható – szöveg, *speech-to-text*, S2T) technikájával nem vagy nem kielégítő módon feldolgozott hanganyagokra, illetve hang és/vagy kép alapján feldolgozott multimédia-anyagokra. A címke mint metaadat származhat egy előre megadott halmazból (például legördülő menüből lehet választani) vagy bármilyen szabadszövegű karaktersorból. Az előbbi előnye a lényegesen jobb kategorizálhatóság, az utóbbié a korlátoktól való mentesség.

A címkézés mint szemantikus annotáció egy különleges megjelenési formája a közösségi helyeken megjelenő közösségi címkézés, azaz közismert nevén a *folksonomy*. Ennek lényege, hogy a web2 portálokon megjelenő szövegeket az olvasók rövid megjegyzésekkel vagy kategóriába sorolással látják el. Ilyen például egy filmnek, könyvnek, egy politikus megnyilvánulásának stb. pontokkal történő értékelése. Szemléletes megjelenítése a címkefelhő (*tag cloud*), amely a betűk méretével és színével jelzi egy hangulatelemzés (*sentiment analysis*) során előforduló kifejezések gyakoriságát és érzelmi töltöttségét.

1.2.8. Formalizált fogalomrendszerek

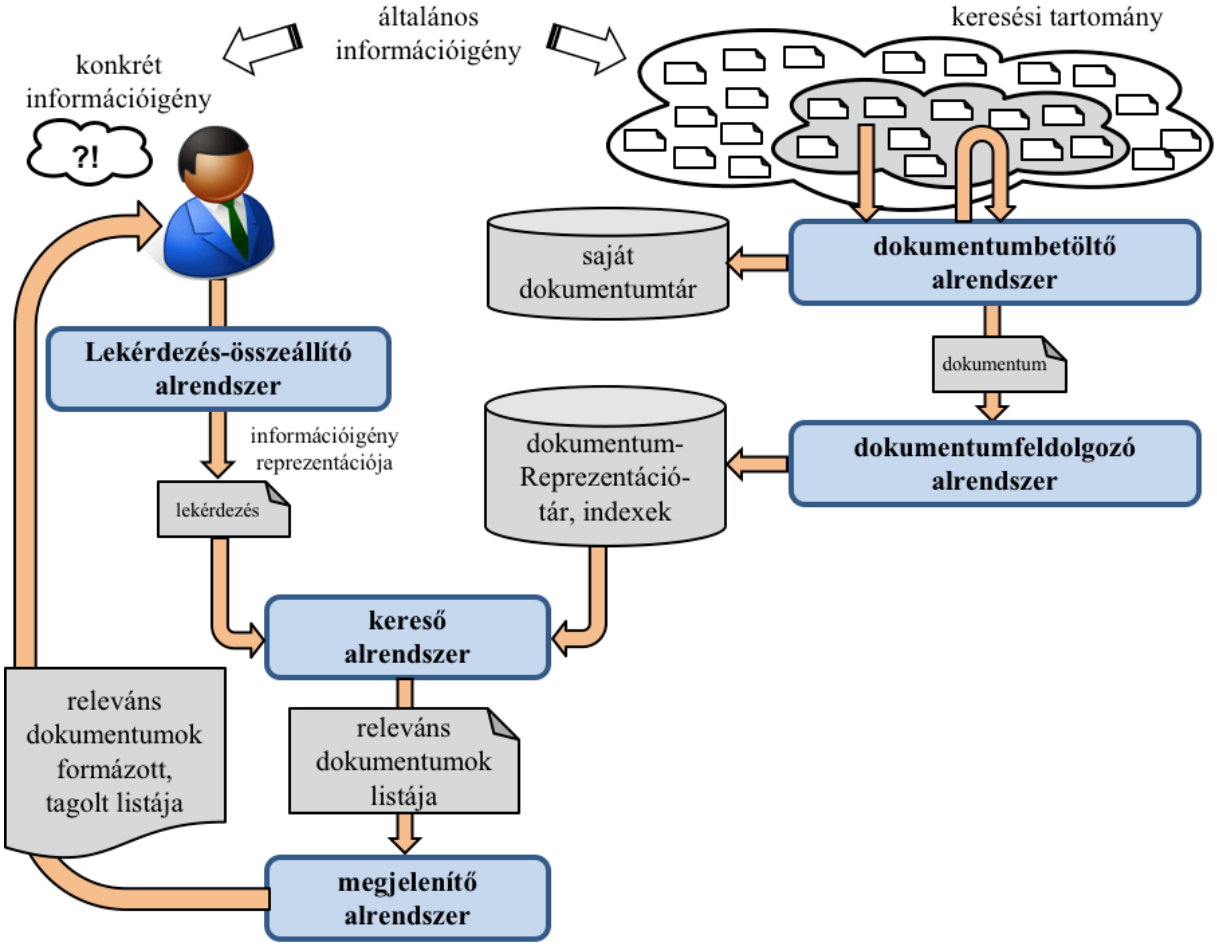
Az annotációk egy-egy fogalomhoz kapcsolódó szavak, jelek. Formájukat tekintve szabadok. Ugyanakkor a bonyolultabb jelentésbeli összefüggésekre formalizált fogalomrendszereket használunk. Ezek a struktúrák a számítógépes nyelvészet elengedhetetlen szemantikai segédeszközei. Szerkezetüket tekintve lehetnek mellérendeltek, amelyek a rokonértelműséget (szinonimák), az azonosalakúságot (homonímiák) vagy az ellentétes értelmet (antinómiák) képezik le, alárendeltséget mutató hierarchikusak (hiperonímiák) vagy rész-egészt mutatók (metonímiák). A szemantikus hálók főnevek, igék, melléknevek és határozószavak akár egész nyelvre kiterjedő sokaságát képezik le. Ilyen a WordNet [17] vagy a Magyar Wordnet [18].

1.2.9. A szemantikus keresés minősége

Az információkeresés hatékonyságát a 4.4 fejezetrészben vizsgálom. Itt viszont ki kell térni a szemantikus keresés minőségének rövid vizsgálatára. A Boole-algebrai operátorokkal történő keresés eredménye bináris, azaz a keresett kulcsszó vagy egyezik a keresési térben levő elemekkel, vagy nem egyezik azzal. Más szóval vagy van találat, vagy nincs. Amikor távolságokkal, valószínűségekkel jellemezzük a keresés eredményét vagy minőségét, akkor a válasz nem egyértelműen igen-nem, hanem a megfelelést valamekkora valószínűség jellemzi. Sok találat esetén, ami a legtöbb keresés eredménye, a találatokat a keresőrendszerek a valószínűségi mutatók, azaz a relevancia alapján rangsorolják. Azok a találatok kerülnek a találati lista elejére, tetejére, amelyeknek a relevanciája nagyobb, vagyis a keresés és a találat

leginkább egyeznek [19], majd a felsorolás a relevancia alapján csökkenő sorrendben folytatódik lefelé. Ez a rangsorolási módszer jellemzi a Google-tól a Lucene alapú keresőrendszerekig a legtöbb modern alkalmazást.

1.2.10. Az információkeresés munkafolyamata, a rendszerek architektúrája



5. ábra: egy információkereső rendszer átfogó felépítése.[10]

Az 5. ábra sematikusan ábrázolja egy információkereső rendszer munkafolyamatait és architektúráját. A szakirodalomban [20], [21] és [22] alapvetően a következő munkafázisokat különítik el. Ezeket megvizsgálom elemenként.

[10] Forrás: [4]

A keresési térből a dokumentumbetöltő alrendszer bejuttatja a saját dokumentumtárba a keresési kritériumok alapján megtalált dokumentumokat. Webes keresés esetén ezt a folyamatot ún. web-crawlerrel[11] végzi el a keresőrobot.

A dokumentumbetöltő alrendszer lehívja a saját dokumentumtárból, és egységes formátumra hozza a különböző forrásokból származó adatokat, majd továbbítja a dokumentum-feldolgozó modulokhoz azokat. A feldolgozás és ezen belül az indexelés után a keresésre már előkészített indexállományt a dokumentumreprezentáció-tárban gyűjti. A dokumentumfeldolgozó alrendszer fő feladata, hogy az egyes dokumentumokhoz a kereső alrendszer számára kezelhető reprezentációt rendeljen.

A lekérdezést összeállító alrendszer állítja elő a felhasználó információigénye szerint a lekérdezést (*query*) a kereső alrendszer számára feldolgozható reprezentációban. Az írott lekérdezést az alrendszer több módon is támogathatja: javasolhatja a lekérdező fogalomkiegészítését (Google) a leggyakoribb hasonló lekérdezések alapján vagy egy több szempontú lekérdezés (*faceted search*) támogatásával (ld. 6. ábra). A lekérdezéseknek öt fajtáját ismerjük: lekérdezés az alkalmazás saját utasításaival, űrlap kitöltésével, menük segítségével, képernyőn mutatóeszközzel történő beavatkozással és szabadszövegesen. Az írott lekérdezést nagy pontossággal tudják a hangalapú megoldások helyettesíteni (ilyen az Apple Siri, az Amazon Echo, a Google Now és a Microsoft Cortana).

A kereső alrendszer az információkereső rendszer központi eleme. Feladata, hogy a lekérdező alrendszer által továbbított lekérdező dokumentumreprezentációkat összevesse a dokumentumreprezentáció-tár állományával, és a releváns dokumentumokat, találatokat továbbítsa a megjelenítő alrendszerhez.

A megjelenítő alrendszer feladata a keresési eredmény bemutatása a felhasználónak. A megjelenítés minősége jelentősen befolyásolhatja a felhasználói élményt (*user experience*, UX), amelynek optimalizálása önmagában is egy iparággá vált. A relevancia foka a megjelenítésnél látható, például a Microfocus IDOL esetében. Az átláthatóság érdekében a megjelenített dokumentumoknak egy rövid kivonata látszik a hozzá tartozó linkkel, amelyre rákattintva a teljes dokumentum láthatóvá válik. Lehetséges a találatok valamely szempont szerinti csoportosítása, amely az átláthatóságot tovább növeli.

[11] Web-crawler: nincs igazán elterjedt magyar megfelelője. Csúszómászó, pók stb.

6. ábra: több szempontú lekérdezés.[12]

A 7. ábra ugyanezt a feldolgozási folyamatot mutatja más szövegbányászati megközelítésben. A dokumentumtárból az előfeldolgozás után a feldolgozáshoz kerülnek, majd feldolgozott formában tárolódnak a lekérdező-kereső számára.

[12] Forrás: [23].

7. ábra: a szövegbányászat modellje.[13]

1.3. Az információkeresés matematikai-informatikai alapjai

1.3.1. Az információkeresésben használt megközelítések

A modern információkeresés oroszlánrésze a XXI. században már informatikai alapon történik akár zárt rendszerekben, akár nyílt forrásban. Függetlenül attól, hogy az adatokat strukturált, félig strukturált vagy strukturálatlan formában tárolják, az adatok kezelése fejlett és rohamosan fejlődő matematikai apparátust használ. A jelen fejezetrészben bemutatom a strukturálatlan és félig strukturált adatok kezeléséhez szükséges matematikai-informatikai háttér alapfogalmait és eljárásait azzal a céllal, hogy a következő fejezetek által bemutatott alkalmazások még érthetőbbé váljanak. Természetesen nem törekszem teljességre. A következőkben ismertetek néhányat a leggyakrabban használt fogalmak közül, amelyeket a területen használnak.

1.3.2. Keresés dokumentumokban

Az információkeresés alapja a mintaillesztés (*pattern matching*), amelynek során egy dokumentumreprezentációt összehasonlítunk egy másikkal.

A legegyszerűbb keresési mód a Boole-algebrai operátorokkal történő karaktersoros (*string*) vagy másképpen kulcsszavas keresés. Itt halmazelméleti megközelítésben a keresőszó és a keresési tér mint halmazok metszetét, unióját vagy ezek valamilyen kombinációját keressük, amely a találati halmazt alkotja. Ennek az egyszerű módszernek a hátránya, hogy ha a keresőkritérium nem pontosan egyezik a keresési tér egyébként fontos elemével, akkor az a találati halmazból kiesik. A felhasználó bevisz a mezőbe egy betűkből és esetleg számokból, jelekből álló karaktersort, amelyet a motor összehasonlít a szövegtest (korpusz) tartalmával, és kimutatja a találatokat. Ennek egy változata a ma leginkább használatos kulcsszavas keresés,

[13] Forrás: [12] 22. oldal

amely során a karaktersor értelmes szavakból vagy azok csonkolt részeiből áll. A karaktersort egyes keresőknél ki lehet egészíteni speciális karakterekkel (pl. *, ?), (*wildcard*), amelyekkel minden olyan találatot kimutat, amelyekben megadott karaktersor előfordul. A karaktersoros keresést kiegészíti a keresett mezők kombinálása, értékek behatárolása, az úgynevezett speciális keresés, amelyet pl. a Google-ban is alkalmaznak. A karaktersorokat Boole-algebrai jelekkel, operátorokkal (AND, OR, NOT) és ezek kombinációival kapcsolhatjuk össze, amelyekkel a találati halmazt pontosíthatjuk. Az eredeti három operátort később kibővítették a szavak távolságának mérésével (NEAR3 például azt jelenti, hogy a két keresett szó ne legyen távolabb egymástól, mint 2 szóköz) [24]. A Boole-algebrai keresés mint a karaktersoros keresés kiterjesztése alkalmas ún. ad hoc lekérdezésekre, de a mindennapi felhasználó igényeit ez a „keresőszavas" módszer csak korlátozottan tudja kielégíteni. A felhasználó elvárja, hogy a rendszer nemcsak azokat a dokumentumokat adja ki, amelyek a konkrét keresőszót tartalmazzák, hanem minden releváns dokumentumot, akár szerepel benne a szó, akár nem. Más szóval **szabadszavas, jelentés alapú keresést vár el**.

Mivel a legtöbb komplex keresés nem pontos találatot eredményez, így szükséges a dokumentumok hasonlóságának matematikai kezelése. Ehhez szükséges a matematikai apparátus, amely a hasonlóságot mérhetővé teszi. Ez az eszköztár a dokumentumok reprezentálásán alapul, amelynek lényegét alább ismertetem. A lekérdezés során a rendszer a hasonlóság mértéke, azaz a relevancia alapján rangsorolja a találatokat [25].

1.3.3. Dokumentumok reprezentálása

A dokumentumok hasonlóságát, amely a műveletek alapja, többféleképen mérhetjük. Az alábbiakban két módszert mutatok be, amelyeket leggyakrabban alkalmaznak egymást nem kizáró módon. Az egyik a kifejezés gyakoriságát vizsgálja, a másik a vektortérmodellt alkalmazza.

Egy kifejezés gyakoriságát (*term frequency,* tf) egy dokumentumon belül az előfordulás számával mérjük. De ez a gyakoriság nem feltétlenül tükrözi vissza a kifejezés fontosságát is. A stopszavak (ld. lejjebb) igen gyakoriak, de jelentéktelenek. Ezt ellensúlyozandó bevezették az **inverz dokumentum gyakoriság fogalmát (*inverse document frequency*, idf), amely azon dokumentumok számával fordítottan arányos, amelyekben az adott kifejezés előfordul**. Könnyen látható, hogy a kettő szorzata, azaz a kifejezés „egyedisége", a tf*idf annál nagyobb, minél többször fordul elő a kifejezés egy adott dokumentumban, és minél ritkábban az egész szövegtestben.

A vektortérmodell a dokumentumokban szereplő összes szót felsorolja egymás mellé, majd minden dokumentumban jelzi, hogy az adott szó abban szerepel-e vagy sem. Egy példával illusztrálom a vektortérmodellt, mivel ez a szövegbányászatban meghatározó jelentőségű. A 2. táblázat oszlopaiban William Shakespeare hat műve szerepel, míg a soraiban az egyes szereplők neve és más fogalmak állnak. A metszéspontokban pedig azt látjuk, hogy az egyes szereplők neve előfordul-e a darabban. Más szavakkal, a dokumentumok kollekciójából, azaz a korpuszból egy fogalomelőfordulás-mátrixot (*term-document incidence matrix*) képeztünk. Természetesen ennek az ún. szózsákmodellnek (*bag of words*) több hátránya is van, például, hogy nem tükrözi a szó helyét, valamint a szavak sorrendjét és számát.

2. táblázat: fogalom előfordulás mátrix[14]

	Antonius és Kleopatra	Julius Ceasar	A vihar	Hamlet	Othello	Macbeth
Antonius	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Ceasar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Kleopatra	1	0	0	0	0	0
Irgalom (mercy)	1	0	1	1	1	1
rosszabb (worser)	1	0	1	1	1	0

Ha fontosságot tulajdonítunk a szavak előfordulása számának, mert az tükrözi valamilyen mértékben azok jelentőségét, akkor a 3. táblázat szerint jelölhetjük azokat is. A szavakat nemcsak binárisan súlyozhatjuk (azaz 1, ha jelen van, és 0, ha nincs), hanem a dokumentumokban történő eloszlásukat vagy a szavak információmennyiségét is figyelembe vehetjük.

[14] Forrás: [26].

3. táblázat: előfordulás-mátrix jelölve a szavak számát a dokumentumokban.[15]

	Antonius és Kleopatra	Julius Ceasar	A vihar	Hamlet	Othello	Macbeth
Antonius	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Ceasar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Kleopatra	57	0	0	0	0	0
Irgalom (mercy)	2	0	3	5	5	1
rosszabb (worser)	2	0	1	1	1	0

Egy új dokumentumnak a már felsoroltakkal történő összehasonlítása a legegyszerűbben úgy történik, hogy az új dokumentum oszlopát összehasonlítjuk a táblázattal, és megállapítjuk, mely szavak szerepelnek benne, és melyek nem. Az összehasonlítást kétféleképpen végezhetjük el. A két vektor által bezárt szög koszinusza jelzi a hasonlóságot, illetve a végpontok közötti szakasz (euklideszi) távolságot. Gyakorlatban ez azt jelenti, hogy egy egyszerű lekérdezés vektorához (*query*) a keresés meg kell találja a leginkább hasonló találatot, azaz azt a vektort, amely a legkisebb szöget zárja be vele. Általában, a lekérdezés (*query*) maga is egy speciális dokumentum, amely a lekérdező szavak, kifejezések kombinációjából áll elő. A lekérdezés során a lekérdezés mátrixát hasonlítják össze a tárgytér mátrixával, és megállapítják a hasonlóságok különböző mértékek alkalmazásával.

Az alábbi roppant egyszerű példával illusztrálható a dokumentumok hasonlóságának mérése. Tekintsük a következő két mondatot mint dokumentumot. Az előfordulás-mátrix alább látható a 4. táblázatban.

D1: A kislány szép.

D2: A kislány okos.

[15] Forrás: [27]

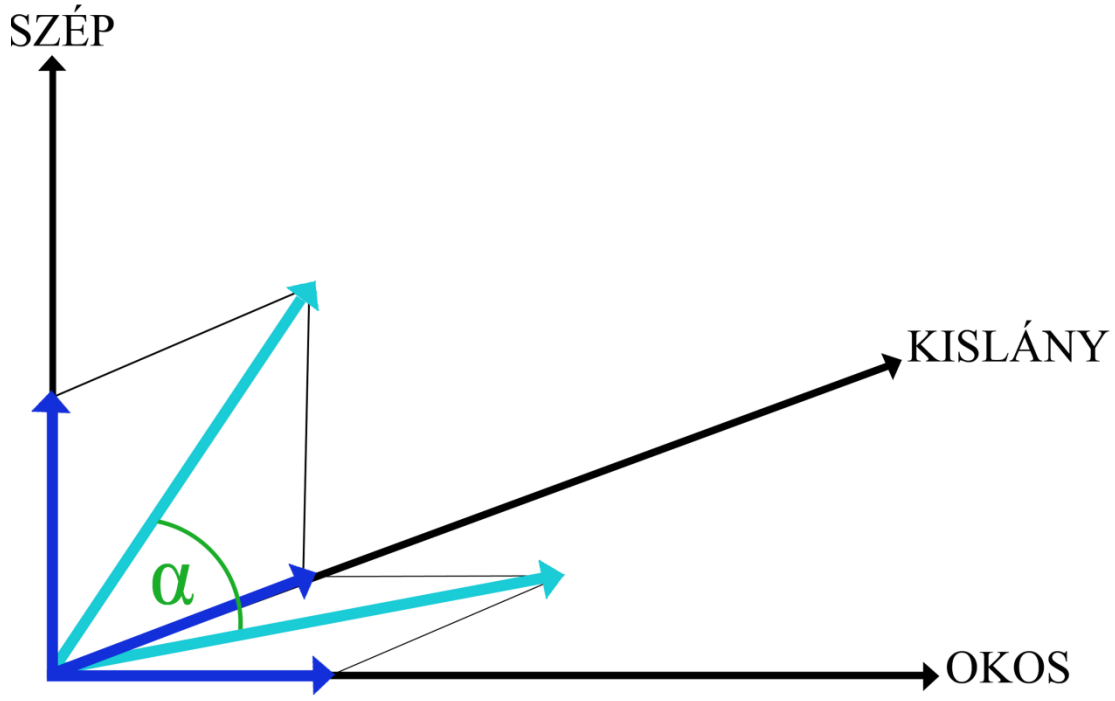
4. táblázat: két egyszerű mondat előfordulás-mátrixa.[16]

	D1	D2
A	1	1
kislány	1	1
szép	1	0
okos	0	1

A 8. ábra szerinti háromdimenziós koordináta-rendszerben ábrázoljuk a dokumentumokat a szavak mint vektorok által kifeszített térben. Az egyszerűsítés jegyében az „A" határozott névelőt mint stopszót (ld. lejjebb) ejtjük. A három szót a három tengely mentén egy-egy vektor reprezentálja. A két eredő (szép kislány, okos kislány) reprezentálja a két mondatot. Ezek hasonlóságát az α szög mutatja, melynek koszinusza a két vektor skalárszorzata. A szög 60°, annak koszinusza ½. Ez nem is meglepő, hiszen a szavak fele azonos (kislány), másik fele nem (szép, ill. okos). Természetesen az életben lényegesen több szót tartalmaznak a vizsgált dokumentumok, ezek előfordulás-mátrixai óriásiak – és ritkák –, amelyeket nem lehet ilyen módon vizualizálni.

Mivel egy szövegtestben nagyszámú dokumentum szerepelhet, és azokon belül is nagyszámú szó vagy kifejezés, amelyek az előfordulások számát, és így a dokumentumot leképező mátrixot hatalmas méretűre duzzasztják, szükséges ezen ún. előfordulás-mátrixoknak a méretét csökkenti a gépi feldolgozás (idő- és tárigény) megkönnyítése végett. A modern információ-visszakereső rendszerekben (*information retrieval, IR*), a dokumentumok besorolása automatikusan történik. Ennek eszközrendszere az indexelés. Az indexelés folyamatát legkönnyebben egy könyv végére lapozva lehet illusztrálni. Az Index rovatban szerepelnek a fontos szavak vagy kifejezések ábécésorrendben, és minden fogalom mellett az oldalszám/ok, ahol a szavak vagy kifejezések előfordulnak.

[16] Forrás: szerző.



8. ábra: két mondat hasonlósága.[17]

Tehát, ha keresek egy kifejezést, akkor nem kell végigböngésznem az egész szövegtestet, hanem az indexhez fordulva csak követnem kell az index mutatóját (*pointer*) a találathoz. Az információkereső rendszerek is ezen az alapon működnek. A robot leindexeli a már előkészített állományt, és így a strukturálatlan szövegtestből kereshető, strukturált állományt, adatbázist készít.

A látens szemantikus indexelés (*latent semantic indexing*, LSI) egy eljárás, amely az egyes kifejezések (*terms*) és kategóriák (*concepts*) között kapcsolatot talál. Tipikus alkalmazási területe a szinonim kifejezések felismerése és csoportba rendezése.

Hibatűrő mintaillesztésre (*fuzzy matching*) akkor van szükség, amikor a keresőkifejezés karaktersorozata nem pontosan egyezik a találati karaktersorral. Például gépelési hiba, helyesírási hiba esetén.

1.3.4. A gépi tanulás

Az osztályozó algoritmusok feladata a dokumentumok osztályba sorolása. Egy osztályba hasonló dokumentumok tartoznak, és lehetőleg csak azok. **A dokumentumok**

[17] Forrás: saját grafika.

osztályozásának lényege, hogy a folyamat eredményeként tanítási folyamat segítségével a hasonló dokumentumok ugyanabba a csoportba kerüljenek, és a csoportok különbözőek legyenek.[18] Vizuálisan ez a folyamat úgy képzelhető el, hogy a feladat az osztályok közötti optimális elválasztó meghatározása. Kétdimenziós térben ez egy vonal. Sokdimenziós térben az elválasztó egy ún. hipersík. Szövegosztályozás esetében a feladat a dokumentumok tartalom szerinti rendezésének automatizálása. Ennek során minden egyes dokumentumhoz hozzárendel a rendszer egy előre definiált kategóriacímkét. A kategóriákat és az azokhoz tartozó mintadokumentumokat emberi erővel állapítják meg (felügyelt tanítás). Ez az ún. tanítókörnyezet. A besorolási művelet automatikusan zajlik. Ez maga az osztályozás. Néhány példa a szövegosztályozásra: spamszűrés, többértelmű szavak egyértelműsítése (*word sense disambiguation*, WSD), automatikus metaadat-készítés, hírügynökségbe befutó jelentések automatikus osztályozása, nyelvfelismerés stb. Az osztályozás feltétele, hogy az egyes eseményeket automatikusan vagy emberi közreműködéssel osztályokba soroljuk, felcímkézzük.

A felügyelet nélküli tanulás népszerű eszköze más, hasonló vagy átfedő elnevezései a szegmentálás vagy a (felügyelet nélküli) osztályozás, csoportosítás[19] (*unsupervised learning*). **A dokumentumok csoportosításának lényege, hogy a folyamat eredményeként tanítási folyamat nélkül a hasonló dokumentumok ugyanabba a csoportba kerüljenek, és a csoportok különbözőek legyenek**[20].

A dokumentumok hasonlóságát ún. klaszterhipotézis alapján határozzuk meg. Eszerint azok a dokumentumok hasonlóak, amelyek szóhasználatukban hasonlóak, vagyis minél több közös szó van bennük, annál hasonlóbbak.[21] Az ismert vektortérmodell segítségével kezelhető a probléma az euklideszi vagy koszinusz távolság segítségével.

Az információ-visszakeresés (*information retrieval*, IR) során strukturálatlan vagy félig strukturált szövegtestekből választunk ki dokumentumokat a kereső személy

[18] [12] 102. oldal

[19] Nevezik még klaszterezésnek is.

[20] [12] 145. oldal

[21] [12] 146. oldal

információigénye szerint.[22] Az információ-visszakeresés célja a felhasználó számára lényeges, releváns információ megtalálása, kiemelése. Információ-visszakeresés a könyvtári ETO rendszer alapján történő keresés, de az az interneten kulcsszavakkal vagy azok kombinációival történő keresés is. A fejlettebb keresőrendszerek előre beépített és karbantartott kontrollált szótárat is alkalmaznak. A találatokat a szemantikus keresés minősége pontban kifejtett szempontok szerint rangsorolják.

Az információkinyerés (*information extraction*, IE) lényege a szövegtestből a felhasználó számára lényeges szövegrészek kinyerése.[23] Más szavakkal szabadszövegű, strukturálatlan korpuszból strukturált információ kinyerése, azaz egy adatbázisba rendezése a lényeges szövegrészeknek. Ilyen például a sajtófigyelés bizonyos témákra, a versenytársfigyelés, jogszabályi változások kiszűrése stb.

1.4. Az információkeresés nyelvészeti alapja

Az információkeresés és ezen belül a szemantikus keresés alkalmazhatóságának megismeréséhez elengedhetetlen az információkeresés számítógépes nyelvészeti alapfogalmainak ismerete. A tudományterület részletesebb vizsgálata nem lehet tárgya a jelen írásnak. Csak egyes részterületeket tudok ismertetni, amelyek a további fejezetek szempontjából lényegesek.

1.4.1. A nyelvi előkészítés folyamata

Az alábbi módszerek csak a leggyakoribbak a számítógépes nyelvészet területén. A témától függően nem mindet alkalmazzák, a szövegtest jellemzőitől függően a hatékonyságuk is különböző. A folyamat lényege a szabad szöveg minél egyszerűbbé tétele a gépi értelmezés és feldolgozás számára. A példák természetesen csak a legegyszerűbb eseteket demonstrálják a megértés céljából. Ezek illusztrálják, milyen alapvető lépésekkel lehetséges a nyers szövegtestet az érdemi folyamatokhoz előkészíteni. A strukturálatlan szövegtesteket strukturált adatbázisokba kell átalakítani, amelyeken matematikai-informatikai módszerekkel már eredményt lehet elérni. Ezt az előkészítő folyamatot nevezik előfeldolgozásnak. Miután a nyers szövegtestet az alábbi és számos más módszerrel előkészítették, megindulhat annak feldolgozása.

---

[22] [12] 63. oldal

[23] [12] 81. oldal

A szöveg számítógépes értelmezése szempontjából a dekompozíció fontos lépése a mondatokra bontás. A legegyszerűbb esetben a mondatzáró (. ! ? stb.) és -nyitó (nagybetű) jelekből lehet következtetni a mondathatárokra. Bonyolult feladat a mondatközi pont vagy a nagybetűvel kezdődő tulajdonnevek kiszűrése: ilyenek például az e-mail-címekben, dátumokban szereplő írásjelek, rövidítések stb. Ezeket kontextusfüggő, programozható szabályok segítségével lehet megoldani. Ilyen szabályok alapulhatnak formailag egységes reguláris kifejezésekre, például a dátum vagy IP-cím formátumú szám- vagy karaktersorokban szereplő pontok nem jelentik a mondat végét.

A tokenizálás lényege, hogy az írásjelek elhagyása után a szöveget szavakra vágjuk a szóközök mentén[24]. Ez a feladat nem magától értetődő, mert például a kötőjellel elválasztott dupla nevek (Catherine Zeta-Jones), a házasság utáni kettős vezetéknevek (Benita Ferrero-Waldner) stb. vagy kereszt- és vezetéknévből álló tulajdonnevek (Johann Wolfgang von Goethe) szétválasztása jelentésvesztést okozhat. Az ilyen szósorokat is egy fogalomnak, egy tokennek tartjuk. Míg a token egy konkrét megjelenése egy adott karaktersornak, **a típus több azonos karaktersor összes megjelenése**. Ennek a szótár készítésénél és az indexelésnél is van jelentősége.

A szófajra bontás (*part-of-speech tagging, POS tagging*) a szövegtestben szereplő szavak szófaji elemzését jelenti.[25]

A parsing vagy szintaktikai elemzés a mondat elemeire bontása[26]. Ennek eredménye egy szóelemző fa, amely a szóelemek egymáshoz való mondattani viszonyát mutatja. Ez a technika segíti a szöveg szemantikai értelmezését, de több okból (a nyelv komplexitása, a gépi feldolgozás határai) csak korlátozott mértékben alkalmazzák, és a 80-as évektől már nem jelenti a nyelvi feldolgozás legfőbb irányát, de egyes alkalmazások kombinálják a statisztikai módszerekkel.

Nyelvfüggetlen folyamat a dokumentumok N hosszúságú karakterekre történő feldarabolása. Különösen előnyös olyan nyelveknél, amelyeknél szóköz nem létezik (például a japán vagy kínai). N-grammokat alkalmaznak például egy szöveg nyelvének felismerése vagy a

[24] [12] 39. oldal

[25] [25] 344. oldal

[26] [25] 107. oldal

dokumentumosztályozás területén. Ez úgy zajlik, hogy minden lehetséges kategóriában (pl. német nyelv, bőrgyógyászat) veszünk egy nagyobb minta dokumentumhalmazt, amelyre részletes N-gramm statisztikai profilt készítünk a leggyakoribb szótöredékekről. A kategorizálás folyamán a dokumentumot is N-grammokra bontjuk, és egy metrika segítségével megkeressük a hasonló profilokat[27].

A dokumentumok reprezentálását megkönnyíti, ha **a tartalom szempontjából lényegtelen szavakat, a stopszavakat** kiszűrjük. Ilyenek a névelők, névmások stb. Ez csökkenti a dokumentummátrix méretét, és így a feldolgozás idő- és tárigényét.

A feldolgozás további lényeges egyszerűsítése végett a szavakból levágjuk a toldalékokat. Ez nyilván tompítja az értelmezést, de a stopszavak kiszűréséhez hasonlóan a dokumentummátrix méretét nagyban redukálja. A szó előfordulásának megléte szempontjából a toldalék lényegtelen, viszont a statisztikai feldolgozás számára így lényegesen kevesebb alakkal kell dolgoznia a rendszernek. Megjegyzendő, hogy míg **a lemmatizálás során a szó szótári alakját állítjuk elő**, a **szótő-visszaállítás (*stemming*)** akár egy önmagában értelmetlen szócsonkot is eredményezhet, ami a feldolgozás szempontjából még egyszerűbb, mint a lemma. A feldolgozás bonyolultsága miatt a szótő-visszaállítás különösen fontos az ún. agglutináló nyelvek számára. Ezek a mondatrészek közötti összefüggéseket elsősorban nem elöljárókkal, hanem toldalékokkal fejezik ki. Agglutináló nyelv a magyar, a török, az észt vagy a koreai.

A névkifejezések felismerése (*named entity recognition*, NER) az állandó formátumú vagy formai jellegzetességgel bíró elemek felismerése és kiemelése további feldolgozás végett.[28] Ilyen egy dátum, hitelkártya-szám, telefonszám vagy egy tulajdonnév, városnév vagy egy bírósági határozatban a vádlott neve, egy hivatkozott törvény száma stb. A gyakorlati jelentősége abban áll, hogy strukturálatlan, szabad szövegből a felismeréssel és kiemeléssel könnyebben lehet strukturált és így könnyebben kereshető állományt készíteni. Ilyen technológiát fejlesztett ki – többek között – nemzetbiztonsági alkalmazásra a Szegedi Tudományegyetem Informatikai Tanszékcsoportja [28]

A jelentés-egyértelműsítés (*word sense disambiguiation,* WSD) az azonos alakú szavak értelem szerinti szétválasztása[29]. Erre egy módszer a szó szövegkörnyezetének vizsgálata.

[27] [12] 39. oldal

[28] [12] 90. oldal

[29] [26] 229. oldal

Például: Javában programoznak, de Java szigetére hajóznak. Ilyen egyértelműsítést kér tőlünk pl. a Google a „*do you mean*” funkcióval vagy a Wikipédia *disambiguation* funkciója.

1.4.2. A tudásreprezentálás nyelvészeti eszközei

Ebben a fejezetrészben bemutatom a tudásreprezentálás legfontosabb eszközeit és módszereit. Az alább ismertetett eszközök és módszerek a tudásmodellezés alapvető eszközei, és mint ilyenek a tudásmenedzsment eszköztárába tartoznak. E szemantikus nyelvészeti eszközök közvetlen felhasználási területe az információkeresés, ugyanis az egyszerű karaktersorokra, kulcsszavakra történő keresés hatékonysága elenyésző a fogalmi kapcsolatokat reprezentáló, jelentésalapú, ún. intelligens kereséssel szemben. Felismeri a formailag különböző, de fogalmilag összetartozó elemeket (szinonimák, hierarchikusan egymás alá tartozó fogalmak stb.), és a látszólag független, de tartalmukban összefüggő dokumentumokat egymáshoz rendeli. Mundie és McItire [29] hat szintjét határozza meg a tudásreprezentációs strukturáknak:

- az ellenőrzött szótár (*controlled vocabulary*) a kiemelt szavak szótára;
- taxonómia;
- a statikus ontológia a statikus megjelenések leírása;
- a dinamikus ontológia a dinamikus események leírása;
- az intencionális ontológia a részvevő szubjektív motivációira épül;
- a metamodell egy keretrendszer, amely a beadott paraméterek alapján ontológiákat készít.

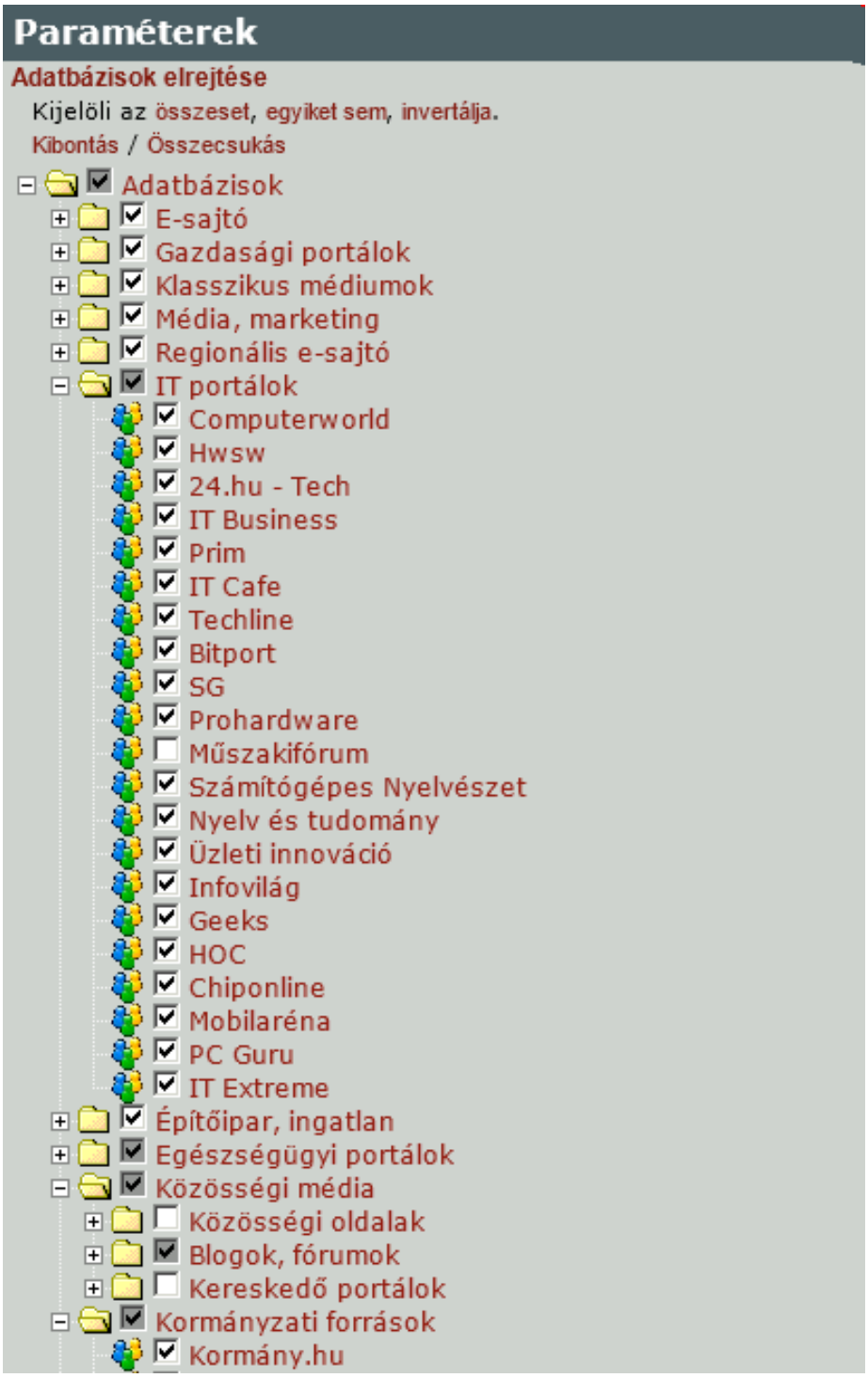
A kontrollált szótár (*controlled vocabulary*), taxonómia, szótár, tezaurusz és ontológia mint tudásmodellezés és -reprezentáció fogalma gyakran keveredik a szakirodalomban. Mindegyik eszköz feladata a gyorsan növekvő jogi szövegmennyiségben történő műveletek támogatása. Ezek a tudásreprezentációs eszközök hivatottak az emberi intelligenciát a gépi intelligencia számára leképezni, de ez a művelet a mai tudásunk szerint nyilván csak megközelítés. Fontos megjegyezni, hogy a fenti struktúrák – normálás után – tetszés szerint bővíthetők, összevonhatók. Használhatóságuk nemcsak a megtalálhatóság (*findability*) javítása, hanem a képernyőn való intelligens navigáció elősegítése is.

A filozófiában az ontológia a léttel és annak legáltalánosabb törvényeivel foglalkozó tudományágat jelenti. A számítógépes nyelvészetben az ontológia definíciója nem kiforrott. Sokszor keveredik a taxonómiával, szótárral, tezaurusszal. Nemcsak a köznyelvben, de még a szakcikkekben is keverednek valamelyest ezek a fogalmak. Ez nem véletlen, a kontúrok eléggé elmosódnak. Az amerikai szakirodalomban a legáltalánosabb tudásreprezentálási eszközként

előfordul a taxonómia fogalma [31], és ennek egy alfaja, az ontológia. Mások a kettő között nem is látnak lényeges különbséget [32]. A Magyarországon általában elfogadott terminológiát követem az alábbiakban. A taxonómia és a tezaurusz mindkettő egyfajta ontológia, de míg az ontológia gazdagabb relációkat is reprezentálhat, a taxonómia csak alárendelt részhalmazt, a tezaurusz mellérendelések sokaságát és átkapcsolódásokat [33]. Az ontológiákat mint a szervezeti tudásmenedzsment fontos eszközeit Kő Andrea mutatja be a doktori értekezésében.

Előnyeit a következőképpen soroljuk fel [34] alapján:

- a tudásmegosztásban játszott szerepét a közös fogalomtár felhasználásával;
- a szakértelem újra felhasználhatóságának támogatását;
- a tudásalapú rendszerek fejlesztésének hatékonyabbá tételét (különös tekintettel a tudás megszerzésére);
- a tudásmenedzsment-rendszerek karbantartásának és továbbfejlesztésének megkönnyítését a doménspecifikus közös tudás felhasználásával;
- a verifikáció támogatását.



9. ábra: részlet egy online témafigyelő alkalmazás taxonómiájából.[30]

Az ontológia a fogalmak nemcsak hierarchikus alárendeltségét mutatja, mint a taxonómia, hanem a fogalmak közötti funkcionális kapcsolatot is leképezi. A TANULÓ tanulója az ISKOLÁ-nak, ahol minden tanuló eleme a TANULÓ halmaznak, és minden iskola eleme az

[30] Forrás: [30]

ISKOLA halmaznak. Ezek a kapcsolatok természetesen lehetnek sokrétűek és bonyolultak. Több ontológiaépítő alkalmazás közül ismert a Protegé, amelyet a stanfordi egyetemen fejlesztettek ki.

Az információkereső tezaurusz természetes nyelven kifejezett fogalmak olyan tartalmilag szabályozott, szükség szerint változtatható szótára, amelyben feltüntetik a legfontosabb fogalmi összefüggéseket.[31] A tezaurusz fő rendeltetése az információk feldolgozása és keresése. Tezaurusz például egy szótár vagy lexikon.

Mivel a gyakorlatban, különösen a magyar minősített kormányzati gyakorlatban tudásreprezentációra a taxonómiák használata terjedt el, ezért ezzel a módszerrel kicsit részletesebben foglalkozom. **A taxonómia dolgok, fogalmak fastruktúrába rendezett halmaza**. Minden alárendelt kategória viseli a fölötte levő tulajdonságait. Ilyen Karl Linné ismert osztályozása az élőlényekről, a könyvtártudományban használt Egyetemes Tizedes Osztályozás (ETO) vagy a bűncselekmények kategorizálása. A taxonómia szinonimájaként a Verity[32] használta a Topic[33] kifejezést, amely olykor megtalálható a szakirodalomban. A taxonómia szó etimológiailag a görög taxisz: elrendezés és a nómosz: törvény, tudomány szavakra vezethető vissza. Szűkebb értelmezésben a taxonómia fastruktúra, azaz egy gyermeknek csak egy szülője lehet. Kicsit tágabb értelmezésben polihierarchikus szerkezetekben egy gyermeknek lehet több szülője is. Például egy könyv egyszerre nyomdatermék és oktatási eszköz is. A polihierarchikus szerkezetek egy hasznos sajátja, hogy osztályok metszeteiben egyszerre több nézet megadásával lehetséges a keresés (*faceted search*). A magyar szaknyelv a taxonómiát a szűkebb értelemben alkalmazza, a bővebb szemantikai struktúrákat nem érti ide. Az információtudomány a késői 90-es évektől alkalmazza egyre nagyobb intenzitással a taxonómiákat. Hazai ipari használatban – már amennyire jelenleg egyáltalán létezik ilyen – leginkább egyszerű taxonómiákat használnak. Ezek többszintű, egymásba ágyazottan hierarchikus szerkezetűek, és egy-egy szinten asszociatív fogalmakat

[31] [31] 9. oldal

[32] A kaliforniai Verityt felvásárolta a brit Autonomy, amelyet felvásárolt az amerikai HP, amelynek az „Autonomy részlegét” felvásárolta a brit MicroFocus. A Topic köznévvé is vált: topic, hasonlóan, mint a dzsip a terepjáróra vagy korábban a xeroxozás a másolásra.

[33] Topicnak nevezte a Verity a keresőmotorját is, amelyet az átvétel után az Autonomy elsorvasztott, és az IDOL nevet futtatta.

vagy szinonimákat tartalmaznak. Ilyen a barkácseszközök alatt a kalapács, fűrész, fúró stb., illetve például egy vállalat neveinek variációi: BKV, B.K.V., Budapesti Közlekedési Vállalat, BESZKÁRT stb. Az ilyen szemantikai kapcsolatok felhasználására jelenleg a nagy internetes keresők, például a Google nem vagy csak kis mértékben alkalmasak. A szűkebben értelmezett taxonómiák formátuma általában XML vagy JSON. A taxonómiákat az információkeresés során három fázisban alkalmazzák.

- Az indexelés során a kapcsolatok nemcsak a fő elemmel épülnek fel, hanem minden, a taxonómia hierarchiájába beletartozó elemmel is. A rendszer így már „tudja", hogy a barkácseszközökhöz milyen szerszámok tartoznak.
- A lekérdezés során nemcsak a keresett karaktersorra kapunk találatot, hanem a taxonómia szerint hozzátartozó fogalmakra is (ld.: megint barkácseszközök).
- A – nem kötelezően – láthatóvá tett taxonómia segít a navigációban. Ilyen segítséget kapunk, ha a törvénytárat megnyitjuk, és a bűncselekménygyökérből eljutunk egészen a kiskorú ellen csoportosan elkövetett rablás kategóriájáig. Akár többféle szempont szerint is kategorizálhatjuk a fogalmi struktúrákat (*faceted taxonomy*). Például lehet újságcikkeket földrajzi hovatartozás, politikai párthoz tartozás, kulcsszemélyek stb. alapján kategorizálni (kollekciók). E vizualizációs képesség hasznosságát nem szükséges dicsérni. A taxonómiára épülő lekérdezéseket tárolni lehet későbbi ismételt felhasználásra. Ez az eszköz alkalmas egy témában való, gyakorlatban sűrűn előforduló célzott keresésre (*drill down*).

Taxonómiák használatosak mind a nyílt interneten történő keresésben, mind pedig a szervezeten belüli keresőrendszerek (*enterprise* [*content*] *search*, ES vagy ECS) során. Egy nagyvállalat taxonómiája hosszú évek alatt épül fel. Tartalmazza az emberi, szervezeti tudás egyfajta leképződését. Megőrzi kivált, nyugdíjazott, beteg stb. dolgozók felhalmozott ismereteit is. Mint ilyenre, olykor úgy vigyáznak, mint egy titkos receptre.

Taxonómiákat lehet készen venni, kapni[34] és ezeket módosítani, illetve készíteni. Bár a taxonómusok zöme a könyvtáros szakmából kerül ki, odasodródnak a területre informatikusok, bölcsészek, tudásmenedzserek és bárki, aki önképzéssel vagy szervezett oktatás keretében megismerkedik a mesterség fogásaival. Taxonómusok lehetnek főállású vagy külsős profik. Taxonómia készítésére léteznek célszoftverek, amelyek egy része nyílt forrású (ingyenes), a

[34] Egy figyelemre érdemes forrás például: http://taxonomywarehouse.com/

zöme pedig kereskedelmi licenc keretében megszerezhető (fizetős).[35] Ezek között léteznek felhőben és offline működő alkalmazások. A taxonómiaépítés egy sajátos, modern és olcsó módja a közösségi helyeken való fogalomgyűjtés (*social tagging*).

Az információkeresés gazdaságosságával a 4.4 fejezetrészben foglalkozom, de ipari tapasztalat alapján itt is megjegyzem, mindenki, akinek elég nagy mennyiségű strukturálatlan szöveget kell feldolgoznia, meg kell fontolja, hogy megéri-e taxonómiaépítésbe beruháznia. Ilyenek többek között a nagyvállalatok, nemzetbiztonsági szolgálatok és rendvédelmi szervek, kiadók és az elektromos média, tartalomszolgáltatók, PR- és marketingügynökségek, könyvtárak, múzeumok stb. Bár érezhető, és a szakirodalom is egyetért abban [31], hogy a taxonómiák javítják a keresés hatékonyságát, elég kevés számszerű adatot találni erre. A felsővezetés számára eldöntendő kérdés: érdemes-e hosszú, fáradtságos emberi munkát igénybe véve taxonómiát készíteni és fenntartani. Konferenciákon állandó téma (SKIP, ISKO, ISS), hogy – néhány kivételtől eltekintve – a nyugati nagyvállalatok is igen csekély forrást (pénzt, időt és drága embert) áldoznak erre a területre. Hazánkban a helyzet talán még siralmasabb. Pedig a hosszú távú beruházás minden bizonnyal megéri. Nemcsak a később lényegesen csökkenő felesleges keresési idő, amit a fehérgalléros gárda elvesztegetni kénytelen, hanem az elfelejtett, meg nem talált adatok csökkenése miatt is. Nem beszélve a dolgozói és ügyfél-elégedettségről. Természetesen a rövid távú érdekek sokszor itt is legyőzik a hosszú távúakat.

Egyes vélemények szerint a taxonómia és általában **a formalizált fogalomrendszerek jelentősége** a Big Datára alkalmazott statisztikai módszerek elterjedésével **elhalványulna**. Én ezzel a felfogással **nem értek egyet**, és a kutatásaim, a konferenciákon hallottak, az ipari tapasztalatok alapján más véleményt képviselek. Elismerem, hogy a nagy internetes keresők szinte kizárólag statisztikai módszerekkel dolgoznak, és a Berners-Lee-féle álom az univerzális ontológiáról nem valósult meg. Ugyanakkor **a vállalati keresőrendszerekben (ECS)**, ahol a felidézés és a pontosság (ld. 4.4 fejezetrész) alapvető igény, **a taxonómiák használata továbbra is fennáll**. Ezt a minőséget statisztikai módszerekkel a jelek szerint messze nem sikerült utolérni. Ugyanakkor azt is el kell ismerjem, hogy hosszas keresés ellenére sem találtam szakirodalmat, amely összehasonlítaná tudományos igénnyel a két módszer hatékonyságát. Egyelőre illúzió automatikus úton a humán minőségét megközelítő taxonómia készítése. A felügyelt gépi tanulás gyorsan fejlődő algoritmusai, a szinte korlátlanul rendelkezésre álló

[35] A legegyszerűbb taxonómiakészítő eszköz egy Excel-fájl. Gyártók többek között TemTres (nyílt forrású), Topic Manager/Mondeca, PoolParty/Semantic Web Company, Ontology Management/SAS, Knowledge Management System/Synaptica stb.

tanítóadat mennyisége, valamint a gépi kapacitások exponenciális növekedése valamennyire pótolni tudja a humán beavatkozást. Talán a többrétegű neurális hálókat alkalmazó algoritmusok egy nap kiszorítják az embert erről a területről is.

1.5. Összefoglalás, részkövetkeztetések

A fejezetben kimutattam, hogy korunkban az elektronikus formában informatikai rendszerek, eszközök által feldolgozható információk mennyisége rohamosan bővül, az informatikai feldolgozási kapacitások korábban elképzelhetetlen feladatok megoldását biztosítják, a korszerű hálózatok egyre szélesebb körben biztosítanak gyors, hatékony hozzáférést az információkat tároló rendszerekhez. Mindebből levonható az a **következtetés, hogy a szervezeti vagy egyéni információigények kielégítésének egyre nagyobb jelentőségű eszköze a már meglévő információk közötti keresés**.

A vonatkozó szakirodalom feldolgozása megerősítette azon személyes szakmai tapasztalataimat, következtetéseimet, hogy a peta-, exa- és zeta-bájtokkal mérhető információtömeg új megoldásokat követel. **Rámutattam, hogy** az informatika által biztosított lehetőségek bővülése következtében egyre több adat, **egy szervezet esetében akár 80-85% keletkezik strukturálatlan (szöveges, multimédia formátumú) és félig strukturált formában**. A heterogén formában rendelkezésre álló információtömegből történő információkinyerésre számos technológia alakult ki. Az információ-visszakeresés és a szövegbányászat egymáshoz kapcsolódó, de számos sajátossággal rendelkező megoldások, amelyek között és amelyek változatai között nincs egyedüli üdvözítő módszer.

A fejezetben foglaltakra, a feldolgozott szakirodalomra épülő alapvető következtetéseim első csoportja a jelentéshez, tudáshoz, tudásreprezentációhoz kapcsolódik. Értekezésem alapvető nézőpontja, hogy az információ az emberi tudás, az ismeretek egy összetevője, az adatok pedig ilyen információk hordozói, és az adatokhoz csak az ember rendel jelentést. Erre építve a legfontosabb következtetés, hogy **a szervezeti vagy egyéni információigényeket kielégítő eredményes információkeresés során valamilyen formában kezelni kell a feldolgozott adatok által hordozott jelentést**. Ez jelentősen eltérő módon kezelhető strukturált, félig strukturált és strukturálatlan (ezen belül elsősorban a szöveges) adatok esetében.

Kimutattam, hogy a jelentésalapú, szemantikus technológiák eltérő megoldásokra épülnek. Ezek egyik, **klasszikus irányát** a jelentéshez kapcsolódó, kiegészítő leíró adatokra (metaadatok, címkék), illetve **a formalizált fogalomrendszerek különböző formáira (ellenőrzött szótárak, taxonómiák, ontológiák) épülő megoldások alkotják**. A másik,

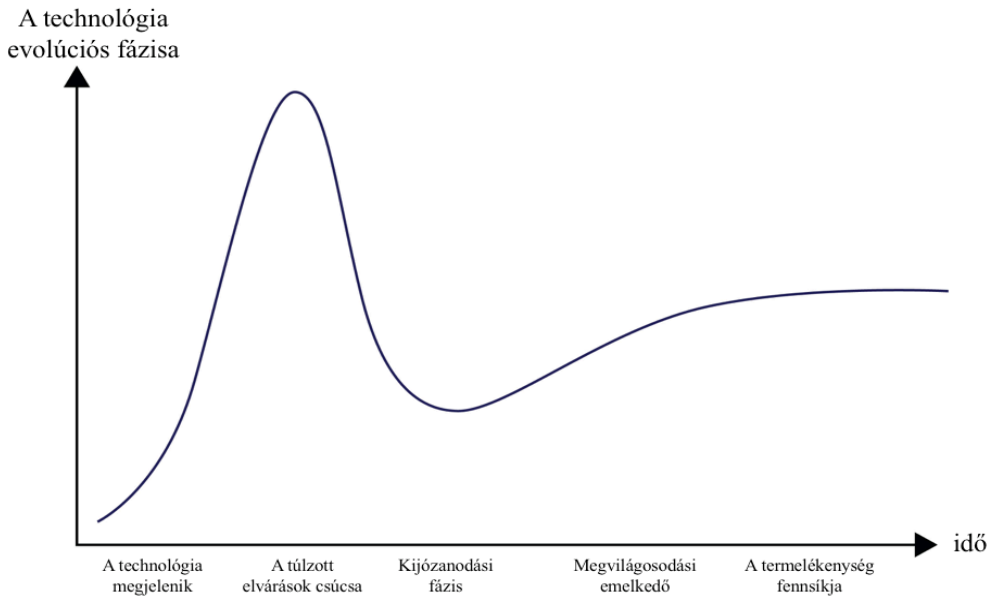
modernebb fő **irány** alapját **a** jelentésnek nagy tömegű adatokra épülő **statisztikai megoldásokkal, gépi tanulási módszerekkel történő megközelítése** képezi. Egyes vélemények szerint az első irány nem vezetett eredményre, azonban ipari tapasztalataim és kutatási részeredményeim alapján **azt a következtetést vontam le**, hogy míg **az internetes keresés során a statisztikai módszerek a dominánsak**, **a vállalati keresőrendszerek esetében** a formalizált fogalomrendszerek, különösen **a taxonómiák használata továbbra is megkerülhetetlen**, és a belátható jövőben az emberi erővel előállított formalizált fogalomrendszereket automatikus megoldások nem fogják kellő minőségben helyettesíteni.

Harmadik alapvető következtetésem, hogy az információkeresés mint az informatika által támogatott funkció, szolgáltatás – sok más informatikai szolgáltatáshoz hasonlóan – számos alap-, és alkalmazott tudományterület eredményeit, módszereit hasznosítja. A szakirodalom ezek közé sorolja elsősorban a matematikát (azon belül a matematikai logikát, a vektoralgebrát, a valószínűségelméletet és a matematikai statisztikát), a nyelvészetet (azon belül a szintaktikát, a szemantikát, a természetesnyelv-feldolgozást). Ezen területek eredményei folyamatosan termékenyítik meg az információkeresés módszereit, és az alkalmazott megközelítések szerepe, jelentősége is folyamatosan változik.

2. FEJEZET: AZ INFORMÁCIÓKERESÉS TECHNOLÓGIÁINAK NÉHÁNY GYAKORI ALKALMAZÁSI TERÜLETE

2.1. Bevezető gondolatok, a fejezet tartalma, célja

Az információkeresés technológiai alapjainak áttekintése után néhány példával illusztrálom a gyakorlati alkalmazási lehetőségeket. A példák kiválasztása önkényes, hiszen szinte végtelen számú lehetőség közül szemeltem ki az alábbi hármat. Egyetlen szempont, ami vezérelt, az ipari gyakorlatban legsűrűbben felmerült témák megszólítása. Mindhárom terület, a szentimentanalízis, a metakeresés és a fúziós központ a Gartner-féle életciklusmodell (lásd 10. ábra) első, emelkedő, korai szakaszában van még Magyarországon. A három terület abban az értelemben nem egyenértékű, hogy míg a metakeresés és a fúziós központok használata inkább technikai eljárás, addig a szentimentelemzés szemantikus alkalmazás. Mindhárom területen az elméleti áttekintés után gyakorlati példákkal, illetve ötletekkel illusztrálom az alkalmazhatóságot. Ezeket a példákat kisebb részben magyar, nagyobb részben nemzetközi forrásból merítettem. Az alkalmazási területek napról napra bővülnek. Nem kérdés, hogy a példák elavulnak-e, csak az, hogy mikor. Iparkodtam hivatkozni magyar kutatási eredményekre, ahol csak találtam forrást.



10. ábra: egy feltörekvő technológia életciklusa.[36]

[36] Forrás: [35] alapján saját szerkesztés.

2.2.Szentimentelemzés[37]

A jelentésalapú technológiai alkalmazások fejlettsége szintje egyre nagyobb mértékben tette lehetővé olyan műveleteket végrehajtását, amelyek emberi kommunikációt emulálni képesek. A korai 2000-es évek óta a szemantikai technológiai megoldások fokozatosan teret nyertek az iparban és a kormányzatban, valamint a mindennapi életben [10]. A 2000-es évek eleje óta az NLP új ága nyert teret, a szemantikus technológiákat alkalmazó szentimentelemzés. A szentimentelemzést a szöveg három különböző szintjén lehet végrehajtani, dokumentum-, mondat-, valamint fragmentumszinten. Még újabb terület az emócióelemzés, amely kiterjeszti a számítógépes nyelvészet felhasználási területét.

A szentimentelemzést széles körben használják a legkülönbözőbb közösségimédia-tartalmak vizsgálatában és értékelésében, például fórumokban, blogbejegyzésekben, üzenőfalakon, a Twitter és a Facebook esetében. A szentimentelemzés alkalmazható strukturálatlan vagy félig strukturált szövegtestekre, így wikikre, e-mailekre, vállalatok vagy szervezetek ügyfélszolgálati központjaira. A véleményen vagy az azt hordozó hangulaton alapuló besorolás hasznos az információ-visszakeresésben, amikor a feladat a bizonyos típusú dokumentumok kiszűrése.

A technológia online és valós idejű alkalmazása kiterjed a marketing- és a PR-ügynökségektől a hírnévelemző és a teljesítményértékelő szakembereken át a politikai kampánymenedzserekig, piaci márkaelemzőkig orvosi felhasználói fórumokon át a pénzügyi elemzőkig. A szentimentelemzés sokat tárgyalt téma a tudományos szakirodalomban és az iparban a 2000-es évek eleje óta. Több mint 7000 publikáció jelent meg a témában azóta. A következő szakaszban meghatározom a szentimentelemzés fogalmát és célját, az osztályozási módszereit és a szemantikus szótár fogalmát.

A szentimentelemzés az a feladat, amelynek során kinyerik a szerzők véleményét kifejező szövegelemeket egy adott a szövegtestből, majd elemzik e szövegelemek polaritását [38]. Más szavakkal, **a szentimentelemzés meghatározza az emberek egy bizonyos céltárggyal vagy célszeméllyel szemben kifejezett véleményét, értékelését, megnyilvánulását, meglátását, benyomását és személyes hozzáállását**. A szentimentelemzés az NLP egyik ága. A 11. ábra bemutatja a szentimentelemzés alkalmazásának felépítését.

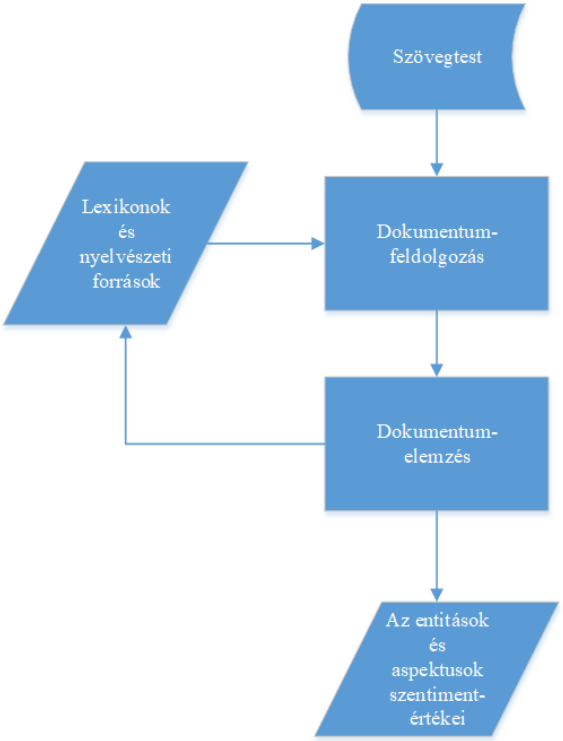
[37] Angolul: sentiment analysis, opinion mining, sentiment mining, opinion extraction, review mining. A magyar nómenklatúrában a szegedi egyetemét követem [36] és [37].

A forrás lehet bármilyen strukturálatlan szöveg, például WORD, PDF (hordozható dokumentumformátum, *Portable Document Format*), HTML (kiemelt szövegű jelölőnyelv, *Hyper Text Markup Language*), XML (kiterjeszthető jelölőnyelv, *Extensible Markup Language)* stb. szövegállomány. Ezeket a szövegállományokat előfeldolgozásnak vetik alá. A szövegállományt szentimentelemzési szakszótárral, lexikonnal, taxonómiával gazdagíthatjuk, illetve egyértelműsíthetjük a jobb teljesítmény végett. A folyamat sarokköve a dokumentum-elemzés, amelynek során az előfeldolgozott dokumentumokat a véleményezési szakszótár elemeivel annotálják. Ezek az annotációk három szinten történhetnek: dokumentum, mondat vagy értékelési szempont kategória, aspektus szintjén. A folyamat végtermékei a kalibrált annotációkkal ellátott dokumentumok.

Döntő fontosságú meghatározni, hogy hány különböző kategóriába kívánjuk besorolni a fenti módon elemzett szöveget. Tipikusan lehet két kategória (pozitív vagy negatív), három (pozitív, negatív vagy semleges), öt (értékelés 1-től 5 csillagig) vagy tíz (ugyanez 10 csillaggal). Általában véve a csoportok számának növekedése csökkenti a pontosságot, növeli a feladat nehézségét, viszont az eredmények sokkal informatívabbak.

A véleményt vagy érzelmi töltést hordozó kifejezések alapvetők egy mondat vagy dokumentum hangulati elemzésében. Ezek a kifejezések strukturált szervezetben alkotják a szentimentlexikont [39]. A szentimentlexikonok az elemzés legfontosabb eszközei. Előállításuk háromféleképpen történhet, természetesen egymást nem kizárva.

- Kézzel (emberi erőforrás segítségével), ami rendkívül munkaigényes és drága, viszont olykor elkerülhetetlen.
- Szótár segítségével, amikor is egy alapszótárt egy lexikális adatbázis segítségével gazdagítunk/dúsítunk/kibővítünk (*enrichment*) [40]. Ilyen adatbázis például a WordNet vagy a Magyar Wordnet.
- Szövegtest alapján, amikor is egy alapszótárat gazdagítunk egy nagy dokumentum-tartomány (*domain*) segítségével. Ez utóbbi történhet gépi úton, ami lényegesen gazdaságosabb, viszont kevésbé pontos.



11. ábra: egy véleményelemző rendszer felépítése.[38]

A sentiment shifterek határozzák meg a tárgy kifejezéspolaritását. A szentimentértékek erősíthetők (roppant izgalmas) vagy gyengíthetők (mérsékelten izgalmas). A jelenlegi kutatások egyik fő témája a sentiment shifterek vizsgálata szemantikus összetételt vizsgáló (*semantic compositional*) szabályok segítségével.

2.2.1. A szentimentelemzés mélysége

A tartalom szubjektivitásától és az alkalmazási területtől függően az elemzés különböző mélységét kell alkalmazni. A legegyszerűbb az egész dokumentumot egyetlen egységként kezelni. Ha a szövegtest tartalma heterogén, akkor a mondatszintű elemzés ajánlott. A legösszetettebb a fragmentum- vagy aspektusszintű elemzés, amely rendkívül számításigényes. Ilyen esetekben a szubjektív és objektív mondatok szétválasztása szükséges lehet.

A dokumentumszintű elemzés a legegyszerűbb. Ez feltételezi, hogy a szerző egyféle érzelmi megközelítést alkalmaz az egész dokumentumban. A tartalom megítélése egységesen vagy

[38] Forrás: [38] alapján saját szerkesztés.

pozitív, vagy negatív. Kétféle módon lehet az elemzést elvégezni. Ha rendelkezésre áll tanítóadat, akkor felügyelt tanítást, azaz osztályozást lehet alkalmazni. A tanítóadatok segítségével az eleddig nem osztályozott dokumentumokat besorolja valamelyik megadott osztályba egy osztályozó algoritmus alkalmazásával, mint a KNN (K-legközelebbi szomszéd, *K-Nearest Neighbors*), SVM (támogatóvektor-gép, *Support Vector Machine*), naïve Bayes stb. Ha viszont nem áll rendelkezésre tanítóadat, akkor felügyelet nélküli tanítást kell alkalmazni a csoportosításhoz. A folyamat során bizonyos kifejezések határozzák meg a szemantikus orientációt. Ezeket a kifejezéseket vagy szemantikus lexikonból vagy szófaji mintákból nyerik ki.

Ha ugyanazt a témát többféle módon közelíti meg a szerző egy dokumentumon belül, akkor az élesebb kép érdekében az elemzést mondatszinten kell elvégezni. Ahhoz, hogy a szubjektív érzelmi töltetű mondatok polaritását meg lehessen határozni, ki kell szűrni az objektív tartalmú mondatokat. [41]. Az objektív mondatok elemzése nehezebb és kevesebb sikerrel kecsegtet. A dokumentumszintű elemzéshez hasonlóan a mondatszintű elemzés is felügyelt tanítást alkalmaz az osztályozáshoz, illetve felügyelet nélkülit a csoportosításhoz. Miután a mondatszintű elemzés elkészült, a mondatszintű eredményeket összesítik.

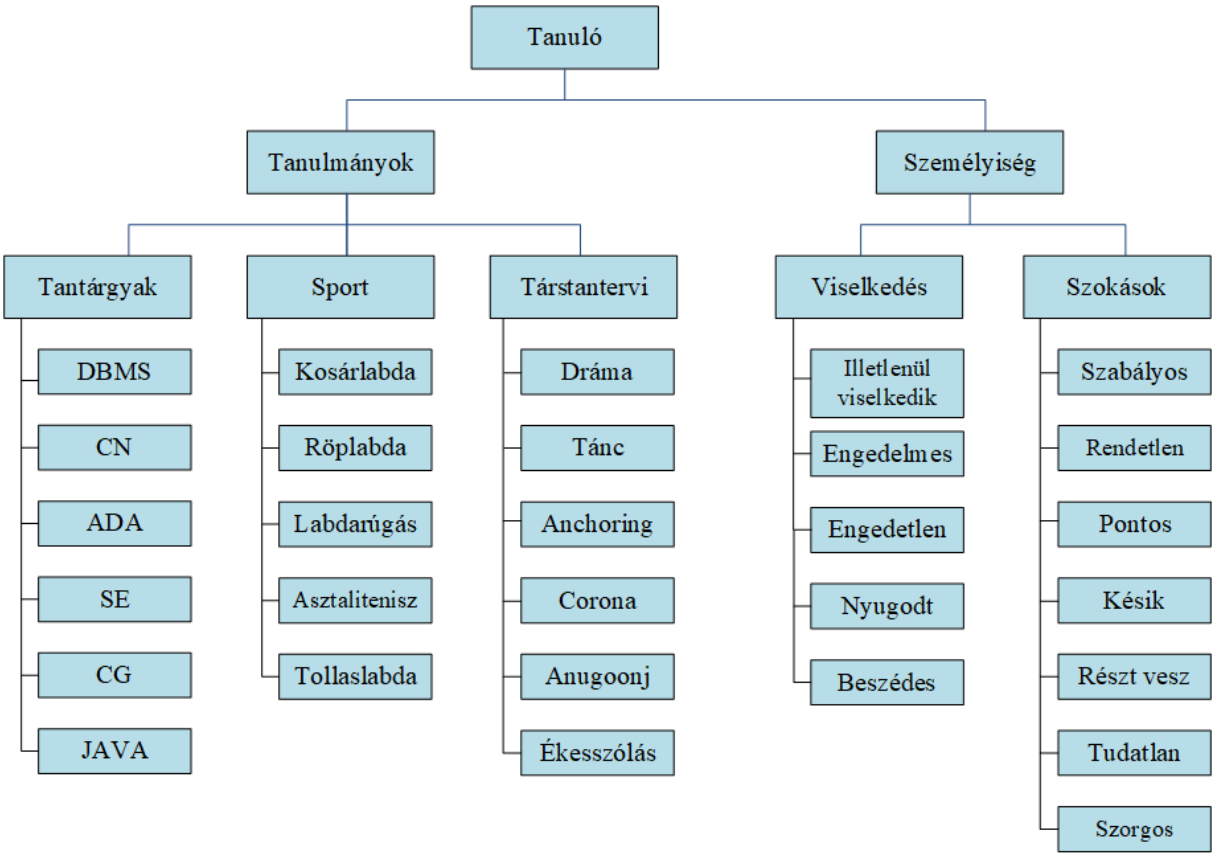
A fenti esetek a valódi szentimentelemzés egyszerűbb formái. A valóságban emberek nemhogy egy dokumentumon, hanem egy mondaton belül is több szempontból fejtik ki a véleményüket egy témáról, vagy ugyanannak a témának több aspektusáról nyilatkoznak. A fragmentumszintű szentimentelemzés ilyen szinten történő bontása sikerrel kecsegtet. A 12. ábra bemutat egy ún. aspektusfát [42]. A példa egy tanár strukturált véleményezését mutatja egy adott diákról. Jól látható a gyökérpozícióban a diák, majd alatta az aspektushierarchia.

2.2.1. A szentimentelemzés magyarul és más nyelvekben

Bár a legtöbb munka a szentimentanalízis területén angol témában, és – a lingua franca használatával – angolul készül, érdemes kitekinteni a más nyelven készült elemzésekre is. A más nyelven történő szentimentanalízisnek két módja van. Az alacsonyan lógó gyümölcs, a nem angol szöveget lefordítani Google vagy más fordítóval angolra, majd a szöveget már angolul az angol eszköztárral feldolgozni. Ez a felhőben történő fordítás természetesen csak nem minősített anyagokkal lehetséges. Minősített anyagokat zárt rendszerben működő fordítószoftverrel kell feldolgozni. Ilyen megoldás a francia–koreai Systran[39] terméke. A másik

[39] www.systransoft.com

megoldás a forrásnyelven történő elemzés. Ez feltételezi, hogy a forrásnyelven rendelkezésre áll a teljes szentimentelemző eszköztár. Feltehetően minél nagyobb a vizsgált nyelv népessége, és minél több pénzügyi forrás áll rendelkezésre a nyelvészeti kutatáshoz, annál fejlettebbek a



12. ábra: aspektusfa.[40]

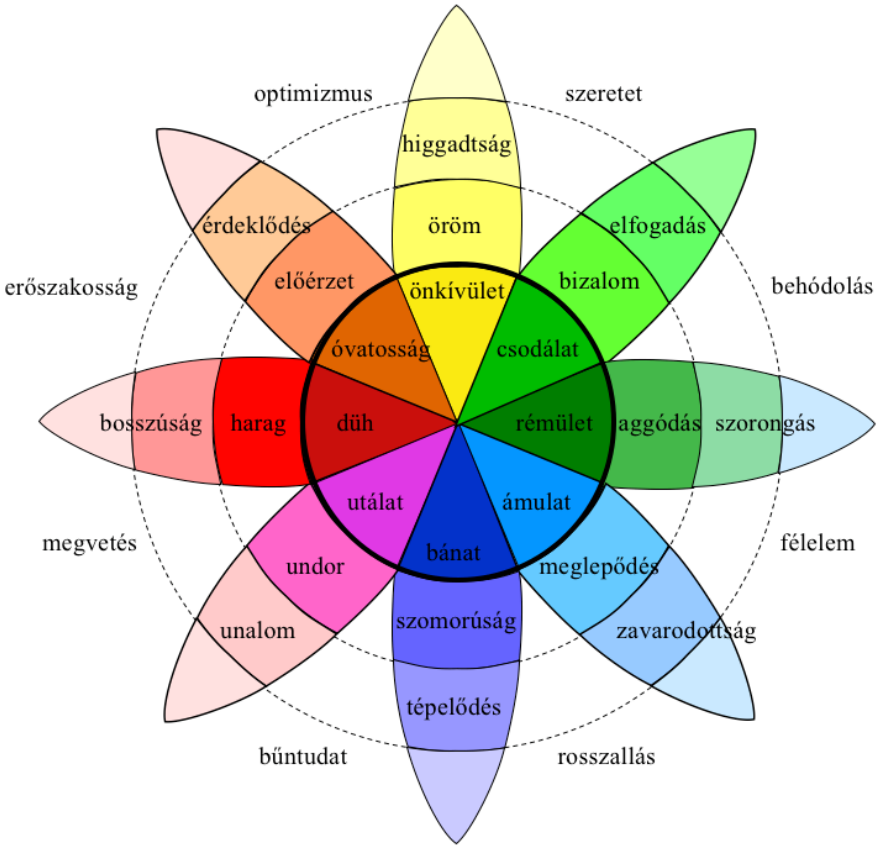
módszerek. Talán kivételt jelentenek az olyan nyelvek, amelyek hírszerző ügynökségek érdeklődésére tartanak számot, de ezek az elemzések nem feltétlenül abban az országban folynak, amelynek a nyelve az elemzés tárgya.

2.2.2. Emócióanalízis

Jóllehet az emócióanalízis és a szentimentanalízis nagyon hasonló számítógépes nyelvészeti módszertant használ, az elemzés tárgyában van egy lényeges különbség. Míg a szentimentanalízis a szerző véleményében kifejezett érzelmi-indulati tartalmat vizsgálja, az emócióanalízis ugyanezt az olvasó, néző, vagyis a kommunikáció célszemélyével teszi. Az emócióelemzés gyakran alkalmazza Paul Ekman [43] kategóriáit (öröm, szomorúság, félelem, undor és düh) vagy Robert Plutchik elméletét az emóciókról [44], akinek az emóciós kerekét a

[40] Forrás: [42] alapján saját szerkesztés.

13. ábra mutatja. Egy közösségi média tartalmát emócióanalízissel vizsgáló alkalmazást ír le Sanjeev Dhawan et al. [45].



13. ábra: a Plutchik-kerék.[41]

2.2.3. A technológia korlátai

A szentimentelemzés nem egzakt tudomány. Nem lehet nagy pontossággal észlelni a számítógépes nyelvészet eszközeivel a szarkazmust és az iróniát. A borzalmas lehet borzalmasan jó vagy borzalmasan elviselhetetlen. A kvantitatív eredmények sem várhatók el ugyanazon a szinten, mint más szövegelemzéseknél. Emberek szubjektíven pontoznak szövegeket, amiért is az eredmények 10-20%-ban eltérőek lehetnek ugyanarra a célszövegre [47]. A sajtó- vagy médiafigyelésnél alkalmazott szentimentanalízis során nem is annyira az abszolút számokat érdemes figyelni, hanem a trendeket. Azt **feltételezve, hogy a szórás időben nagyjából állandó, az eredményekből felállított idősor informatív**.

A szentimentelemzés katonai és rendvédelmi területen széles körben alkalmazható. Ilyen a civil-katonai együttműködés (*CIvil-MIlitary Cooperation,* CIMIC), a katonaság és a civilek

[41] Forrás: [46] alapján Munk Sándor szerkesztése.

együttműködése, ahol elkerülhetetlen a célszemélyek és -csoportok véleményének, hangulatának követése.

2.3. Metakeresés

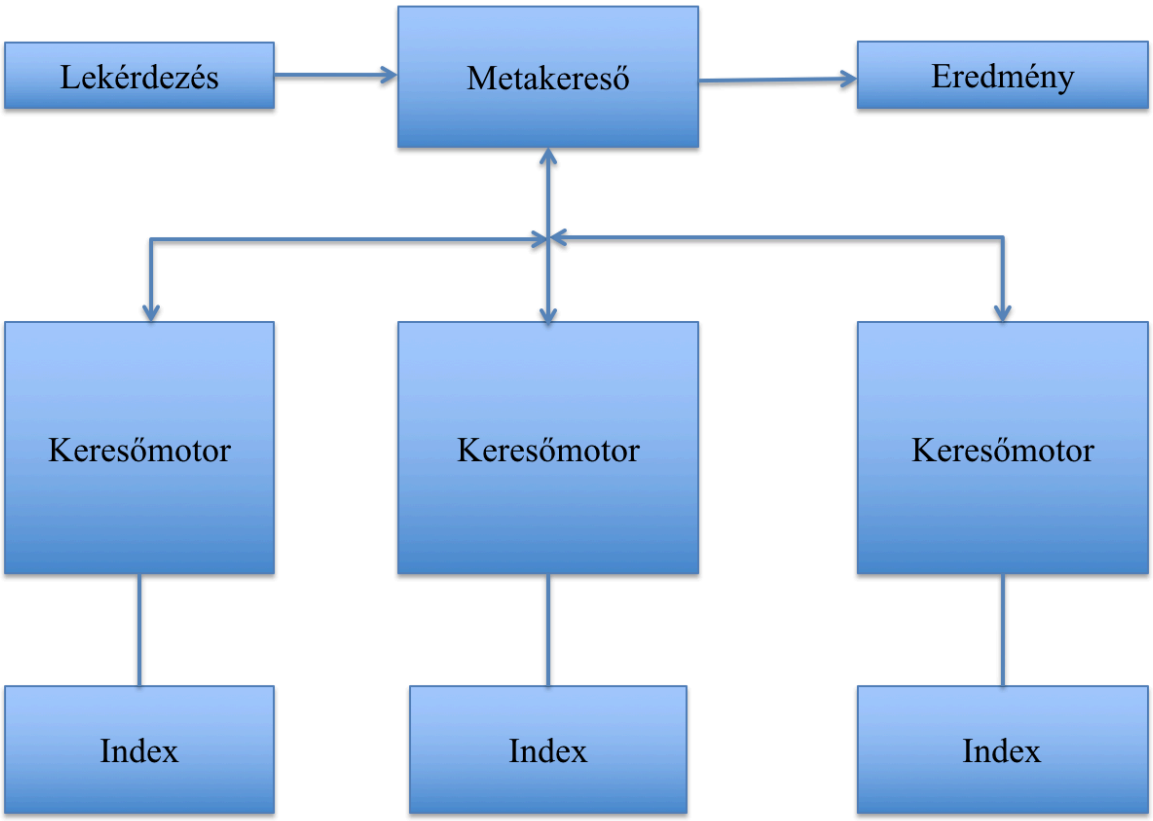
A metakeresés[42] egy információkinyerési technológia, amely több egymásra szuperponált keresővel lehetővé teszi az egyidejű keresést több kereshető forrásban [48]. A 14. ábrán látható módon a felhasználó lekérdezését a rendszer megfelelő szintaxissal továbbítja különféle keresőkhöz, amelyek a saját keresési műveleteiket elvégzik, eredményeiket visszaküldik, és azokat a metakereső egységesítve, a lehetőség szerinti legkevesebb duplikációval megjeleníti a felhasználó igény szerinti sorba rendezett formában. Ez a felhasználó számára egy egységes keresési felületet jelent, amelyen keresztül kommunikálja a keresési igényét. E felület mögött a metakereső a felhasználó számára nem látható módon elvégzi a keresési feladatokat a különböző adatforrásokban, majd az ezekből származó eredményeket egységes találati formában jeleníti meg a felhasználói felületen [49].

A metakeresés alkalmazása egyrészt időt takarít meg a felhasználónak, kényelmesebbé teszi a munkáját, a találati eredményességet, a felidézést és a pontosságot javítja, másrészt olyan feladatok elvégzését teszi lehetővé, amelyek e technológia alkalmazása nélkül az adatok mennyisége és rendezetlensége miatt gyakorlatilag más módszerrel megoldhatatlanok lennének [50], [51].

Az időmegtakarítás ténye magától értetődő. Ha keresek könyveket egy témakörben több könyvtárban is, akkor a metakeresővel nem kell minden egyes könyvtárat végigböngészni, hiszen csak egyszer kell bevinni a kérdést, a robot minden megadott könyvtárban keres, viszont a sok forrásból származó válasz csak egy helyen érkezik be [52]. Ha internetes óriáskeresőt futtatok a saját nyelvi technológiát alkalmazó metakeresővel, akkor közvetve – mint az az alábbi példában is látható – akár források millióiból is kinyerhető találat. Ugyanakkor a nagy internetes közösségi helyek, mint a Facebook, Twitter és az internetes keresők – mint akár maga a Google – metakeresőkkel eredményesen futtathatók.

---

[42] Használt angol szinonimái: *federated search, federated information retrieval, distributed information retrieval*, *multiple database searching, polysearching.* Bár a metasearch és a federated search eredetileg nem szinonim fogalmak, a mai szaknyelvben már felcserélhetőként használják.



14. ábra: egy metakereső architektúrája.[43]

2.3.1. A metakeresés fajtái

A szakirodalomban **csoportosítás**t nem találtam, ezért megpróbáltam ezt a feladatot saját magam elvégezni. Szerkezetét vizsgálva alapvetően háromféle metakeresés terjedt el.

Zárt kereső futtat zárt keresőt

A zárt rendszereken futó keresőhierarchiák előre megadott adatbázisokban működnek. Ilyenek például a könyvtárak, egészségügyi adatbázisok egyedi keresőit futtató központi metakeresők. A források száma, a dokumentumok formátuma előre ismert és időben többé-kevésbé állandó. A központi kereső követi az egyedi keresők változásait, a sok helyről befutó találatokat egységes formátumra hozza, és ezt az egyébként gyorsan változó adatforrás-tömeget időről időre újra indexeli, majd a felhasználó számára hozzáférhetővé teszi.

Internetes kereső futtat internetes keresőt

[43] Forrás: saját szerkesztés.

Számos internetes kereső futtat más internetes keresőket. Ilyen metakeresők az IXquick, Dogpile, AIIPlus, Surfwax, Yippy, Mamma, Metacca, Beacoup, MetaCrawler, Search.com, Findelio, Info.com. Ezek jellegükben alig különböznek egymástól. A találatok mellett feltüntetik a forrást. Súlyozásuk szerint nem feltétlenül ugyanazt találják meg, és nem ugyanabban a sorrendben közlik a találataikat. Konkrét alkalmazási területek többek között például hotel-, repülőjárat-, HR-keresők stb.

Testre szabott, egyéni kereső futtat internetes keresőt

Az eddig elvégzett elemzésekből az szűröm le, hogy **ahol a keresett adatok helye nem ismert, mert azok bárhol lehetnek, ott a metakeresőnek nyílt forrásra tartó internetes keresőket (Google, Bing, Yahoo, Yandex, Gigablast stb.) előnyös futtatnia**. Megjegyzendő, hogy az ilyen internetes keresők nem dolgozzák fel (indexelik le) a mély web (*deep web*) tartalmát, így nyilván az egy célrendszerhez képest a találati eredményeik is sekélyesebbek lehetnek a zárt rendszereken futó célkeresőkénél [53].

2.3.2. Metakereső versus Google Scholar

A könyvtárak működésében a 90-es évektől gomba módra szaporodtak a metakeresők. Céljuk, hogy a kutató egy felületen a változásokkal egyidejűleg gyorsan és könnyen megtaláljon minden releváns dokumentumot. Számos különböző formátumú adatbázist kellett keresni egységes szintaktikára hozva. Egy-egy ilyen rendszer felállítása évekig tartó projekteket igényelt jelentős ráfordítással. A metakeresés mint módszer a Google megjelenésével és kiterjedésével részben elsorvadt, részben átalakult, új értelmet nyert [54]. Ezt legjobban megint csak a könyvtári rendszereken lehet demonstrálni. A Google Scholarnak az addig elképzelhetetlen tárolókapacitása és a különböző formátumokat kezelni képessége miatt szükségtelenné vált a keresőláncok felépítése, hiszen a Google szinte azonnal és egyszerre leindexelte az összes könyvtári állományt, amelyhez jogot kapott. Itt megint csak meg kell jegyezni, hogy megkerülhetetlen tudományos adatbázisokhoz, mint az Elsevier, a Google nem fér hozzá. És a találati minőséggel szemben is megfogalmazódtak észrevételek. Ennek ellenére a gyors, olcsó és elsőre felületes keresési feladatok terén a Google Scholar a 2000-es évek közepe óta átvette a vezető szerepet [55].

2.3.3. Esettanulmány a metakeresés alkalmazására

Az alábbi esettanulmány egy kísérlet eredményeképpen született [56]. A kísérlet lényege az volt, hogy a NAV, OEP vagy az ORFK részére egy egyszerűen kezelhető eszközt mutassak be az illlegális gyógyszer-kereskedelem feltárására. A cél: nyílt forrásból hozzájutni az interneten

illegálisan ajánlott, másképpen fogalmazva feketekereskedelemben árult gyógyszerajánlatokhoz az ajánló azonosítójával (ez mobiltelefon, e-mail-cím vagy egy portál) a nyomozás megkönnyítése végett. A technológia később továbbfejlődött a csempészáruk és az illegális drogkereskedelem feltárása felé. Kiemelem, hogy egyrészt nem tudni, az illegális kereskedők hol hirdetik a termékeket (ilyen az Apronet, Vatera stb., de lehetnek egészen szétszórt helyeken is, különösen a – később feltárt – szintetikus drogok esetében), másrészt a keresési kritériumok olyan komplexek (különösen a később indított csempészáruk esetében), hogy azok meghatározása a Google segítségével nem lehetséges.

Első lépésben bemutatom a manuális keresés korlátait. Előbb megpróbálok lekérdezni egy mintatételt, értékelem a kapott eredményt, elemzem a hiányosságait. Majd kombinálom a saját egyéni keresőt az arra szuperponált internetes óriáskeresővel, és látható lesz, mennyivel jobb és a felhasználó számára barátságosabb eredményre jutok. Ez utóbbi, vizsgált metakeresést egy konkrét példával illusztrálom. A számtalan alkalmazási lehetőségből az illegális gyógyszer-kereskedelmet ragadtam ki. Ez abból a szempontból könnyű terület, hogy a magyar lakosság zöme nincs is tisztában a tevékenység törvénytelen voltával, illetve nemigen üldözik, ezért nem túl bonyolult nyelvészeti technológiával lehet bőséges eredményre jutni.

Megjegyzem, három kategóriát lehet megkülönböztetni a csalás áldozatának szempontjából.

- A betegeket, ha rossz minőségű készítményhez jutnak.
- A gyártókat, ha szabadalommal védett készítményt másolnak.
- A biztosítót (OEP[44]), ha a felhasználás nem rendeltetésszerű, pl. nem jogosultnak adják a készítményt (határon túlra viszik, állatnak adják stb.).

Ha bevisszük a Google-ba például a „Viagra eladó” kifejezést, 12.600 találatot jelez a rendszer[45]. Ellenben a 12. oldal alján a 15. ábrán olvasható üzenet fogad**.**

A leginkább releváns eredmények megjelenítése érdekében kihagytunk néhány, a(z) 120 megjelenítetthez nagyon hasonló bejegyzést.Ha szeretné, megismételheti a keresést a kihagyott eredmények belefoglalásával.

[44] Az OEP (Országos Egészségügyi Pénztár) neve 2017. január 1-től NEAK (Nemzeti Egészségbiztosítási Alapkezelő), ami az OEP feladatkörének zömét átvette.

[45] (letöltés ideje: 2017.11.26.)

15. ábra: képernyőletöltés a fenti keresés alapján.[46]

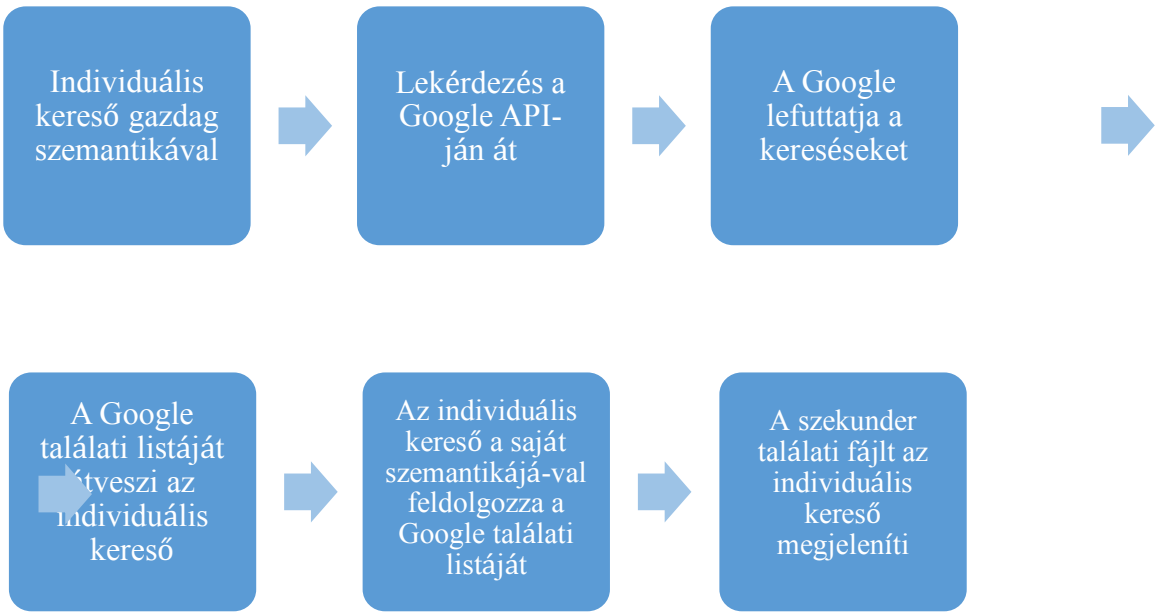
Ha az aláhúzott szövegre kattintunk, több találat jelenik meg, de így is csak egy töredéke a jelzett találati számnak. Tehát megállapítható, az ingyenes online keresés elenyésző felidézést eredményezett. Megjegyzem, hogy nagy mennyiségű dokumentum keresésére vonatkozó árlista elérhető a Google honlapján [57], de fizetős szolgáltatás esetén a kereső álcázása nehezebben megoldható (lásd 4.3.), valamint a szemantikai korlátok továbbra is fennálnak.

A Google speciális keresésének nyelvészeti lehetőségei sem terjednek túl a Boole-algebrai keresőoperátorokon. ÉS és VAGY alkalmazható, valamint a kizárás és a szó szerinti keresés. Továbbá az operátorokat nem lehet csoportosítani, azaz zárójelezni: „(A OR B) AND C" kombinációt nem értelmezi „(A AND B) OR (A AND C)" -ként. A Google tudja a szavak közelségét is mint operátort kezelni, de ez jelen idő szerint nem publikált képesség, így ez is korlátozottságnak számít, hiszen a felhasználók zöme nem tud erről. Szintén nem tudja súlyozni a keresett kifejezéseket, és bonyolultabb zárójeles mondatot, logikai kifejezések írását sem támogatja. Szótárak, tudásfák stb. alkalmazása nem lehetséges.

A manuális keresés előbbi bemutatása után bemutatom az összetett keresést. Az összetett keresés lényege, hogy a felhasználó saját keresője a megadott kulcsszavakat felbővíti a saját rendszer tudásfájával, és ezt a felbővített lekérdezést küldi át az internetes kereső API-jának (felhasználói programozói felület, *application programming interface*) mint annak a lekérdezését. Ha például a gyógyszer szóra keres, akkor az orvosságot, medicinát, tablettát, orvosi készítményt stb. is továbbítja, hiszen a saját rendszer tudásfája tartalmazza a szinonimákat.

Az internetes kereső a kibővített lekérdezést feldolgozza, és a találati listáját – amely lehet emberi erővel feldolgozhatatlan méretű – visszaküldi a saját keresőbe, amelyet az ismét felhasználva a saját nyelvészeti technikáját feldolgozza, majd a saját – szekunder – találati listáját a felhasználó számára kényelmes módon elérhetővé teszi. A folyamatot a 16. ábra mutatja be.

[46] Forrás: Google.



16. ábra: az egyéni keresőre szuperponált internetes metakereső architektúrája.[47]

A tudásfa (taxonómia) alapvetően két ágból tevődik össze, amelyet a 17. ábra mutat be. Az illegális kereskedelemre utalnak a teljesség igénye nélkül a következő kifejezések: *eladó, olcsó, olcsón, Vatera, online, vény nélkül, eBay, teszvesz, megfizethető áron, eredeti, pénzvisszafizetési garanciával, rendkívül kedvező áron, online gyógyszertár, vennék* stb.

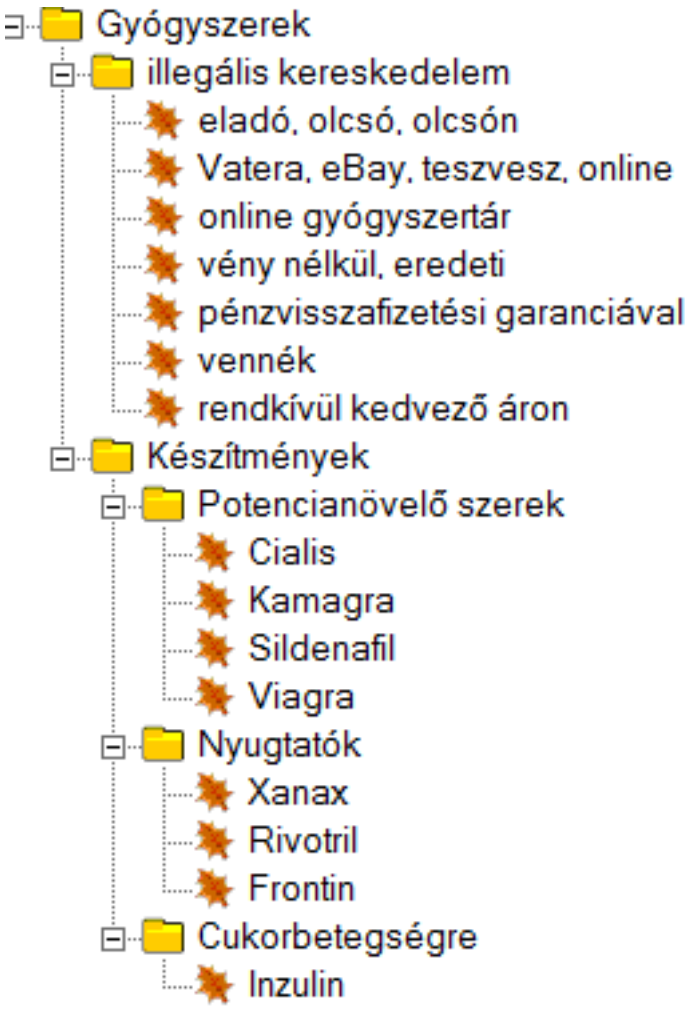
Maguk a készítmények hatásuk szerint besorolhatóak. Néhány példa a következőkben.

Potencianövelő szerek: *Viagra, Cialis, Sildenafil, Kamagra*.

Nyugtatók: *Xanax, Rivotril, Frontin*.

Cukorbetegségre: *Inzulin*.

[47] Forrás: saját szerkesztés.



17. ábra: a keresett gyógyszerek hármas mélységű tudásfája.[48]

A 5. táblázat a keresés végeredményét jelentő lista sok ezer eleméből egy illusztratív részlet. Csak néhány napra szűkítettem a túl nagy mennyiségű találat elkerülése végett. Megjegyzendő, sok esetben az eladó nem e-mail-címet ad meg, hanem honlapot (ilyen például a www.viagraelado.me), amelyre be kell jelentkezni, és az eladó lép majdan kapcsolatba az érdeklődővel. Ritkábban – feltehetően eldobható – mobilszámot adnak meg a kapcsolatfelvételhez.

5. táblázat: a találati lista.[49]

Készít-mény	Dózis	Mennyi-ség	Nick-name	e-mail	venne/elad-na	Dátum

[48] Forrás: szerző.

[49] Forrás: a metakereső letöltéseiből. A futtatás ideje: 2013.09.25.

Xanax		30db	nikike1987	nikike1987@freemail.hu	eladna	2013.09.25.
Frontin		30db	nikike1987	nikike1987@freemail.hu	eladna	2013.09.25.
Xanax	0,5 mg	100db	baboca		eladna	2013.09.22.
Frontin	0,5 mg	100db	baboca		eladna	2013.09.22.
Xanax	0,5 mg	100db	random	randik@freemail.hu	eladna	2013.09.22.
Frontin	0,5 mg		lijja	lijja0214@freemail.hu	vevő	2013.09.21.

A keresőt természetesen személyre szabott profil alapján parametrizálhatja minden felhasználó. A keresőparamétereket a rendszer tárolni tudja, így azokat nem kell minden alkalommal bevinni, de szükség szerint módosíthatók.

2.3.4. További alkalmazási példák

Egy tipikus feladat, amelyet metakereséssel lehet megoldani, további közösségi oldalak lekérdezése. Konkrét példa, hogy a fórumokat, blogokat rendszeresen vizsgáló kereső rákérdez a „korrupt zsaru” fogalomra. A találatok között ott szerepelnek a hencegő gyorshajtók, akik pontos hellyel és idővel megjelölve számolnak be a „probléma gyors, papírmentes kezeléséről”. Innen a hatóságnak már csak azt kell vizsgálnia, hogy ott és akkor éppen ki volt szolgálatban.

A metakeresés természetesen alkalmazható bármilyen más témára is, például az illegális fegyver-, ember-, kábítószer-kereskedelem, hírszerzési és elhárítási érdeklődésre számot tartó virágnyelven megfogalmazott üzenetek kiszűrése, csapdák, félrevezetések felderítése (*deception detection*) stb. Minél kényesebb a téma, annál kifinomultabb nyelvi technológiákat kell alkalmazni, ami az értékelő-elemző fontos előkészítő és karbantartó munkáját igényli.

A metakeresés, mint bármely jelentésalapú keresés a természeténél fogva nyelvfüggő. A fogalmi összefüggések ábrázolása a tudásfákban és más megjelenítési formákban a keresés nyelvén történik. Ezek automatikus portálhatóságáról nincs tudomásom. Bevett gyakorlat a

keresést a forrásnyelven végrehajtani, az eredményt pedig egy fordítóprogrammal áttenni a kívánt célnyelvre.

Még egyszer utalok arra, hogy minősített környezetben futó alkalmazásoknál elkerülhetetlen a kereső azonosíthatatlansága. Az álcázásról bővebben a 4.3.2 fejezetrészben írok.

2.4. Fúziós központok

2.4.1. Az adatfúzió meghatározása

Az adatfúzió egy objektumról leképezett többféle adat- és/vagy tudásállomány összeolvasztása egy egységes, pontos és használható állományba. Az egységesített állománynak informatívabbnak kell lennie, mint a forrásoknak. Vagy másképpen: **az adatfúzió egy olyan folyamat, melynek során a több forrásból és szenzorból bejövő adatokat a rendszer automatikusan szűri, osztályozza, kivonatolja és megjeleníti** [58].

A 2001. szeptember 11-i események ismert módon rámutattak arra a tényre, hogy megfelelő mennyiségű és minőségű adat állt rendelkezésre a merénylet előrejelzésére, csak azok nem találkoztak megfelelő aggregációs szinten az értékelő-elemzők számára. Ezt követően elképzelhetetlen összegek és erőforrások koncentrálódtak a Homeland Security égisze alatti fejlesztésekre. Ezek természetesen nemcsak strukturált adatok, hanem strukturálatlan szövegek feldolgozását is jelentették [59]. Európai adatfúziós központ létrehozását szorgalmazza az Interpol keretében Gruszczak [60]. A fúziós központok keresési módszertana a gyártók egyik legféltettebb titka. Egyetlen gyártó sem publikálja az algoritmusait, nyelvi technológiájának módszertanát, a találatok rendezési elveit. A külön-külön silókban tárolt heterogén adattartalomra épülő keresőrendszerek – ha vannak ilyenek – összevezetésének, egységes feldolgozásának, a rejtett összefüggések emberi erővel lehetetlen feltárásának és a mesterséges intelligenciával történő prediktív analízisének fundamentuma a fúziós központok hatékony működtetése.

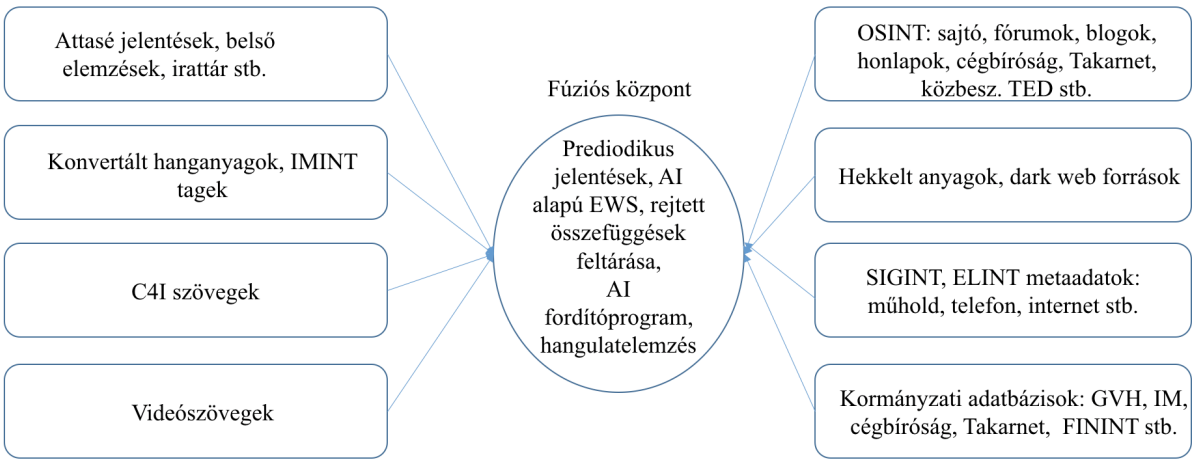
2.4.2. Szöveges adatfúziós központok

A legkorábbi fúziós központok az elektronikus hadviselés igényei szerint épültek ki. A különböző forrásokból szenzorokon át bejutott elektromágneses jeleket normálták, majd összevezették kiértékelés végett. Az elektromágneses jelek feldolgozásának mintájára fejlesztették ki a szöveges források központi kezelésére a szemantikus keresőrendszerek fúziós központját. A továbbiakban nem foglalkozom az elektromágneses jeleket feldolgozó fúziós központokkal, csak a szöveges tartalmakat feldolgozókkal, amelyekbe beleértem a képi és videótartalmak metaadatait, valamint az írott szöveggé alakított hanganyagokat.

Egyesek arra panaszkodnak, hogy a döntéshozók nem látnak az OSINT-ban különösebb hozzáadott értéket, hiszen mindent megtalálnak a sajtófigyelők jelentéseiben is. Ez tévedés több szempontból is. Egyrészt egy jól megkonstruált keresőrendszer a sebességénél és a komplexitásánál fogva képes olyan tömegű feladatot megoldani, amelyeket manuális úton lehetetlen lenne. Másrészt a ráfordított időt töredékére csökkenti, ami lehetőséget teremt több produktív munkára. De egy, az OSINT adatforrásait egyesítő és feldolgozó fúziós központ a keresések egymásra építésével a többdimenziós strukturált és strukturálatlan adattömegből képes kiszűrni keresztezett keresésekkel (*cross search*) a lényeget. Lévay Gábor már 2004-ben felvázolt egy adatközpontot egy nemzeti OSINT formájában [61]. Béres János és Kenedli Tamás leír egy ilyen architektúrát [62], amely az IBM-i2 Analyst's Notebookjára épül. A cikkhez kiegészítésül meg kell jegyezni, hogy bár az Analyst's Notebook mind az EU-ban, mind a NATO-ban széles körben alkalmazott analitikai eszköz, feladata nem nagy tömegű strukturálatlan vagy félig strukturált szövegtestekből szemantikus keresést végrehajtani, inkább az ilyen keresők outputját feldolgozni. Erre számos példát ismerünk, ilyen a cseh Tovek terméke, amely Magyarországon nem ismeretlen. **Ipari gyakorlatom alapján egy félreértést tisztázandó, megállapítom: egy Analyst's Notebook, Sentinel vagy más – akár nyílt forráskódú – fúziós alkalmazás nem helyettesít, hanem kiegészít egy jelentésalapú keresőrendszert.**

Az általam javasolt architektúra szerint egy „külső" OSINT-alkalmazás (közösségi helyeket, műholdas forgalmat, drónok gyűjtötte adatokat stb. figyelő) és egy „belső", semmilyen külső kapcsolattal nem rendelkező fúziós központ kombinációja tipikus architektúra nagybiztonságú adatfeldolgozó rendszerre [63]. A fenti alkalmazásokat felfűzhetjük egy rendszeresen ismétlődő, működő figyelőrendszerre. Ennek lényege, hogy a robot időszakonként – ez lehet hetente vagy akár percenként is – egy pillanatfelvételt készít a vizsgálandó állományról, majd ezt a megelőző felvétellel összehasonlítja. Az összehasonlításból kiemeli az előre meghatározott, érdeklődésre számot tartó változásokat, és ezeket automatikusan, proaktív módon továbbítja – például egy e-mail formájában (*alert*) – a felhasználónak.

A 18. ábra szemlélteti az általam javasolt lehetséges hírszerző fúziós központ architektúráját. A nyolc „csáp" kapcsolja be a lehetséges adatforrásokat. A fúziós központban történik az adatok feldolgozása. Természetesen lehetséges az egyes adatforrásközpontokban saját belső keresőrendszert kiépíteni, amelyek kiszűrt tartalmát – akár airgappel – továbbítani lehet a központi feldolgozóba.



18. ábra: Egy lehetséges fúziós központ architektúrája.[50]

Az ábrában használt rövidítések feloldása: IMINT (képi hírszerzés, *Image Intelligence*), C4I (parancs, ellenőrzés, kommunikáció, számítógépek, hírszerzés) Command, Control, Communication, Computers, Intelligence), AI (mesterséges intelligencia, *Artificial Intelligence*), EWS (korai előrejelző rendszer, *Early Warning System*), TED (napi elektronikus tender(figyelő), *Tenders Electronic Daily*), SIGINT (jelhírszerzés, *Signal Intelligence*), ELINT (elektronikus hírszerzés, *Electronic Intelligence*), GVH (Gazdasági Versenyhivatal), IM (Igazságügyi Minisztérium), FININT (pénzügyi hírszerzés, *Financial Intelligence*).

Egy valódi fúziós központ nemcsak egyirányú információmozgást jelent, ami a 18. ábra esetében a „csápoktól" a „fejrészig" befelé, illetve a fejrészfeldolgozó központból jelentések formájában kifelé, hanem egy oda-vissza történő információmozgást az információforrások és a központ közötti „idegpályákon".

2.5.Összefoglalás, részkövetkeztetések

Az információkereséshez mint technológiához a gyakorlatban eltérő rendeltetésű alkalmazási területek, illetve összetett alkalmazási megoldások is kapcsolódnak. Szakmai tapasztalataim alapján ezek között kiemelt szerepet játszik, illetve játszhat többek között a szentimentelemzés, a metakeresés és a szöveges adatfúzió.

Az első fejezetrészben **meghatároztam a szentimentelemzés és az emócióelemzés fogalmait**, és **különbséget tettem a két** hasonló, de mégis eltérő **fogalom jelentése között**. Rámutattam, hogy **az információkeresés és a szentimentelemzés technológiai alapjai jelentős részben**

[50] Forrás: a szerző.

megegyeznek, eltérés inkább a kinyerendő információ, illetve a releváns információhordozók kiválasztására szolgáló információk szintjében van (elemi információ, illetve annak besorolása). Mivel a szentimentelemzés alapvetően épít a véleményt, érzelmi töltést hordozó, valamint az azt módosító kifejezések és azok jellemzőinek listájára, **következtetésként kell megfogalmazni a forrásnyelvi eszköztár kiemelt jelentőségét, szükségességét**. Szentimentelemzés elvégezhető egy másik, ilyen eszköztárral már rendelkező nyelvre történő fordítással is. De **rámutattam, hogy a fordított szövegen végzett szentimentelemzés torzíthatja az eredeti szöveg érzelmi töltetét**. További alkalmazott kutatási irány lehet a szentimentelemzés felhasználási lehetőségeinek vizsgálata az információs hadviselés, a pszichológiai műveletek és a civil-katonai kapcsolatok területén a célszemélyek és -csoportok véleményének, hangulatának követésére.

Áttekintettem a metakeresésre vonatkozó lényeges szakirodalmat. Ennek alapján megállapítottam, hogy a metakeresés hozzáadott értékével növeli a keresés eredményességét, hatékonyságát, kényelmességét, javítja pontosságát. **Saját szempontjaim szerint rendszereztem és osztályoztam a metakeresés különböző változatait** az elsődleges keresők, illetve a második szintű kereső nyílt vagy zárt típusa szerint.

Rámutattam, hogy a metakeresés a nyílt forrású információk feldolgozása során eredményesen használható a nemzetbiztonsági és rendvédelmi területen. Ezt kutatásaim során egy illegális gyógyszer-kereskedelemre vonatkozó információk feltárását célzó **esettanulmánnyal a gyakorlatban igazoltam**. **Kidolgoztam egy módszert, amely segítségével egy vállalati keresőalkalmazás futtat Google keresőalkalmazásokat. Bizonyítottam, hogy egy egyéni metakereső és egy nagy internetes kereső, mint például a Google kombinált alkalmazásával ki lehet használni mindkét keresőrendszer előnyeit.** Példával is illusztráltam egy konkrét, megvalósított metakereső alkalmazást, amely a NAV bűnügyi nyomozóinak, az ORFK kábítószer- és illegális gyógyszer-kereskedelem elleni nyomozóinak és az OEP illetékes főosztályának hivatott nyílt forrásból azonnali akciót támogató adatdesztillátumot szolgálni. A trendek alapján nagy biztonsággal megjósoltam, hogy a publikus internetes keresők idővel egyre több nyelvészeti technikát fognak beépíteni a rendszereikbe. Ezzel a felhasználó saját keresője elvben veszíthet a jelentőségéből. Mai tudásunk szerint ez a tökéletes szemantikus web, Tim Berners-Lee víziója még sokáig nem valósul meg a formátumok kaotikus volta és a nyelvek bábeli sokasága miatt. E beteljesülésig a bemutatott metakeresők létjogosultsága nemhogy csökkenni, hanem nőni fog.

Megvilágítottam, hogy az óriási feldolgozandó információtömeg és korunk égető nemzetbiztonsági és rendvédelmi igényei miatt kerül egyre inkább előtérbe a különböző forrásokból származó adatok szűrésének, osztályozásának, feldolgozásának, kivonatolásának és megjelenítésének eszköze, a fúziós központ. **Ipari gyakorlat alapján bemutattam egy működő fúziós központ architektrúráját, amely nyílt forrású találatokat periodikusan átemel és feldolgoz egy sokforrású belső feldolgozó rendszerbe. Rámutattam arra, hogy az adatelemző és -vizualizáló termékek nem helyettesítik a fúziós központok szemantikus keresőrendszereit, hanem kiegészítik azokat.**

3. FEJEZET: AZ INFORMÁCIÓKERESÉS ALKALMAZÁSI IGÉNYEI, LEHETŐSÉGEI EGYES ALKALMAZÁSI TERÜLETEKEN

3.1. Bevezető gondolatok, a fejezet tartalma, célja

Az előző fejezetben három példával megvilágítottam az első fejezetben tárgyalt elméleti alapokra épülő technológiai lehetőségeket, röviden utalva megvalósítható alkalmazásukra. Ebben a fejezetben a mindennapi élet három területén (nemzetbiztonság és rendvédelem, közigazgatás és a gazdaság) vizsgálom meg az információkeresés alkalmazhatóságát. Az alkalmazási területek kiválasztása itt is önkényes. A katonai-nemzetbiztonsági-rendvédelmi hármasból itt „külső" keresést, az OSINT-ot vizsgálom, a „belső" keresést, az ECS-t a második fejezetben, a Fúziós központok alatt tárgyalom. A közigazgatás felhasználási területe is jóval tágabb, mint a jogi alkalmazások, de Magyarországon ez a leginkább kiérett, és a saját ipari és oktatói gyakorlatom is itt engedett legjobban elmélyülni. A gazdasági életben használt keresőmegoldásokra a napi gyakorlaton túl a SCIP (stratégai és üzleti hírszerző szakemberek, *Strategic and Competitive Intelligence Professionals*, SCIP)[51], tagságban szerzett tapasztalatom segített. Ebben az amerikai központú, de nemzetközi agytrösztben volt igazán látható, hogyan váltott az információkeresés évről évre egyre inkább a klasszikus módszerekről az informatikaiakra. További segítséget nyújtott az ISKO (Nemzetközi Társaság a Tudásszervezésért, *International Society for Knowledge Organisation*)[52] brit részlegében szerzett tagságom.

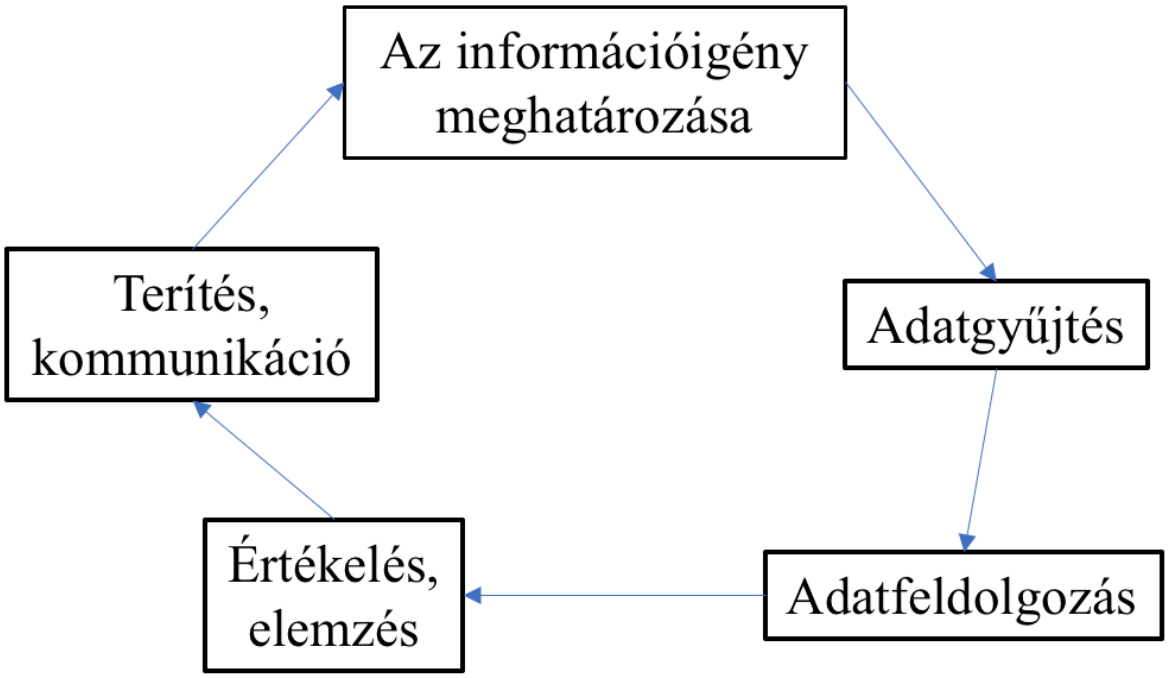
Hírszerzés történhet folyamatos, rutinszerű működésben (*continuous intelligence*), és alkalomszerűen (*discreet intelligence*). A hír titkosítási minősítése alapján lehet szigorúan titkos (*top secret*), minősített (*classified*), zárt körű védett (*closed proprietary*), nyílt védett (*open proprietary*) és nyílt (*open source*) [64].

A szakirodalomban [65] és [66] gyakorlatilag egységesen 5 lépésre bontják a hírszerzési folyamatot. Ezek az ismert lépések az 19. ábrán láthatóak. Fontos, hogy a keresés mint technológia három direkt (adatgyűjtés és -feldolgozás valamint terítés, kommunikáció), két indirekt (információigény megfogalmazása és értékelés, elemzés), tehát a ciklus minden

[51] www.scip.org

[52] http://www.isko.org

lépésében fontos szerepet játszik. Alapkövetelmény az újszerűség, az időszerűség, a feldolgozottság foka, a hitelesség és a rendelkezésre állás[53]



19. ábra: a hírszerzés folyamata.[54]

A nemzetbiztonsági szolgálatoknál és a rendvédelmi szerveknél az információkeresés legszélesebb körben a nyílt forrású hírszerzés (OSINT) és a belső keresőrendszerek (ECS), különösen a fúziós központok keretei között kerül alkalmazásra. A következőkben ismertetem a nyílt forrású hírszerzés fogalmi alapjait, elemzem az információkeresés szempontjából lényeges jellemzőit, majd értékelem az információkeresés jelenlegi helyzetét és jövőbeni alkalmazásának irányait, lehetőségeit.

3.2. Információkeresés a védelmi szférában

3.2.1. A nyílt forrású hírszerzés alapjai

Könyvtárnyi irodalom elemzi az OSINT hasznosságát. A vélemények meglehetősen eltérőek. A II. világháború vége óta kering a szám, hogy a megszerzett információk 80%-a nyílt forrásból származtatható. Ez eredetileg Allen Dullestől származik, a Szenátus Hadügyi Bizottságának meghallgatásán hangzott el 1947-ben [68]. A Gibson által idézett beszélgetés szerint a 90-es

[53] [66] 104. oldal

[54] Forrás: [65] 44. oldal alapján saját szerkesztés.

évek közepén a CIA (Központi Hírszerző Ügynökség, *Central Intelligence Agency*) Közösségi Nyílt Forrású Programja (*Community Open Source Programme)* hivatalosan 40%-ra becsülte a nyílt forrású információ szerepét általában, jóllehet a forrás rejtettségétől függően ez lehet 10% vagy akár 90% is [69]. Az mindenképpen leszögezhető, hogy a nyíltan hozzáférhető információk és források mennyisége exponenciálisan nő.[55] Akármilyen fontosságot is tulajdonítunk ennek a diszciplínának, kétségtelen, hogy **az információk jelentős része csak nyílt forrásból érhető el, nincs is rejtett megjelenítésük**. Sokszor nyílt információ utal rejtett tartalomra. Erre a közösségi média tárgyalásánál mutatok példát.

A következőkben áttekintem a nyílt forrású hírszerzés történetét, definícióit és az azok körül kialakult bizonytalanságokat. Elhelyezem az OSINT-ot az összadatforrású hírszerzés rendszerében. Végül felsorolom az OSINT legismertebb szakmai és technikai célterületeit.

A nyílt forrású hírszerzés története az ókorig nyúlik vissza. II. Ramszesz (i.e. 1302 után) a hadműveleteit előkészítendő királyi útikalauzokat, *mohar*okat küldött a célterületre, akik ebben az álcában felmérték az ellenség erejét, és jelentették a tapasztaltakat. Józsué könyvében a következő áll (Józs. 2:1) [70].

„Józsue, Nun fia, Sittimből titokban két hírszerzőt küldött előre ezzel a megbízatással: »Menjetek, fürkésszétek ki Jerikó földjét!« Előre mentek, s egy Rácháb nevű rossz hírű nő házába tértek be, aztán nyugalomra tértek."

A magyar történelemből ismert Julianus barát akarva-akaratlan nyílt forrású hírszerzést végzett, amikor időben előre jelezte IV. Béla udvarának a tatárok közeledését [71]. Sir Francis Walsingham kalmároknak öltözött kémeket küldött Spanyolországba az Armada készültségének felmérésére. Ezen információk birtokában tudta Sir Francis Drake Cádizt bombázni, amivel az angolok időt nyertek a felkészülésre [72]. 1826-ban Henry Brougham a radikális Whig[56] politikus megalapította a „Hasznos tudás terjesztésére való társaság"-ot (*Society for the Diffusion of Useful Knowledge*) Angliában azzal a céllal, hogy a tudást a társadalom minden osztályához eljuttassa. Ez alig különbözik a Google mai jelmondatától.[57]

A legújabb kori történelemben fellelhető érdekes példa: az angolszászok a franciaországi utak, vasútvonalak és hidak bombázásainak eredményét a narancsok árának változásaiból is mérték

[55] [67] 102. oldal

[56] A mai liberális párt elődje

[57] Az első három példa definíciótól függően HUMINT-kategóriába is sorolható.

a párizsi piacokon. A német fizikusok a Big Ben élő harangozásának hangját elemezve segítettek a Luftwaffe meteorológusainak a tervezéshez, amíg az angolok rá nem jöttek, és felvételről kezdték az ismert dallamot sugározni. Reinhard Gehlen vezérőrnagy, a Wehrmacht Fremde Heere Ost hírszerzési igazgatója nyílt forrásban megtalálható statisztikai adatok alapján jelezte előre, hogy a szovjet hadsereg potenciálja korábbi véleményekkel szemben nem az élőerő kifogyása miatt gyengül meg, hanem a szén, acél stb. hiánya miatt. Ez egészen addig tartott, amíg az angolszász segítség meg nem érkezett.

Az OSINT nagy mértékű kiépülése a II. világháború utáni időszakra tehető. A lejjebb ismertetendő klasszikus módszerek mellett igazi paradigmaváltást az 1989 után rohamosan terjedő internet hozott. Mivel idővel szinte minden adat olcsón és gyorsan elérhető lett az interneten, a feladat a könyvtárakból átterelődött a számítógépekhez. Az igazi kihívás ekkor keletkezett, mivel az adatok szénakazlában meg kellett találni a lényeget szimbolizáló gombostűt. Az adat- és szövegbányászat, a számítógépes nyelvészet, a statisztika és számos más tudományág segítette az adatok megszerzését, kinyerését és feldolgozását felhasználható információvá.

Lévay Gábor, a nyílt forrású hírszerzés hazai szakértője 2006-ban az OSINT-ot a következőképpen határozta meg.

„Az OSINT a katonai felderítés és a hírszerzés rendszerén kívül létező, a publikum (tehát minden egyén) számára nyilvánosan, legális eszközökkel megszerezhető, vagy korlátozott körben terjesztett, de nem minősített adatok szakmai szempontok alapján történő felkutatását, gyűjtését, szelektálását, elemzését-értékelését és felhasználását jelenti” [73].

Lévay Gábor és más elődök munkáját tiszteletben tartva rámutatok a definíció újragondolásának szükségességére. Az angol hírszerzési enciklopédia szerint az **OSINT: hírszerzés a nyilvánosság számára elérhető forrásból** [74]. Lényegében ugyanezt írja Robert Steele, az OSINT amerikai evangélistája is.[58] Az alapvető különbség a „nyilvános”, „publikus” stb. és a „törvényes”, „legális” között van. Bár a NATO kézikönyve [76] – 2001-ben – még kifejezetten jogszerű információszerzésről ír, általában elmondható, hogy a modern angolszász szakirodalom inkább hajlik a „nyilvánosan elérhető”, míg a kontinentális inkább a „törvényes” felé. Egyfajta vízválasztó lehet a beavatkozás aktív (pl. *hacking*) és passzív volta. A források jelentős része maga is meghatározza, hogy az adatainak mely felhasználásai jogosak és melyek

[58] [75] 129. oldal

nem. Az sem egyértelmű, hogy ami törvényes, az melyik ország törvényei szerint az. Általában elmondható, hogy az – egyébként önmagában sem egységes – az EU szigorúbban védi a személyiségi jogokat, míg az USA és Svájc közvéleménye és jogrendszere ebben engedékenyebb a biztonság javára. Az OSINT alkalmazójának tisztában kell lennie működésének mindenkori törvényes kereteivel, és azokat be kell tartania. A hazai szakirodalom egy része a vállalati célú üzleti hírszerzést is az OSINT kategóriájába sorolja.[59] E felfogás szerint az OSINT katonai és rendvédelmi tevékenység, amelynek polgári ága az üzleti hírszerzés. Mindkét felfogás él a nemzetközi szakirodalomban is. Megoszlanak a vélemények arról, hogy az adatok, információk, amelyek valaha titkosak voltak, de legális vagy illegális módon nyilvánosságra kerültek – például feltették ezeket a világhálóra –, már nyílt forrásúnak minősülnek-e. Ilyenek például a Snowden vagy a Wikileaks anyagok. Az OSINT mint legális, passzív módszer és a hacking mint illegális, aktív technológia közötti határ sem mindig éles. Számos nyílt forrású eszköz alkalmas például sebezhetőségek feltárására, és így illegális adatszerzésre. Ilyen például a Google-dorking vagy a Shodan.

A fentiek alapján újra definiálom a nyílt forrású keresés fogalmát. **Az OSINT a katonai felderítés és a hírszerzés által nyilvánosan megszerezhető adatok felkutatását, gyűjtését, szelektálását, elemzését-értékelését és felhasználását jelenti.** Az OSINT mint a fegyveres erők eszközrendszere elkülönül az üzleti hírszerzéstől (lásd lejjebb), modern felfogásában már nem feltétlenül ragaszkodik a legális módszerekhez, de nem foglalja magában az offenzív eszköztárat, így a hekkelést stb.

Az OSINT fogalma egyre inkább azonosul a nyílt forrású keresőrendszerekkel. Az eddig elvégzett elemzésekből azt szűröm le, hogy a megfelelően kidolgozott és karbantartott fogalomtárral ellátott keresőrendszerek rendszeres futtatása, majd a találatok egy fúziós központban történő összevetése és kiértékelése nemcsak lényegesen megkönnyíti és lerövidíti az értékelő-elemző munkát, hanem nagyobb adatmennyiség esetén mai tudásunk szerint mint módszer megkerülhetetlen.

Érdemes értékelni az OSINT előnyeit és hátrányait a felhasználó szempontjából.

- Összehasonlítva más hírszerzési módszerekkel az OSINT lényegesen olcsóbb.
- Mivel az OSINT definíció szerint csak legális módszerekkel vagy legalábbis nyílt forrásból működhet, így alkalmazása nem jelent fizikai veszélyt a személyi állományra vagy a működtető szervezetre jogi vagy politikai szempontból.

[59] [67] 103. oldal

- Megfelelő technológiával szinte tökéletesen álcázható.
- Mentes az engedélyezésekhez kötődő bürokráciától. Ha például ki kell deríteni egy célszemély helyét gyorsan, akkor sokkal egyszerűbb a legutolsó Twitteren készült bejegyzésének (ha ilyen van) földrajzi koordinátáit kinyerni, mint engedélyezések után a cellainformációkat megszerezni.
- Kisebb a szakirányú erőforrás igénye. Az értékelő-elemzőknek sokszor várakozniuk kell a képzett informatikusokra a feladatok torlódása miatt. Így sokkal gyorsabban jutnak eredményre az OSINT eszközeivel saját maguk.

Ugyanakkor az OSINT hátrányait sem szabad figyelmen kívül hagyni.

- Védett adatokhoz nem fér hozzá. Biztonságtudatos személyek esetében szinte lehetetlen nyílt forrásból adathoz jutni. Ugyanez igaz mélyfedésű hírszerzők rejtett identitására. Ehhez HUMINT, hacking vagy más nem nyílt módszer szükséges. Ezért az OSINT használhatósága erősen korlátozott.
- A talált adatok hitelességét különösen ellenőrizni kell, hiszen a nyílt forrásokban rengeteg felesleges, téves vagy szándékosan megtévesztő információ szerepel. A megszerzett adatokból szűrőkkel, kiemelő algoritmusokkal desztilláljuk ki a minket érdeklő lényeget. A téves vagy szándékosan megtévesztő információk kiszűrése erre specializált szakértők feladata. Ennek algoritmizálása nem oldható meg tökéletesen.
- A nyílt adatforrású keresés a technológia nyíltsága miatt nemcsak a katonai és rendvédelmi szervek számára elérhető, hanem mindenki részére, így a bűnözőknek is, akik a tevékenységüket olykor nagyobb erőforrásokkal tudják támogatni, mint az állami szervek.

Mivel a nyílt forrású hírszerzés a lényegéből fakadóan aránytalanul sok felesleges, téves vagy szándékosan megtévesztő adatot tartalmaz (lásd előnyök és hátrányok), fontos Robert Steele csoportosítását felsorolni. Steele az információt négy kategóriába sorolja az alábbiak szerint.[60]

- Nyílt forrású adat (*Open Source Data OSD*). Az adat nyers formátumú, mint nyomdai termék, kereskedelmi műhold képe, szóbeli tájékoztatás, műsorsugárzás, hangfelvétel stb.
- Nyílt forrású információ (*Open Source Information OSIF*). OSIF előkészített, tömörített adat, amelyet a szerkesztő valamennyire megszűrt, hitelesített és terítésre előkészített. Ilyen például egy napi vagy heti jelentés.

[60] [75] 129–147. oldal

- Nyílt forrású híranyag (*Open Source Intelligence OSINT*).
- Hitelesített nyílt forrású híranyag (*Validated OSINT OSINT-V*). OSINT-V magas fokon hitelesített, gyakorlatilag biztosnak tekinthető hírszerzési információ.

A keresés jóval az internet elterjedése előtt már használatos volt, így a meglévő tradicionális módszereket fejlesztették tovább a weben futó alkalmazásokhoz. Alapvetően a technológia ugyanaz, de fontos megemlíteni a lényeges különbségeket. A weben tárolt szövegek gyorsan változhatnak, és ezt a rendszernek követnie kell. Ezek formátuma roppant változatos, hiszen semmi sem kényszeríti a szerzőket egységesítésre. Fontos megjegyezni, hogy míg a saját anyagokhoz a hozzáférést szigorúan szabályozni lehet, a weben tároltak mindenki számára eleve hozzáférhetőek. Ezt azért fontos megjegyezni, mert az érdeklődők sokszor nem értik, minek egy külön belső keresőrendszer, amikor bármit meg lehet a Google segítségével találni. Egyrészt kényes adatokat senki nem szeretne a köz számára a Google-on keresztül hozzáférhetővé tenni, másrészt a nagy keresőrendszerek sem érnek el mindent, így például a fizetős adatbázisokat sem.

Bár egy OSINT-alkalmazás keresési témái roppant sokfélék lehetnek, érdemes néhányat kiemelni a nemzetbiztonság és a rendvédelem területéről: terrorelhárítás, pénzmosás, adócsalás, csempészet, fegyverkereskedelem, emberkereskedelem, pedofília, kábítószer-kereskedelem, bandaháború, kémelhárítás, kartellesetek, kormányzati döntések előkészítése, kormányzati döntések visszhangjának mérése, váratlan eseményekről való gyors – HUMINT-ot megelőző – tájékozódás.

A személyeken kívül fontos terület a személyekhez kötődő szervezetek, vállalatok, intézmények figyelése. A gazdasági szervezetek nyílt forrású figyelése és elemzése az üzleti hírszerzés témakörébe tartozik, amelyet bővebben a következő fejezetrészben külön elemzek.

Az OSINT információforrásait a keresés szempontjából két kategóriába sorolhatjuk[61]:

- klasszikus források: könyvtárak, irattárak, nyomtatott sajtó, tv, rádió, vásárok, konferenciák, folyóiratok, tudományos publikációk, interjúk;
- elektronikus források: a klasszikus források digitális formában elérhető változatai, ingyenes és fizetős adatbázisok, közösségi helyek (Facebook, LinkedIn, Twitter, Instagram, blogok, fórumok, nyilvános csevegőhelyek), honlapok, műholdak adatai, online keresőrendszerek, online térképek, közösségi helyek forgalmi adatai, online

[61] [67] 105. oldal

közösségek, online dokumentumtárolók, online fényképtárak, online videótárak, online cím- és telefonszám-adatbázisok, IP- és doméncímlelőhelyek, speciális metakeresők (pl. emberekre), kormányzati adattárolók, rádiófrekvenciás figyelők.

Az adatforrások lehetnek szövegesek, hang- és kép- vagy videóformátumúak. Mindegyik formátum feldolgozása más és más technológiát igényel. Az elektronikus források természetesen nem mind szöveges tartalmúak. Mivel ezekre nem alkalmazhatóak a szemantikus keresés eszközei, így a részletesebb vizsgálódásom területéről kijjebb esnek. Az internetes adatforrásokat – mivel a klasszikus HTML-formátumokat már nemigen tartják be –célzott egyedi lekérdezőkkel (*scraping*) lehet kinyerni. Ezek a robotok általában az emberi lekérdezést emulálják, szükség esetén külön módszereket alkalmazva a célforrás robotot észlelő algoritmusainak megtévesztésére. Iparággá nőtt az ilyen lekérdezés („scrapelés") szolgáltatói darabszámhoz kötött díjért nagy volumenben töltik le az internetes tartalmakat a megadott helyekről a megrendelő számára.

A szöveges hírek általában több nyelven kerülnek be az adatgyűjtő mechanizmusba. Mivel az értékelő-elemzők nyelvtudása általában korlátozott, ezért a kevésbé ismert nyelveken megjelenő híreket gépi úton lefordítják, majd a fontosnak tartott elemeket alaposabb vizsgálatnak vetik alá. A nyílt források a világ összes nyelvén megjelennek. Fordításra alkalmas például a Google-fordító, amely a felhőben működik, és így minősített szövegek feldolgozására alkalmatlan, valamint az offline állapotban működő fordítók, mint például a Systran termékei. A hanganyagokat gépi úton írott szöveggé alakítják (S2T), majd a többé-kevésbé tökéletes szövegtestet szövegbányászati eszközökkel vizsgálják. Érdekes téma esetén emberi erőforrással részletesen feldolgozzák.

Kutatásaim szerint a nyílt forrású keresés a magyar jogban ugyan használt (lásd bővebben 4.2 fejezetrész), de korántsem definiált fogalom. Nemcsak az nem egyértelmű, hogy mi a nyílt forrás, hanem az sem, hogy a nyílt forrásból kinyert információ kezelésére milyen szabályok vonatkoznak. Egyesek szerint, ami nyílt forrásból származik, az szabadon felhasználható, hiszen mindenki ennek tudatában tette nyilvánossá, míg mások a nyílt forrásból származó adatokból készül desztillátumok készítését engedélyhez kötnék. Ebben az Általános Adatvédelmi Rendelet (*General Data Protection Regulation*, GDPR) [77] valamennyire egyértelműsít a civil szféra számára. E sorok írásakor (kicsivel több mint egy hónappal a bevezetés előtt) a módosított Infotv. még nem jelent meg, így egyértelmű következtetéseket levonni sem lehet.

3.2.2. Az információkeresés alkalmazása a nyílt forrású hírszerzésben

Az alábbiakban ismertetek néhány területet, amelyek a nyílt forrásra való szemantikus keresés alkalmazhatóságát hivatottak bemutatni. Mindegyik esetben kizárólag nyílt, mindenki számára elérhető forrásból dolgozhatunk, és alkalmazzuk a számítógépes nyelvészet széles eszköztárát. Fontos viszont megjegyezni, hogy az alkalmazott keresőtechnológiák felhasználása többszörös. A 19. ábrát tekintve egyrészt alkalmazzák az adatgyűjtésében, internetes keresőrendszerek formájában (*web search*, *webmining* stb.), majd a már feldolgozott, strukturált információ terítésében, jelen esetben a lekérdezésekben (*enterpise content search*).

Kiemelem a közösségi média (*Social Media,* SocMed) figyelését, ami ma már a nyílt forrású hírszerzés alapvető műveleti ága [78]. A közösségi média üzleti logikája arra épül, hogy emberek, cégek, szervezetek önként tárnak fel minél többet az életükről, érdeklődésükről, működésükről, kapcsolataikról. Ez valódi aranybánya az érdeklődőnek, aki erre kifejlesztett keresők segítségével nemcsak naprakész adatokhoz jut, hanem azokat feldolgozva rejtett összefüggéseket is feltárhat. A Facebook, LinkedIn, Instagram, számtalan blog, fórum, csevegőfelület stb. vizsgálata az OSINT egyik zászlóshajója lett.

A közösségi média keresőalkalmazásokkal történő figyelése általában úgy történik, hogy egy vagy néhány paraméter ismeretében találjuk meg a többi releváns adatot. Adat- illetve szövegbányász eszközökkel ezekből – akár rejtett – összefüggéseket tárhatunk fel, és előrejelzéseket készíthetünk. Az ismert paraméter lehet egy földrajzi pont, annak egy bizonyos sugarú környezete, egy személy vagy szervezet valamely adata mint név, felhasználói név, e-mail-cím, IP-cím, doménnév, telefonszám stb., egy téma kulcsszó vagy hashtag formájában, időpont vagy -szakasz stb. A kiszűrt, elemzett adatállományok külön eszközökkel vizualizálhatók, exportálhatók más rendszerekbe, archiválhatók későbbi felhasználásra. A szakirodalom által rendkívül intenzíven tárgyalt területből néhány alkalmazási lehetőséget sorolok fel.

Anonim szerző beazonosítása szemantikus nyelvi elemzéssel

A nyelvi sajátosságok (szórend, stílus, tipikus hibák, regiszter stb.) elemzése alkalmas a szerző beazonosítására vagy éppen a kizárására, hangulatának, földrajzi helyének megállapítására.

Internetes fórumokon történő álcázott üzenetváltás felismerése

Ismert tény, hogy a rövid és legkevésbé feltűnő nem testközeli kommunikációs módszerek egyike a felhőben, elsősorban fórumokon való üzenés. Az eljárás lényege, hogy az üzenő felek egy előre megbeszélt fórumon ugyancsak előre egyeztetett felhasználói néven bejelentkeznek, és egy adott témához hozzászólnak. A hozzászólásnak egyrészt feltűnésmentesnek kell lennie, másrészt pedig az előre definiált virágnyelven a kívánt üzenetet továbbítania kell. A feladat egy

tipikus OSINT-probléma kezelése szemantikus keresési módszerrel. Meg kell találni a fórumos hozzászólások szénakazlában azt a tűt, amely egy ilyen üzenetváltás mintáját, típusát magán hordozza. Az alább felsorolt szempontok alapján a keresési mintákat egy nemzetbiztonsági vagy rendvédelmi szakértő és egy erre a területre specializálódott nyelvész együtt határozhatják meg.

Vállalatok kartellező tevékenységének felismerése

A kartellező vállalatok törvényellenes viselkedésmintái jól elkülöníthető kategóriákba osztályozhatók. Egy nagyon egyszerű példa, amikor egy személy három vállalat tulajdonosa. E három vállalat felváltva ad ajánlatot közbeszerzésen oly módon, hogy az egyik mindig olcsó, kettő pedig drága. Így mindig ugyanaz a tulajdonos nyer. Természetesen ennél lényegesen bonyolultabb módszerek is tipizálhatók. Egy közbeszerzési értesítő strukturálatlan, azaz szabadszövegű PDF-állományaiból entitásfelismeréssel kiemelhetők a vizsgálat szempontjából fontos szövegelemek. Ilyen a beszerzés tárgya, a nyertes és vesztes cégek neve, címe, az ár stb. Az egyes közbeszerzések adatcsoportjait összehasonlítjuk gépi úton az előre ismert kartellezési mintákkal, és kiszűrjük a kartellgyanús eseteket, majd ezeket egy adatbázisban tároljuk. Innen átkérdezünk az előre elkészített céginformációs adatbázisba, és kinyerjük a vizsgált cégek további adatait, úgy mint tulajdonos, banki információ, kapcsolt vállalatok, összefüggésbe hozható személyek. A teljes adatállomány grafikusan ábrázolható, ahol a gráfok élei a különböző kapcsolatok (tulajdonos, ügyvezető stb.), a csomópontok pedig maguk a vállalatok [79], [80].

Deviáns viselkedésre való hajlam detektálása szemantikus OSINT segítségével

Gritzalis két OSINT-alkalmazást is ismertet [81], amelyek segítségével a társadalomra, illetve egy adott szervezetre ellenséges magatartást, illetve arra való hajlamot lehet kimutatni. A vizsgált személyek szöveges megnyilvánulásait elemzik, majd deviáns viselkedésmintákhoz hasonlítva gépi úton osztályozzák. Így kiszűrik azokat, akikről feltételezhető, hogy a környezetükre káros módon viselkednek vagy terveznek viselkedni, különösen a rendvédelmi szervekkel szemben. Egy másik kísérlettel az előbbihez hasonló módon detektált nárcizmusból következtetnek arra a hajlamra, hogy az őket alkalmazó szervezetnek ártsanak. A végzett pszichológiai elemzéseket a szerző a társaival több fórumon is ismerteti. A jelen írás szempontjából a legérdekesebb, ahogy a szemantikus módszereket alkalmazzák adott minták alapján prediktív analízisre.

Bűn- és terrorcselekmények előrejelzése

A nyílt forrású keresés egyik legfontosabb alkalmazási területe a bűn- és terrorcselekmények előrejelzése. A keresés fő forrása a közösségi média, amelyben nyíltan vagy burkoltan erős

vagy gyenge jelek érzékelhetők egy cselekmény előrejelzésére. Ilyen lehet az utalás egy fegyverkereskedelmi eseményre [78], készülés egy robbantásra [82], felhívás zavargásra [83], egy magányos farkas „önvallomása" az öngyilkos merénylete előtti indoklásként [84] stb. Az ilyen és ehhez hasonló jelek időben történő érzékelése és kiértékelése alkalmas megelőző intézkedések meghozatalára.

Vélemény- és hangulatfigyelés

A 2.2 fejezetrészben tárgyal szentimentelemzés lényege, hogy szemantikus eszközökkel mérjük a szerző véleményét az általa alkotott szöveg alapján egy bizonyos témában. E mérési eredményeket idősorba helyezhetjük, és így az alkalmazás képes a trendek figyelésére. A vizsgált dokumentumok ritkábban egy szervezetben belül keletkeznek (elsősorban e-mailek, wikik, dokumentumok, fórumok), leggyakrabban pedig a weben (közösségi helyek, mint Facebook, LinkedIn, fórumok, blogok stb.). A hangulatfigyelés klasszikus alkalmazási területei például az egyes árukról és szolgáltatásokról kialakult vevői vélemény mérése, politikusokat, pártokat övező hangulat elemzése, a tőzsde mutatóinak előrejelzése a pszichológiai jellemzők alapján [85], [86].

Polgári-katonai együttműködés

Katonai területen a polgári-katonai együttműködés (*civilian-military co-operation*, CIMIC) az OSINT egyik felhasználási területe. [87] A civil közösség véleménye, információi, esetleges készülődése, tevékenysége nyilvánvalóan fontos területe a hírszerzésnek.

Crowdsourcing[62]

A crowdsourcing a mi szövegösszefüggésünkben **az a tevékenység, amelynek során a közösségi médián keresztül megismerjük, befolyásoljuk a közösség véleményét egy cél érdekében.** A crowdsourcing segítségével képesek vagyunk a tömegből információt szerezni, a tömeggel bizonyos problémákat megoldatni, illetve a tömeg véleményét befolyásolni, akár manipulálni. Ilyen feladatok lehetnek a kollektív pénzgyűjtéstől a közösségépítésen át egy meteorológiai krízishelyzet kezeléséig [88].

A továbbiakban a nyílt forrású technológiákat vizsgálom. Az OSINT keresési eszköztár igen gyorsan változik. Alkalmazások, cégek szűnnek meg máról holnapra. Igyekeztem minden, az irodalom által tárgyalt kategóriából legalább mintát venni a jobb áttekinthetőség kedvéért. Így

[62] Jelenleg még nincs a magyar szakirodalomban rá megfelelő szó, így el kell fogadjam idegen szóként. Közösségi ötletbörze.

az itt leírtak részleteiben bármikor módosulhatnak. Ezért egy OSINT-technikát használó szakembernek folyamatosan követnie kell a változásokat.

Mivel a nagy internetes keresők közül tapasztalat alapján a Google adja a legnagyobb pontosságot és felidézést magyar nyelvi környezetben, érdemes a többi ismert nagy internetes kereső (Bing, ASK, Yandex, Baidu stb.) előtt a Google-t kipróbálni. Ennek ellenére igaz, hogy nincs „legjobb" kereső. Tématól függően más és más eszköz adja a legjobb eredményt [89].

Dark Web, Deep Web keresése

Deep Web, Deepnet, Invisible Web, Hidden Web stb. [90], [91] alatt az internet azon részét értjük, amit az ismert nagy keresők (Google, Bing, Yandex stb.) nem indexelnek. A legtöbb becslés ezt a teljes WWW 90%-nál nagyobb részének becsüli, sőt 99%-nál nagyobb részének becsüli Sui et al. [92]. A Dark Web, Darknet, Dark Internet a Deep Web illegális világa. Mint ilyen, kevésbé egzakt fogalom, pejoratív elnevezését a törvénytelen tartalom után kapta. Nemrégen feltárt Dark Web kereskedőhely (piachely, *marketplace*) a Silk Road, amelynek vezetőjét az FBI letartóztatta. Ismert és az írás pillanatában működő darkwebek a TOR, az I2P és a Freenet. A Dark Web alkalmazások architektúrája és titkosítási technikái alkalmassá teszik bizonyos mértékig a rejtett kommunikációra. Ezt használják oknyomozó újságírók, whistleblowerek[63], emberjogi aktivisták, elnyomó rendszerek szűrőprogramjainak „radarernyője" alatt kommunikáló állampolgárok stb. De ezt használják a drog- és fegyverkereskedők, pedofilok, hekkerek, bérverőemberek, útlevéllel, bankkártyával üzletelők, terroristák stb. Az említett három darknet közül a legtöbbet használt a TOR. A TOR-t csak különleges böngészővel (TOR Browser) és azon belül keresővel (Ahmia, Torch, DuckDuckGo stb.) lehet olvasni. A TOR-ban történő kereséshez linkfarmokat vesznek igénybe. Ilyenek a Grams, deepweblinks, The Hidden Wiki, TOR Links stb. Ezek tárolják az ajánlatok URL-jét (webcím, *Uniform Resource Locator*). A nemzetbiztonsági szolgálatok és rendvédelmi szervek célja az anonim helyek tulajdonosainak megismerése. A feladat korántsem egyszerű. Az informatikai megoldások mellett gyakran alkalmaznak csapdákat, vagyis rendelnek árut, szolgáltatást, és követik a szállítás vagy megjelenés útját [93]. Különlegesen nehéz a felderítés, ha a TOR-t VPN-nel kombinálva használják, és önmagában is nehezen fejthető titkosítást alkalmaznak (pl. a PGP-t vagy annak valamelyik mutációját). A TOR-felhasználók száma kb. 1,5-2 millió, a proxy szerverek száma jóval 6000 feletti [94]. A felderítést megkönnyíti, hogy

[63] Jelenleg még nincs a magyar szakirodalomban rá megfelelő szó, így el kell fogadjam idegen szóként. Közérdekű bejelentő.

a proxy szerverek nem elhanyagolható részét maguk a nemzetbiztonsági szolgálatok üzemeltetik.

Információkeresés online közösségi oldalakon

A Facebook Magyarországon messze a leginkább használt közösségi hely. A Facebook a Graph segítségével elég széles körű keresési lehetőséget biztosít. Ennek feltétele amerikai angolként való felhasználói fellépés. A kereshető attribútumok a név, hely, barátok, kedvencek, telefonszám, e-mail, munkahely, nem, rokonok stb. A felületen való keresés mellett alkalmazható a programmal való keresés is, ami a metakeresés egy válfaja. A Facebook állandóan változó (nem kis mértékben politikai és személyiségi jogi lobbisták nyomására) biztonsági előírásai erősen korlátozzák a lekérdezhetőséget. A pillanatnyi lehetőségek ismerete komplex feladat, amelyet naprakészen kell tartani.

A Twitter Magyarországon jóval kevésbé használatos, mint az USA-ban, bár egyes vélemények szerint gyorsan terjed. Így a Twitterre kidolgozott eszköztár értelemszerűen kevésbé alkalmazható. A Twitterre kereső alkalmazások általában egy *tweet* valamelyik adatának ismeretében keres ezzel összefüggőkre. Ilyen adat a hely, szerző, címzett, téma (*hashtag*), nyelv, szavak a szövegben, e-mail-cím. A felületen való keresés mellett alkalmazható a programmal való keresés is, ami a metakeresés egy válfaja A Twitter elemzésének jelentősége nemcsak a múlt vizsgálatában merül ki, hanem események megjóslásában is. A predikció alapjául egy földrajzi pont körül sűrűn megjelenő hasonló témájú üzenetek szolgálnak. Ezek előre jelezhetnek terrorista akciót, terjedő járványt stb.

A LinkedIn célzottan és következetesen a professzionális réteg szakmai információinak nyilvántartására és a tagok egymás közötti kapcsolatának menedzselésére készült. Az alkalmazás architektúrája elsősorban a tagok önéletrajzából áll, amelyekhez a különböző szakmai fórumok, illetve az oktatási intézmények és a foglalkoztató cégek keresztreferenciái tartoznak. Felhasználói az állás- és kapcsolatkeresőkön túl a fejvadászok és bárki, aki valamilyen szakmai háttérrel embert vagy céget keres. Rohamosan terjed, és alapjaiban változtatja meg az emberi erőforrás szakmát. A Google indexeli, és a saját keresőjén túl Google-trükkökkel is lekérdezhető. Amióta a Microsoft megvásárolta, a keresési lehetőségeket lényegesen korlátozták.

Az online közösségek Magyarországon lényegesen kevésbé elterjedtek, mint Amerikában. A legismertebb Craigslist sem mutat semmilyen releváns eseményt Budapesten. Az eBay.com vagy a vatera.hu alkalmas például illegálisan megszerzett ingóságok megtalálásához. Attól

függően, hogy a robots.txt mit engedélyez, partnerkereső honlapokon lehet személyre szóló információhoz jutni manuálisan vagy keresőmotorral. Például 2014. december 14-én a Badoo.com engedett keresést, a viszony.hu pedig nem. A prostituáltakat hirdető helyek általában kereshetők.

Számos kereső figyeli a közösségi média forgalmát (*social traffic*) [95]. Ezek előnye, hogy egyben láttatja a találatokat (Facebook, Twitter, LinkedIn, Foursquare, blogok stb. ld. lejjebb), hátránya, hogy a találati lista sokszor nem teljes. Ilyen alkalmazás az UVRX, az Ice Rocket a Meltwatertől (Magyarországon is fenntart fejlesztőközpontot) vagy a Delicious. Sajnos több ilyen kombinált kereső erősen fókuszál az amerikai forrásokra kisebb magyar relevanciával.

Az Avatar[64] menedzsment nem szigorúan technológiai téma, de az OSINT-ban gyakran alkalmazzák. Ahhoz, hogy a közösségi médiában valakihez vagy valakin keresztül valamihez közel lehessen kerülni, az illető ismerősévé kell válni. Ahhoz pedig, hogy ez álcázott módon történjen, egy fiktív személyiséget, ún. avatárt kell kreálni és futtatni. Minden nagyobb gyártó és intézmény futtat avatárokat mint nyílt hírszerző médiumot. A keresőprogramokat és keresőeszközöket stb. az avatáron keresztül mozgatják. A legenda kialakítása és menedzselése nem témája a jelen írásnak.

3.2.3. A nyílt forrású keresés fejlődésének várható jövőbeli irányai

A nemzetbiztonsági és rendvédelmi szektorban történő információkeresés jövője néhány évre nagy biztonsággal megjósolható. Tovább nő a mobil alkalmazások szerepe, illetve a meglévő alkalmazások mobil platformra történő adaptálása vagy migrálása. A keresési források vonatkozásában tovább nő a közösségi média szerepe. A gép-ember interfész egyre inkább felhasználóbarát lesz: a billentyűzetet felváltja a hangvezérlés, a megjelenítésben szerepet kap a virtuális valóság (*virtual reality*, VR), a holografikus módszerek és az egyre fejlettebb monitoron megjeleníthető vizualizációs technikák. A számítógépes nyelvészeti alkalmazások megoldják az emberi beszéd gyakorlatilag tökéletes írott szöveggé történő alakítását, valamint bármely nyelvről bármely nyelvre történő automatikus fordítást. A gépi tanulásos algoritmusok, különösen a mélytanulásos módszerek lehetővé teszik a sok forrásból bejutó adatok intelligens feldolgozását és az azokból történő előrejelzéseket. A nyílt forrású és belső tartalmat feldolgozó keresőrendszerek a mesterséges intelligenciával karöltve lehetővé teszik a felhasználó számára a teljes körű információszerzést, az automatikus feldolgozást, az intelligens lekérdezést

[64] A szó eredetije a szanszkrit (ava-,“le”) és a (tāra,“megmentő”) szavakból áll össze. Visnu földi inkarnációja. A Cameron film címe a kettősséget szimbolizálja.

felhasználóbarát formában, valamint a meglévő és frissen befutott adatok alapján az előrejelzést.

3.3.Információkeresés a közigazgatásban

3.3.1. A jogi információkeresés alapjai

Az információkeresés jelen korunk talán legaktuálisabb témája az adatbiztonság mellett. Nemcsak azért, mert az exponenciálisan növekvő adatmennyiségben a keresett lényegi elemek megtalálása létfontosságú a gazdaság, államigazgatás és az élet más területein, hanem azért is, mert a legutóbbi időkben – inkább hónapokra, mint évekre visszamenőleg – sorban megjelenő, mélytanuláson (*deep learning*) alapuló mesterséges intelligencia megoldások forradalmasítják az életünket. Így van ez a tumordiagnosztikától a vezető nélküli járműveken át a pénzügyi műveletekig.

Ahogy a közigazgatás egyre inkább átáll a papíralapú adatkezelésről az elektronikus alapúra, úgy nő a gépi információkeresés jelentősége is. Nem célom e fejezetrészben a magyar kormányzati információtechnológiai stratégia elemzése, még kevésbé a bírálata. Az mindenképpen megállapítható, hogy egyszerre működtek centralizáló törekvések időben változó – olykor egymással versengő – erőcentrumok részéről (ilyenek a teljesség igénye nélkül a Magyar Posta, az Informatikai és Hírközlési Minisztérium, a Miniszterelnöki Hivatal, a Belügyminisztérium és azon belül a NISZ (Nemzeti Infokommunikációs Szolgáltató Zrt.), a Nemzeti Fejlesztési Minisztérium stb.), és – időnkét túlburjánzó – önálló törekvések az egyes intézmények, szervezetek, állami tulajdonú vállalatok részéről. A kialakult fejlesztések és szolgáltatások tükrözték a momentán erőviszonyokat, az egységesítés és gazdaságosság szempontjait, olykor a parciális érdekek mögé helyezve. Az információkeresésre és így a megtalálhatóságra (*findability*) való igényt többek között a működési hatékonyságra és eredményességre törekvés, ugyanakkor az információ monopolizálásának igénye, valamint a túlzott átláthatóságtól való ódzkodás jellemezte. A közigazgatást még mindig egy erősen hibrid adatkezelés és adatfeldolgozás jellemzi. Lokális gépi feldolgozás, kinyomtatás utáni papíron történő hivatalos kommunikáció, beszkennelés (OCR-rel vagy éppen anélkül), és papíralapú, képi formátumú vagy OCR-rel gépi olvashatóságra alkalmassá tett formátumban történő tárolás a kétségtelenül növekvő arányú, eredendően gépi olvasásra alkalmas formátumok (Word, PDF, e-mail stb.) mellett. Ugyanakkor észlelhető a privát felhőalapú technológiára épülő központosított adatfeldolgozásra törekvés elsősorban a NISZ részéről, ami az információkeresést, összefüggések feltárását – talán olykor túlzott mértékben is – megkönnyíti. A vizsgálatomnak nem tárgya a nem szövegtípusú adatforrások keresése. Ilyenek a képek,

hangok, videók, amelyeknek fontos szerepe van a közigazgatásban, de a relevanciájuk a szövegkeresés szempontjából másodlagos.

Az információkeresést a közigazgatás több területén alkalmazzák, ilyen az e-kormányzat, a különböző közigazgatási adatbázisok és portálok stb. Ide tartoznak a közigazgatási nyilvántartások, a közigazgatásban egyedi ügyekben keletkezett ügyiratok, normatív, nem egyedi ügyiratok (például belső utasítások, módszertanok) stb. Ezek felhasználói egyrészt a legkülönbözőbb professzionális állami alkalmazottak, másrészt az informatikai szempontból literátus lakosság. Egy különleges felhasználói terület a közösségi információt feldolgozó és továbbszolgáltató (*public sector information reuse*, PSI) információbrókerek. Talán a legigényesebb terület a jogi keresőrendszereké. Mint lejjebb bemutatásra kerül, a jogi keresőrendszerek – szemben a kulcsszavakra kereső adatbázis-kezelőkkel és szabadszavas keresőkkel – nemcsak karaktersorozatra, hanem fejlett szemantikát alkalmazó jelentésalapú keresésre, valamint mesterséges intelligenciára épülő elemzésre és így prediktív analízisre is alkalmasak lehetnek. A következőkben ismertetem a jogi informatika sajátságait, elemzem az információkeresés szempontjából lényeges jellemzőit, majd értékelem az információkeresés jelenlegi helyzetét és jövőbeni alkalmazásának irányait, lehetőségeit. Az első részben áttekintem a jogi információkeresés alapjait és alkalmazott technológiáit. A második részben konkrét példákon mutatom be a klasszikus jogi keresőrendszereket, míg a harmadik részben kitekintést adok a legújabb, mesterséges intelligenciára épülő alkalmazásokra.

A jogi információkeresés elsődleges forrása értelemszerűen minden jogi természetű szöveg, amely nyílt, korlátozott vagy zárt formában informatikai keresőeszközzel elérhető. Ilyen a teljesség igénye nélkül a törvényeket tartalmazó jogtár, a bírósági döntéseket tartalmazó döntvénytár, az Alkotmánybíróság, a Gazdasági és Versenyhivatal, a Magyar Nemzeti Bank, a Közbeszerzési Hatóság határozatai, a Kúria állásfoglalásai, minisztériumok, országos hatáskörű szervek rendeletei stb. Ezen adatforrások egy része közérdekű adat, amelyekhez az ingyenes hozzáférés állampolgári jog. Ilyenek például az anonimizált bírósági határozatok. Nyilvánosan hozzáférhetők, de fizetősek például a Kúria által kiadott elvi jelentőségű egyedi döntések, az ún. BH-k [96]. A nyilvánosság számára nem elérhető a bírósági határozatok eredeti, anonimizálás előtti formája. Természetesen ugyancsak titkosak a rendvédelmi és nemzetbiztonsági szervezetek, az ügyészség stb. jogi anyagai. A jogi anyagok mint szövegtestek általában strukturálatlan Word- vagy PDF-formátumúak, de lehetnek strukturált adatbázisba rendezettek. A PDF-formátumúak általában olvashatóak gépi úton, de lehetnek – akár szándékosan is – képi formátumúak, amelyet a gép csak OCR-rel tud olvasni, vagy úgy

sem, mert tartalmazhatnak véletlen módszerrel generált, láthatatlan karaktersorokat a gépi olvasás megnehezítése végett. Ilyen esetben szükséges ezek kriptoanalitikai eszközzel történő feloldása és az eredeti szöveg olvashatóvá tétele.

Az Európai Parlament és a Tanács 2003/98/EK [97] számú irányelve a közszféra információinak további felhasználásáról kimondottan a közadatok feldolgozását és újrahasznosítását hivatott elősegíteni. Ezt az irányelvet vette át nem hamarabb, mint 2012-ben a magyar törvénykezés a 2012. évi LXIII. törvényben [98] a közadatok újrahasznosításáról. Az EU a 2003-as irányelvet 2013-ban módosította a 2013/37/EK irányelvben [99], amelyet a Parlament 2015-ben vett át a 2015. évi XCVI. törvényben (Infotv.) [100]. A törvény elemzése nem feladata a jelen írásnak, de néhány elemet feltétlenül ki kell emelni az információkeresés szempontjából.

- Minden jogi vagy természetes személy jogosult közadat igénylésére továbbhasznosítás végett (ilyen például az információbróker).
- Az adatszolgáltató a közadatot ingyen vagy minden igénylő számára ugyanolyan és átlátható feltételek (díjazás ellenében) mellett köteles szolgáltatni.
- A díjazás legfeljebb 5%-kal haladhatja meg az önköltséget.
- A közadatot géppel olvasható formátumban (*machine readable*) kell rendelkezésre bocsátani.
- Nem megfelelő eljárás esetén az igénylő bírósághoz fordulhat.

Ugyancsak nem feladatom annak elemzése, hogy a fenti törvényeket – szankciók hiányában – mennyire tartja be a magyar közigazgatás. **Ugyanakkor feltétlenül kiemelem, hogy a törvény be nem tartása vagy látszólagos betartása az információkeresés szempontjából döntő fontosságú**. Az adatgazdák „szemérmessége” a kereshetőséget jelentősen megnehezítheti vagy éppen el is lehetetlenítheti. A módszerek többek között a következők:

- az információt nem jelenítik meg;
- az információt nagy késéssel jelenítik meg, amikorra az már elavul;
- az információt gépi úton kereshetetlen (képi vagy láthatatlan karakterekkel „megszórt” [*salted*]) formában jelenítik meg;
- az információt a könyvtári hierarchiában mélyen elrejtve jelenítik meg.

Meglátásom szerint feltétlenül szükséges lenne az átláthatóság végett a fenti törvényt két irányban módosítani:

- **a közérdekű adatokat könnyen elérhetővé és gépi úton könnyen olvashatóvá kell tenni;**
- **szankcionálni kell a törvény be nem tartását;**
- **meg kell alkotni a törvény végrehajtási rendeleteit.**

3.3.2. Az információkeresés alkalmazása a jogi informatikában

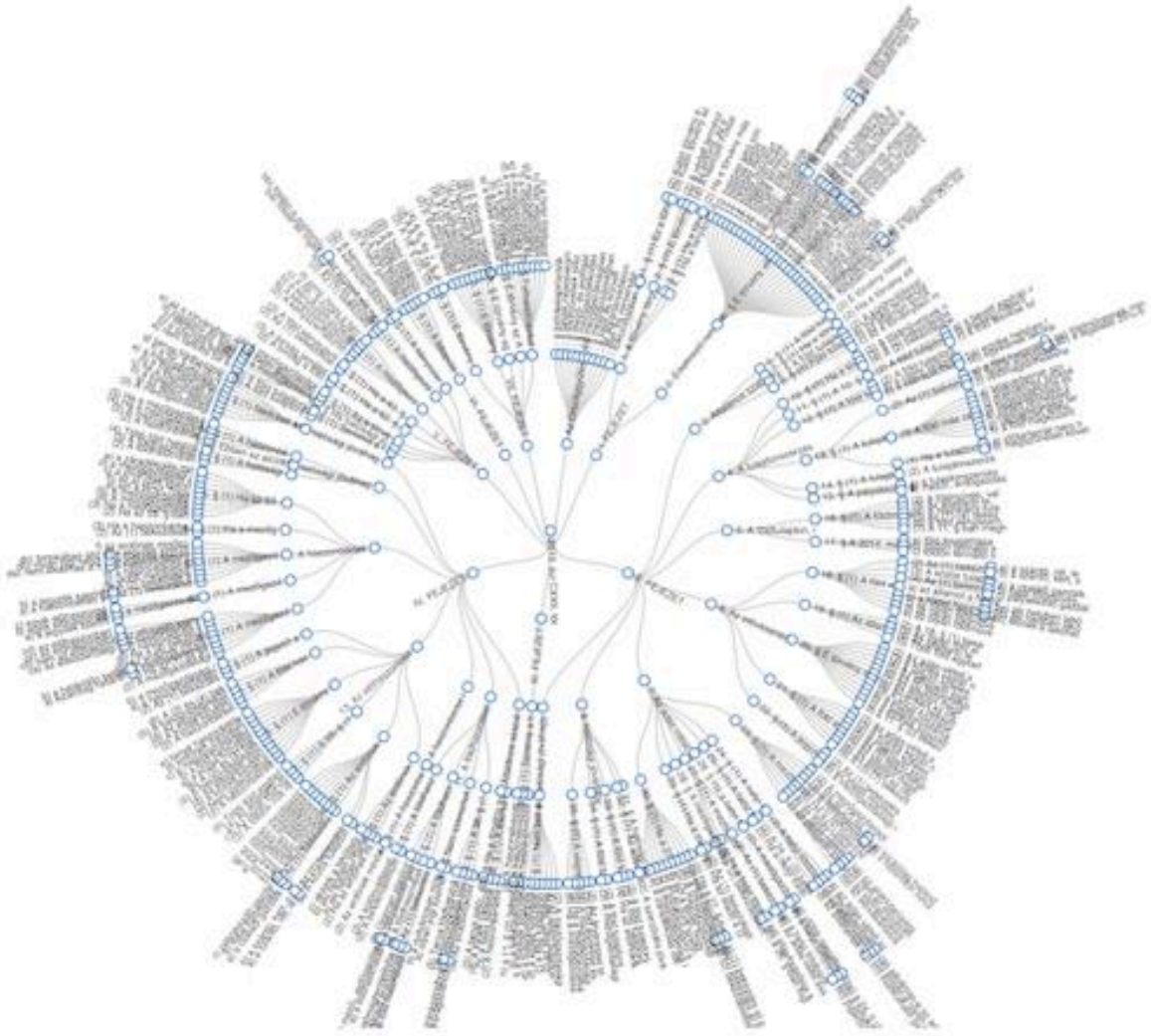
Mint azt az első fejezetben kimutattam, az ontológiák magasabb absztrakcióval képesek a jelentéstartalmú tudásreprezentációra, mint a taxonómiák. Az ontológiák a leggyakrabban használt hierarchikus elrendeződés és a szinonimaviszonyok leképezésén túl folyamatok, cselekmények reprezentálására is alkalmasak. Egyes jogi szakterületek számára ún. doménontológiákat dolgoznak ki [104]. Leképezésük több formalizmust használhat, ilyen az XML, RDF, RDFS (*RDF Schema*), SPARQL (*Simple Protocol and RDF Query Language*) és az OWL.

A mesterséges intelligencia hőskorából ismert szakértői rendszerek megtalálták a helyüket a jogi információkeresés területén is. Ezek az alkalmazások – durván leegyszerűsítve – számos döntési helyzetet tartalmaztak (if-then), amelyek az emberi gondolkodást igyekeztek szimulálni. Mivel a jog – legalább annyira, mint pl. az orvostudomány – csak nagyon vázlatosan modellálható ilyen bináris döntési pontok sokaságával, az ilyen szakértői rendszerek – hasonlóan a más területen megalkotottakhoz – többé-kevésbé elenyésztek a fejlődés során.

Bár a vizualizáció csak indirekt tartozik az információkeresés szemantikai arzenáljához, a szerepe a jogi keresésben nem elhanyagolható. Például szolgáljon a törvények egymásra hivatkozásának hálója, amit tudományos igénnyel Hamp Gábor és Markovics Réka dolgoztott fel [105] a 20. ábrán látható módon. A törvények azonosítója (keletkezés éve és a római száma), illetve a hivatkozás ténye entitáskiemeléssel kinyerhető. Majd a törvényeket mint csúcsokat irányított élekkel (ez a hivatkozás) összekötve kimutatható a teljes háló. Ennek a – Magyarországon megvalósítatlan – módszernek kulcsszerepe van a kodifikáció folyamatában. A hivatkozások nem teljességét jó láthatjuk az ábrán, hiszen csak a fastruktúra látható, a kereszthivatkozások nem. A bírósági ítéletekre való hivatkozások vizualizációs lehetőségei jól láthatóak Ződi Zsolt tanulmányában [106]. Az osztrák RIS [107] jogszabályait vizualizáló módszert tárgyalja Nabizai és Fill [108], valamint Staudegger [109].

Az elektronikus jogi információkeresés legelső és a mai napig legfontosabb forrásai a jogi adatbázisok. A jogi adatbázisok általában törvényeket, rendeleteket, bírósági ítéleteket, kommentárokat, állásfoglalásokat tartalmaznak. A nyers közigazgatási forrásuk általában

ingyenes, de a kereshetőségük és a felhasználói kényelmük olykor nem elégítenek ki minden igényt. A hivatalos forrásokra épülnek rá fizetős, ezért a korlátozott hozzáférést biztosító szolgáltatások, amelyek lényegesen jobb kereshetőséget és komfortot biztosítanak. Ez a hozzáadott értékű szolgáltatás erősen profitábilis üzlet, akár iparágnak is nevezhető. Még a kicsiny magyar piac mérete is 1,5-3 milliárd forintra becsülhető.



20. ábra: a 2013. évi CXXII. törvény szerkezete.[65]

Az uniós jog megismerését és gyakorlását az Unió szervei által létrehozott általános és szakosodott jogi alkalmazások segítik. Ezek túlnyomó része ingyenes, de vannak fizetős alkalmazások is. A hivatalos jogi alkalmazások mellett egyéb nemzeti, illetve üzleti alapú alkalmazások is fellelhetőek, amelyekben nagyobb számban vannak a fizetős szolgáltatások.

[65] Forrás: [105]. Sajnos nem sikerült olyan bontásban megszerezni az ábrát, hogy a szövegek olvashatók legyenek, de a struktúra így is jól kivehető.

Az Európai Unió intézményei (az Európai Parlament, az Európai Unió Tanácsa, az Európai Bizottság, az Európai Bíróság, az Európai Számvevőszék, az Európai Központi Bank) is üzemeltetnek saját honlapot. Ezek segítségével az uniós joggal kapcsolatos tevékenységük megismerhető, de az Európai Unió az uniós jog ismereteinek hatékony elterjesztése érdekében ezek mellett működtet EUR-Lex néven ingyenesen egy önálló, komplex jogi adatbázist, amelyen az összes releváns adat, információ egy helyen naprakészen elérhető.

A jogalkalmazók számára az előző alkalmazáshoz közel azonos fontossággal bír az Európai Bíróság – az EU legfelsőbb igazságszolgáltatási szerve – által Curia néven, ingyenesen működtetett, saját szakosodott jogi adatbázisa, amely az Európai Bíróság működésének, az ítélkezési gyakorlatának és az uniós joghoz kapcsolódó nemzeti és nemzetközi ítélkezési gyakorlatnak a megismerését is segíti. E két kiemelkedő fontosságú alkalmazás mellett más, az Európai Bizottság és az Európai Unió más szakosodott szervei által működtetett jogi alkalmazások is fontosnak tekinthetők. Az 1. számú Mellékletben felsorolom – teljesség igénye nélkül az ismertebb és fontosabb külföldi szolgáltatásokat.

A magyar jogszabályokat (törvényeket és rendeleteket), nemzetközi szerződéseket, az Országgyűlés és a Köztársasági Elnök határozatai és a jogi iránymutatásai, a Legfelsőbb Bíróság jogegységi határozatai, valamint a személyi kérdésekben hozott döntések hivatalos forrása az állami tulajdonú Magyar Közlöny- és Lapkiadó Kft. által kiadott Magyar Közlöny. A PDF-formátumú megjelenések gépi úton olvashatók, és ez a hatályosítás automatizálása szempontjából lényeges. Ennek a jelentősége az ún. „salátatörvények” esetében kiemelendő. A bírósági döntéseket az Országos Bírói Hivatal jeleníti meg. Ez tartalmazza a törvényszékek, ítélőtáblák, a Kúria ítéleteit, az OBH közleményeit stb. Az Alkotmánybíróság állásfoglalásait a honlapja ismerteti. A versenyjogi döntéseket a Gazdasági és Versenyhivatal honlapja közli. A közbeszerzési döntőbíróság határozatait a Közbeszerzési Hatóság, a Pénzintézeti Felügyelet döntéseit a Magyar Nemzeti Bank honlapján lehet elolvasni.

Kevés kivétellel elmondható, hogy a jogi adatbázisok által nyújtott felhasználói kényelem nem mindig tart lépést a korszellemmel. Ezen segítenek a lekérdező szolgáltatások, amelyek a hozzáadott értékükkel – némi túlzással – egy iparágat alkotnak. A kényelmi funkciók a teljesség igénye nélkül a következők.

- Több adatforrásból történő egyidejű keresés.
- Kollekciók alkalmazása, azaz adatforrás-alapú keresés, több szempontú (*faceted*) lekérdezés.

- A találatok kiemelése.
- Csoportmunka-lehetőségek.
- Kiemelések, átemelések egyéb munkadokumentumokba.
- Többnyelvűség.
- Taxonómia-alapú keresés.
- Metaadat-alapú keresés.
- Mentett keresés.
- Relevancia szerinti besorolás.
- Kapcsolati összefüggések feltárása és vizualizációja.

Az alábbiakban felsorolom az ismert magyarországi szolgáltatásokat.

Az MKLK (Magyar Közlöny Lap- és Könyvkiadó Kft.) által biztosított felhasználói kényelem szintje fennállása óta teret biztosít privát tulajdonban levő brókereknek a minőségi szolgáltatásra. Ezek között toronymagasan piacvezető a holland tulajdonú, multinacionális Wolters Kluwer (régi nevén Complex). Felhasználói kényelme alapján többé-kevésbé standard lett a jogalkalmazók körében. Tartalmában kiváló, felhasználói kényelmét tekintve inkább közepes szolgáltatást nyújt a HVG-ORAC, amely a jelen írás keletkezésekor az országosan is meghatározó befolyású Budapesti Ügyvédi Kamara, és ezen keresztül a budapesti jogi irodák egyetlen hivatalos beszállítója. 2017 elején az Opten átadta a jogi szolgáltatási csomagját a Wolters Kluwernek, így ebből a piacból kiszállt. Meg kell említeni a Menedzser Paxis által biztosított LexPraxis [110] szolgáltatást, amely tartalmában nem tér el lényegesen a Wolters Kluwerétől. Piaci résben mozog a Montana Lexpert nevű terméke, amely elsősorban a bíróságok és más hatóságok határozatait publikálja komplex taxonómiára épülő, számítógépes nyelvészeti technológiák és mesterséges intelligencia segítségével felhasználóbarát módon. 2017 tavaszán átadásra került, de piacra még nem vitte a HMEI (Honvédelmi Minisztérium Elektronikai, Logisztikai és Vagyonkezelő Zrt.) a Justeus [111] nevű lekérdező alkalmazását, amely az EU jogi adatbázisait teszi felhasználóbarát módon elérhetővé.

Megállapítható, hogy a jogi informatika és azon belül az információkeresés fejlődése a 2000-es évek elejére elért egyfajta platót, amelyről nemigen tudott bő egy évtizedig tovább emelkedni. A konferenciákon újabb és újabb ontológiákat mutattak be, de nagy horderejű áttörést egyik sem eredményezett. Nem látszik jele annak, hogy bármelyik kifejlesztett ontológia széles körű elfogadottságot vívott volna ki magának. Ez annak is betudható, hogy a jog szigorúan nemzeti hatáskörben maradt, és feltehetően belátható ideig ez nem is fog

megváltozni. A nemzetközi és még inkább a hazai kitekintés azt mutatja, hogy még az ismert és elfogadott számítógépes nyelvészeti technológiák sem tudtak a keresőrendszerekben széles körben meghonosodni. Nemcsak a felsorolt magyar jogi adatbázisok nem alkalmaznak szemantikai megoldásokat, de a Lexpert és a Justeus kivételével még a rájuk épülő keresőrendszerek sem igen jutottak túl a kulcsszavas, karaktersoros lekérdezéseken. **Meglátásom szerint új irány a jogi – és nem csak a jogi – keresőrendszerek fejlődésében a mélytanulásra épülő megoldások alkalmazása**. Ahogy gépi tanulórendszerek forradalmasítják a közlekedéstől a gyógyászatig a modern szolgáltatásokat, úgy idővel a jog területén is **várható egy inflexiós pont a fejlődésben**.

3.3.3. A jogi információkeresés fejlődésének várható jövőbeli irányai

A következőkben áttekintem a modern jogi információkeresés technológiai hátterét. A keresés tárgyai változatlanul zömében törvények, rendeletek, bírósági határozatok, állásfoglalások, indoklások. Adattípus tekintetében kisebb részben strukturált adatbázisok, és zömében strukturálatlan vagy gyengén strukturált szabad szövegű szövegtestek. Ezek feldolgozása modern számítógépes nyelvészeti, szemantikai eszköztárat igényel. Ezeket a technológiákat az első fejezetben részletesen ismertettem. Némi figyelmet érdemel a beolvasandó fájltípusok kérdése. Nagy gondot okoz a régebbi papíralapú anyagok beolvasása. Ezek beszkennelt képi (tiff, jpg stb.) változata (és bizonyos PDF típusú fájlok is) gépi úton nem értelmezhető, csak miután az OCR-rel olvashatóvá változtatták. A szkennelési és az azt követő hibajavítási munka költséges, időrabló, és nagy humán erőfeszítést igényel.

A korszerű információkeresési módszerek és technológiák alkalmazását a jog világában két szempont szerint igyekszem megvilágítani. Egyrészt bemutatom azokat a nagyobb európai kutatásfejlesztési projekteket, amelyek mérföldkövek a fejlődés útján, másrészt ismertetek már elterjedt vagy éppen most áttörést jelentő termékeket a legutóbbi időkből. A válogatás természetesen szubjektív, de próbálja követni a nagy jogi informatikai konferenciák programját. A negyedik ipari forradalom a jog területét sem hagyja érintetlenül. Nemcsak a jog, hanem más professzionális szakterületeket is vizsgál Richard Susskind, az Oxford Egyetem professzora a The Future of Professions című munkájában [112]. Ide vonatkozó megállapítása egyrészt, hogy a professzionális munkák tekintélyes részét át fogják venni a robotok, másrészt, hogy ez különösen vonatkozik a jogi munkára. Végigveszi azon jogi munkafolyamatokat, amelyeket a mesterséges intelligencia eszközeivel már megoldottak vagy várhatóan a közeli jövőben meg fognak oldani. E folyamat a jogi munka áramvonalasítását és nem utolsósorban olcsóbbá tételét eredményezi. A Deloitte nemrég megjelent tanulmánya [113] szerint kb.

114.000 fővel csökken a jogi területen dolgozók száma az Egyesült Királyságban a következő 20 évben.

A – kétségtelenül önkényesen – korszerűnek titulált technológiák egyrészt az NLP legújabb eredményeinek felhasználását, másrészt a gépi tanulás és azon belül a neurális hálókat alkalmazó mélytanulás alkalmazását jelentik. A klasszikus Boole-algebrai operátorokkal történő keresést messze meghaladja a nem bináris döntési eseteket kezelő elmosódott halmazok logikáján operáló (*fuzzy logic*) fuzzy keresés (*fuzzy search*). Az irány egyértelmű. A sok időt – és így pénzt – igénylő munkát minél inkább robotokkal kell elvégeztetni. Felsorolok néhány, a fenti technológiára épülő alkalmazási területet, Stephen Wolfram csoportosítását felhasználva [114]: döntéstámogatás, kivonatolás, érvelési modellek kidolgozása, esetalapú érvelés, a jogi érvelés formális gépi modellezése, jogi érvelés multiágens rendszerekkel (*multi agent-systems*, MAS), tudásábrázolás, érvelés, következtető alkalmazások, szerződésírás és -vizsgálat, dokumentumosztályozás.

Az alábbiakban felsorolom az általam összegyűjtött legismertebb, mesterséges intelligenciára épülő jogi kereső- és döntéstámogató rendszereket.

A széles területet lefedő, de felépítésében nem mély EUROVOC-nak komoly versenytársa az EU kutatási keretében kifejlesztett LOIS (Lexical ontologies for legal information sharing) [115]. A projektet alaposan elemzi Paolo Curtoni et al. [116] és Daniela Tiscornia [117]. Ismert, hogy a MetaLex nevű ontológia lett az ESTRELLA [118] nevű EU-finanszírozott projekt direkt eredménye. A projekt különlegessége, hogy a Corvinus Egyetem és az APEH személyében magyar részvevője is volt.

A **Lex Machina Legal Analytics Platform** [119] nevű terméke a szabadalmi és szellemi termékekre vonatkozó bírósági döntéseket dolgozza fel gépi tanulásos módszerrel. Az ítéleteket kategorizálja, és számos szempont szerint elemzi. Ilyen kiemelt szempont például a bíró habitusa, milyen jogszabályokra szokott hivatkozni, hasonló esetek összehasonlítása, az ítélkezés konzisztenciájának vizsgálata. Az eredményeket, összefüggéseket vizualizálja, és gazdagon illusztrálja. A cég a Stanford Egyetem projektjeként indult 2006-ban, majd később megvásárolta a LexisNexis. Lényegében csak amerikai eseteket dolgoz fel, európai felhasználónak csak amerikai szellemi tulajdoni kérdésekben hasznos.

A **Ravel** [120] hasonló feladatot lát el, mint a Lex Machina, de az innovációs ereje miatt mindenképpen érdemes külön is megemlíteni.

A **Ravn** [121] egy hétéves angol jogra specializálódott szoftverfejlesztő cég, amely a mélytanulásos rendszerével a Rolls-Royce korrupciós botrány felgöngyölítésében nyújtott nélkülözhetetlen segítséget az angol Csalás Elleni Hivatalnak (*Serious Fraud Office*, SFO) [122]. A rendszer nemcsak osztályozta, rendszerezte a több mint 30 millió dokumentumot, hanem kivonatokat is készített a lényeget felismerő képességével, valamint felismert képi alakzatokat a dokumentumtömegben, mint például útleveleket stb.

A **Luminance** [123] az Autonomy alapítójaként ismert Mike Lynch segítségével jött létre. Küldetése a vállalati bevizsgálás (*due diligence*), a megfelelőségi audit (*complience*), rendellenességek, csalások, anomáliák feltárása (*fraud, anomaly detection*) vagy egyszerűen a szerződésmenedzsment munkájának megkönnyítése. A gépi tanulás eszközével automatikusan osztályozza a vizsgálandó dokumentumokat, felismeri a vizsgálat szempontjából releváns fogalmakat, azokhoz ugyancsak automatikusan előhívja a releváns jogszabályokat. Ezzel hatalmas mennyiségű rutinmunkát takarít meg a személyzetnek, akik a felszabadult időt gazdaságosan produktív tevékenységre tudják fordítani.

Az **IBM ROSS** [124] nevű alkalmazása az egyik elsőnek tartott robotjogász, amelyet az IBM Watson nevű framworkje üzemeltet. Segítségével a New York-i Baker & Hostetler iroda csődeseteket dolgoz fel. A robot igyekszik a szabadszövegű kérdéseket megérteni, azokra ugyancsak szabadszövegű válaszokat adni, hipotéziseket állít fel egy-egy eset kapcsán, és követi a peres folyamatot.

3.4. Információkeresés a gazdasági életben

3.4.1. A gazdasági hírszerzés alapjai

A II. világháború óta az országok küzdelme a jelentős lokális fegyveres konfliktusok mellett egyre növekvő mértékben a gazdasági verseny síkján folyik. Az erőviszonyokat nemcsak a fegyverkezési potenciál elrettentő ereje, hanem az ipari technológiák, pénzügyi manipulációk ismerete, a piaci információk időben történő megszerzése, kiaknázása állami és főleg vállalati szinten határozza meg. Az ipari kémkedés eseteiről könyvtárnyi irodalom jelent meg. Ezekben a művekben általában minden szerző más országok bűneit elemzi. A számtalan esetből felidézek kettőt, igazán baráti államok és azok cégei között. 1996-ban a Siemens (ICE) elvesztette a Dél-Korea által kiírt 4 milliárd márkás gyorsvasúttendert a francia Alcatel-Alstommal (TGV) szemben. Hosszas kutakodás után sikerült beazonosítani a szivárgás helyét: a München–Szöul-kábel Marseille mellett merül a tenger alá, ahol azt a franciák

„meggyűrűzték” [125].[66] Edward Snowden jóvoltából pedig megtudtuk, hogy a kanadai CSEC (Kommunikációbiztonsági Intéznény, Kanada, *Communications Security Establishment Canada)*, az NSA kanadai megfelelője meghekkelte a brazil energiaügyi minisztérium és a Petrobras szervereit [126].

Az ipari kémkedés jámborabb rokona az üzleti hírszerzés, amely a definíció szerint csak legális eszközökkel működhet. Mint önálló diszciplína a 80-as években alakult ki elsősorban klasszikus, manuális és asztali számítógépes alkalmazásokat és nagygépes belső kereső eszközöket alkalmazva. 1989 után robbanásszerűen nőtt az internetes keresés szerepe az üzleti hírszerzésben. Ha megvizsgáljuk a SCIP vagy az ICI (Üzleti Hírszerzés Intézet, *Institute of Competitive Intelligence*) [127] éves konferenciáinak napirendjét, láthatjuk, hogy az internetes keresés évről évre egyre nagyobb részt hasít ki magának a tematikából. Ahogy az információk nyíltan és olcsón tárolhatóvá és elérhetővé válnak az interneten, úgy azok keresése is meghatározó részévé válik a teljes gazdasági információszerzésnek.

Ebben a fejezetpontban áttekintem a gazdasági hírszerzéshez kapcsolódó terminológiát. Megvizsgálom az üzleti hírszerzés folyamatát, forrásait, céljait és sajátosságait. Feltárom a gazdasági – és ezen belül az üzleti – hírszerzés és a keresőrendszerekre alapuló információszerzés összefüggéseit. Majd néhány példával illusztrálom az alkalmazási területüket.

Az üzleti hírszerzés mint önálló diszciplína lemaradt Magyarországon az USA-hoz és Nyugat-Európához, de még a környező országokhoz képest is. A megjelent művek inkább katonai vagy polgári hírszerző háttérrel rendelkező szerzők tollából származnak [128], [129], akik tudásukat transzponálták a gazdasági területre. „civil” kezdeményezés igen ritka.

A következőkben kiélesítem a kontúrokat a fogalmi tartományok között. Az alkalmazott besorolás részben önkényes, hiszen még a nemzetközileg alkalmazott terminológia sem egyértelmű. Amennyire lehet, igyekszem támaszkodni a már elfogadott angolszász és magyar fogalomrendszerre. A megértést segíti a 21. ábra, amely a gazdasági hírszerzés szegmenseit illusztrálja határoló vonalakkal. A definíciók, ha nem hivatkozom meg külön, tőlem származnak, természetesen a felhasznált irodalom segítségével.

Az angol *intelligence* fogalom a latin *intelligo*-ból származik, amelynek jelentése: megértem, felfogom, rájövök, megismerem, tudomásom van (valamiről). A szónak két angol jelentése is van. Az értelmi felfogóképesség, ítélőképesség és a hírszerző szervezet (*intelligence agency*)

[66] A Siemens később erőteljesen küzdött a francia cég megvásárlásáért.

által produkált elemzés. Fontos megjegyezni, hogy a hírszerzés és a kémkedés szinonim fogalmak, az eltérés mindössze hangulati jellegű. Az MTA Magyar értelmező kéziszótára a hírszerzést semlegesnek, a kémkedést pejoratív fogalomnak határozza meg.

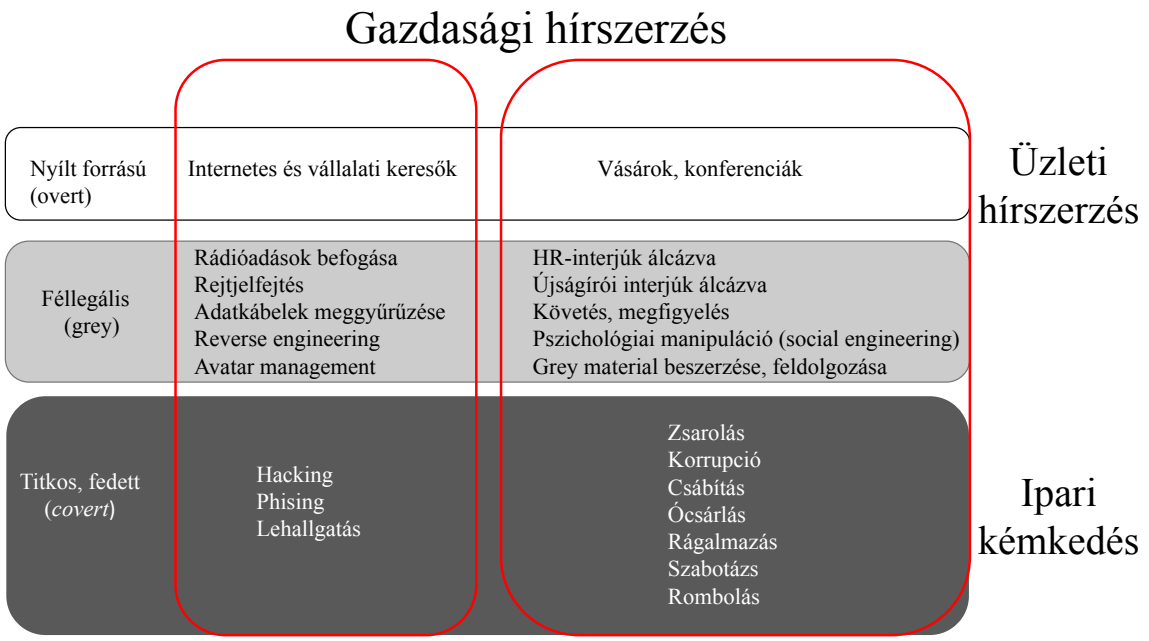
Saját definícióm szerint **a gazdasági hírszerzés[24] a következőket magában foglaló tevékenység: termékek, vevők, versenytársak, tőke- és HR-piaci szereplők, a jogi környezet és bármely más környezeti elemről szóló híranyag meghatározása, megszerzése, kiértékelése és terítése a döntéshozók számára.** Az értelmezésemben a gazdasági hírszerzés felöleli az információszerzés nyílt forrású, (*overt*), legális; féllegális (*grey*) és fedett, titkos, illegális (*covert*) ágát is, ahogy ez a függőleges tengely mentén fentről lefelé haladva bal oldalon látható. A vízszintes tengely mentén a két nagy csoport látható: a technikai és a humán módszerek. Természetesen a humán tevékenység is vehet igénybe technikai eszközöket, és egyes technikai eszközök alkalmazása is átnyúlhat a fehér-szürke vagy a szürke-fekete határon. A féllegális fogalom sem egyértelműen definiált. **Én Larry Kahaner alapján úgy értelmezem, hogy olyan tevékenység, ami nem törvénytelen, de nem is etikus. Más szavakkal: „a bíró nem ítél el érte, de nem szeretnéd, ha az újság írna róla."[67]** Az illegalitás természetesen viszonylagos fogalom abban az értelemben, hogy az ipari kémkedést mint a gazdasági hírszerző tevékenység egy fajtáját végző állami szervezet a saját rendszerén belül általában törvényesen vagy legalábbis ellenőrzötten dolgozik, de a műveleti területen a tevékenysége törvénytelen. Tehát a most tárgyalt definíciórendszer szerint az üzleti hírszerzés a gazdasági hírszerzés szerves része.

Az üzleti hírszerzés (competitive intelligence, CI) a gazdasági hírszerzés legális ága. Kizárólag nyílt vagy engedélyezett forrásokból dolgozhat törvényes eszközökkel. Az üzleti hírszerzés interdiszciplináris terület. Határos a vállalati stratégiával, marketinggel, pszichológiával, számvitellel, kontrollinggal, a tudásmenedzsmenttel, vezetéstudománnyal, döntéselmélettel, statisztikával és egyre inkább az informatikával.

A magyar üzleti hírszerzés, ipari hírszerzés, gazdasági hírszerzés, versenypiaci hírszerzés, vállalati hírszerzés (*corporate intelligence*) stb. fogalmak egyrészt összekeverednek, másrészt kevésbé szerencsések, mert a hírszerzés összemosódik a kémkedés fogalmával, és ez a civil szervezetek számára illegális tevékenységet sugall. Jellemző, hogy még a nehezen honosító német nyelvben is megmaradt az amerikaiból jött *competitive intelligence* szóösszetétel.

---

[67] [65] 241. oldal



21. ábra: gazdasági hírszerzés.[68]

Benjamin Gilad, az üzleti hírszerzés egyik meghatározó személyisége szívesen használja az EWS (*early warning system*) [130] kifejezést, ami a tevékenység célját jól tükrözi, bár ez sem pontos definíció, hiszen sok más korai előrejelző rendszer is létezik, amely a gazdasági hírszerzéstől távol esik. A továbbiakban mégis a leginkább elterjedt üzleti hírszerzés vagy CI megnevezéseket használom.

Az ipari kémkedés illegális, titkos, kényszerítő vagy félrevezető eszközök alkalmazása a privát szektorban gazdasági információk gyűjtése és piaci hátrány okozása céljából [131]. Törvény szerint ipari kémkedést csak az arra felhatalmazott nemzetbiztonsági szolgálatok végezhetnek. A műveletek természetesen csak a saját országuk törvényei szerint jogszerűek, a műveleti területükön értelemszerűen nem. Ez nem jelenti azt, hogy a korrupció, a zsarolás, a kényszerítés, lehallgatás, elcsábítás (*honey trap*) stb. ismeretlen lenne a civil ipari világban. Magyarországon előszeretettel alkalmazott technika a rágalmazás és ennek egy kifinomultabb változata: lefizetett vagy más módon befolyásolt sajtó, illetve gerillamarketing technikával a másik befeketítése, karakterrombolása és az ócsárlás.

A vállalatok üzleti (felső) szinten értelmezhető és használható információkat szolgáltató rendszere, ill. azon megközelítések és eszközök összessége, amelyekkel ilyen rendszereket

[68] Forrás: a szerző.

lehet készíteni.[69] A széles körben elterjedt üzleti intelligencia (*business intelligence*, BI) elég jól lefedné a szó szerint értett fogalmat. Nagyon gyakran használják is az üzleti hírszerzés szinonimájaként. Mégsem tartható alkalmasnak, hiszen a BI az üzleti problémákat feldolgozó adatbányászat egy jól körülhatárolható területe, és így a használata fogalmi zavart okozhat.

Az OSINT definíciója a 3.2 fejezetrészben került meghatározásra. Ez egyértelműen a katonai terminológiából származik. Jóllehet sem módszerei sem pedig vizsgálódási területe nem határolódik el élesen az üzleti hírszerzésétől, mindkét szempontból érezhetőek árnyalatnyi különbségek. Míg az üzleti hírszerzés szigorúan a nemzet- és vállalatgazdasági területen operál, az OSINT – mint az összadatforrású felderítés egyik ága – a katonai és polgári hírszerzés eszköze [132]. Természetesen a katonai és a polgári hírszerzés is vizsgál makro- és mikrogazdasági adatokat, nem kevésbé virágzó célterülete az üzleti hírszerzésnek a hadiipar és biztonsági piac. Bár a jelen fejezetrész fő témája nem az OSINT, még kevésbé a katonai felderítés, mégis érdemes egy bekezdés erejéig a fő témát ebben az összefüggésben is elhelyezni. Az üzleti hírszerzés egyértelműen a katonai és polgári hírszerzés, valamint a stratégiaalkotás és a marketing módszertanában gyökerezik. Nagy művelői és oktatói közül sokan jöttek át az állami hírszerzés világából. Katonai terminológiával élve mindkettő célja és feladata az információs fölény és uralom megszerzése és megtartása. Míg a katonai logika inkább kétdimenziósan (barát-ellenség) gondolkodik, a CI natív környezete egy sokdimenziós, sokszereplős, egyszerre versenyző és együttműködő világ. A katonai és polgári hírszerzés módszertana nagyon sok tekintetben hasonló az üzleti hírszerzéséhez, bár az utóbbi a törvényesség határait betartva értelemszerűen szűkebb. Mivel Magyarországon jóval többen kaptak kiképzést katonai, bűnügyi vagy polgári területen, mint üzletin, érdemes a párhuzamot elmélyíteni.

Az üzleti hírszerzés információforrásait az adatforrások szempontjából három kategóriába sorolhatjuk az alábbiak szerint.

Nyílt, közérdekű adatforrások:

- cégbírósági nyilvántartások, belföldi és külföldi adatbázisok;
- statisztikai hivatalok közleményei;
- közbeszerzési hirdetmények: kiírások és eredmények;
- adóhivatali közlemények;

[69] [3] 406. oldal

- bírósági döntvénytár (anonimizálva);
- alkotmánybírósági határozatok tára;
- versenyhivatali (GVH) határozatok;
- szabadalmi bejegyzések.

Nyílt, privát nyilvántartások, amelyek mindenki számára hozzáférhetőek:

- HR-hirdetések;
- közösségi helyek (pl.: fotók GPS-adatokkal);
- aukciók eredményei;
- műkincs-kereskedelemre vonatkozó megjelenések;
- tudományos publikációs helyek: konferenciák, folyóiratok, adatbázisok;
- elektronikus sajtó;
- vállalati honlapok;
- társadalmi, szakmai szervezetek, klubok, jótékonysági események stb. nyilvántartása (szabadkőműves páholyok, lovagrendek stb.).

Nem nyilvános, állami nyilvántartás, de arra felhatalmazottak – ügyvédek, bankok, brókerek stb. – segítségével megszerezhető információk:

- földhivatali nyilvántartás;
- bankközi adósnyilvántartás.

Bár az alábbiak határozottan nem tekinthetők az üzleti hírszerzés adatforrásainak, de az ipari kémkedésben nagyon is használatosak. Ezek a zárt[70] nyilvántartások. A hozzáférésük az erre törvényi úton felhatalmazott szervezetek és személyek kivételével illegális. Az alábbi példákban felsorolt és más adatbázisokhoz anonimizált formában tudományos vagy statisztikai céllal részben hozzá lehet férni:

- gépjárművek nyilvántartása;
- egészségügyi nyilvántartások (OEP, kórházi adatbázisok);
- bűnügyi nyilvántartások;
- utaslisták;
- híváslisták, cellainformációk a telekommunikációs szolgáltatóktól;
- banki információk;

[70] nem nyilvános, jogi és információbiztonsági értelemben védett

- biztosítási nyilvántartások.

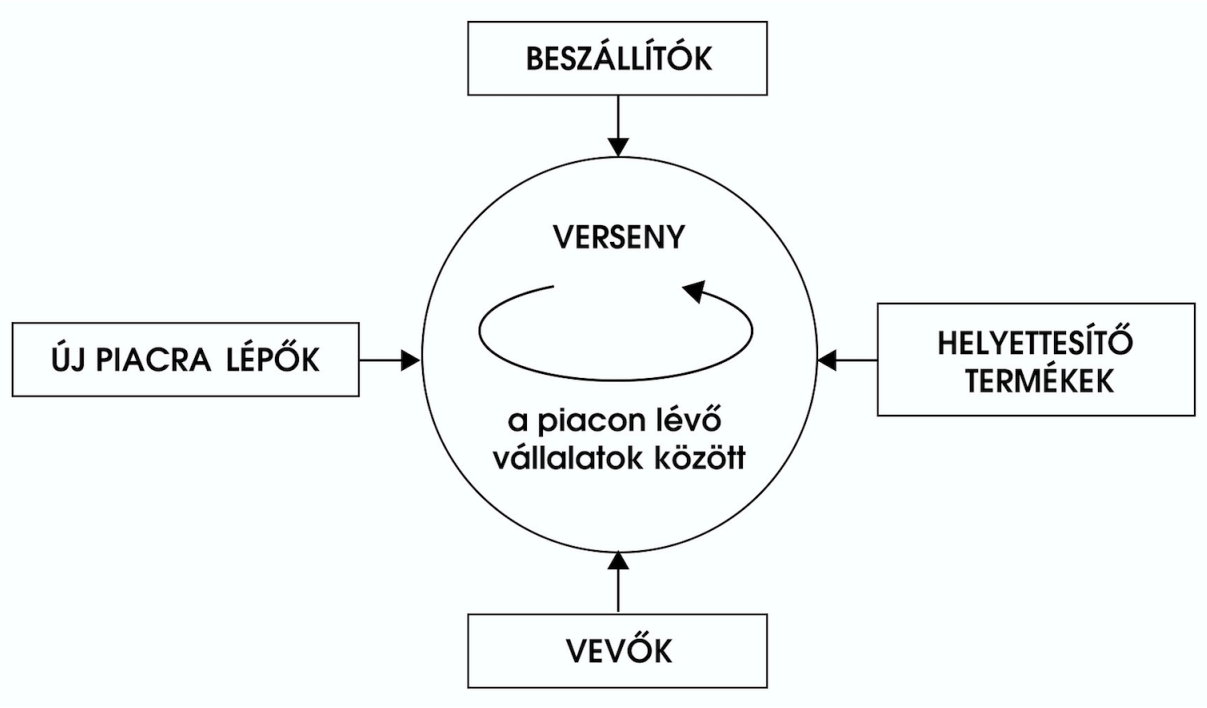
Az üzleti hírszerzés ciklusa lényegében megegyezik az OSINT-éval csak a tárgya és alkalmazott módszerei mások. Az üzleti hírszerzés célja a vevőkről, szállítókról, versenytársakról, pénzpiaci és jogi környezetről, technológiai fejlesztésekről, a HR-piacról törvényes úton információt szerezni, azt feldolgozni és eljuttatni a döntéshozókhoz. Az üzleti hírszerzés meghatározó versenyelőnyhelyzetébe juttatja az adott vállalatot.

Az üzleti hírszerzés az elmúlt két évtizedben önálló diszciplínává vált. Több helyütt egyetemi tantárgy, nagyobb országokban akadémiákat működtetnek nagy sikerrel. Ilyen a Fuld Gilad Herring Academy of Competitive Intelligence, Cambridge, Massachussetts vagy az ICI, Bad Nauheim, Hessen, Németország. Nagyobb multinacionális vállalatok önálló CI-osztályt működtetnek. Ilyenek a Siemens, Shell, IBM, DuPont, BP stb.

Egy értékelő-elemző célja egy-egy tulajdonosról, vezetőről, szakemberről, hozzátartozóról stb. minél teljesebb szakmai és személyiségprofilt készíteni, feltérképezni a személyes, társadalmi és gazdasági kapcsolatait. Az információk ismeretében megkísérli megjósolni a döntéseit, és támogatni az erre szakosodott kollégáit, kolléganőit ezen döntések befolyásolásában. A megszerzett adatok birtokában felrajzolja a vizsgált személy kapcsolati hálóját, amely alapján következtet a vizsgált személy lehetséges lépéseire.

Egy vállalat működését a környezete nyilvánvalóan befolyásolja. Környezet alatt érthető a város, az ország, a régió vagy akár maga a világgazdaság. Egy-egy beruházás előtt, vagy akár a folyamatos működés során, a vállalat méretétől függő mélységben a vezetés megvizsgálja mindazokat a tényezőket, amelyek hatással lehetnek a tevékenységre. A jelentés mérete terjedhet a prezentációtól a vaskos tanulmányokig. Egy multinacionális cég környezettanulmányának összes szempontját (*checklist*) lehetetlen itt felsorolni, de a főbb területek alább láthatóak: politikai környezet, jogi környezet, technológiai trendek, ágazati trendek a világgazdaságban, ágazati trendek a piaci környezetben, munkaerőpiaci trendek, oktatás színvonala és az infrastruktúra fejlettsége.

A fenti makrogazdasági környezet mellett elengedhetetlen a vállalatvezetés számára a vállalati környezet alapos vizsgálata. Michael Porter elmélete [133] a 22. ábra szerint 5 erő vizsgálatára épül. Ennek a részletes elemzése nem feladat itt, de szinte minden korszerű marketing-, stratégia- vagy versenytárselemzés-tankönyvben fellelhető.



22. ábra: a Porter-féle 5 erő.[71]

A cél mind a makro-, mind a mikrogazdasági elemzésnél a lehetséges változásokat jó előre látni a szükséges lépések megtétele végett. Az alább felsorolt szempontok szerint figyelni kell az adott cég beszállítóit, vevőit és versenytársait: árazási stratégiák, lehetséges ügyfelek, ügyfelek aktivitása, versenytársak üzletpolitikája, beszállítók aktivitása, stratégiai partnerek, marketingprogramok, vezető kereskedők nevei, K+F tervek, viszonteladói hálózat, megvásárlandó cégek, szabadalmi beadványok, elégedetlen ügyfelek, termékgyengeségek, peres esetek.

Az üzleti hírszerzés számos klasszikus módszerrel dolgozik. Klasszikus alatt az internet elterjedése előtti időszakot értem, amikor a keresőrendszerekre még nem lehetett támaszkodni. Kiemelten kezelem azokat a módszereket, amelyeket az internetes keresőrendszerek automatizáltak a 90-es évektől kezdve: a befolyásoló tényezők vizsgálata (STEEP/PEST), társadalmi környezet (*S*ocial), technológiai környezet (*T*echnological), gazdasági környezet (*E*conomic), környezetvédelmi (*E*cological), politikai, jogi környezet (*P*olitical) [134], [135].

A sorrendet az amerikai angolban annyira szeretett akronímák miatt tartottam meg. A kettő közötti különbség a második „E”, a környezetvédelem. Megjegyzendő, hogy az erőltetett akroníma miatt a politikai és jogi kategória egybeesik.

[71] Forrás: [133] 4. oldal

Az összes szempont szerint történik az elemzés a globális, az országos és a vállalati környezetben. A módszert alkalmazzák a Porter-féle 5 erő vizsgálatával kombináltan.

Üzleti vakfolt alatt azt értjük, amikor egy vagy több döntéshozó képtelen megszabadulni a saját prekoncepcióitól, és ez gátolja abban, hogy előre meglásson fontos fejleményeket. Klasszikus példa erre az Eastman Kodak esete. A vállalat vezetése nem tudta vagy nem akarta előre látni a CMOS (komplementer fém-oxid félvezető, *Complementary Metal-Oxyd Semiconductor*) technológia fejlődésének hatását a digitális fotózásra. A fényképek vegyi alapú rögzítése kevés kivételtől eltekintve kiment a divatból, a Kodak pedig csődbe ment [136]. Az üzleti vakfoltok problémakörével bővebben foglalkozom a 4.3.1 fejezetrészben.

A magyar jogrendszer a jelen írás keletkezésének idején nem bünteti az ipari kémkedést, mert ezt a fogalmat nem ismeri. Bűncselekménynek mindössze a vállalati titoktartási kötelezettség megszegését, illetve az üzleti titkok kiadását fogadja el [137].

Fontos kérdés, mennyit segít az állami hírszerzés a gazdaság civil szereplőinek. Ipari titkok megszerzése, beszerzési döntések befolyásolása, versenytárspiaci szereplők és szolgálatok tevékenységének feltárása és elhárítása. A nyílt szakirodalomban jószerével csak a legutóbbi témában találni utalást. Könyvtárnyi irodalom, cikk, riport jelent meg arról, hogy mások hogyan tevékenykednek egy adott ország érdekei ellen. Gondoljunk a nem olyan régen Németországból kiutasított amerikai diplomatára, a franciák által az Air France New York–Párizs járatának első osztályú üléseibe szerelt lehallgatóberendezésekre, amelyek segítettek felkészíteni a francia partnert az amerikaiakkal való másnapi tárgyalásra stb. A Snowden-ügy, amely ebből a szempontból is megérdemel egy önálló elemzést, megvilágította nemcsak az ipari kémkedés módszereit, hanem alkalmazásának még a szakmát is meglepő széleskörűségét. A példák végtelenéből érzékelhető a gazdasági diplomácia és a gazdasági hírszerzés összefonódása.

Az természetesen nyilvánvaló, hogy minden ország foglalkozik gazdasági hírszerzéssel. Az igazi kérdés, hogy a megszerzett információt továbbadja-e a gazdaság szereplőinek, és ha igen – nyilván igen –, melyiknek és milyen mértékben. Milyen politikát, gyakorlatot folytat egy-egy ország a saját vállalataival, már amennyire a „saját” fogalma egyáltalán meghatározható. Erre nyílt irodalmat nem találtam. Melyik saját országbeli versenytárscégnek segítenek, melyiknek nem? Csak a nagyoknak, stratégiai fontosságúaknak, vagy kisebbeknek is? Melyik szektornak igen, és melyiknek nem? Melyik vállalat számít jónak és melyik nem? És milyen típusú információt adnak át? Stratégiai elemzéseket és becsléseket, vagy egy adott projekthez kapcsolódó taktikai információkat? És végül, de korántsem utolsósorban: ki dönti el ezeket a

kérdéseket?[72] Kevésbé rejtélyes, de említésre méltó téma, hogy az állami szervek milyen és mennyi feladatot helyeznek ki privát vállalatokhoz.

3.4.2. Az információkeresés alkalmazása a gazdasági hírszerzésben

Az üzleti hírszerzésben alkalmazott keresőrendszerek kétféle technológiát alkalmaznak. A strukturált adatok feldolgozására szánt adatbányászatot és a strukturálatlan szövegek feldolgozására szánt szövegbányászatot. A jelen részben a szövegbányászati alkalmazásokra összpontosítok, de lejjebb röviden ismertetek egy adatbányászati megoldást is, amely az interneten fellelhető azonos termékek árait hasonlítja össze.

A keresőrendszerek működhetnek külső adatforrásokra (*web search*) és belső adatforrásokra (ECS). Bár a technológia alapjai hasonlóak, funkcionalitásuk sok szempontból eltérő. A leglényegesebb különbség a felhasználó szempontjából az adatbiztonság kezelése. Egy belső kereső esetében a hozzáférés szigorúan szabályozott, ez a hírszerzési munkakörnyezetben elengedhetetlen. Erről többet a Humán és biztonság fejezetpontban írok.

Keresési módszertan szempontjából lényegtelen, hogy a feldolgozandó szövegtestet legális vagy illegális úton és milyen technológiával szerezték, maga a nyersszöveg-feldolgozása már legális körülmények között zajlik egy védett vagy legalábbis privát objektumban, adatközpontban, irodában. Csak az adatgyűjtés az, ami törvényes szempontból kritikus, végrehajtás szempontjából kockázatos lehet. Megjegyzendő, hogy bizonyos információk feldolgozását és tárolását is korlátozhatják a személyiség jogokra vonatkozó törvények és rendelkezések.

A sok – akár külső, akár belső – forrásból származó adat feldolgozott formában rögzül már egységes, strukturált (adatbázis) formában. Ez analóg a feljebb megismert OSINT-nál alkalmazott katonai adatfúziós központ működésével. Az értékelő-elemző számára az egységes formátum alkalmas az egységes lekérdezések indítására. Elengedhetetlen a formalizált fogalomrendszer szerepének kiemelése. A hatékony, jelentésalapú, azaz szemantikus keresőrendszerek mögött komplex számítógépes nyelvészeti, tudásreprezentációs háttér áll.

[72] Idevonatkozó gondolatok az amerikai haditengerészettől: Intelligence support for the private sector, http://web.nps.navy.mil/~relooney/4141_56.htm (a korábbi letöltés papíron megmaradt, de a megadott helyen az anyag már nem található, talán a téma kényes volta miatt távolították el).

A felhasználói felület (*user interface*, UI), valamint a képernyők közötti navigációs technika elsőrendű fontosságú abból a szempontból, hogy egy – adott esetben új, szokatlan, egyesek számára félelmet keltő – alkalmazást milyen gyorsan fogad el a felhasználói kör. Gyakori igény, különösen kapcsolati hálók esetében, ezek grafikus ábrázolása. Itt a gráfok csúcsai a személyek, cégek stb., élei pedig maguk a kapcsolatok. Ezek sokfélék lehetnek, „felhívta", „leányvállalata", „rokona" stb. Az értékelő-elemző számára a vizualizálás a keresési munka eredményének egyik természetes megjelenítése, a csoportosítások, indirekt kapcsolatok felismerésének eszköze.

A közösségi honlapok figyelése (*SocMed analysis*) ma már az üzleti – és nem csak az üzleti – hírszerzés alapvető műveleti ága. A SocMed üzleti logikája arra épül, hogy emberek, cégek, szervezetek önként tárnak fel minél többet az életükről, érdeklődésükről, működésükről, kapcsolataikról. Ez valódi aranybánya az érdeklődőnek, aki erre kifejlesztett keresők segítségével nemcsak naprakész adatokhoz jut, hanem azokat feldolgozva rejtett összefüggéseket is feltárhat. A Facebook, a LinkedIn, számtalan blog, fórum, csevegőfelület vizsgálata a modern üzleti hírszerzés egyik zászlóshajója lett.

A 2.2 fejezetrészben tárgyalt szentimentelemzés lényege, hogy szemantikus eszközökkel mérjük a szerző véleményét az általa alkotott szöveg alapján egy bizonyos témában. E mérési eredményeket idősorba helyezhetjük, és így az alkalmazás képes a trendek figyelésére. A vizsgált dokumentumok ritkábban egy szervezetben belül keletkeznek (elsősorban e-mailek, de wikik, dokumentumok, fórumok), leggyakrabban pedig a weben (közösségi helyek, mint Facebook, LinkedIn, fórumok, blogok stb.). A hangulatfigyelés klasszikus alkalmazási területei például az egyes árukról és szolgáltatásokról kialakult vevői vélemény mérése, politikusokat, pártokat övező hangulat elemzése, a tőzsde mutatóinak előrejelzése a pszichológiai jellemzők alapján.

Az ár-összehasonlító alkalmazások lényege, hogy azonos termékekről szóló – elsősorban árra vonatkozó – adatokat talál meg az interneten, és ezeket összehasonlítja. A probléma megoldása, hacsak nincs az ár-összehasonlító honlapok és az eladó között együttműködés, amely megteremti a strukturált adatátadás feltételeit, két szempontból sem triviális. Egyrészt meg kell találni a termék forgalmazóit. Ezt vagy egy óriáskereső futtatásával érheti el, ha nem ismeri a forgalmazók URL-jét, vagy direkt bányászhatja a honlapokat, ha ismeri az URL-jüket. Másrészt egy kereskedelmi honlap kusza, reklámokkal, bannerekkel megtűzdelt, erősen strukturálatlan szerkezetéből ki kell tudni emelni az adott termékre vonatkozó árinformációt.

A vállalati hírnév vizsgálata (*corporate reputation analysis*, CRA) módszerének a lényege, hogy a fent megismert technikai eszközökkel folyamatosan figyelemmel kísérik egy adott vállalat imázsát, képét a közfelfogásban. Vizsgálják a véleményeket, megjelenéseket a közösségi helyeken, a hirdetéseket stb. Ehhez hozzáfűzhetnek metrikát is a hangulatelemzés módszertanával.

A releváns események idősorba állítása (*event and timeline analysis*, E&T) megvilágíthatja a megszokottól eltérő megnyilvánulásokat, és következtetésekre, esetleges problémák, nehézségek észlelésére vagy ezek előrejelzésére ad lehetőséget.

A több mint 700 nyilvántartott módszer [134], illetve rendszer között egyre nagyobb szerepet játszanak a vállalati belső információs forrásokra támaszkodó, és az interneten, illetve fizetős adatbázisokban fellelhető adatokat begyűjtő, kiszűrő és kiértékelő szemantikus keresőalkalmazások. A figyelt témák a szakterülettől függően változatosak, de szinte mindegyik felöleli a politikai és jogi környezet üzleti szempontból releváns mozgásait, a főbb technológiai trendeket, a versenytársak aktivitását, a vevők helyzetét, kiemelve azok pénzügyi stabilitását, esetlegeses befektetéseit vagy tőkebevonását, a HR-piac eseményeit, a beszállítók jellemzőit, különös tekintettel azokra, amelyektől az adott vállalat erősen függ, és a kapcsolat megváltozása kritikus lehet.

A sajtófigyelés a médiafigyelés azon részterülete, amely az írott elektronikus megjelenésekre keres rá. Az elektronikus megjelenés alatt gépbe olvasható formátumú szöveganyagokat értünk (pl. Word, PDF, HTML stb.). Tipikus források a honlapok, portálok, fizetős szöveges adatbázisok, pl. a hírügynökségekéi. A még nagyon gyakori, de már éppen a kereshetőség miatt lassan kikopó képi formátumok és papíralapú megjelenések OCR-es feldolgozás, beszkennelés, illetve az azt megelőző beszkennelés nélkül természetesen nem kereshetőek. Tipikus figyelési mód, amelyet szinte minden ilyen alkalmazás használ, az RSS (kb. bőséges helyi összefoglaló, *Rich Site Summary*) csatorna beolvasása. A médiafigyelés fontos területei még a rádióadások (hangalapú) és tv-adások (hang- és képalapú) figyelése. A hangból S2T technikával többé-kevésbé érthető szöveget lehet generálni, amely már szövegbányászati eszközökkel feljavítható és kereshető.

3.4.3. Az üzleti hirszerzés fejlődésének várható jövőbeli irányai

Technológiai szempontból az üzleti hírszerzés várható irányai megítélésem szerint nem különböznek a nyílt forrású hírszerzés várható irányaitól.

3.5. Összefoglalás, részkövetkeztetések

A fejezetben foglaltak alapján **megállapítható, hogy mind a három választott alkalmazási területen** (védelmi szféra, közigazgatás és a gazdasági élet) **van olyan jelentős tevékenység, amely az információkeresésre épül.** A nemzetbiztonsági szolgálatok és rendvédelmi szervek esetében az információkeresés kiemelkedő alkalmazási területe a nyílt forrású hírszerzés, de hasonlóan jelentős szerepet játszik a szervezeti belső keresőrendszerek és a fúziós központok esetében is. **A témakörhöz kapcsolódó legfontosabb következtetés, hogy a nyílt forrású hírszerzésen belül a hagyományos információszerzés mellett elsődlegessé, dominánssá vált az informatikai eszközökkel támogatott információkeresés.**

Bár a tevékenység régi, a nyílt forrású keresés fogalma általában sem rendelkezik egyértelmű definícióval. Az én értelmezésem szerint a korábbival ellentétben **a keresés fő kritériuma nem a jogszerűség, hanem a nyilvános elérhetőség** (lásd 1. tudományos eredmény). A modern angolszász szakirodalom ezt az értelmezést használja.

A szakirodalom alapján **összegeztem a nyílt forrású hírszerzés** és az ezt támogató nyílt forrású keresés **előnyeit és hátrányait**. Megállapítottam, hogy **a hírszerzés öt fázisából az információkeresés háromban közvetlen, kettőben pedig közvetett szerepet játszik**.

Szakmai tapasztalataim, következtetéseim alapján a nyílt forrású keresés kiemelt jelentőségű alkalmazási területei közé a következők tartoznak: a közösségi média figyelése; szereplők azonosítása nyelvi sajátosságok alapján; álcázott üzenetváltás felismerése internetes fórumokon; bűn- és terrorcselekmények előrejelzése; deviáns viselkedésre való hajlam detektálása; célcsoport vélemény- és hangulatfigyelése.

A nyílt forrású keresés jövőbeni fejlődési irányai meglátásaim szerint a következők lesznek: a mobil platform szerepének növekedése; a közösségi média mint információforrás szerepének további növekedése; az ember-gép interfész új megoldásainak elterjedése; a beszéd-szöveg átalakítás, a gépi fordítás lehetőségeinek jelentős bővülése; a gépi tanulásra épülő megoldások elterjedése, különösen a prediktív analitika területén.

A közigazgatásban az információkeresés kiemelt jelentőségű alkalmazási területe a jogi keresés, de jelentős szerepet játszik a közigazgatási nyilvántartásokban, egyedi ügyekben keletkezett közigazgatási ügyiratokban, illetve normatív ügyiratokban (belső utasítások, eljárásrendek) történő keresés is. A jogi keresés forrásait a törvényeket tartalmazó jogtár, a bírósági döntvénytár, a központi közigazgatás szerveinek, a minisztériumoknak, országos hatáskörű szerveknek a rendeletei, valamint a Kúria állásfoglalásai képezik.

Megállapítottam, hogy a közszféra információinak további felhasználásáról szóló EU-irányelv előírásainak érvényesülése alapvetően befolyásolja a közérdekű adatokban történő keresés lehetőségeit. **Feltártam a 2015. évi XCVI. törvény hiányosságait, és javaslatokat tettem annak módosítására. Így a közérdekű adatokat könnyen elérhetővé és gépi úton könnyen olvashatóvá kell tenni; szankcionálni kell a törvény be nem tartását; és meg kell alkotni a törvény végrehajtási rendeleteit** (lásd 2. tudományos eredmény).

A jogi információkeresés alapvető forrásait a magyar és nemzetközi jogi adatbázisok képezik. Ezek egy része – az alapvető jogszabályokat tartalmazók – ingyenesen elérhető, de léteznek térítés ellenében használható szakosított adatbázisok is. Következtetéseim szerint az ingyenesen elérhető adatbázisok kereshetősége, keresési szolgáltatásaik kényelmessége, hatékonysága sok igényt nem elégít ki. Ezért alakultak ki jobb szolgáltatást nyújtó, térítéses keresőszolgáltatások ezekre épülően is. A hozzáadott értéket nyújtó keresőszolgáltatások alapvető szolgáltatásai közé tartoznak: keresés több forrásból; több szempontú keresés; többnyelvű keresés; metaadat alapján történő keresés; taxonómiaalapú keresés; relevancia szerinti besorolás.

Kutatásaim eredményeként fogalmazható meg az a következtetés, hogy **a hatékony jogi információkeresésnek elengedhetetlen eszköze a megfelelő jogi taxonómia, a jogi ontológia.** Ezek teszik lehetővé a joganyagok visszakeresését tartalom és jelentésalapú összefüggések alapján. Ilyen jogi taxonómiák, ontológiák elérhetőek ingyenesen vagy térítés ellenében. Ezek alkalmazása – bár számos új fejlesztés jelent meg – a 2000-es évek eleje óta nem bővült, a szemantikus technológiák napjainkig nem nyertek tért a jogi keresőrendszerekben.

Kutatásaim során megállapítottam, hogy a gazdasági hírszerzés, és az ehhez kapcsolódó fogalmak keverednek, tartalmuk nem egyértelmű. Következtetéseim szerint a gazdasági hírszerzés három alapvető összetevőjét a legális, nyílt forrású üzleti hírszerzés, az illegális ipari kémkedés, illetve a kettő között elhelyezkedő féllegális gazdasági információszerzés képezi. **Újradefináltam a gazdasági hírszerzés fogalmát**, illetve annak nyílt forrású változatát, és az üzleti hírszerzés fogalmát (lásd 3. tudományos eredmény). Megállapítottam, hogy a védelmi szféra nyílt forrású hírszerzéséhez hasonlóan 1990 után robbanásszerűen megnőtt az internetes keresés szerepe az üzleti hírszerzésben.

Az üzleti hírszerzés információforrásai közé tartoznak: nyílt, közérdekű források, nyílt, magánnyilvántartások; nem nyilvános, de arra felhatalmazottak által elérhető állami

nyilvántartások. Ezekenkívül vannak zárt nyilvántartások, amelyekhez legálisan nem lehet hozzáférni, de tudományos vagy statisztikai céllal anonimizált formában igen.

Következtetéseim szerint az információkeresés gazdasági életben történő alkalmazásának alapvető területei, formái: a közösségi média figyelése; árukkal, szolgáltatásokkal kapcsolatos vevői vélemény- és hangulatfigyelés; ár-összehasonlítás; vállalati hírnév vizsgálata; esemény- és trendfigyelés, üzleti előrejelzések és a sajtófigyelés.

4. FEJEZET: AZ INFORMÁCIÓKERESÉS KERETEI, ÉRTÉKELÉSE

4.1. Bevezető gondolatok, a fejezet tartalma, célja

Az előző fejezetekben áttekintettem az információkeresés elméleti alapjait, néhány alkalmazott technológiát és három gyakorlati alkalmazását. Nem lenne teljes a kép anélkül, hogy ne vizsgálnám az információkeresést az ember alkotta mikro- és makrokörnyezetben.

A következőkben megvizsgálom a magyar jogrendszert az információkeresés szempontjából. Kísérletezem egy rövid kitekintéssel az Egyesült Királyságban végbemenő változásokra a releváns jogi környezetben, különösen szem előtt tartva négy fő vizsgálati szempontot. Nincs kétségem afelől, hogy ez a fejezetrész vitát generálhat, amit elismerésnek tekintenék. Célom a konstruktív vita indítása, nem a végső igazság kijelentése.

Megvizsgálom azokat a tényezőket, amelyek az információkeresés sikerét vagy kudarcát okozhatják egy nagyon is humán környezetben. Álszerénység nélkül állíthatom, hogy működő szervezetekkel szerzett évtizedes ipari tapasztalatom kellő alapot biztosít ezeknek a megállapításoknak a megtételére. A fejezetrészben némi áttekintést adok az információkeresés biztonsági aspektusának megítélésére is. Fontos ez már csak azért is, mert az ismeretek hiánya okozta – olykor babonás – félelem nem egy sikeres projekt elvetéséhez vezetett.

Végül áttekintem az információkeresés mint ügyviteli folyamat gazdaságossági szempontjait. Egy korlátozott, de használható modellt mutatok be a beruházás megtérülésének kiszámítására. A kvantitatív mutatók mellett áttekintem a kvalitatív mutatókat is az információ-visszakeresés értékelési módszertanából.

4.2. Az információkeresés jogi keretei

4.2.1. A legfontosabb technológiák, amelyeket a jogi környezet korlátozhat

Az alábbiakban összegzem a nemzetbiztonsági szolgálatok és rendvédelmi szervek korszerű információkereső rendszerekkel szemben támasztott igényeit. Ennek forrásai – a hivatkozott szakirodalmon túl – személyes interjúk[73] és kiegészítések vezető beosztású szakértőkkel, valamint a sokéves személyes tapasztalatom a nemzetbiztonsági szolgálatok és rendvédelmi szervek részére szolgáltató szoftverfejlesztő cég vezetőjeként. A nemzetbiztonsági szolgálatok

[73] Bár érthető, hogy a tudományos hitelességet erősítené, ha itt a forrásokra név szerint hivatkozhatnék, de erre sajnos egyetlen nemzetbiztonsági vagy rendvédelmi interjúalany sem adott engedélyt.

és rendvédelmi szervek természetes elvárása, hogy az adathoz minél teljes körűbben, minél gyorsabban, lehetőleg valós időben, minél kevesebb korlátozással, és a legtöbb nyilvánvaló és rejtett összefüggést feltárva jussanak hozzá. Ezek a szempontok az információéhség irányában aligha igényelnek magyarázatot, a hozzáférés korlátai annál inkább.

A továbbiakban bemutatom azokat a technológiákat, amelyek az utóbbi másfél évtizedben olyan mértékben fejlődtek ki vagy tovább, hogy azok használata a nemzetbiztonsági szolgálatok és rendvédelmi szervek számára elkerülhetetlenné vált. Az elkerülhetetlenségnek több oka van. Az információtechnológiai infrastruktúra óriási mértékben fejlődött. Az adat- és szövegbányászati technológiák robbanásszerű fejlődését látjuk. Ilyen terület a gépi tanulási technológiák ipari alkalmazhatósága, különösen a neurális hálók prediktív képességeinek megbízhatósága, a videó- és képfelismerés 98% feletti megbízhatósága, a nem relációs adatbázis-technológiák elterjedése, vagy a szemantikus nyelvi technológiák tökéletesedése.

A terrorizmus és a szervezett bűnözés köreiben széleskörűen elterjedtek a modern információtechnológiai alkalmazások, így a rejtőzködés és az álcázás, a titkosítás; a közösségi média használata drog-, ember-, fegyverkereskedelemre, pedofil tartalmak közlésére, tömegpusztító eszközök csempészésére, terrorista csoportok toborzására vagy a terrorcselekményre magányosan készülő személyek (*lone wolfs*) magukat igazoló előzetes megnyilatkozásaira. Személyiségi jogokkal keveset törődő autokratikus rezsimek tömegesen és szervezetten álhíreket gyártanak a közvélemény befolyásolására, vagy katonai és hírszerző műveleteket támogatnak információtechnológiai eszközökkel.
A nemzetbiztonsági szolgálatok és rendvédelmi szervek tényleges igényei nem ismernek földrajzi határokat. Természetesen különböző országok más és más formában szabályozzák azokat a jogi kereteket, amelyeken belül a nemzetbiztonsági szolgálatoknak és rendvédelmi szerveknek működniük szabad. Hangsúlyozni kell, hogy nincs legjobb vagy egyetlen eszköz a nemzetbiztonsági szolgálatok és rendvédelmi szervek kezében. A működéshez a rendelkezésre álló módszerek kombinációjára van szükség [138].

A hazai és nemzetközi szakirodalom tanulmányozása alapján arra a következtetésre jutottam, hogy a kritikus technológiák, amelyek alkalmazásának szabadsága nélkülözhetetlen a nemzetbiztonsági szolgálatok és rendvédelmi szervek működésében, a következők:

- **az információkeresés célhoz kötöttségének feloldása;**
- **az adatbázisok korlátozások nélküli összekapcsolhatósága;**

- **az adatállomány időkorlát nélküli tárolhatósága;**
- **a profilalkotás szabadsága.**

A legfontosabb igény a korlátozások nélküli információkeresés, azaz a célhoz kötöttség feloldása. Bár **a tömeges információkeresés** (*bulk search*) nem egy világosan definiált, mindenki által ugyanúgy értelmezett fogalom[74], a **lényege, hogy a keresés nem egy konkrét célszemélyre, szervezetre történik, hanem valamely kritériumok szerint egy nagy adathalmazt gyűjt be legtöbbször külső forrásokból (is), majd azt feldolgozva az így keletkezett közbenső adatállományból** (pl. indexfájlból) **szűri ki lekérdezésekkel a kívánt konkrét tartalmat**. A tömeges információkeresés „ellentettje" a célirányos információkeresés (*targeted search*). A – törvényes keretek között gyakorolt – tömeges információkeresés nem azonos a korlátozás nélküli megfigyeléssel (*mass surveillance*), ezeket a személyiségi jogokért küzdők hajlamosak összekeverni. Az egyik ugyanis engedélyhez kötött, szoros felügyelet mellett zajlik, és a végrehajtó szervezetek számonkérhetőek, míg a másik esetében mindez nem igaz. A számonkérhetőségről lejjebb még bővebben írok.

Rá kell világítsak a törvény elvárásainak és a mindennapi gyakorlatnak az ellenétére. A tömeges információkeresés igénye alapvető a nemzetbiztonsági szolgálatok és rendvédelmi szervek részéről. Nagyon sok esetben nem tudják, hogy pontosan mit és hol keressenek. Egyszerű példa a terroristagyanús személyek figyelése. Több mindent tudnak róluk, de sok mindent nem. A személyi profiljuk alapján figyelhetők a közösségi médiában, az utasnyilvántartás, rendszámmegjelenés az autópályákon, telefon- és e-mail-forgalom, internetes lekérdezések formájában stb. A profil és a fellelt viselkedésminták valamekkora valószínűséggel történő egyezése esetén további műveletek határozhatók meg. Ezt kizárólag célirányos kereséssel nem lehet megvalósítani, mert a cél mint olyan, pontosan ismeretlen. A célhoz kötöttség feloldásának igénye értelemszerűen azokra az információkeresésekre vonatkozik, amelyeket a törvény tilt. A célhoz kötöttség mellőzése természetes igény még a legegyszerűbb bűnüldözési feladatok esetében is. A határőr a dolgok természetéből fakadóan minden autót megnéz a sorban, és így szűri ki a gyanúsakat. A kábítószer-kereső kutya is minden csomagot megszagol, és jelzi a drogtartalmúakat. A posta is minden levelet megvizsgál, és kiszűri az

[74] Nagy mennyiségű adattömeg engedélyezett gyűjtése, amelyhez a hozzáférés megkülönböztető szűrés (pl.: különleges azonosító, kiválasztó feltétel stb.) nélkül történik (*authorized collection of large quantities of ... data which ... is acquired without the use of discriminants [e.g. specific identifiers, selection terms, etc.]*) [139]

anthraxtartalmúakat. Ez az általános megközelítés természetesen nem vonatkozik arra az esetre, ha „fülest” kaptak, és pontosan tudják, kit vagy mit kell kiszűrni.

Az adatbázisok összekapcsolása a kért adatok más számára történő rendelkezésre bocsátását jelenti. Technikai értelemben különböző adatbázisok adataiból egy új adatbázist hozunk létre. [140]. Megjegyzendő, hogy a később tárgyalandó kritikus 15/1991. (IV. 13.) Alkotmánybírósági határozat [141] (továbbiakban: „1991-es AB határozat”) nem tesz különbséget külső (civil) és hatósági (többek között nemzetbiztonsági szolgálatok és rendvédelmi szervek által kezelt) adatbázis között, így a megfogalmazás vonatkozik mindkettőre és a hibrid rendszerekre is. Az izolált silókban való külön-külön keresés hatékonysága összemérhetetlenül kisebb, mint egy, az összefüggéseket automatikusan feltáró keresőrendszer esetében. Egy klasszikus példa erre az utazási ügynökségek ügyfél-nyilvántartásának összevezetése az adóhivatal jövedelem-nyilvántartásával. Korábban az APEH végzett ilyen összevetéseket, amelyek földrengésszerű eredményekhez vezettek [142]. A gyakorlatot hamar leállították, feltehetően a civil jogvédők és az utazási irodák tiltakozása miatt.

Példákkal világítok rá arra, hogy a nemzetbiztonsági szolgálatok és rendvédelmi szervek közelmúlt történelmének számos ismert példája bizonyítja, hogy a különböző adatsilókban meglévő adatok gyors összevezetésével tragédiákat kerülhettek volna el. Közismert, hogy a 9/11 vizsgálata kimutatta, a megelőzéshez szükséges részadatok a különböző nemzetbiztonsági szolgálatok és rendvédelmi szervek adatbázisaiban utólag mind fellelhetőknek bizonyultak, csak éppen a rejtett összefüggéseket nem tárta fel időben humán erőforrás. Ugyanezt láthattuk a párizsi merénylet utólagos vizsgálatakor. Hasonló helyzetet jelentett a sajtó a brüsszeli robbantásokkal összefüggésben. A terrorista lakcíme a rendőrség kezében volt, csak nem került a nemzetbiztonsági szolgálatok és rendvédelmi szervek által feldolgozásra. A feladatot egy mesterséges intelligenciára épülő robot automatikusan és időben elvégezte volna. Az adatintegráció hiánya hátráltatja az összefüggések hatékony feltárását, vagy akár lehetetlenné is teszi azt. Természetesen egyik rendszer sem tud minden bűncselekményt előre jelezni, nyilván csak azokat, amelyekre a tanítóadatok alapján valamiféle utalást talál. Ha nem volt korábban adat arra, hogy ki tanul pilótatanfolyamon felszállást, de leszállást nem, akkor erre sem a robot, sem pedig ember nem figyelhetett.

Rá kell mutassak arra, hogy a szemantikus keresésen messze túlnyúló, felügyelt tanításra épülő mesterségesintelligencia-alkalmazások, nemcsak kirívó eseteket tárnak fel, hanem a korábbi ismeretek birtokában – hasonló jellegű cselekmények előkészületére vonatkozó adatok

feltárásával – előrejelzésekre is képesek. Ennek pedig elengedhetetlen feltétele minél több korábbi adat tárolása és korlátlanul kereshetővé tétele, ugyanis ezek a tanítóadatok az alapjai a gépi tanulásnak. Megállapítom: **a korábbi tapasztalati adatok törlése erősen behatárolja a korszerű prediktív módszerek alkalmazását, valamint az esetek gyors, utólagos vizsgálatát és megoldását, pedig a nemzetbiztonsági szolgálatok és rendvédelmi szervek hatékonyságának lemaradásához vezet.** A gépi tanulásnak elengedhetetlen feltételei a tanító- és tesztadatok. Ezek pedig értelemszerűen a korábbi esetekből kerülhetnek ki, amelyek eltűnnek, ha törlik azokat. Azt pedig már az előző részben is kimutattam, hogy a modern, intelligens információkeresés nem nélkülözheti a mesterséges intelligencia vívmányait. Itt nem elemzem, hogy mekkora veszteséget jelenthet pl. egy bűnügyi nyomozásnál, ha korábbi esetek személyes adatait a vonatkozó jogszabályok szerint meg kell semmisíteni, majd azokat újra be kell gyűjteni, ha ez egyáltalán lehetséges.

Külön kiemelem itt is a nemzetbiztonsági szolgálatok és rendvédelmi szervek kezelésében működő **fúziós központ**ok jelentőségét. Ezek a mesterséges intelligencia eszköztárával működő szöveg- és adatbankok tárolják és dolgozzák fel a különböző forrásokból beérkező adatállományokat. Prediktív analitikai módszerekkel jelzik előre a várható eseményeket. Korszerű bűnüldözési vagy nemzetbiztonsági intézmény számára a fúziós központ használata megkerülhetetlen. Ilyen feladatot kapott a TIBEK (Terrorelhárítási Információs és Bűnügyi Elemző Központ). A releváns törvények áttekintése után meg kell állapítsam, az adattárolásra és -törlésre vonatkozó jogszabályok nem egyenszilárdságúak a nemzetbiztonsági szolgálatokra és rendvédelmi szervekre nézve.

Ahhoz, hogy egy személyt beazonosítsanak, legyen az bűnöző, kém, terrorista stb., **személyiségprofil**t kell róla alkotni. E nélkül a hozzá kapcsolódó információkeresés lehetetlen. A tárolt személyiségprofilt mint mintát így a mesterséges intelligencia eszköztárával lehet összevetni az előforduló mintákkal beazonosítás végett.

Összefoglalásképpen kijelenthetjük, hogy **a modern technika adta kihívásokra és lehetőségekre a nemzetbiztonsági szolgálatok és rendvédelmi szervek képtelenek lesznek megfelelő választ adni a szükséges adatkezelési módszerek és technológiák hiányában. Ezen anomália feloldására javasolom, hogy egyrészt a 15/1991. (IV.13.) Alkotmánybírósági határozatot vonják vissza és dolgozzák át.**

4.2.2. Az információkeresés magyar jogi környezete

Az alábbiakban elemzem az információkeresés magyar jogi környezetét. Hangsúlyoznom kell, hogy a kutatás a jogi korlátok miatt természetesen csak a nyilvánosan elérhető jogszabályokra és egyes nyílt közjogi szervezetszabályozó eszközök vizsgálatára korlátozódott. Mivel sem a rendvédelmi szervek, sem pedig a nemzetbiztonsági szolgálatok belső használatra készült, illetve minősített adatot tartalmazó belső szabályozóikhoz értelemszerűen nem volt hozzáférésem, az idevonatkozó megállapításaimat dr. Séllei Márton főhadnagy úrral, az AH (Alkotmányvédelmi Hivatal) jogászával tettem [143], akinek a munkáját dr. Lévay Szabolcs, az AH vezető jogásza felügyelte. Mint feljebb, egy lábjegyzetben említettem, a többi interjúalany nem engedte meg a név szerinti hivatkozást. A kijelentéseim érvényességét korlátozhatja, hogy a helyzetből fakadóan nem látom át, hogy a mindennapi munkájában melyik szervezet mennyire tartja be ténylegesen a jogi korlátokat.

Az információkeresés magyar jogi környezetének vizsgálata komplex feladat. Az Alkotmány, illetve később az Magyarország Alaptörvénye [144] (továbbiakban: „Alaptörvény") az információs önrendelkezési jogot határozza meg mint alapvető emberi jogot (VI. cikk. (2) és (3) bekezdései). Mindkettő összhangban van a Római Egyezménnyel [145] (8. cikk 2.) és az ENSZ Közgyűlése XXI. ülésszakán elfogadott Polgári és Politikai Jogok Nemzetközi Egyezségokmányával [146]. Az idevonatkozó törvények a személyes adathoz fűződő jogokat innen vezetik le. A Római Egyezmény joganyagát az Európai Emberi Jogok Bírósága (EJEB) viszi át a gyakorlatba.

Természetesen információkeresés bármilyen forrásra történhet, így bizalmasra, minősítettre stb. is. Ezen forrásokra való keresést más és más jogi keretek szabályozzák, pl. banktitokra, üzleti titokra stb. vonatkoznak. A jelen részletes vizsgálat tárgya az információkeresés a személyes adatokra, mert a technológiai haladás és a jogi kötöttségek antagonizmusát itt látom a legmeghatározóbbnak. Mielőtt azonban a törvényeket vizsgálnám, meg kell említeni az 1991-es AB határozatot, amely az adatkezelésre a mai napig korlátozó hatással van.

- A célhoz kötöttséget előírja, hogy „*személyes adatot feldolgozni csak pontosan meghatározott és jogszerű célra szabad."* [II. 5. bek.] Ez a tömeges, konkrét céllal nem rendelkező információkeresést („halászást") a (értelemszerűen) zárt adatbázisokban kizárja. Nyílt forrású keresésre nem tér ki. Ugyancsak nem tesz különbséget civil és rendvédelmi, nemzetbiztonsági felhasználók között.
- Adatbázisok összekapcsolását csak akkor engedi, „*ha minden egyes adat vonatkozásában az adattovábbítást megengedő összes feltétel teljesült."* [II. 6. bek.]

Tehát az egyes állami vagy civil adatsilókban fellelhető adatok közötti összefüggések feltárását erősen korlátozza.

- Az adatok cél nélküli tárolását tiltja: „*meghatározott cél nélküli, készletre, előre nem meghatározott jövőbeni felhasználásra való adatgyűjtés és -tárolás alkotmányellenes”*. [II. 5. bek.] Így a modern gépi tanulórendszerek tanítóadatait csak erősen korlátozott mértékben teszi hozzáférhetővé.
- A profilalkotást tiltja, mert az „*nagy valószínűséggel torz is.”* [III. 2.4 pont] Nem állítja senki a gépi profilalkotás tévedhetetlenségét, de mai neurális hálókra, ill. mélytanulásra (*deep learning*) épülő algoritmusokkal ez meghaladhatja a 90%-os pontosságot. Nagy luxus erről az eszközről a nemzetbiztonsági szolgálatok és rendvédelmi szervek munkájában lemondani.
- Indoklásként nem lehet nem idézni a teljes bekezdést. *„Az államigazgatás hatékonysága különösen nem lehet ilyen érdek, mert nem bizonyítható, hogy az információs önrendelkezési jog súlyos sérelmével járó adatfeldolgozási mód az egyedül lehetséges útja a hatékonyan működő államigazgatásnak.”* [III. 6. bek.]

A jelen írásban azt tárom fel, hogy az elmúlt 26 év alatt ez **a megközelítés anakronisztikussá vált**, **és az információs önrendelkezési jog nagyon is fontos tiszteletben tartásához más eszköztárat kell találni**. **A terrorizmus és a szervezett bűnözés világában az állam nem engedheti meg magának, hogy ilyen mértékben kösse meg a saját kezét.**

Ahhoz, hogy eljussunk a nemzetbiztonsági szolgálatok és rendvédelmi szervek működését korlátozó jogszabályok kérdésköréhez, bizonyos alapfogalmakat át kell látnunk. Ismertnek tételezem fel a 2011. évi CXII. törvény az információs önrendelkezési jogról és az információszabadságról [147] (a továbbiakban: „Infotörvény”) fogalmait, amelyeket az összes releváns törvény és rendelkezés használ vagy utal rá. Ezek a következők: érintett, személyes adat, különleges adat, bűnügyi személyes adat, közérdekű adat, közérdekből nyilvános adat, adatkezelő, adatkezelés, adattovábbítás, adattörlés, adatmegsemmisítés, adatfeldolgozás, adatbiztonság.

A törvényi szabályozás nem egységesített (Infotv., Rendőrségi tv. és Nbtv.) (ld. lejjebb): „a szétszórt magyar szabályozás nemcsak az eszközök felsorolása tekintetében gyengíti a szabályozás egyenszilárdságát, de az egyéb követelmények (pl. előreláthatóság, arányosság, fokozatosság) rendszerszerű biztosítását is megnehezíti.” [148]

A vizsgált törvények információkeresési szempontból csak részben relevánsak. A külső forrásból történő (itt nem a személyes adatok tárolására hivatott állami és civil adatsilókról van

szó, amelyek összekötése nem értelmezhető külső forrásnak) adatszerzésre (a hírszerzési ciklus második eleme, *data capture*) vonatkozó korlátozások (külső[75] és belső[76] engedélyeztetés) a vizsgált módszereket nem érintik. Ugyanis az vagy nem információkeresés, hanem titkos információgyűjtés, vagy titkos adatszerzés, ami nem tárgya a vizsgálódásnak, vagy pedig nyílt forrású[77] keresés, amit a jogszabályok nem korlátoznak. Az adatkezelésre, -feldolgozásra, -törlésre és -megsemmisítésre (mint a hírszerzési ciklus harmadik, negyedik és ötödik elemére, adatfeldolgozás, *data processing*, értékelés, elemzés, *analysis and evaluation*, terítés, *dissemination*) viszont nagyon is vonatkoznak a vizsgált törvények. A vizsgálat szempontjából értelemszerűen irreleváns az is, hogy az adatszerzés a nyomozás elrendelése előtti fázisra vonatkozik (titkos információgyűjtés, TIGY) vagy a nyomozás elrendelése utáni fázisra (titkos adatszerzés, TASZ), ezen módszerek bevezetése egyébként is vitatott. Itt feltétlenül utalnom kell a Szabó és Vissy kontra Magyarország ügyre [149], amelynek ítéletében az EJEB részletesen kibontja, hogy kinek és hogyan kell engedélyt kiadni[78] megfigyelésre.

A jogi környezet vizsgálata, alkotmánybírósági értelmezése sokkal inkább a titkos információgyűjtés vagy titkos adatszerzés vonatkozásában történt meg, mint az információkeresés és ezen belül adatkezelés vonatkozásában. A 2/2007 (I.24.) AB határozat AB határozat taxatíve felsorolja [150] azokat a – külső engedélyhez kötött és nem kötött – titkos információgyűjtő vagy titkos adatszerző eszközöket, amelyekre a megkötések vonatkoznak. Az információkeresés négy vizsgált területe csak marginálisan érintett ezekben.

A személyes adat kezelésére vonatkozó öt alapelvet alább sorolom fel. Ezeket [148] a következőképpen foglalja össze az Infotv., a 95/46/EK adatvédelmi irányelv [151] 6. cikk alapján. A későbbi megjelent GDPR [77] is részben releváns, bár az a rendvédelmi és nemzetbiztonsági adatkezelést nem tárgyalja, sőt **a Preambulum 16 paragrafusa**

---

[75] Bírói vagy igazságügyi-miniszteri engedélyhez kötött, a magánéletbe jelentősen beavatkozó tevékenység: lehallgatás, levél felbontása, „online házkutatás”, hekkelés stb.

[76] Az illetékes nemzetbiztonsági szolgálatok és rendvédelmi szervek felső vezetője által engedélyezhető, a magánéletbe kevésbe beavatkozó tevékenység: követés, híváslista-lekérés stb.

[77] Ld. a Nyílt forrású keresés paragrafust

[78] Tudomásom szerint az ítélet végrehajtása még nem fejeződött be.

egyértelműen kijelenti, hogy a nemzetbiztonsági szolgálatokra a GDPR nem vonatkozik. A jogkorlátozás akkor alkotmányos, ha

- arra egy másik alapjog vagy szabadság védelme érdekében kerül sor (célhoz kötöttség),
- csak ha a korlátozás elkerülhetetlenül szükséges (szükségesség),
- az alkalmazni kívánt eszköz a cél elérésére alkalmas (alkalmasság),
- az adott cél eléréséhez szükséges legenyhébb eszközt alkalmazza (fokozatosság),
- az elérni kívánt céllal a jogkorlátozás arányos (arányosság) [152].

Nem lehet egységes megállapításokat tenni a nemzetbiztonsági szolgálatok és rendvédelmi szervek adatkezelésére vonatkozóan, mert a bűnüldöző szervezetekre más törvények vonatkoznak, mint a nemzetbiztonsági szolgálatokra. Ez a distinkció végigkövethető az ágazati szabályokon. Ezért a két szervezettípust külön vizsgálom.

A nemzetbiztonsági szolgálatok és rendvédelmi szervek működési körének bővítése az információs önrendelkezési jog mint korlát törvényi tágítására épül. A nemzetbiztonsági szolgálatok és rendvédelmi szervek ennek az alapjogi rendelkezésnek megfelelően gyakorolnak különleges jogokat az adatkezelés területén. **Tehát az Alaptörvény, az 1991-es AB határozat és az Infotörvény megtiltja az információkeresés szempontjából kulcsfontosságú műveleteket, de a következő paragrafus mégis különleges engedélyekre ad felhatalmazást**.

"Személyes adat akkor kezelhető, ha… a nemzetbiztonság, a bűncselekmények megelőzése vagy üldözése érdekében vagy honvédelmi érdekből törvény elrendeli…" [Infotörvény, 5.§ (2) bekezdés b) pontja].

A közigazgatási adatbázisokban és hatósági nyilvántartásokban történő keresést korlátozza a célhoz kötöttség elve [153].

A nemzetbiztonsági szolgálatok tevékenységét szabályozó 1995. évi CXXV. törvény a nemzetbiztonsági szolgálatokról [154] (továbbiakban: „Nbtv.") 40.§ (1) és (1a)] bekezdése szabályozza az adatkérés célirányosságát. Az egyedi adatkérés céljának megjelölése és dokumentálása kötelező. Ugyanakkor a 2016/681/EK parlamenti és tanács irányelv [155] szerint működő utasadat-információs rendszerből már terrorizmusgyanús személyek után történhet adatlekérés (*pull*) és adattovábbítás (*push*) alapú keresés a TIBEK-ben, ami nyilvánvalóan nem egy konkrét személy célirányos keresését jelenti, hanem a teljes utasállomány végigpásztázását, és egy megadott kritériumrendszer alapján a gyanús elemek

kiszűrését az Nbtv., 52/I.§-a szerint. Ez értelemszerűen az 1991-es AB határozat szigorának felpuhulását, és a korszerű technológiákhoz való alkalmazkodást jelenti.

Figyelemmel a korábban a titkos információgyűjtés és titkos adatszerzés közötti különbségek bemutatására is a szolgálatok – szemben a nyomozóhatóságok által alkalmazott titkos adatszerzés által beszerzett adatokkal – nem feltétlenül használják fel a titkos információgyűjtés során keletkezett információkat a büntetőeljárás során. Ilyen esetben, ha nemzetbiztonsági célú titkos információgyűjtés kerül alkalmazásra (vagyis nem egy konkrét bűncselekménnyel kapcsolatos cél kerül meghatározásra), nem sérül a célhoz kötöttség elve.

A rendőrség adatkezelésében a célhoz kötöttséget általában az Infotörvény szabályozza. Itt azonban különbséget kell tenni a különböző adatok kezelése között. Ugyanis nem mindegy, hogy titkos adatszerzésből származik-e az adat, és azt egy konkrét nyomozás során akarja a nyomozóhatóság felhasználni, vagy pusztán egy szabálysértési ügyben keletkezik adat. Az Infotörvény kizárólag keretet nyújt, de az ágazati jogszabály, az 1994. évi XXXIV. törvény a rendőrségről [156] (továbbiakban: Rtv.) 84. § (1) szabályozza az egyedi adatátadás dokumentáltságát. A továbbiakban a Rtv. részletesen szabályozza a bűnügyi, rendészeti stb. adatkezelést, amelynek a vizsgálata a négy fő szempontból kevésbé lényeges, mert minden adattovábbítás egyedi és dokumentált.

Az adatbázisok összekapcsolására vonatkozó jogi előírások, korlátozások az adatbázisokban foglalt személyes adatok védelméhez kapcsolódnak. A jogi korlátozások alapvető rendeltetése, hogy biztosítsák a személyes adatok védelmét az összekapcsolás eredményeként bekövetkező jogosulatlan megismerés, illetve felhasználás ellen. Emellett meghatározzák a jogszerű összekapcsolás azon feltételeit, amelyek fenntartják a megfelelő szintű adatvédelmet az Infotörvény 7.§ (4.) szerint. Az Nbtv. 39. § (1.) bekezdés d) pontja teremti meg a lehetőséget a szolgálatok részére egyes adatbázisok összekapcsolására, amely a gyakorlatban azt jelenti, hogy meghatározott esetekben a szolgálatok által működtetett belső adatbázisokba eltárolni szánt adatok beszerzésének egy része külső adatkezelő által működtetett, külső adatbázisokból közvetlen elektronikus adatkapcsolat kiépítésével történik meg. Az Nbtv. 40. § (1) bekezdése határozza meg azokat az eseteket, ahol ez az adatkapcsolat kialakítható. Ez a közvetlen adatkérés célhoz kötött és dokumentált. Tehát tömeges keresésre nem alkalmas. Az Nbtv. 47. § (1) bekezdésében rögzítettek lehetőséget nyújtanak arra, hogy a nemzetbiztonsági szolgálatok adatbázisait néhány kivételtől eltekintve egymással, illetve más állami rendszerekkel összekössék.

Magyarországon adatfúziós központi funkciót lényegében a TIBEK lát el. A 2016. évi LXIX. törvény [157] (továbbiakban: Antiterrortv.) hozta életre Magyarország legújabb nemzetbiztonsági szolgálatát, a TIBEK-et, amely feladatellátásában lényegesen eltér a többi szolgálattól. A TIBEK-et lényegében a többi, operatív feladatokat ellátó szolgálatok által begyűjtött adatok és információk elemzésére és szintetizálására hozták létre a SZEBKK (Szervezett Bűnözés elleni Koordinációs Központ) utódszervezeteként. Az Nbtv. 47. § (4.) lehetőséget biztosít a TIBEK-nek az állandó adatkapcsolatra az Infotörvény keretei között, de másik irányban nem. Ezek a 2016-os változtatások érezhető enyhülést jelentenek a korábbi szigorral szemben. Az Antiterrortv. 8/A § kibővítette az Rtv.-t a TIBEK-re vonatkozó részekkel, amelyek új elemeket tartalmaznak az információkeresés szempontjából. Megengedi az adattovábbítást és adatkérést más nemzetbiztonsági szolgálatok és rendvédelmi szervekkel egyedi, tehát nem tömeges, automatikus formában.

A Rtv. 88.§ és 91/H§ részletesen szabályozza az adattovábbítást a különböző adatsilók között, illetve az adatbázisok korlátozás nélküli összekapcsolását és az azokban történő szabad keresést. A rendőrség a törvényben meghatározott bűnüldözési feladatai teljesítése érdekében, az adott bűncselekmény felderítése és nyomozása során, bűnüldözési adatállományait a közigazgatási adatállományaival, illetőleg más bűnüldözési adatállománnyal összekapcsolva egyedi adatkezelést végezhet. Az összekapcsolást az adott üggyel kapcsolatos eljárás befejezése után meg kell szüntetni. Az összekapcsolás során keletkezett olyan új adatokat, amelyek a büntetőeljárás során nem kerülnek felhasználásra, haladéktalanul törölni kell.

A rendőrség külföldi társszervekkel való kapcsolattartását hivatott részlegének, a NEBEK-nek az adatkezeléséről a 2002. évi LIV. törvény [158] (a bűnüldöző szervek nemzetközi együttműködéséről), valamint a Rtv. 86.§ és 87.§ rendelkezik. Az adatcsere csak célhoz kötött, egyedi lekérdezésekkel oldható meg. A nemzeti adatbázisok más országok általi közvetlen elérését és az adatcserét az ún. prümi irányelv [159] szabályozza.

A 2011. évi CLXIII. törvény az ügyészségről [160] (továbbiakban: „Ütv.") 32.§ szerint a legfőbb ügyész engedélyezheti az „ügyészség által kezelt személyes adatoknak más adatkezelésekkel történő összekapcsolását", vagyis – feltehetőleg – az adatbázisok összekötését.

Az adatok megőrzésére (*data retention*) vonatkozó leglényegesebb korlátozás arról szól, hogy mennyi idő múlva kell törölni a tárolt adatokat. **A nemzetbiztonsági szolgálatok a 2016-ban módosított törvény értelmében** tulajdonképpen **bármilyen adatot megtarthatnak, amely**

„szükséges" (!). Ez mérlegelési jogkört biztosít a szolgálat vezetőjének a célhoz kötöttség, a szükségesség és arányosság elvei mentén.

Az összekapcsolást a konkrét nemzetbiztonsági feladat elvégzését követően meg kell szüntetni, az összekapcsolás során keletkezett adatállományt az eljárás befejezését követően törölni kell. Nem kell törölni az összekapcsolás eredményeként keletkezett azon adatot, amely a nemzetbiztonsági szolgálatok feladatainak ellátásához szükséges. (Antiterrortv, 47.§ 2.)

Az adatok akármely, a rendőrség által végzett feladat (bűnüldözés, lakcímnyilvántartás) ellátása érdekében történő vezetése, tárolása időben jogilag korlátozott. Az adatok köre ugyanúgy szabályozott, mint a folyamatos frissítésükre vonatkozó tevékenység. A kapcsolódó feladat megszűntekor az adatokat archiválni, törölni kell. Előbbi egy passzív állapotban tárolja az adatokat, utóbbi már a konkrét személyhez köthetőségét is tiltja, csak úgynevezett metaadatok maradhatnak, ezek a vizsgálat szempontjából értékelhető adatkört már nem tartalmaznak.

A 2016. évi CXVI. törvény az egyes belügyi tárgyú törvények módosításáról [161] lényegesen kibővítette az adattárolásra vonatkozó korlátokat. A bűncselekmény típusától függően 30-50 évig megengedi az adatok tárolását. Ez nyilvánvalóan már nem tükrözi az AB határozat szellemét.

Az adattárolással kapcsolatos magyar szabályozás egyik legfontosabb ágazati szabályozója a 2003. évi C. törvény az elektronikus hírközlésről [162], amely a távközlési szolgáltatást végző szervezetek részére bizonyos adattárolási időintervallumokat állapít meg. Ennek lényege, hogy minden adat vagy hangalapú kommunikáció metaadatait 1 évig meg kell őrizni, és a hatóságok rendelkezésére kell bocsátani.

A releváns törvényekben nem találtam a profilalkotásra vonatkozó korlátozásokat. A GDPR is csak az üzleti célú profilalkotást korlátozza [163].

Az Antiterrortv 52/I.§ megengedi az utasadat-nyilvántartó rendszerekben történő információkeresést. A teljes utasállomány végigpásztázása csak előre elkészített profilok alapján lehetséges. Bár a módszer minősített volta miatt nincs rálátásom a konkrét keresési kritériumokra, de feltételezhető, hogy egy Khartoumból berepülő, különböző hatósági adatbázisokban nyilvántartott személy nagyobb figyelmet kap, mint egy Tel-Avivból érkező nagymama a kisunokájával.

Fontos kérdés a nemzetbiztonsági szolgálatok által megszerzett információk büntetőeljárásban történő felhasználása. Itt különösen a titkos információgyűjtés során megszerzett információk büntetőeljárásban történő felhasználása és alkalmazása okozhat gondot. Erre vonatkozóan

iránymutatást az 1998. évi XIX. törvény a büntetőeljárásról [164] (továbbiakban: „Be.”) 206/A. § (1) bekezdése nyújt, amely kimondja, hogy a külső – tehát bírói vagy miniszteri – engedélyhez kötött titkos információgyűjtés eredménye büntetőeljárásban bizonyítékként csak akkor használható fel, ha a nyomozóhatóság által a már megindított nyomozás során alkalmazandó titkos adatszerzés feltételei is fennállnak, továbbá, ha a titkos információgyűjtést alkalmazandó szerv a kérdéses információ beszerzését követően nyomban elrendelte a nyomozást, vagy eleget tett feljelentési kötelezettségének. Természetesen ki kell emelni, hogy ezen szabály alkalmazása kizárólag a megjelölt adatforrás alkalmazása során merül fel, az OSINT útján megszerzett információk tekintetében nem. Megjegyzendő, hogy a nemzetbiztonsági szolgálatok nemcsak a Be. szerinti bűncselekményhez köthető nyomozás során gyűjtenek adatot, hanem megelőző jelleggel is.

Függetlenül attól, hogy az információkeresés eredménye felhasználható-e bizonyítékként vagy sem, az Nbtv. 41. § alapján a szolgálatok alapjában véve minden elektronikus forráshoz az Infotörvény feljebb tárgyalt korlátain belül hozzáférnek.

A rendőrség titkos adatszerzését a Be. 200–206. § szabályozzák. Ez az adatszerzési mód lényegében egy, már korábban elkövetett bűncselekmény után alkalmazható adatszerzési mód, amely a korábban elkövetett bűncselekménnyel kapcsolatos nyomozás során az eljáró nyomozó nemzetbiztonsági szolgálatok és rendvédelmi szervek által lefolytatott, meghatározott „jellemzők” vizsgálatára és rekonstruálására terjed ki. Ilyen „jellemző” többek között a bűncselekmény elkövetője kilétének, tartózkodási helyének megállapítása, az elkövető elfogása, valamint a releváns bizonyítási eszköz megszerzése, vagyis a titkos adatszerzés eredményének ezeket a célokat kell megvalósítania.

Fentiekkel szemben a titkos információgyűjtés a nyomozó jogkörrel nem rendelkező nemzetbiztonsági szolgálatok (és bizonyos esetekben a TEK) privilégiuma. Fontos különbség, hogy a titkos információgyűjtés nem feltétlenül egy konkrét, befejezett bűncselekményhez kapcsolódik, hanem a meghatározott bűncselekmény elkövetését megelőző felderítő és adott esetben elhárító munkához. Az Nbtv. által az adott nemzetbiztonsági szolgálat konkrét feladatkörének megfelelően a titkos információgyűjtés kevésbé irányzott, mint a titkos adatszerzés, hiszen nem egy, már elkövetett konkrét bűncselekmény nyomozásához szolgáltat adatokat, hanem több, potenciális bűncselekmény elkövetésének megelőzésére is szolgál.

A Rtv. 84. § taxative szabályozza a felhasználható adatforrásokat.

Az Ütv. 36.§ 3 (a) szerint az ügyészség gyakorlatilag bármilyen közigazgatási, bűnügyi és NAV-os adatforráshoz hozzáférhet egyedi, dokumentált adatkérés során.

A nyílt forrású keresést a magyar törvények nem korlátozzák. Itt megismétlem, amit az OSINT fejezetrészben már állítottam, nem tudok arról, hogy a „nyílt forrás" fogalmát a magyar jog meghatározta volna. Meg kell állapítsam, **a nyílt forrású keresés fogalma még szakmai körökben sem egyértelmű**, **pontos jogi definíciója pedig a magyar jogban egyáltalán nem létezik.** Meg kell viszont említeni, hogy a nagy mennyiségű, nyílt forrású, egészében áttekinthetetlen adattömegből készült „desztillátum" tartalmazhat már ebben a formában nagyon kényes, személyre vonatkozó információt. A nyílt források természetesen lehetnek ingyenesek vagy fizetősek. Nemzetbiztonsági szolgálatok és rendvédelmi szervek tisztjeivel történt személyes beszélgetések során nehezményezték, hogy nem egyértelmű a szabályozás sem, hogy mihez kell már bírói engedély, mihez nem. Az OSINT fő forrásai is folyamatosan változnak. Sokszor a szolgáltatók maguktól vagy nyomásra szigorítanak az adatokhoz való hozzáférésen (pl. a Facebook változásai), így egy korábbi forrás elapadhat. Mások eltűnhetnek (pl. IWIW), újak jelennek meg, amelyek népszerűek lesznek. Ismereteim szerint a nyílt forrású adatok keresése nem célhoz kötötten kell történjen. Más szavakkal, nem tiltott a tömeges információkeresés („halászás") (*bulk search*). A bűnüldözésnek része lehet a nyilvános adatbázisok ellenőrzése (pl. vagyon elleni bűncselekményt felderítő nemzetbiztonsági szolgálatok és rendvédelmi szervek ellenőrzik az orgazda-tevékenység eredményeként árult termékeket hirdető nyilvános honlapokat).

6. táblázat: a civil szervezetek, nemzetbiztonsági szolgálatok és rendvédelmi szervek jogosultságai. Forrás: a szerző.

	Civil szolgáltatók[79]	Közigazgatási alkalmazott	Ügyészség	Rendőrség, TEK, NVSZ	NAV	Nemzetbiztonsági szolgálatok	TIBEK
Nyílt forrás	igen	igen	igen	igen	igen	igen	igen
Profilalkotás	nem	nem	igen	igen	igen	Igen	igen
Fúziós központ	nem	nem	nem	nem	nem	nem	igen

[79] Távközlési szolgáltatást végző szervezetek, biztosítók, pénzintézetek

Adatmegőrzés	igen	igen	igen	igen	igen	igen	igen
Célhozkötöttség	igen	igen	igen	igen	igen	igen	igen
Adatbázisok összekapcsolása	igen	igen	igen	igen	igen	igen	igen

A 2000-es évektől egyre elterjedtebb, és ma már a hétköznapi és üzleti életet ezer szállal átszövő felhőalapú kommunikációs alkalmazások (levelezőrendszerek, üzenőalkalmazások) a nemzetbiztonsági szolgálatok és rendvédelmi szervek számára létfontosságú adatforrások. 2001. évi CVIII. törvény az elektronikus kereskedelmi szolgáltatások, valamint az információs társadalommal összefüggő szolgáltatások egyes kérdéseiről [165] 3/B, 13/B § szabályozza a hozzáférést. Ugyanakkor nyilvánvaló, hogy ezekre nem a magyar jog vonatkozik, így a hozzáférés erősen korlátozott, lassú, nehézkes, és a legtöbb esetben a szolgáltató a kért adatokat nem adja át. További nehézséget jelentenek a felhőalapú egyéb szolgáltatások (játékok stb.), amelyek üzemeltetőitől a gyors adatszerzés gyakorlatilag lehetetlen [166]. A felhőalapú rendszereknél is keményebb feladat a dark weben történő adatok elérhetősége. Ezek egészen különleges technológiákat igényelnek, és csak korlátozott sikerrel kecsegtetnek.

4.2.3. Az Anderson-jelentés

Sok víz lefolyt a Dunán az 1991-es alkotmánybírósági állásfoglalás óta: a személyes adatok kezelésének szigora valamelyest hozzáidomult a technológia fejlődéséhez, és a fejlett világban is elmozdultak a szabadság és biztonság erőinek frontvonalai. Feljebb felsoroltam a hat országot, amelyről tudjuk, hogy a törvényi keretek újragondolása folyamatban van. Az Egyesült Királyságban keletkezett az a tanulmány [167], amely feladatául tűzte ki a probléma vizsgálatát mindkét szemszögből. Az anyag politikai súlyát fémjelezi, hogy a David Anderson Q.C.[80] vezette szerzői négyest maga a királynő nevezte ki. Másik három tagja titkosszolgálati, jogi, illetve technikai háttérrel rendelkezett. A tanulmány a parlament által hevesen vitatott titkosszolgálati törvény tervezetéhez (*Investigatory Powers Bill*)[81] készült, és természetesen nemcsak a szigorúan vett információkeresésről szól, hanem a tömeges megfigyelés más

[80]Queen's Counsellor, a királynőnek közvetlenül alárendelt jogi méltóság.

[81] A polgári jogi aktivisták által gúnyosan Snooper's Charter néven kritizált.

területeit is lefedi. Így a tömeges lehallgatást[82] (*bulk interception*), a számítógépekbe, telefonokba és hálózatokba történő tömeges behatolást (hekkelés, *bulk equipment interference, EI*), a telekommunikációs szolgáltatók adatainak feldolgozását (*bulk acquisition of communications data*) és a tömeges adatkezelést (*bulk personal datasets, BPD*).

A bizottság mindkét oldallal együtt dolgozott. A három titkosszolgálat az MI5 (Katonai Hírszerzés 5 [osztály], *Military Intelligence 5*, Az Egyesült Királyság [a nevével ellentétben ma már] polgári elhárító szolgálata), MI6 (Katonai Hírszerzés 6 [osztály], Military Intelligence 6, az Egyesült Királyság [a nevével ellentétben ma már] polgári hírszerzése), a GCHQ számos szakértőjével, de a szabadságjogi szervezetekkel is (többek között Liberty, Big Brother Watch, Open Rights Group, Affiliates of Open Democracy, Amnesty International, Electronic Frontier Foundation, Centre for Investigative Journalism, World Wide Web Foundation) együtt dolgozott.

A tanulmány elemezi a négy fenti módszer munkafolyamatait és technikai részleteit, amelyeket részleteiben nem taglalok. Az információkeresés szempontjából az első három annyiban releváns, amennyiben a megszerzett adatokat nyelvi eszközökkel elemezhető szövegtestekké alakítják, és tömeges kiértékelésnek vetik alá. A negyedik módszer felhasználási területe ettől eltér. Itt elsősorban személyiségprofilok mint minták gyűjtéséről és tárolásáról van szó, amelyeket a mesterségesintelligencia-alapú keresőrobotok dolgoznak fel.

A tanulmány hosszan taglalja a fenti négy módszer előnyeit, amelyeket a megkérdezett nemzetbiztonsági szakemberek méltatnak. A kifejtés sajnos pont a lényeget mellőzi a tények minősített volta miatt. Ami az információkeresés szempontjából nyomatékosan kiemelendő, az az adatok dúsítására (*enrichment*) történő utalás.[83] Ezekből az utalásokból teljesen **egyértelműen kitűnik, hogy a UK titkosszolgálatai a meglévő adatokat tanító, illetve tesztadatként gépi tanulásos klaszterezésre alapuló információkeresésre használják**.

Ami a legfontosabb, Anderson csoportja óvatosan és korlátozásokkal, de megállapítja, hogy a személyes adatokban való tömeges keresés kisebb beavatkozást jelent az emberek magánszférájába, mint a csak célhoz kötöttségen alapuló keresés. A kissé paradoxnak tűnő kijelentés mögött tapasztalati logika áll. Ugyanis a tömeges keresésre épülő előszűrés sokkal

[82] Itt lehallgatás alatt a hang-, szöveg-, kép- és videótartalmú üzenetek befogása, feldolgozása és kiértékelése értendő. Üzenet alatt a tartalom és a metaadatok egyaránt értendők.

[83] [167] 3.31., 4.7., 8.5., 8.35., 9.14., Annex 4., Annex 6., Annex 7.

kevésbé erőszakos behatolás a magánszférába, mint a célszemély alapos bevizsgálása. Másképpen fogalmazva: a tömeges információkeresést alkalmazva csak azokat a személyeket teszik ki alaposabb vizsgálatnak – képletesen fogalmazva a T betű függőleges szára mentén –, akiket az előszűrés – a T betű vízszintes szára mentén – már kiemelt. Az Anderson-jelentés alapján **bizonyítottnak látom, hogy a magyar jogrendszerben a tömeges keresés tiltását és a célhoz kötött keresés kizárólagosságát meg kell szüntetni**.

4.2.4. A nemzetbiztonsági szolgálatok és rendvédelmi szervek információkereséshez kapcsolódó számonkérhetőségének jogi keretei

A szabadság és a biztonság antagonizmusa évezredek óta foglalkoztatja gondolkodókat, jogászokat, államférfiakat (-nőket újabban)[84], írókat, filozófusokat, politikusokat. Bár a kérdés, hogy „Ki őrzi az őrzőket?" eredetileg Decimus Junius Juvenalis 6. Szatírájának (Szatíra a nők ellen) [168] címében szerepel – és nem a ma használatos kontextusban, hanem a hanyatló női erényekkel összefüggésben –, a felelős állam ellenőrzése Platon Az állam című művétől Dan Brown A digitális erőd című művéig újra és újra felmerülő téma.

A témakört szűkítve utalni kell a számtalan példa közül a 27 éve feltárt Echelon [169] rendszerre, a nemrég nyilvánosságra került Snowden-aktákra [170] vagy éppen az adok-kapokra az amerikai NSA (Nemzetbiztonsági Ügynökség, *National Security Agency*) vagy a GCHQ (Kormányzati Kommunikációs Központ/Főhadiszállás, *Government Communication Headquarters*) adathalászási jogkörével kapcsolatban. Könyvtárnyi irodalom foglalkozik a személyi szabadságjokat féltő, egy digitális orwelli állapotot vizionáló érvelőkről az egyik oldalon, míg a bűnüldözés, a hírszerzés, a terror- és kémelhárítás által a legmodernebb technológiák mellett lobbizókról a másik oldalon. A személyiségi jogokat féltő társadalmi szervezetek, illetve a terrorizmus és a szervezett bűnözés által alkalmazott egyre fejlettebb információtechnológiai módszerek ellen küzdő nemzetbiztonsági szolgálatok és rendvédelmi szervek fokozódó egymásnak feszülése a jogrendszer újragondolását eredményezi több fejlett

[84] Erre a magyar nyelv sajnos még nem honosított kifejezést a gyakran használt angol-amerikai *stateswoman*nek, a német *Staatsfrau*nak vagy a lényegesen ritkábban előforduló francia *madame d'etat*-nak megfelelően. A honatya női megfelelője, a honanya inkább törvényhozói, mint végrehajtói szerepre utal.

demokráciában, így az USA-ban, az EU-n belül pedig az Egyesült Királyságban[85], Németországban, Franciaországban, Hollandiában és Svédországban.[86]

Nem tartom feladatomnak a vitában állást foglalni. Véleményem szerint mindkét félnek igaza van egy bizonyos pontig, mert a nagyon is létező túlkapások nemcsak átlépik a törvényesség határait, hanem aláássák az állami nemzetbiztonsági szolgálatok és rendvédelmi szervekbe vetett bizalmat, de a szükséges eszközök megvonása a szolgálatoktól nehezíti az eredményességüket, hatékonyságukat. Péterfalvi Attila a disszertációjában hosszan érvel – Hans Peter Bullt, az első német adatvédelmi biztost idézve [171] – a szabadság és az átláthatóság együttes értelmezése mellett [172]. Megkísérlem viszont megvizsgálni mindkét oldal érveit, és kísérletet teszek az antagonizmus feloldására olyan módszerek megvilágításával, amelyekkel kitágulhat a nemzetbiztonsági szolgálatok és rendvédelmi szervek műveleti tartománya anélkül, hogy a személyiségi jogok csorbulnának. Mivel a témában érvelők általában vagy az egyik, vagy a másik oldalt képviselik erőteljesebben, igyekeztem kiegyensúlyozottan mérlegelni.

A vizsgálatom leszűkül az információkeresés területére. Természetesen a szabadság versus biztonság kérdése az információkeresés területénél jóval szélesebb területet fed le. Jelen írásban nem foglalkozom a lehallgatás, hekkelés és az adatszerzés más, bírósági vagy igazságügyi miniszteri engedélyhez kötött módszereivel, amelyek egyébként kínálják magukat egy szélesebb körű kutatás tárgyául.

Külön figyelmet érdemel a juvenalisi kérdésre keresendő válasz: milyen jogi és technikai eszközökkel ellenőrzik a hatóságok tevékenységét a vizsgált technikai területeken? El kell fogadjuk, hogy egyrészt szükség van a lehető leghatékonyabb eszköztárra a nemzetbiztonsági szolgálatok és rendvédelmi szervek részére, másrészt szélsőséges esetben a demokrácia alapjait rengethetné meg egy ilyen hatékony digitális arzenál kontroll nélküli használata. Fontosnak tartom megemlíteni, hogy még a Snowden-akták sem tárták fel a személyiségi jogok nagyobb horderejű, tényleges megsértését. A legsúlyosabb esetek arról szóltak, hogy NSA-alkalmazottak féltékenységtől vezérelve rádolgoztak a barátnőikre [173]. Ezeket nevezte el a szakma némi öniróniával LOVEINT-nek[87]. Bár e túlkapások is kétségtelenül törvénysértők,

[85] Jelen pillanatban az Egyesül Királyság az EU tagállama (2016.10.16.)

[86] [167] 20–24. oldal

[87] vö. HUMINT, SIGINT, OSINT stb.

súlyuk nem összemérhető a terrortámadások kockázatával. A kiegyensúlyozott válasz érdekében meg kell vizsgálni az ellenőrző mechanizmusokat és azok hatékonyságát. Nincsenek illúzióim, nem hiszek egy tökéletes megoldásban.

A következőkben rendszerezem az elérhető magyar és angolszász szakirodalomból a nemzetbiztonsági szolgálatok és rendvédelmi szervek ellenőrizhetőségének, számonkérhetőségének (*accountability*) a jogi, szervezeti és technikai kereteit. A probléma rendkívül kényes, sőt feltehetőleg mindkét oldalról erősen átpolitizált. A kérdés kiemelt érzékenységét mi sem illusztrálja jobban, mint a Belügyminisztérium 2010 óta végbement két próbálkozása egy fúziós központ létrehozatalára. Tudott, hogy a NIBEK (Nemzeti Információs és Bűnügyi Elemző Központ) törvénytervezet 2011-ben nem kapott elegendő támogatást a parlamentben, és a 2016-ban megszavazott TIBEK-törvénynek is egy erősen visszafogott verziója került elfogadásra. A kritikus pont minden esetben az adatbázisok összekapcsolhatóságának mikéntje volt, másképpen fogalmazva, kinek mit és mennyire automatikusan vagy kontrolláltan kell továbbítania a TIBEK fúziós központjába.

A nemzetbiztonsági szolgálatok és rendvédelmi szervek számonkérhetőségéről bőséges irodalom áll rendelkezésre. Ennek teljes áttekintése és elemzése messze szétfeszítené a rendelkezésre álló keretet. A szakirodalom természetesen nem hegyezi ki a számonkérhetőséget az információkeresésre, hanem a nemzetbiztonsági szolgálatok és rendvédelmi szervek tevékenységét egészében vizsgálja [174], [175] és [176]. **A számonkérhetőség problémáját nagyon egyszerűen meg lehet fogalmazni: hogyan gyakoroljanak demokratikus ellenőrzést olyan szervezetek felett, amelyeknek a működése az állam szempontjából létfontosságú, miközben a működésük eredendően titkos.** Az antagonizmus egyértelmű: az ellenőrző mechanizmusok szeretnének minél többet megtudni, miközben a nemzetbiztonsági szolgálatok és rendvédelmi szervek minél kevésbé akarnak kitárulkozni. Mindkét oldal érvelése legitim. Hogyan felügyeljenek intézményeket, ha nem látnak bele a működésükbe? És hogyan működjenek, ha minden, esetleges kiszivárgás a műveletek sikerét, hosszú munkával felépített struktúrák fennmaradását vagy akár emberéleteket kockáztat? Különösen igaz ez olyan műveletekre, amelyek a végrehajtási területen illegálisak. Az ellenőrzés alapja a fékek és ellensúlyok megteremtése. A demokráciákban alapvetően két megoldás alakult ki a probléma megoldására. Egyrészt egyensúlyozni a jogokat és kötelességeket a nemzetbiztonsági szolgálatok és rendvédelmi szervek és az azokat ellenőrző intézmények között. Másrészt ellenőrző mechanizmusokat kiépíteni a végrehajtó szervezeteken kívül.

Az alapvető veszély országtól függetlenül az, hogy a politika úgy és olyan mértékben szól bele a nemzetbiztonsági szolgálatok és rendvédelmi szervek működésébe, hogy azok szakmai függetlenségét, demokratikus céljait veszélyezteti. Ilyen politikai cél lehet többek között egy-egy példával érzékeltetve:

- ellenzéki politikai pártok, mozgalmak befolyásolása, ilyen az Öcalan-ügy [177];
- saját vagy szövetséges párt tagjainak megfigyelése, ilyen a Watergate-ügy [178];
- civil személyek vagy szervezetek elleni fellépés, ilyen a Politovszkaja-ügy [179];
- újságírók megfigyelése, például francia újságírók megfigyelése a források végett [180];
- bennfentes informátorok *(whistleblower*ek) elleni fellépés, ilyen a Mordechai Vanunu-ügy [181];
- politikai döntések szakmai alátámasztása, ilyen a Valerie Plame-ügy [182].

Áttekintve azokat a jogi és szervezeti mechanizmusokat, amelyek a fentebb idézett fékeket és ellensúlyokat biztosítják, megállapítható, hogy semmilyen, formailag tökéletes ellenőrző mechanizmus sem működőképes, ha a személyi feltételek nem biztosítottak. Ha minden, látszólag függetlenséget igénylő posztra egy, a kulisszák mögötti megállapodás alapján mégis más irányból befolyásolható és befolyásolt embert neveznek ki, akkor a fékek és ellensúlyok valójában nem működnek. Ilyen volt például a sztálini rendszer, a nemzetiszocialista állam 1933 utáni *state capture*-e, vagy a II. világháború utáni Nyugat-Németországban terebélyesedő ODESSA (a korábbi SS-hez tartozó személyek szervezete, *Organisation der ehemalingen SS Angehörigen*) szervezet. Az alábbiakban felsorolok néhány eszközt és intézményt, amelyeket a „fékek és ellensúlyok" módszereként tart számon a szakirodalom.

- A szolgálatok egymást figyelik.
- Főigazgatók kinevezéseihez parlamenti jóváhagyás szükséges. Így a végrehajtó hatalmat személyi feltételeken át ellenőrzi az Országgyűlés Nemzetbiztonsági Bizottsága.
- A szolgálatok és szervek belső ellenőrzésének lehetősége és kötelessége a visszaélések felderítése.
- Parlamenti bizottságok (Magyarországon honvédelmi és nemzetbizottsági) beszámoltatják a nemzetbiztonsági szolgálatok és rendvédelmi szervek vezetőit. Az, hogy milyen mélységben láthat bele egy parlamenti bizottság a nemzetbiztonsági szolgálatok és rendvédelmi szervek ügyeibe, ország szerint különböző. Van, ahol csak

stratégiai szinten, és van, ahol konkrét műveletekbe is belelátnak a nemzetbiztonsági szolgálatok jogellenes tevékenység esetén. Az sem mindegy, hogy általában milyen minősített információkhoz jutnak hozzá a bizottsági tagok.

- Különleges parlamenti bizottságokat nevezhet ki a törvényhozás egy-egy konkrét ügy kivizsgálására.
- A nemzetbiztonsági szolgálatok és rendvédelmi szervek munkáját egy felelős miniszter felügyeli, akinek országonként különböző mértékben van joga beleszólni az operatív ügyekbe. Magyarországon ez nem megengedett.
- Az igazságügyi miniszter engedélyezhet olyan műveleteket, amelyet a személyi adatok különleges kezelésére is kiterjednek.
- Bírói döntés engedélyezhet a legtöbb országban, így Magyarországon is a személyiségi jogokat korlátozó műveleteket, így adatszerzést és -kezelést is.
- A NAIH (Nemzeti Adatvédelmi és Információszabadság Hatóság) felügyelet ellenőrzi a nemzetbiztonsági szolgálatok és rendvédelmi szervek adatkezelését az Infotörvény alapján. Ugyancsak a NAIH-hoz fordulhat egy állampolgár panasszal, amelyet a NAIH kivizsgál. [183].
- Az Európai Emberi Jogi Bíróság (*European Court of Justice*) fórumához fordulhat bármelyik EU-állampolgár, aki nem találta meg az igazát a hazájában.
- Civil szervezetek futtathatnak véleményformáló fórumokat, szervezhetnek polgári megnyilvánulásokat.
- Agytrösztök (*thinktanks*) kiemelkedő szakértői tudásukkal politikai beállítottságuk függvényében figyelemmel kísérnek eseményeket, és ezek alapján akár a nyilvánosság igénybevételével is befolyásolnak folyamatokat.
- A szabad sajtó feltárhat visszaéléseket.
- A közösségi média lehet a szabad kritikus véleménynyilvánítás platformja, akár anonim formában is.
- Köztiszteletben álló, független emberekből álló bizottságok vizsgálódhatnak.

A jogi garanciákon túl rendelkezésre állnak a számonkérhetőség technikai lehetőségei. A tevékenység naplózása és a dokumentáltság az ellenőrzés egyik leghatékonyabb módja. Lényege, hogy minden tranzakció (megtekintés, változtatás, törlés vagy ezek sikertelen megkísérlése), minden felhasználó a rendszergazdák számára elérhetetlen és megváltoztathatatlan helyen rögzítésre kerül a későbbi ellenőrzés vagy audit végett. Zárt rendszerben (közigazgatási, állami végrehajtására rendszeresített adatbázisban, adattárban)

keresés csak naplózás mellett engedélyezett és lehetséges. A nem naplózott keresés technikailag lehetetlen kell legyen.

Néhány, a civil szervezetek által publikált eset és módszer szemlélteti a személyiségi jogokat biztosító törvények megkerülésének lehetőségét. Ezek természetükből fakadóan kevésbé ellenőrizhetőek, de vizsgálatuk mindenképpen megfontolandó.

A „kéz kezet mos" modell, amelynek lényege, hogy ami tilos az egyik országban, az nem tilos egy másikban. Ezt kihasználva kikerülhetők a nemzeti törvények. Eklatáns példa az NSA és a GCHQ együttműködése. Mindkettő számára erősen korlátozott a saját állampolgárok figyelése, de nem tilos a másik ország állampolgárait figyelni, hiszen azokra mint külföldiekre, a nemzeti törvények nem vonatkoznak. Az adatok kicserélése pedig lehetséges. [184].

A feladatok privát szervezetekbe való erőforrás-kihelyezése (*outsourcing*) nem ismeretlen a nemzetbiztonsági szolgálatok és rendvédelmi szervek köreiben [185]. Egy külföldi alvállalkozó tevékenységét elég nehéz hivatalosan ellenőrizni. Ilyen szervezeteket lehet olyan kényes feladatokkal megbízni, amelyekről nem lenne kellemes egy parlamenti bizottság előtt számot adni [186].

4.3. Az információkeresés humán és biztonsági keretei

4.3.1. A humán oldal

A következő fejezetrészben jórészt saját tapasztalat alapján ismertetem és foglalom össze az információkeresés humán aspektusait. A tapasztalat szerint egy komplex keresőrendszer technikai megvalósítása néhány hét vagy hónap alatt megtörténhet, míg a szervezetben való tényleges abszorbeálása lényegesen hosszabb időt vesz igénybe. Személyes ipari tapasztalatból is ismeretesek példák hatalmas beruházások elhalására a folyamatok szabályozása és a képzés elmaradása miatt.

Az információkeresési rendszerrel kapcsolatos szerepkörök azonosítása a működés sikeréhez elengedhetetlen. A szerepkörök meghatározásához nemcsak a rendszer sajátosságait, hanem az azt használni készülők munkafolyamatait és információs igényeit is figyelembe kell venni. Ezek a szerepkörök a munkafolyamat során jól elkülöníthető, más és más képzésszintet igénylő feladatcsoportokhoz kapcsolódnak.

- Alkalmi felhasználó

Alkalmi felhasználónak nevezzük azt a munkatársat, aki az internet által biztosított keresőfelületeteket használja bizonyos rendszerességgel, de automatizált információszolgáltatást nem vesz igénybe.

- Folyamatos felhasználó

A rendszer folyamatos felhasználója az, aki a keresőfelület mellett igénybe veszi a különböző értesítési szolgáltatásokat. Olyan felülettel rendelkezik, amelyen a frissülő értesítési információk mindig testreszabottan jelennek meg.

- Kulcsfelhasználó

A kulcsfelhasználó amellett, hogy a rendszer folyamatos használója, kapcsolatban van a tudásmenedzserrel, akinek javaslatot ad a rendszer tartalmi oldalával kapcsolatosan, illetve támogatja szakmai oldalról.

- Tudásmenedzser

A tudásmenedzser felelős a rendszer tartalmi karbantartásáért, bővítéséért, teszteléséért. A feladatok ellátásához elsősorban a kulcsfelhasználókkal történő egyeztetés a célszerű, de folyamatosan kapcsolatban kell lenni minden felhasználóval. Mivel a rendszer nagy mennyiségű és folyamatosan változó információtömeg szintén változó szempontok alapján történő feldolgozását végzi, a jó működéshez elengedhetetlenül fontos a tudásmenedzser megléte.

- Üzemeltető

Az üzemeltetői szerepkör jelenti a rendszer informatikai szempontú kezelését. Természetesen a rendszer jellegéből adódóan több feladat is van, amelyet a tudásmenedzseri szerepkörrel közösen, konzultálva kell elvégezni.

- Informatikus

Az informatikusok végzik a keresőrendszerek szoftverének fejlesztését és karbantartását. Ők adnak tanácsot a különböző keresőeszközök (*tool*ok) használatához, és a kiegészítő modulok (*add-on app*ok) illesztéséhez és beüzemeléséhez. Természetesen, mint minden informatikai alkalmazás, a keresőrendszerek is igénylik az infrastruktúra (hardver, hálózat, alapszoftverek stb.) támogatását.

Az információkeresést egy szervezetben többen, több szinten művelik. Ezekhez a különböző szintű tevékenységi szintekhez más és más szintű képzettség szükséges. A 7. táblázat mutatja a szükséges képzés szintjét a foglalkozás és a keresési alkalmazások használatának intenzitása szerint. A felhasználók a képzés szempontjából öt kategóriába sorolhatók. Az első oszlopban a

felhasználók láthatók foglalkozás szerinti csoportban. A második oszlopban a felhasználói szint, a harmadikban a képzés szintje, tartalma olvasható. Külön kiemelendő a felsővezetők képzése. Abban a gyakori esetben, amikor a felsővezető nincs tisztában fogalmi szinten sem az információkereső alkalmazások képességeivel, sem a felhasználhatóságaival, a technológiai fejlődés és megújulás lelassulhat, vagy akár el is maradhat az ismeretlentől való félelem okozta vonakodás miatt.

A tudás megosztásának csak egyik oszlopa a megfelelő informatikai alkalmazásra épülő informatikai tudásmenedzsment-rendszer. Legalább akkora feladat egy ilyet bevezetni egy felhasználói közösségbe, mint elfogadtatni azt. Ezt a folyamatot többféle akadály nehezíti. Igyekszem felsorolni az évtizedek alatt megismert akadályokat, és azok leküzdésének lehetséges módjait.

7. táblázat: a szükséges képzési szintek.[88]

Felsővezetők	Nem aktív felhasználók	Koncepcionális ismeretek. Érteniük kell a képességeket és azok felhasználhatóságát.
Alkalmi felhasználók	Csak keresőben és közösségi médiában keresnek (Avatárral).	Alapkiképzés: keresők, operátorok, közösségi média lekérdezései, a hírigény dekompozíciója.
Elemző-értékelők	Napi szinten keresnek az interneten	Alapkiképzés, *Tools, API*, *APPs,* metakeresők,
Professzionális keresők	Ők az idejük nagy részében információkereséssel foglalkoznak. Munkájuk során felhasználói szinten a teljes rendelkezésre álló arzenált alkalmazzák.	A használt eszköztár nóvumokkal, delicatesse-ekkel történő kiegészítése. Programok használata.

[88] Forrás: a szerző.

Informatikusok	Ők fejlesztik és támogatják a belső alkalmazásokat rendszerszinten	Rendszerszintű ismeretek, programozás, parametrizálás.

Az egyik ilyen akadály az ismeretlentől való félelem. Azok a felhasználók, akik koruknál, helyzetüknél, képzettségüknél, tapasztalatuknál fogva kevésbé érzik magukat otthonosan egy fejlett technológiai környezetben, vonakodásukkal ellenállást fognak tanúsítani a bevezetéskor. Az oktatás, türelmes meggyőzés, valamint az informatikailag képzett – sokszor fiatalabb – kolleginák vagy kollégák *train-the-trainer* alapon való alkalmazása segít áthidalni a nehézségeket.

A másik gyakori akadály a tudás megosztásától való félelem. Minden kollegina vagy kolléga, kisebb-nagyobb munkacsoport a helyzetének és biztonságának zálogát látja abban, hogy monopolizálhatja az általa birtokolt tudást. Plasztikusan fogalmazva úgy érzi magát, mint egy félbevágott citrom, amelynek – a tudás átadásakor és egy tudásmenedzsment-rendszerbe bevitelekor – kifacsarják a levét, és utána akár el is dobhatják. A megosztás-megtartás antagonizmusa jól ismert a tudásmenedzsment minden területén. Ezt a típusú ellenállást a legnehezebb leküzdeni, ugyanis többé-kevésbé megalapozott. Az egyénekkel, csoportokkal azt kell megértetni, hogy a parciális és rövid távú érdekeik a fejlődés akadályai, így a globális és hosszú távú szempontokat kell, hogy figyelembe vegyék. A biztonságérzetüket különböző HR-módszerekkel lehet javítani. Végső esetben a vezetés kevésbé demokratikus eszközök használatára kényszerülhet. Ezek taglalása nem tárgya a jelen írásnak. Rendvédelmi és nemzetbiztonsági körökben külön veszélyérzetet okoz a forrás felfedésétől való félelem. Ha egy keresőrendszer rálát a nyers ügynöki jelentésekre, egy tapasztalt értékelő-elemző az adatmozaikokból idővel össze tudja rakni a teljes képet, és következtetni tud a hírszerzési forrás kilétére, ami szivárgás esetén hosszú idő alatt és költséges módon felépített műveleteket dekonspirálhat.

A harmadik akadály az állás elvesztésétől való félelem. Ez az első géprombolók óta ismert jelenség. A dolgozók attól félnek, hogy az automatizálás feleslegessé teszi a munkájukat, aminek következtében a munkahelyük megszűnik, őket pedig elbocsátják. Ezt a – nem feltétlenül alaptalan – félelmet időben történő átszervezéssel és átképzéssel lehet megszüntetni vagy legalábbis csökkenteni. Szakképzett értékelő-elemzők, fordítók esetében a munkahelyek megszűnése teljesen valószínűtlen. A robotok bevezetése átalakítja a munkafolyamatot, és lényegesen nagyobb áteresztőképességet eredményez. De a mai viszonyok között

elképzelhetetlen a tökéletesen automatizált munka, a folyamat végén mindenképpen szükséges az eredmény tökéletesítése és finomhangolása, ami belátható időn belül csak humán erőforrással végezhető el. Példaként álljon egy automatizált S2T lehallgatórendszer gépi fordítással. A S2T robot 70-80%-os pontossággal átalakítja a hangot szöveggé, azt a gépi fordító 60-80%-os pontossággal lefordítja ismert nyelvre. Ezt a szövegtestet egy szövegbányász-alkalmazás feldolgozza, és kiszűri a figyelmet igénylő elemeket. Ezeket viszont humán erőforrással fel kell dolgozni a nagyobb pontosság érdekében. A munka jellege valamennyire megváltozik, a teljesítmény nő, de az igény az emberi kapacitásokra nem csökken.

További akadályai az információk terítésének (*dissemination*) az ún. vezetői vakfoltok. A vezetői vakfoltokkal foglalkoztam az Információkeresés a gazdasági életben fejezetrészben. A vezetői vakfolt lényege, hogy hiába kapják meg a döntéshozók az értékelő-elemzőktől a szükséges elemzéseket és előrejelzéseket, azokat nem veszik figyelembe, mert nem azt látják, ami van, hanem azt, amit látni szeretnének [187]. Vagy fordítva, nem látják meg azt, amit nem akarnak meglátni, mert ellenkezik a prekoncepcióikkal. Egy hadtörténeti példa a harmadik arab–izraeli háború. 1973 őszén Izraelt meglepetésszerűen megtámadta Egyiptom és Szíria. Vajon mi történhetett az izraeli hadvezetés és politikai elit köreiben, hogy nem számítottak erre az eseményre? Kérdezzük ezt függetlenül attól, hogy a háború harmadik napján nem kis véráldozattal sikerült visszaverniük annak a hadseregnek a támadását, amelyik 1967-ben olyan fergeteges győzelmet aratott három arab ország hadserege felett. A Yom Kippur háború 10.000 fiatal izraeli életét követelte. A későbbi vizsgálatok eredménye abban összegezhető, hogy nem számítottak ilyen támadásra. Nem voltak képesek elhinni, hogy a ’67-es győzelem után ilyesmi bekövetkezhet. Vajon a legendás izraeli titkosszolgálatok mondtak csődöt, vagy a vezetők hittek az ösztöneiknek, tapasztalatuknak, vagyis a megszokásnak? Vakon hittek a legyőzhetetlenségükben. Ezt a jelenséget hívják vakfoltnak. A képesség hiánya – hogy valaki lássa a valós helyzetet, elutasítsa a realitást, mert nem tud vagy akar szembenézni azzal –, szervezetpszichológiai vagy -szociológiai terminussal három eredőre vezethető vissza.

- a megtámadhatatlan feltételezések,
- a szervezeti mítoszok, legendák, amelyekben bűn nem hinni,
- a tabuk, amelyeket tilos megérinteni is.

A taxonómiát részletesen tárgyaltam Az információkeresés nyelvészeti alapjai fejezetrészben. Ebben a fejezetrészben csak felhasználói és szervezeti szempontból vizsgálom. Számos esetben lohadt le a felhasználók lelkesedése egy keresőrendszer iránt azért, mert idővel a felidézés és a

pontosság gyengült. Ennek a leggyakoribb oka az, hogy a taxonómiát nem tartották karban. Az a felfogás, hogy egy keresőrendszernek ugyanúgy naprakésznek kell lennie minden felhasználói beavatkozás nélkül, mint egy könyvelési csomagnak, teljesen hibás. A tudásmenedzsernek figyelnie kell a felhasználók lekérdezéseit, visszajelzéseit. Túl kevés vagy éppen nem létező találat esetén, továbbá túl sok hibás találat esetén a taxonómiát módosítania kell. Ez a szemantikai segítséggel működő keresőrendszerek sajátsága, ami bizonyos értelemben hátránynak is tekinthető. A nyelvészeti fejezetrészben utaltam arra, hogy léteznek öntanuló algoritmusok, amelyek a kézi beavatkozást bizonyos mértékig helyettesítik. De a technológia mai fejlettsége mellett a humán tanítás megkerülhetetlen.

Ugyancsak a vezetés feladata elégséges humán erőforrást rendelkezésre bocsátani az értékelő-elemző tevékenység elvégzéséhez. Hiába képes a robot hatalmas mennyiségű külső és belső információt feldolgozni, összefüggéseket feltárni, ha ezeket avatott szakértők nem elemzik, és nem továbbítják az eredményeket a felsővezetés felé olyan formában, tömörítésben – lényegi kiemelésekkel, vizualizálva az összefüggéseket –, hogy a felsővezetés számára mindez könnyen érthető és feldolgozható legyen.

4.3.2. Biztonsági megfontolások

Az adatbiztonság feltárt és a szakmai véleményformálókat is intenzíven foglalkoztató terület. Ugyanakkor az információkeresés biztonsági oldaláról keveset olvasni. A következőkben rendszerezem a téma lényeges aspektusait. Mivel egy jól működő keresőrendszer lényegéből fakadóan „sokat lát, és főleg egyben”, ezért a mindenkori vezetőséget és biztonsági felelősöket intenzíven foglalkoztatják a biztonsági megfontolások.

Egy korszerű keresőrendszer – gyakran a Microsoft *Active Directory*jára épülve – átveszi az indexelt forrás hozzáférési jogosultságát akár rekordszinten is. Más szavakkal: a felhasználó a keresőrendszer (az index adatbázis) lekérdezésekor pontosan ugyanazt láthatja és módosíthatja, mint az eredeti (leindexelt) adatállományban. Ha például a fájlszerveren egy dokumentumot nem láthatott vagy nem nyithatott meg, akkor a keresőrendszer lekérdezésekor sem láthatja vagy nyithatja meg a kérdéses állományt.

A jogosultság négy szinten lehetséges:

- nem látja a dokumentum létezését sem;
- látja a dokumentum létezését, de nem nyithatja meg;
- megnyithatja a dokumentumot, de nem változtathatja meg a tartalmát;
- megváltoztathatja a dokumentum tartalmát.

Ha a felhasználó láthatja a dokumentum létezését a lekérdezéskor, de abba nem tekinthet bele, akkor kérheti annak engedélyezését.

A Snowden-ügy óta senkinek sem lehet kétsége afelől, hogy egy rendszergazda mekkora veszélyt jelenthet egy szervezet adatbiztonságára. Egy klasszikus értelemben vett (tehát a teljes infrastruktúrát felügyelő) rendszergazdának nem kell hozzáférése legyen a keresőrendszer indexadatbázisához, és lekérdezési jogokkal sem kell rendelkeznie. Egy kiszervezett üzemeltető, aki a keresőrendszer működéséért felel, természetesen hozzáfér a teljes állományhoz. Egy biztonsági szempontból igényes keresőrendszer automatikusan naplózza a benne előforduló eseményeket, lekérdezéseket. Más szavakkal: visszakereshető, hogy ki, mikor és mit nézett meg. Egyes csoportos felhasználók határozottan megtiltják a naplózást éppen azért, hogy az üzemeltető ne lássa, hogy ők mire kíváncsiak.

Egy szervezet (külön cég, kiszervezett vagy belső IT-részleg stb.) a keresőrendszerhez háromféleképpen tud hozzáférést szolgáltatni.

- Önálló, teljes infrastruktúrát épít ki, és a teljes üzemeltetést a felhasználókra bízza. Ebben az esetben csak másodlagos támogatást (*second line support*) biztosít. Ilyenkor külön tesztkörnyezetet létesíthetnek ugyanazzal az alkalmazással, de – biztonsági szempontból – érdektelen tesztadatokkal. A tesztkörnyezetben végrehajtott módosításokat az élő rendszerben „belső” embereknek meg kell ismételniük.
- A teljes infrastruktúra az IT-részlegben fut, a felhasználók csak a keresőfelülettel állnak kapcsolatban.
- A külső internetes helyeken és a belső forrásokon az indexelést az IT-részleg végzi el, majd a teljes indexadatbázist eljuttatja a felhasználókhoz, akik a lekérdezéseket úgy folytathatják a saját környezetükben, hogy annak naplózását az IT-részleg nem látja.

A taxonómia és ontológia biztonságának vizsgálatakor nem klasszikus értelemben vett IT-biztonságról van szó, de a keresőrendszer igazi egyediségét adó taxonómia vagy ontológia különleges figyelmet érdemel. Világcégek igazán nagy alkalmazásai nem is eladók, de egy egyszerűsített változat is kerülhet dollárszázezrekbe vagy -milliókba. Egy ilyen tudásreprezentáció szakértők emberéveinek ezreit testesíti meg, ezért az adatállomány védelmét mindenképpen biztosítani kell.

A need-to-share versus need-to-know[89] minden biztonságtudatos szervezet egyik alapvető problémája. A felhatalmazás (*empowerment*) és csak a legszükségesebbek megosztása (*need-to-know*) paradoxona éppúgy feloldhatatlan, mint a szabadság kontra biztonság (*liberty versus security*) ellentéte. A spektrum egyik végén a liberális szemléletű vezetési filozófia áll, amely minél több információt kíván megosztani a dolgozókkal attól a meggyőződéstől vezérelve, hogy a munka minősége javul, ha a személyzet tájékozott, és birtokában van minden szükséges részletnek. A felelősség növelésével, az információk demokratikus megosztásával alkotói energiák szabadulnak fel, amelyek a tudás átadását és a hatékonyságot sokkal jobban szolgálják, mint a hagyományos hierarchikus struktúrákban. Sokak között emellett érvel Blanchard is [188] is. A spektrum másik végén az olykor paranoiába hajló biztonságtudatú, erősen hierarchikusan szervezett, parancsvezérelt szervezetek állnak, amelyekben a biztonság mint szempont mindenek felett áll. Ezekben mindenki csak annyi információhoz jut hozzá, amennyihez elengedhetetlenül szükséges, és a tudása – saját magát beleértve – ideális esetben olyan könnyen cserélhető és pótolható, mint egy meghibásodott kártya egy számítógépben. Nyilvánvaló, hogy az optimumot minden szervezetnek saját magának kell megállapítania. Itt nem vizsgálom az egyébként erősen parancsuralmi rendszer égbekiáltó fiaskóit, mint a Chelsea Manning- [189] vagy az Edward Snowden-féle adatszivárgást [190]. Manning közlegény korlátlanul hozzáfért háromnegyedmillió minősített adathoz, míg Snowden rendszergazdaként az NSA szinte teljes adatállományához. A szemantikus keresőrendszerek – mint fent láttuk – alkalmasak a hozzáférési jogok kívánság szerint korlátozásához, és ezzel a szükséges biztonság megteremtéséhez feltéve, ha emberi hiba nem okozza az ellenkezőjét.

Keresési tevékenységet minősített vagy kényes helyzetekben nem szabad nyíltan végezni. Más szavakkal, a kereső azonosságát álcázni szükséges. Ilyen körülmények között kell dolgoznia egy rendvédelmi vagy nemzetbiztonsági tevékenységet végző embernek, fedett forrásnak, oknyomozó újságírónak vagy whistleblowernek. Az álcázás (*anonimity, ingonito search*) szinte egy önálló diszciplína, amelynek néhány olyan elemét foglalom össze, amelyek a keresés során alkalmazhatóak. Fontos megemlíteni, hogy – mint sok más terület – ez is folytonosan változik. A módszereket az érdeklődőnek állandóan követnie és frissítenie kell.

[89] A két fogalom idegen szóként elég élénken él a magyar szakmai köztudatban, ezért – kivételesen – meghagytam az eredeti formában. Szabad fordításban: a (széles körű) megosztási kényszer/igény szemben azzal, hogy csak annyit tudjon a dolgozó, amennyit szükséges.

Külső operációs rendszer használata

Legismertebb a TAILS (*The Amnesic Incognito Live System*) nevű nyílt forrású, Linux-alapú operációs rendszer. A keresést kizárólag a TOR-on keresztül engedélyezi, kizárva a nem anonimizált kapcsolatokat. Külső adathordozóról (DVD vagy USB) kell indítani, és nem hagy nyomot a számítógép merevlemezén, csak a gyorsítótárban. 2009 óta használatos. VPN-nel kombinálva elég erős védelmet nyújt.

Proxy kiszolgálók használata

Leggyakoribb eszköze a TOR (*The Onion Router*). Lényege, hogy a kommunikáció összesen 11, földrajzilag távoli szerveren át folyik véletlenszerű útválasztással. Ezek mindegyike külön titkosítja a tartalmat egy 128 bit hosszúságú szimmetrikus kulcsolással, amely egy aszimmetrikus kézfogással (*handshake*) jön létre. A TOR titkosítása két ismert módon támadható. Egyrészt lehetséges figyelni a teljes internetforgalmat, így látni az egy ponton bejövő, még nem titkosított, majd a végponton kilépő, már nem titkosított forgalmat. Másrészt lehetséges, sőt gyakori ál proxy szerverek alkalmazása, amelyek mint átmeneti állomás kinyerik az átmenő tartalmat. Mindkét módszerhez komoly, államilag támogatott apparátus szükséges.

VPN használata

A virtuális magánhálózatok (*Virtual Private Network, VPN*) egy létező számítógép-hálózatra ráépült hálózatok, amelyek általában titkosítottak. Egyes VPN-ek a titkosításon túl bizonyos mértékű anonimitást is biztosítanak. VPN alkalmazása esetén a hálózati szolgáltató (*internet service provider*, ISP) nem látja a felhasználó valódi IP-címét, csak a VPN-szolgáltatóét, ami a különböző országokban működő szerverek miatt könnyen változtatható, és a felderítés szempontjából semmitmondó. Lényeges azonban, hogy a VPN-szolgáltató látja a felhasználó IP-címét, és naplózhatja a felhasználó kereséseit. A fizetős VPN-szolgáltatóknál a fizetésnek is nyoma van. Sok VPN-szolgáltató állítja magáról, hogy nem naplózza a felhasználó kereséseit, amit a függetlennek ismert fórumok vagy visszaigazolnak, vagy sem.

Internetkávézók használata

Mivel az anonimitás relatív, és teljes biztonságot egyetlen eszköz sem ad, a legbiztosabb keresési mód eltűnni a tömegben egy távoli internetkávézóban, ahol lehetőleg nincs kamera a bejárattal szemben.

Böngészőben tárolt nyomok

Nem elég az IP-címet álcázni, a böngésző egy sor egyéni azonosításra alkalmas nyomot tartalmaz, amelyek eltűntetése szinte lehetetlen. Hasznos, ha angol nyelvre beállított gépen keresünk, angol billentyűzettel, böngészőbővítmények nélkül, csak egyféle betűtípussal. El kell kerülni, hogy ugyanazzal a böngészővel keressünk anonim módban, mint amelyiket hétköznapi körülmények között használunk. Bővebb információt az Electronic Frontier Foundation által kezelt Panopticlick honlapján találunk. A Panopticlick egyébként alkalmas a kitettség tesztelésére.

Hardverkitettség

Ennek legfontosabb lépése – a felhasználás hálózati környezetétől függően – az IP-cím (internetprotokoll, *internet protocol*) és a hálózatkártya MAC (médiaelérés-szabályzó, *Media Access Control*) címének megváltoztatása. Még kényesebb helyzetekben a keresést futtató gép további jellemzőit, mint a klaviatúra, nyelvi azonosító stb. is álcázni kell. Nem térek ki, csak utalok itt a MAC-cím bemérhetőségére és az egér-, illetve billentyűzet-használati statisztika kiértékelhetőségére [191].

Sütik

A sütik (*cookies*) kimondottan azzal a céllal készülnek, hogy a felhasználó tevékenységét figyeljék. Ma már a legtöbb honlap feltételként szabja a sütik használatának elfogadását. Természetüknél fogva lehetnek kevésbé ártalmasak, és csak a meglátogatott honlapokat rögzítik, de lehetnek kimondottan rosszindulatú kódok, amelyek minden billentyűleütést rögzítenek és továbbítanak (*keylogger* funkció).

Aktív és passzív támadás elleni védelem

Fontos megjegyezni, hogy a fenti technikai megoldások feltételezik, hogy a felhasználó tevékenységét csak „kívülről” figyelik. Nem biztosítanak anonimitást akkor, ha valamilyen aktív eszközzel „belülről” követik a tevékenységét. Más szavakkal: keyloggerek, vírusok, trójaiak stb. ellen nem jelentenek védelmet. Ezeket kiszűrik a vírus- és kémprogramirtók, hacsak nem repülnek a radarernyő alatt, mert a gyártó megállapodott a titkosszolgálatokkal [192]. Ide tartozó „balesetről” számol be a Kaspersky Lab a honlapján [193].

Az anonimizálás két szempontból is kapcsolódik a szemantikus keresés biztonságának témájához. Egyrészt, mert erőteljesen használ szemantikus eszközöket, másrészt, mert a személyiségi jogi feltételeket (államigazgatási, igazságügyi, egészségügyi stb. alkalmazásokban) meg is követeli az adatvédelmi törvény. **Az anonimizálás során az adatokat úgy alakítják át, hogy azok az átalakítás után már ne legyenek természetes**

személyhez köthetőek. Az átalakítás történhet a személyes részek titkosításával vagy eltávolításával. Ezeket a megtisztított adatokat az adatkezelők már minden korlátozás nélkül feldolgozhatják. Az anonimizálás történhet manuálisan és gépi támogatással. A gépi támogatás célja, hogy a nyilvánvalóan drága és hibára hajlamos emberi munkát minél inkább kiküszöbölje. Megjegyzendő, hogy a legfejlettebb automatizált eszköz sem végez tökéletes munkát, és szükséges az emberi ellenőrzés. Az anonimizálás technológiája az entitáskiemelésre épül. A technológia lényege, hogy bizonyos szótípusokat (földrajzi nevek, vezeték- és keresztnevek stb.), alakra felismerhető egységeket (telefonszám, hitelkártyaszám, irányítószám stb.) vagy egy szabadszövegű szövegtestben jellemző tartalmi elemeket (vád, ítélet, kórisme stb.) automatikusan felismerjen, osztályozzon, és más műveletekre (törlés, titkosítás, adatbázisba átemelés stb.) továbbítson. Az azonosítás történhet alaki azonosság felismerésével (*Regular Expression*, RegEx vagy Regexp) vagy statisztikára épülő gépi tanulás segítségével. A tanítóadatok frissítése, a hibák korrigálása, a helyesbítések újratöltése ugyancsak emberi beavatkozást igényel.

4.4. Az információkeresés értékelése

Az információkeresés az informatika interdiszciplináris területe, amely határos a statisztikával, a lineáris algebrával, a számítógépes nyelvészettel, valamint számos célterülettel, amelyre alkalmazzák: a nemzetbiztonság, a rendvédelem, üzleti szféra, jog, közigazgatás, egészségügy stb. Alkalmazása akár a weben történő keresés, akár egy szervezet belső strukturálatlan szövegeiben történik, beruházást és folyó költségeket is jelent, amelyről a döntéshozóknak határozniuk kell. Viszonylag csekély szakirodalom áll rendelkezésre a számítógépes alkalmazások gazdaságosságáról és szervezeten belül mérhető hasznosságáról, a keresőrendszerek gazdaságosságáról még annál is kevesebb, jóllehet az exponenciálisan növekvő szövegállományokban való kereshetőség létkérdés a XXI. században.

Miközben könyvtárnyi cikk jelent meg főleg angol nyelven e témában tudósoknak és mérnököknek, az alkalmazói terület, valamint az üzleti és egyéb nem szakirányú felhasználók számára a témakör szinte terra incognita maradt. Üzleti, beruházási döntéseiket nemcsak azért nem tudják meghozni, mert nem ismerik a megtérülés idejét, hanem sokszor azért sem, mert magukat a fogalmakat sem ismerik.

Gazdasági vállalkozásoknál vagy kormányzati szerveknél felmerül a kérdés, mennyire éri meg egy keresőrendszerbe történő beruházás, mik a megtérülési mutatók. Mennyire elégíti ki egy keresőrendszer a hírigényeket? Külön figyelmet érdemel egy szemantikus keresőrendszer megalkotásának és fenntartásának a kérdése. Mennyivel javítja a működést egy nem csak

statisztikai módszerekkel készített ontológia alkalmazása? Egy taxonómia elkészítése és karbantartása a felhasználók számára sokszor ijesztő és felesleges feladatnak tűnik, ami miatt egy alkalmazás bevezetésével szemben a szervezet kelletlenséget, sőt ellenállást is gerjeszthet. Az ellenállással szemben világos becslésekre alapuló megtérülési mutatókat kell felsorakoztatni a vezetői döntés alátámasztásának érdekében.

Egy keresőrendszer működésének értékelésére – akárcsak más informatikai alkalmazáséra – nincs egyértelmű objektív módszer. Plasztikusan szemléltetve: teljes kép helyett különböző metszeteket lehet készíteni, és ezek vizsgálatával következtetéseket levonni. A szakirodalom három nagy részre csoportosítható: informatikai-matematikai-információelméleti szempontok, számviteli-kontrolling szempontok, és a felhasználók többé-kevésbé szubjektív, de korántsem elhanyagolható szempontjai. Ez utóbbiakat lehet az egyén és a szervezet szempontjából vizsgálni, ezeket nevezhetjük szervezetpszichológiai és szervezetszociológia szempontoknak. Miközben mindhárom megközelítést a szakirodalom bőségesen feldolgozta, magáról a gazdaságosság méréséről igen kevés található, jóllehet, talán pont erre lenne a legnagyobb igény. Egy felsővezetőnek vagy gazdasági szakembernek az informatikai vagy szervezeti mutatók nem sokat mondanak. Ők azt akarják tudni, hogy a rendszerre elköltött pénz mikor és mekkora haszonnal térül meg. Ugyanúgy, mint bármely más beruházásnál, mint egy gép vásárlása, tőkeemelés egy leányvállalatban, vagy a személyi állomány bővítése. Egy informatikai beruházás tényleges megtérülésének mérésére ritkán találni módszert a szakirodalomban [194], keresőrendszerekre, különösen információ-visszakeresőkre vonatkozóan pedig még annál is ritkábban. A továbbiakban ezekre a kérdésekre keresek választ, megoldást.

Az információkeresés teljesítménymméréséhez (eredményességének és hatékonyságának méréséhez) alapvető feladat a keresési igényeknek történő megfelelést jellemző, valamint a keresés megvalósításához felhasznált erőforrásokat leíró teljesítménymutatók meghatározása. A keresési igényeknek történő megfelelést jellemző teljesítménymutatók esetében célul tűzhető ki általános tartalmú, alkalmazásiterület-specifikus, illetve konkrét alkalmazás, ill. szervezetspecifikus mutatók meghatározása. Az utóbbiak gyakorlati tapasztalatok, vélemények, felmérések, interjúk alapján, az előbbiek pedig ezeket, valamint elméleti megfontolásokat is felhasználva határozhatók meg.

Az eredményesség és hatékonyság mérésének fogalmi alapjait, kereteit a szervezeti vagy üzleti teljesítménymérés (*performance measurement*) elméletére és módszertanára kell építeni. **A teljesítménymérés egyik széles körben elfogadott definíciója: valamely tevékenység**

eredményességének (*effectiveness*) és hatékonyságának (*efficiency*) mérési folyamata. Ezen belül:

- **az eredményesség a tervezett cél megvalósulásának, a kívánt eredmény elérésének, az előírt igényeknek történő megfelelésnek a mértéke, azaz az elért eredmény összevetve a kívánt eredménnyel**, ami a teljesítmény „külső" megítélése, jellemzője (ennek során nem foglalkozik a ráfordításokkal), és általában nehezen mérhető;
- **a hatékonyság a tervezett célok megvalósulásához, eredmények eléréséhez, igényeknek történő megfeleléshez szükséges erőforrás-felhasználás mértéke**, azaz az elért eredmény viszonyítva a ráfordításhoz, ami a teljesítmény „belső" megítélése, jellemzője, és általában könnyebben mérhető;

A teljesítménymérés alapját teljesítménymutatók (*performance indicator*), ezen belül kulcsfontosságú teljesítménymutatók (*key performance indicator*) képezik. Ezekre épülhet egy teljesítménymérési rendszer, módszer. A teljesítménymutatók különböző típusokba sorolhatók, és jelentős részük a megítélendő tevékenységre specifikus.

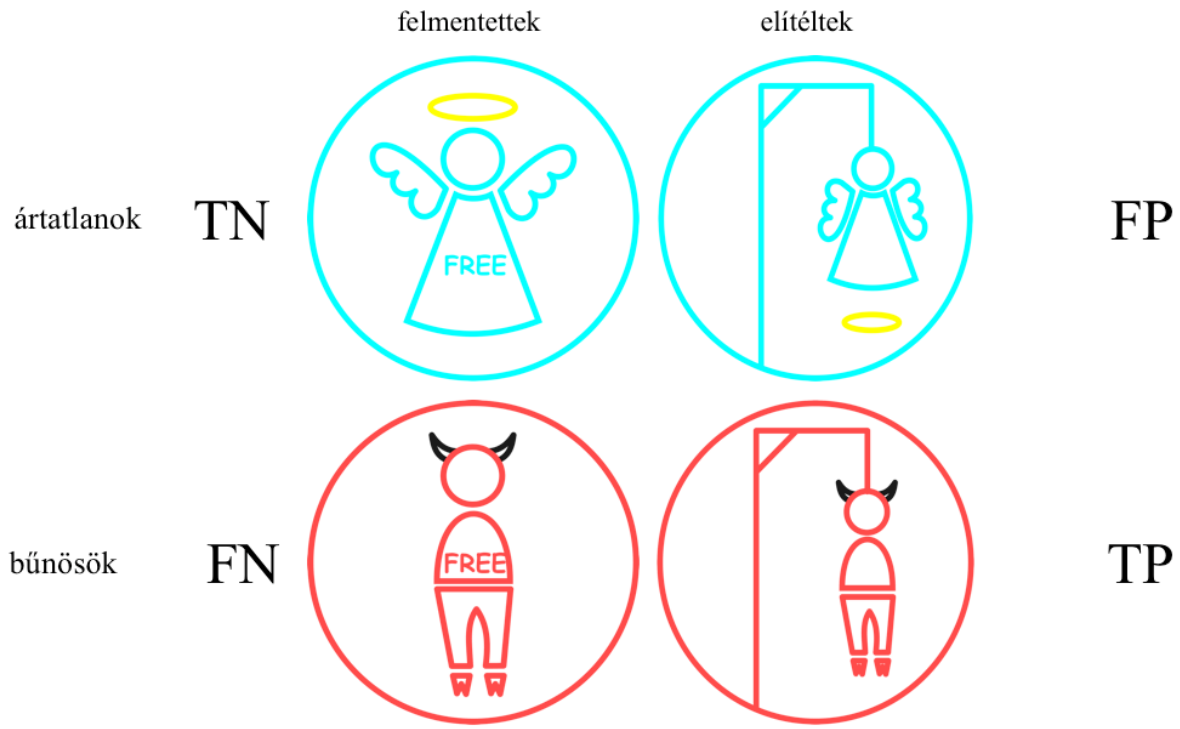
4.4.1. Az információkeresés eredményességének informatikai megközelítése

Amikor az információkeresés eredményességéről vagy hatékonyságáról beszélünk, akkor pontosan az információ-visszakeresést mint a szakirodalomban egyértelmű fogalmat vizsgáljuk. Az alábbiakban rendszerezem azokat a paramétereket, amelyekkel az információ-visszakeresés hatékonyságát mérik.[90]

A 23. ábra a fellelhető dokumentumok négy kategóriáját igyekszik illusztrálni egy rendvédelmi-igazságszolgáltatási kontextusban. A cél nyilván az, hogy minél több bűnöst elítéljenek, és az elítéltek között minél kevesebb legyen az ártatlan. Természetesen az illusztrációnak magához az igazságszolgáltatás hatékonyságához semmi köze, kizárólag a lehetséges kimeneteket szemlélteti:

- TP (*true positive*): a helyes találat, azaz a bűnöst elítélik;
- FT (*false positive*): a helytelen találat, azaz az ártatlant elítélik;
- TN (*true negative*): a helyes kihagyás, azaz az ártatlant felmentik;
- FN (*false negative*): a helytelen kihagyás, azaz a bűnöst felmentik.

[90] [12] 69. oldal



23. ábra: a találati hatékonyság mérése.[91]

A fedés vagy felidézés (*recall*) azt méri, hogy az összes dokumentumban megtaláltak között mennyi a helyes találat. Az alábbi egyenlet könnyen érthető, hiszen a megtalálható releváns (megtalálandó) dokumentumok két halmazból állnak: a helyesen megtaláltak és a helytelenül meg nem találtak. Igazságügyi hasonlatunkkal élve: a helyesen felmentettek arányát nézzük az összes felmentetthez mérve.

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

A pontosság (*precision*) azt méri, hogy a megtalált dokumentumok között mennyi a helyes találat. Az alábbi egyenlet is könnyen érthető, hiszen a megtalált dokumentumok két halmazból állnak: a helyesen megtaláltak, és a helytelenül megtaláltak. Igazságügyi hasonlatunkkal élve: a helyesen felmentett ártatlanok arányát nézzük az összes ártatlanhoz (tehát az ártatlanul elítéltekhez is) viszonyítva.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Az F-mérték (F-measure) a pontosság és a fedés harmonikus közepe, azaz

[91] Forrás: szerző.

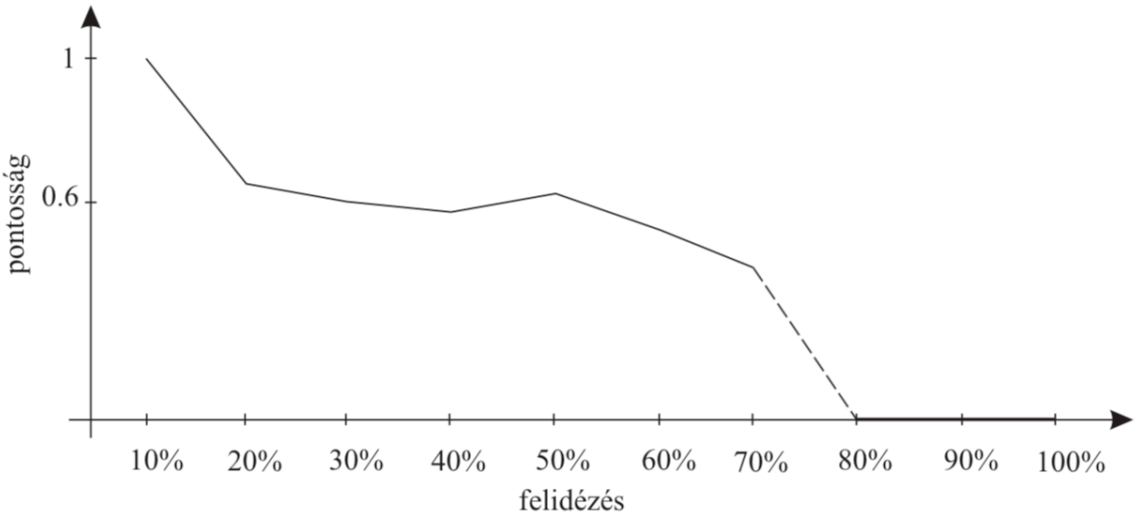
$$F = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

Előnye, hogy két paraméter helyett csak egyet mér.

A fedés és a pontosság közötti összefüggést érdemes értelmezni, ugyanis van gyakorlati jelentősége a keresés hatékonyságának a hétköznapi megítélésében. Mint az 1. grafikonon is jól látható, a két paraméter bizonyos értelemben egymással fordítottan arányos. Ha ugyanis minden dokumentumot találatként értékelünk, akkor a fedés teljes, de a pontosság nyilván alacsony, hiszen sok hamis találat is lesz (másképpen kifejezve nagy a zaj). Ez a függvény jobb széle. Ha akár csak egyetlen helyes találatunk van, akkor a pontosság teljes, viszont nagyon sok jó találatot kihagytunk, azaz alacsony a fedés. A mindennapi életben ezt úgy éljük meg, hogy ha szigorúra állítjuk be a keresőrendszert, akkor kevés „szemét jön be", de elmulaszthatunk értékes lehetőségeket. Ha viszont nagyon lazára kalibráljuk, akkor „bejön majdnem minden", de sok hibás találat is. Az optimális arányt egyénileg szokták megtalálni.

A felhasználó szempontjából csak indirekten fontos, de a keresés hatékonysága szempontjából lényeges a keresés memóriafelhasználása és a keresés sebessége. E két paraméter egymással ellentétesen működik, és a mai hardverárak figyelembevételével általában a memóriát bővítik a sebesség fokozásának érdekében.

Jelen dolgozat szempontjából az egyik legizgalmasabb kérdés, hogy mennyivel növeli a hatékonyságot egy szemantikus keresőrendszer használata egy pusztán karaktersorokra vagy kulcsszavakra történő kereséssel. Sajnos meg kell állapítani, hogy szemben a találati eredmények feldolgozásával, erre nem találtam kutatási eredményeket a szakirodalomban. Itt a kérdés az, hogy mennyivel javulnak a keresési eredmények, ha szemantikus eszközöket (taxonómiát vagy ontológiát) alkalmazunk a keresőrendszerben. További megválaszolatlan kérdés, hogy ezeknek a korábban tárgyalt szemantikus eszközöknek a manuális előállítása mennyivel hoz jobb eredményeket, mint azok gépi úton történő gyártása.



1. grafikon: pontosság versus felidézés.[92]

4.4.2. Az információkeresés eredményességének számviteli-kontrolling megközelítése

Az informatikai-matematikai paramétereknél jobban érdeklik egy gazdaságossági beruházást elbíráló döntéshozót a klasszikus mutatók, amelyek a keresőrendszer mint beruházás megtérülését jellemzik. Ilyen értelemben a gazdaságosság vizsgálata nem tér el más beruházásokétól, és így a számszerűsítésre a vállalati pénzügy mint szakterület teljes arzenálját alkalmazni lehet [196]. Természetesen nem feladata a jelen írásnak ezen módszertanok ismertetése, kizárólag néhány eszköz alkalmazására szorítkozom a keresőrendszerek szempontjából.

A klasszikus pénzügyi modellek lényege, hogy a megszerzett tudást, elért eredményt igyekeznek leképezni a számvitel vagy üzemgazdaság nyelvére, vagy más szavakkal, pénzben kifejezni a ráfordítandó vagy ráfordított költséget és a keletkezett eredményt. A beruházás költségoldalának (*cost*) kiszámítása sem egyszerű feladat, de lényegesen kidolgozottabb a módszertana, mint a várható előnyök, vagy még inkább a pénzben kifejezhető haszon (*benefit*) kézzelfoghatóvá tétele.

Az alább taglalt módszerek mindegyike alapvetően legfeljebb két elemből áll: a költségek és a pénzügyi előnyök (*financial benefits*) valamilyen összevetéséből. Pénzügyi előny kategóriába a következők sorolhatók [197]:

[92] Forrás: [12], 71. oldal

- megnövekedett bevétel;
- költségcsökkentés;
- költségek elkerülése;
- a beruházási igény csökkentése;
- beruházási kényszer elkerülése.

Az egyszeri (beruházás) és futó költségek megállapítása viszonylag egyszerű. Sokkal összetettebb feladat a megnövekedett „bevétel" megállapítása egy keresőrendszer esetében. Ez utóbbi még egy gazdálkodó szervnél sem könnyű, ahol a bevétel és annak növekedése objektív mércével mérhető. Még nehezebb egy államigazgatási, nemzetbiztonsági szolgálati vagy rendvédelmi szerv esetében, ahol a „bevétel" értelmezhetetlen. Itt nyilván más mérőszámokat kell alkalmazni. Gazdaságossági elemzésnél nyilván csak azokat az elemeket lehet figyelembe venni, amelyek számszerűsíthetők. Nehezebben számszerűsíthető paraméterek esetében kulcsfontosságú teljesítménymutatókat (*key performance indicator*, KPI) kell vizsgálnunk. Az alábbiakban ezeket mutatom be, majd egy esettanulmánnyal illusztrálom a bevezetett fogalmak alkalmazhatóságát.

A legalapvetőbb paraméter az alkalmazás használatának teljes költsége (*total cost of ownership*, TCO). Ezt a mérőszámot a Gartner Group vezette be 1987-ben, és azóta is az egyik nagyon gyakran használt kontrolling típusú mutató. A TCO kimutatja egy informatikai beruházás összes költségét, legyen az a kezdeti beruházásból fakadó vagy a folyamatos üzemeltetéshez köthető.

A payback megmutatja, hogy a beruházott költség mennyi idő alatt térül meg. Ha a pénzáramlást kumuláltan vizsgáljuk (*cumulated cash-flow*), grafikusan is jól ábrázolhatóan kimutatható a megtérülés hónapról hónapra. Különleges a törési pont, amely azt a pillanatot mutatja, amikor a kumulált pénzáram eléri a 0 szintet, vagy más szavakkal, a beruházás megtérül. Innen – jó esetben – a beruházás már nyereséges.

A teljes beruházás megtérülése (*return on investment*, ROI) az egyik leggyakrabban használt megtérülést mérő mutató [198]. Az arányszám megmutatja a megszerzett pénzösszeg nagyságának arányát a befektetett pénzösszeg nagyságához. Általában évekre lebontva százalékban mutatják ki.

4.4.3. Esettanulmány egy keresőrendszer mint beruházás bemutatására

A következőkben egy esettanulmány segítségével egy keresőrendszert mint beruházást mutatok be. Az általam vizsgált probléma az, hogy mennyi idő alatt térül meg egy szemantikus

keresőrendszer beruházása a magyar bírói kar számára. [199]. A felmérés még 2011-ben történt a Fővárosi Ítélőtábla bírói körében dr. Lukács Zsuzsanna elnök asszony és dr. Mohácsy Zsuzsanna elnökhelyettes asszony vezetésével. Tudomásom szerint ilyen típusú felmérés nem készült azóta sem a magyar bírói szervezetben.[93] A táblázatokban szereplő adatok a bírói kar saját becslései, amelyeket az elnök- és elnökhelyettes asszonyok verifikáltak 2012. július 5-én. Az esettanulmány modellje nem tekinthető mindenre alkalmazható bölcsek kövének, de arra alkalmas, hogy mintát mutasson egy keresőrendszerbe történő beruházás minimális megtérülésének kiszámítására. A számítás azért tekinthető konzervatívnak, mert kizárólag az időt veszi figyelembe, és más tényezőt, mint nyomtató, papír, tintapatron, felesleges gépi amortizáció stb., nem. Konkrét példát nem megjelölve ipari tapasztatból állítom, hogy ez egy nagyobb szervezetnél százmillió forintos nagyságrendet is elérhet évente. A béreket és költségeket meghagytam a 2011-es becslés szintjén, ami a mai viszonyok között már nyilván elavult, de a megváltoztatása torzítaná az eredeti arányokat, viszont a modell egy másik, konkrét felhasználása során könnyen módosíthatók.

A 8. táblázat mutatja egy bíró munkaidejének átlagos megoszlását. A sorokban láthatók a tevékenységek típusonként csoportosítva. Az oszlopok az időmegoszlást mutatják százalékos bontásban. Minden tevékenységcsoportot tovább bontottam típusonként százalékos megosztásban. Jól látható, hogy minden alcsoport elemeinek összege kiadja a 100%-ot. Például az első oszlopban (Fő tevékenységek...) 25%+30%+5%+5%+25%+5%+5% (=100%) stb. A keresésre fordított időtartamok az ötödik oszlopban (Összes keresésre...) látszanak (pl.: 3. sor: bíróelemzési szempontrendszer... 25%*55%*45%=4% kerekítve), amelyek összege a táblázat alján látható, kerekítve 21%. Ez az a kalkulált időarány, amennyit a becslések alapján egy bíró információkeresési munkával tölt el.

A 9. táblázat a teljes bírói kar költséganalízisét mutatja. Manuális kereséssel telik el az előbbiek alapján a teljes munkaidő 21%-a. Iparági becslések alapján a keresőrendszer alkalmazása a keresési időt 1:5, illetve 1:10 arányban rövidíti le [200]. A teljes 2011-es bírói létszámot (2936) beszorozva az akkori havi átlagfizetés teljes költségének (350 ezer Ft) 12-szeresével megkapjuk

[93] 2017. decemberében írt ki az Országos Bírói Hivatal (OBH) egy közbeszerzési felhívást a vizsgált témában „A bírósági határozatkereső számára öntanuló mesterséges intelligencia korlátlan ideig használható, 1 db öröklicenc beszerzése és kapcsolódó szolgáltatások 100217” címmel, amelyet 2018 áprilisában a Montana Tudásmenedzsment Kft. és az MTA SZTAKI konzorciuma nyert meg.

a teljes személyi költséget (konzervatívan nem számolva jutalmat, költségtérítést stb.), 12 331 200 Ft. Így azonal látszik, hogy a manuális keresésre fordított idő költsége ennek 21%-a, 2 620 380 000 Ft, míg ugyanez keresőrendszerrel a tizede, azaz 262 038 000 Ft, másképpen fogalmazva, a bevezetett rendszer okán a kettő különbségével (az összes keresési költség 90%-ával), azaz 2 358 342 000 forinttal kerül kevesebbe a bírói időráfordítás a keresésre. Mivel nyilván nem cél bírók elbocsátása, így a felszabadult 562 emberévet (összes megtakarítás osztva egy bíró éves költségével, (2 358 342 000/(350*12)) produktív munkára, ítélkezésre lehet fordítani, ami kerekítve 19%-os (562 emberév/2936 fő) javulását jelentené a bírói rendszer hatékonyságának.

A 10. táblázat a hatékonyságnövelést mutatja ügyfélszámban. Ha a befejezett ügyek oszlopát tekintjük, és az éves összes befejezett ügy számát (564 765) megemeljük az előbbi 19%-kal, akkor azt látjuk, hogy éves szinten átlag még 108 011 ügyet tud a bírói apparátus lezárni, ha egy hatékony információkereső rendszert alkalmaz.

A 11. táblázat már a megtérüléseket mutatja. A táblázat felső részében a bírók már ismert bérköltségén kívül 125 tudásmenedzserrel is számol. Alatta látszanak az egyszeri beruházási költségek, amelyek részletes elemzésétől eltekintek, mert az összegszerűségük a modell szempontjából lényegtelen. A következő részben évre lebontva látszanak a költségek és a korábban kiszámolt megtakarítások virtuális bevételként. A táblázat harmadik rekeszében a kumulált költségek, az éves eredmény és a megtérülési mutatók láthatók ugyancsak éves bontásban.

8. táblázat: egy bíró munkaidejének átlagos megoszlása.[94]

Feladatok	Fő tevékenységek a munkaidő százalékában	Fő tevékenységek alábontása részletes feladatokra	Feladatok további alábontása	A keresésre fordított teljes idő	Összes elemzésre fordított idő	Egyéb tevékenységre fordított idő
Tárgyalások előkészítése						
Peranyagok elemzése						
elemzési szempontrendszer kialakítása és e szerinti entitások keresése		55%	30%	4%		
összefüggések feltárása (keresés)			45%	4%		
peranyagok megismerése, összefüggések elemzése			25%		3%	
Konkrét ügyhöz hasonló fellebbviteli bírósági határozatok, gyakorlatok áttekintése						
ügyek keresése	**25%**	15%	40%	2%		
ügyek elemzése			60%	0%	2%	
Szakértői vélemények elemzése						
szakértői anyagok keresése		20%	30%	2%		
összehasonlító elemzések			70%		4%	
A konkrét eset szempontjából releváns jogszabályok keresése						
keresési idő		10%	20%	0,5%		
mérlegelés			80%		2%	
Tárgyalások lefolytatása	**30%**					**30%**
Értekezletek	**5%**					**5%**
Szakmai fórumokon való részvétel	**5%**					**5%**
Dokumentációk elkészítése, ellenőrzése						
Dokumentálandó tények, entitások összeállítása	**25%**	20%		5%		
Dokumentáció elkészítése		70%			18%	
Dokumentáció ellenőrzése		10%				3%
Statisztikák készítése						
Információk keresése	**5%**	40%		2%		
Információk kiértékelése		40%			2%	
Statisztikák, kimutatások összeállítása		20%				1%
Anonimizálás ellenőrzése						
Anonimizálandó entitások keresése	**5%**	50%		3%		
Anonimizálás elvégzése, ellenőrzése		50%			3%	
Összesen	**100%**			**21,25%**	**33,19%**	**43,50%**

[94] Forrás: a szerző.

9. táblázat: a teljes bírói kar időráfordításának költséganalízise.[95]

A teljes bírói kar időráfordításának költséganalízise		
	jelenleg, manuális kereséssel	**szemantikus keresőrendszerrel**
keresésre fordított idő munkaidő %-ban	21%	2,1%
bírók létszáma (fő)	2 936	
havi átlag bruttó bér/bíró (Ft)	350 000	
bírók bruttó éves összes személyi költsége (Ft)	12 331 200 000	
éves keresésre fordított összes költség (Ft)	2 620 380 000	262 038 000
megtakarítás (Ft)		**2 358 342 000**
megtakarítás bírói munka tekintetében (emberévben)		562
hatékonyságnövekedés mértéke		**19,13%**

10. táblázat: a bíróságok ügymenetének leképezése számokban.[96]

A bíróságok ügymenetének leképezése számokban			
	érkezett	**befejezett**	**nyitva maradt**
helyi bíróságoknál			
helyi bíróságok polgári	161 335	164 702	64 807
helyi bíróságok gazdasági	13 881	15 414	5 596
nem peres polgári és gazdasági	64 328	66 087	2 820
büntetőperes	77 980	82 676	48 752
közvádas büntető	61 510	66 337	42 894
magánvádas büntető	15 569	15 472	5 196
szabálysértés	107 276	106 910	16 366
megyei (fellebviteli) bíróságok			
polgári	18 850	18 587	5 489
gazdasági	2 393	2 192	852
nem peres polgári és gazdasági	13 697	13 212	2 245
közvádas büntető	12 472	12 033	5 592
büntető magánvádas	676	566	288
szabálysértés	1 159	1 143	46
összesen	**550 450**	**564 765**	**200 655**
éves országos átlag/ bíró	187	192	68
hatékonyságnövekedés ügyszámban		108 011	
hatékonyságnövekedés ügyszámban %		**19,13%**	

[95] Forrás: a szerző.

[96] Forrás: a szerző.

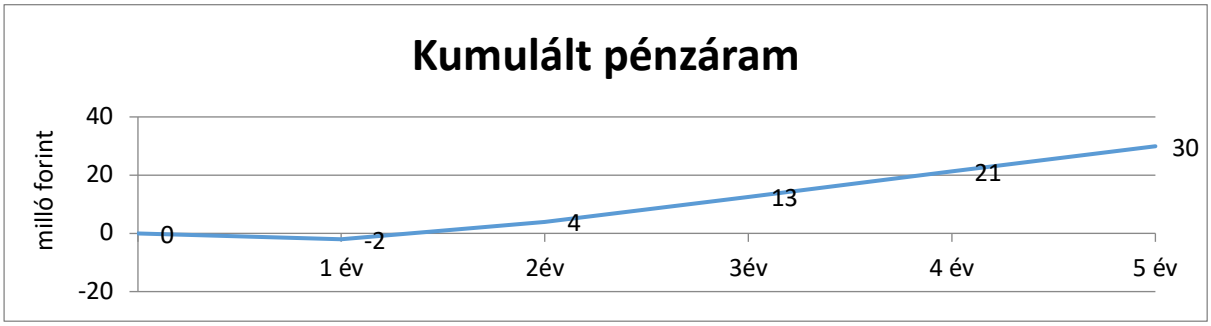
11. táblázat: bruttó költségekkel számított megtérülés áfa-visszaigénylés nélkül. Forrás: a szerző.

Bruttó költségekkel számított megtérülés áfa-visszaigénylés nélkül					
Meghatározások	*Érték*	*Mértékegység*	*Arány*		
Bírók összlétszáma	2 936	fő			
Szükséges tudásmenedzseri létszám	125	fő			
Tudásmenedzser havi bére	150 000	Ft/év			
Tudásmenedzser bruttó kiegészítő bére	225 000 000	Ft/év			
A tudásmenedzsment okán keletkezett megtakarítás	2 358 342 000	Ft/év			
Eszközök (infrastruktúra) költségei	292 100 000	Ft			
Szolgáltatási költségek	374 282 500	Ft			
Beruházási összköltség	666 382 500	Ft			
Költségek és megtakarításból származó virtuális bevétel	**1 év**	**2 év**	**3 év**	**4 év**	**5 év**
Bírói munka racionalizálásából származó megtakarítás	**0**	**2 358 342 000**	**2 358 342 000**	**2 358 342 000**	**2 358 342 000**
Tudásmenedzserek bére	-225 000 000	-225 000 000	-225 000 000	-225 000 000	-225 000 000
Egyéb adminisztrációs költségek	-90 000 000	-90 000 000	-90 000 000	-90 000 000	-90 000 000
Szoftverkövetés	0	0	33 020 000	33 020 000	33 020 000
Üzemeltetés, támogatás	0	0	-60 960 000	-60 960 000	-60 960 000
Beruházás költsége	-663 040 000	-3 342 500			
Amortizáció		-222 127 500	-222 127 500	-222 127 500	
Bevezető marketingköltség		-37 500 000	-31 250 000		
Összes költség	**-978 040 000**	**-577 970 000**	**-596 317 500**	**-565 067 500**	**-342 940 000**
Megtérülés	**1 év**	**2év**	**3év**	**4 év**	**5 év**
Kumulált költségek	-978 040 000	-1 556 010 000	-2 152 327 500	-2 717 395 000	-3 060 335 000
Éves eredmény	-978 040 000	1 780 372 000	1 762 024 500	1 793 274 500	2 015 402 000
Kumulált pénzáram (CF)	-978 040 000	802 332 000	2 564 356 500	4 357 631 000	6 373 033 000
ROI	-100%	52%	119%	160%	208%

Az elemzés során látható, hogy mind a ROI (2. grafikon), mind pedig a kumulált pénzáram (3. grafikon) az első év vége felé vált át pozitívba. Jóllehet egy valós beruházásnál sok egyéb

szempontot is figyelembe kell venni, például a kockázatelemzést, az időben változó paramétereket stb. A tapasztalt vezetők nagyon szkeptikusak az ilyen „hokiütőszerű” grafikonokkal, pedig megcáfolhatatlanul látszik, mennyire érdemes beruházni egy vállalati keresőrendszerbe.

2. grafikon: a beruházás megtérülése százalékban (ROI).[97]



3. grafikon: kumulált pénzáram becslése egy bírói keresőrendszer bevezetéséhez.[98]

4.4.4. Az információkeresés mint tudástőke-növelés értékelése

Az információigény vagy más szóval hírigény meghatározása sokkal kevésbé magától értetődő feladat, mint azt a felhasználók gondolnák. A legtöbb esetben az információ végfelhasználója, akinél a napi munkában használható tudásnak képződnie kell vagy kellene, nem képes pontosan megfogalmazni a hírigényét olyan szinten és formában, amilyet egy keresőrendszer számára szemantikai eszközökkel egzaktul meg lehetne formálni. Az elképzelések tartalmilag homályosak, sejtésekkel tarkítottak, fogalmilag pontatlanok. A végfelhasználónak továbbá sokszor fogalma sincs, hogy az egyébként sem világosan megfogalmazott kérdésekre hol lehet választ találni, pedig ez a keresőrendszerek számára létfontosságú iránymutatás. Ilyenkor a tudásmenedzser feladata a hírigény feldolgozása, a körülbelüli megfogalmazás dekompozíciója és olyan nyelvi formába öntése, amely szemantikai eszközökkel kifejezve a keresőrendszer

[97] Forrás: a szerző.

[98] Forrás: a szerző.

számára érthető és feldolgozható lesz. Egy egyszerű példa szolgáljon a piackutatás területéről. Meg kell találni azokat a kifejezéseket és azok előforduló szinonimáit, amelyek egy adott terméket, szolgáltatást, gyártót, kereskedőt stb. jellemeznek, és a céltárgy vagy -személy említésekor előfordulhatnak. Általában ismerni kell továbbá a helyeket a külső vagy belső térben, ahol a keresést meg akarjuk valósítani, hacsak nem óriáskeresőt (Google, Bing, Yandex stb.) alkalmazunk direkt vagy metakeresés során. Ezeket a paramétereket a végfelhasználó nem feltétlenül ismeri, vagy nincs ideje megfogalmazni, ami további kutatást igényelhet. Nem ritka, hogy a hírigény specifikációja során komoly emberi feszültségek keletkeznek, amelyek kezelése a tudásmenedzser számára kihívást jelenthet. A következő részben a tudás, és ezen belül a keresőrendszerek által generált plusz tudás értékelését vizsgálom.

Az eszközök klasszikus értékelése a kettős könyvelés eszközök oldalán történik. Ilyen „kézzelfogható" (*tangible*) eszköz a raktári áru, állóeszköz stb. „Kvázi kézzelfogható" eszköz a készpénz, követelés, értékpapír stb. Nem „kézzelfogható" (*intangible*), immateriális eszközök a jogok, licencek stb. Karl Sveiby [201], Polányi Mihály [202], [203] és mások munkássága mutatta ki azt a jelenséget, miszerint a vállalatok piaci értéke meghaladhatja, vagy messze meghaladhatja a könyv szerinti értéküket, amelyet az előzőek figyelembe vételével a mérleg fejez ki. Ez a különbség elsősorban a tudásalapú vállalatoknál volt észlelhető a 90-es években. Ilyenek voltak jellemzően a tanácsadó cégek, a gyógyszeripari vállalatok, a szoftverfejlesztők stb. Az eltérést az is okozta, hogy a humán tőkébe történt beruházást, így a bért, képzést, szervezetfejlesztést, toborzást stb. a számviteli törvények nem engedték aktiválni, hanem költségként kellett elszámolni. Tehát, miközben egy új számítógép vagy bútor vételára növelte a társaság saját tőkéjét, hiszen állóeszközként aktiválni lehetett vagy éppen kellett, a szakemberek megszerzésére, megtartására és képzésére fordított költségeket nem lehetett. Boda György [204] hosszan elemzi humán tudás – mint tőke – értékének kiszámítási módszereit. Ezek lényege egyrészt a költségek összevonása látens tőkeként (direkt módszer), másrészt a piaci és a könyv szerinti érték különbségének kimutatása (indirekt módszer).

A Sveiby és követői által kimutatott tudástőke főleg a humán erőforrásban mint értéktermelő potenciálban volt található. Mivel a gépi tudásba, így a klasszikus és a mesterségesintellgiencia-alapú keresőrendszerekbe fektetett tőke teljesen analóg módon kezelhető, mint a humán tőke, a fenti gondolatot időszerű kiegészíteni egy újabb elemmel. **Ez alapján állítom, hogy az elmúlt 20 évben a vállalatokban és állami szervezetekben felépített, informatikai eszközökben tárolt és kereshető tudásvagyon a humán tudással összemérhető és egyre növekvő arányú**

értéket jelent, amelyet önállóan kell értékelni és kimutatni. Az értékelés módszertana hasonló az emberi tudástőke értékeléséhez.

Az előző két fejezetrészben összegyűjtöttem az interaktív információ-visszakeresés (*interactive information retrieval*, IIR) hatékonyságának mérhetőségét matematikai-informatikai, kontrolling és számviteli szempontból. De vannak olyan korántsem elhanyagolható szempontok, amelyeket nem lehet mindig direkt paraméterek alapján mérni. Ha nincs lehetőség egzakt mérésre, általában kérdőíves vagy közösségi ötletbörzemódszerek (*crowdsourcing*) alkalmazhatók. A kulcsfontosságú teljesítménymutatók (*key performance indicators*, KPI):

- nyílt forrású vagy gyártói tulajdonban van-e a technológia (a nyílt forrás nagyobb szabadságot ad a képességek kiválasztásánál és kiépítésénél, de ugyanakkor nagyobb a szakemberigénye is a megvalósításnak és karbantartásnak);
- belső és külső szakértelem igénye;
- milyen technológiára épül a rendszer;
- skálázhatóság, amely függ az adatforrások számától, a rekordszámtól, a feltöltések (*updates*), lekérdezések számától, és ettől függ a rendszer konfigurációja;
- a konnektorok elérhetősége, illetve kifejlesztési költségei; a konnektorok összekötik a keresőt az adatforrásokkal, és biztosítják az elérhetőséget, az adatminőséget és a biztonságot;
- kész taxonómiák elérhetősége, illetve kifejlesztési költsége;
- az indexelés sebessége és minősége (új rekordok esetén újra kell-e indexelni a teljes állományt vagy sem);
- a lekérdezés minősége kulcsszóra, Boole-operátorokon kívül milyen művelet lehetséges, taxonómia megjelenítése, metszetek (*facets*) használhatósága, két- vagy többértelműség feloldása (*disambiguation*), megjelenítés a relevancia sorrendjében stb.;
- logelemző funkciók használata (lekérdezés, eredmény, kattintás);
- biztonság (belépés autentikációja, siló-, dokumentum- vagy mezőszintű hozzáférés).
- az átkattintások száma a keresett cél eléréséig;
- felhasználói elégedettség (*user experience*, UX);
- ügyfél-elégedettség;
- a keresési idő a cél eléréséig;
- ergonómia;
- a navigáció minősége;

- eredménytelen keresések aránya;
- interoperabilitás.

4.5. Összefoglalás, részkövetkeztetések

Kutatási eredményeim alapján megfogalmaztam azt a következtetést, hogy az információkereséshez kapcsolódóan számos kérdésben **szemben állnak egymással a nemzetbiztonsági szolgálatok és rendvédelmi szervek alaprendeltetésükből fakadó igényei, és az érvényben lévő jogi keretek**. Megvilágítottam, hogy ez utóbbiak alapvetően a személyiségi jogokra, az információs önrendelkezés jogára és a személyi adatok védelmére alapulnak. Vizsgálataim alapján arra a következtetésre jutottam, hogy az információkeresés szempontjából a legfontosabb **korlátok a következőek: az információkeresés célhoz kötöttségének előírása, az adatbázisok összekapcsolásának korlátai, a megszerzett adatok tárolásának időkorlátai, valamint a profilalkotás korlátai**. Rávilágítottam arra, hogy **a felesleges célhoz kötött keresés nagyobb beavatkozást jelent az emberek magánszférájába, mint a keresési folyamat előkészítő fázisaként végrehajtott tömeges keresés**. Az információkeresési technológiákból kiindulva arra a következtetésre jutottam, hogy **a korábbi keresések adatainak**, illetve az ezek alapján kialakított mintáknak (akár személyiségprofiloknak) **megőrzése nélkül a korszerű, gépi tanulásra épülő keresési módszerek szinte nem alkalmazhatóak**. Kimutattam, hogy **a jogi korlátok nemzeti jellegénél fogva azok érvényesítése komoly nehézségekbe ütközhet**.

Az érvényben lévő **magyar szabályozás**t elemezve megállapítottam, hogy az **nem követte a technológiai változásokat**, és ez nehezíti a nemzetbiztonsági szolgálatok és rendvédelmi szervek alaprendeltetés szerinti tevékenységét. A jogi környezet változtatásának fő irányai javaslatom szerint a következőek lehetnek. **Az Alkotmánybíróság** az 15/1991. (IV.30.) határozatát vissza kellene vonja, és újra kell fogalmazza. **Ki kell mondja, hogy a feltárt négy technológia használata törvényes keretek között megengedett**. **Szükségesnek tartom a nyílt forrás és a nyílt forrású keresés jogi definíciójának megalkotását és elfogadását** (lásd 4. új tudományos eredmény).

A jogi keretek elemzésének részeként szakmai tapasztalataim és a szakirodalom feldolgozása alapján kimutattam, hogy **a fékek és ellensúlyok rendszere nem gyakorol kellő ellenőrzést a nemzetbiztonsági szolgálatok és rendvédelmi szervek felett, ha az ellenőrzést gyakorló szervezetekben a személyi feltételek nem adottak**. Ugyanakkor **arra a következtetésre jutottam, hogy lehet olyan megoldást találni, amely növeli az állampolgári bizalmat,**

valamint a nemzetbiztonsági szolgálatok és rendvédelmi szervek munkájának hatékonyságát. Rámutattam, hogy **a tömeges keresés kisebb mértékben avatkozik be a magánszférába, mint a célirányos keresés**. Az információkeresés esetében ez a fennálló korlátozások enyhítésére, de emellett a számonkérhetőség jogi kereteinek, eljárásrendjének bővítésére épül. Ugyanakkor alkotmányos garanciákat kell biztosítani a visszaélés megakadályozására. **Az én javaslatom a meglévő garanciák mellé egy független, széles körben elfogadott integritású személyekből álló civil kontroll bevezetése**. Konkrét **technikai megoldásokat javasoltam** az ellenőrzés megvalósítására (lásd 5. új tudományos eredmény).

Kutatásom során megfogalmazódott az a következtetés is, hogy **az információkeresés eredményes és hatékony szervezeti célú alkalmazásához személyi és biztonsági feltételek szükségesek**. A személyi feltételek közé tartozik elsősorban az információkereséshez kapcsolódó szervezeti szerepkörök azonosítása, és az egyes szerepkörökhöz illeszkedő felkészítés végrehajtása. A vizsgálataim alapján levonható az a következtetés is, hogy az eredményes alkalmazás legfontosabb – feloldandó, csökkentendő – **akadályai** lehetnek: az ismeretlentől való félelem; az ellenérdekeltség, a tudás másokkal történő megosztására; az állás elvesztésétől való félelem; valamint az úgynevezett vezetői vakfoltok, amelyeknek lényege az információkeresés és az erre épülő elemzések, előrejelzések eredményeinek prekoncepciók miatti elvetése. **Minden akadály leküzdésére javasoltam ipari tapasztalatból származó konkrét módszereket** (lásd 6. új tudományos eredmény).

A szervezeti információkereséshez kapcsolódó biztonsági kérdések a keresés alapjául szolgáló információkhoz, valamint a keresőrendszer által használt belső információkhoz való hozzáférés köré csoportosulnak. Fontos következtetésként fogalmazható meg, hogy egy keresőrendszer egyediségét adó, minőségét meghatározó taxonómia/ontológia kiemelt védelmet igényel. Rámutattam, hogy **az információkeresési tevékenység rejtése, álcázása** is speciális biztonsági kérdéskör. Ennek jelentős része **megegyezik az informatikai vagy az interneten folytatott tevékenység rejtésével, álcázásával**. Végül **következtetés**ként fogalmazódott meg, hogy az információkeresés során használt egyik technológiai funkció (**entitáskiemelés) eredményesen használható fel az automatizált anonimizálás** (természetes személyhez köthetőség megszüntetésének) **támogatására**.

Kutatásaim fontos **következtetése**, hogy **az információkeresés** szélesebb körű **elterjedésének feltétele a szervezeti működésre gyakorolt hatás, illetve gazdaságosság értékelési lehetősége**. A keresőrendszerek működésének értékelésére nem áll rendelkezésre egységes

értékelő módszer. Következtetéseim szerint az értékelési szempontok főbb típusai három csoportba sorolhatóak: informatikai, számviteli és szubjektív felhasználói szempontok. Ezek értékelési módszerei – elsősorban a teljesítménymérés módszertanára épülve – a szakirodalomban feldolgozottnak tekinthetőek, azonban **az információkeresés speciális gazdaságossági mérései lényegében hiányoznak**.

Kutatásaim során **meghatároztam** az információkereséshez kapcsolódó **kezdeti és folyó költségként figyelembe vehető tételeket**. **Megalkottam egy általános modellt az információkeresés gazdaságosságának értékelésére**, és ezt kísérleti vizsgálattal ellenőriztem a bírói tevékenység esetében (lásd 7. új tudományos eredmény). A modell eredményei azt bizonyították, hogy korszerű információkeresési módszerek, eszközök bevezetésének beruházása 9-12 hónap alatt megtérülne. Egyben bizonyítottam, hogy egy szemantikus keresőrendszer bevezetése a bírói munka hatékonyságát drasztikusan növelné azáltal, hogy a lecsökkent manuális keresési időt a bírók érdemi tevékenységre tudnák fordítani.

ÖSSZEGZETT KÖVETKEZTETÉSEK

A XXI. század elejére az információtechnológia és annak felhasználása paradigmaváltáson megy keresztül. Szinte minden új adat elektronikus formában keletkezik, ezek tárolása és kommunikációja globálisan és költséghatékonyan megoldott. A IV. ipari forradalom hajnalán az egyik legnagyobb kihívás az adatok óceánjából kiszűrni a felhasználó számára releváns részt, és tenni ezt gyorsan, hatékonyan, lehetőleg minél kevesebb emberi munkával. Kutatásom célja az e kihívások megoldására kialakított technológiák alkalmazhatóságának feltárása a nemzetbiztonság és rendvédelem, az államigazgatás és az üzleti szféra területén. Célom az információkeresés gyökereitől kezdve feltárni az alkalmazáshoz szükséges legmodernebb technológiák ismertetését, így a számítógépes nyelvészeti és az matematikai-informatikai hátteret, majd bemutatni ezen technológiák néhány kurrens alkalmazását előbb technológiai, majd alkalmazói szemszögből. Ezt követve felhasználói oldalról közelítem meg a témát nemzetbiztonsági és rendvédelmi, államigazgatási szempontból, különös figyelemmel a jogra, végül az üzleti élet aspektusából. Végül az ágazati megközelítés után a szervezeti oldalról vizsgálom az információkeresést jogi, üzemgazdasági és szervezetbiztonsági szempontból.

Az értekezésemet négy alapvető fejezetre bontom.

Az értekezés első fejezetében az információkeresés elméleti hátterét, a tudásreprezentáció eszköztárát, a további kifejtésekhez szükséges számítógépes nyelvészeti és matematikai-informatikai eszköztárat mutatom be, utalva a legújabb mélytanulásos módszerek diszruptív hatásaira. **Rámutattam arra, hogy az exponenciálisan növekvő, 80-85%-ban strukturálatlan szövegmennyiség kereshetősége megkerülhetetlen technológiai kihívás**. A tudásreprezentálás eszközeit végigvizsgálva **arra a következtetésre jutottam, hogy** bár a klasszikus, morfológiai eszközöket és kontrollált szótárakat alkalmazó technológiák háttérbe szorultak a modern statisztikai és mélytanulásos módszerekkel szemben az internetes keresések esetében, **a vállalati keresőknél az emberi munkával generált taxonómiák és ontológiák még sokáig szerepet kapnak a nagyobb pontosság és felidézés elérése érdekében**.

A második fejezetben három alapvető technológián keresztül mutattam be az információkeresés alkalmazhatóságát. **A három technológia** karakterében eltérő, de közös bennük a mindennapi életben való **felhasználhatóság**uk, amit mindhárom esetben **példákkal illusztráltam**. A metakeresés fejezetrészben osztályoztam a technológia fajtáit, majd egy esettanulmányban bemutattam az alkalmazhatóságát az illegális gyógyszerkereskedelem feltárásása. Utaltam arra,

hogy a módszer alkalmazható kábítószer, csempészáruk internetes hirdetéseinek begyűjtésére is, ezt a módszert korábban konferenciákon is bemutattam.

A harmadik fejezetben az információkeresést a felhasználói területek szempontjából vizsgáltam. Ezek a védelmi szféra, a közigazgatás és az gazdasági élet. Az OSINT elemzése során a modern angolszász szakirodalom alapján **újraértelmeztem a nyílt forrású keresés fogalmát**. Értelmezésem szerint az OSINT lényege ma már nem feltétlenül a jogszerűség, hanem a mindenki számára való elérhetőség. **Éles határt vontam az aktív és a passzív beavatkozások közé**. Tisztában vagyok azzal, hogy ezzel a felfogással nem minden iskola ért egyet, de állítom, hogy bárki, aki a korszerű technológiák fejlődését figyeli és gyakorolja, az érzékeli a súlypont eltolódását.

Ugyancsak a harmadik fejezetben vizsgáltam a magyar jogi környezet hiányosságait a közadatok kereshetőségének vonatkozásában. **Feltártam a jogi hiányosságokat**, amelyek miatt nem teljes mértékben kerülnek átvételre az EU-irányelvek a magyar törvényekben. **Rámutattam a végrehajtási utasítások hiányára**, valamint a törvényt végre nem hajtók szankcionálásának elmaradására. **Ajánlást dolgoztam ki a szükséges változtatásokra**.

Újradefiniáltam a gazdasági hírszerzés és az üzleti hírszerzés fogalmát. Tisztáztam a fogalmi zavarokat az előbbi kettő, valamint az ipari kémkedés és az üzleti intelligencia vonatkozásában. Rendszereztem az információkeresés alkalmazható technológiáit a gazdasági hírszerzés területén.

A negyedik fejezetben részletesen **megvizsgáltam a magyar jogi környezetet az információkeresés szempontjából**. **A vizsgálat négy, eddig a szakirodalomban nem vagy csak részben tárgyalt szempont szerint történt**: a célhoz kötöttség szükségessége, az adatbázisok összekapcsolhatóságának engedélyezése vagy tiltása, az adatállomány tárolásának időbeli korlátozása, valamint a profilalkotás szabadsága. **Bebizonyítottam, hogy a magyar jogrendszer lényeges elmaradásban van a korszerű technológiai elvárásoktól**. Feltártam, hogy radikális változás híján a nemzetbiztonsági szolgálatok és rendvédelmi szervek rövid időn belül képtelenek lesznek az elvárt feladataik teljesítésére, ha a magyar jogrendszer nem követi a nemzetközi trendeket, és nem felel meg a technológiai fejlődés adta kihívásoknak. **Ajánlásokat dolgoztam ki a szükséges változtatásokra**.

A jogi környezet hazai feltárása után alapos elemzésnek vetettem alá az ún. Anderson-jelentést, amely az Egyesült Királyság helyzetét vizsgálja részben az információkeresés szempontjából. A leglényegesebb **megállapítás az, hogy a tömeges keresés kisebb mértékben sérti a**

személyiségi jogokat, mint a célhoz kötött keresés, mert – bár sokakat vizsgál felszínesen – a privát szférába csak ott hatol be mélyen, ahol arra a tömeges keresés alapján erős gyanú ébredt, így nem zaklat feleslegesen olyanokat, akiknek semmi közük a vizsgált ügyhöz.

Végül a nemzetközi tudományos szakirodalom alapján **rendszereztem azokat a módszereket, amelyekkel a nemzetbiztonsági szolgálatok és rendvédelmi szervek számonkérhetőségét javítani lehet** avégett, hogy ne a technikai képességeik visszafogásával próbálják a személyiségi jogok biztosítását elérni. **Javaslatot tettem a magyar viszonyoknak megfelelő változtatásokra**.

Új eredménynek tartom, hogy feltártam és rendszereztem azokat a faktorokat, amelyek egy keresőrendszer bevezetésének és egy szervezet által történő elfogadásának a korlátai. Megállapítottam, hogy **a siker fő akadályai elsősorban nem a technológiai korlátok, hanem az emberi tényezők**. **Hosszú ipari tapasztalatom alapján javaslatot tettem minden egyes tényező esetében az akadály leküzdésére.**

A keresőrendszerek gazdaságosságának vizsgálata során használható **modellt mutattam be egy keresőrendszer mint beruházás megtérülésének kiszámítására** a megtakarított professzionális idő produktív felhasználásának figyelembe vételével. Mivel egy tudásalapú szervezetben, mint egy jogi iroda, bíróság, ügyészség, piackutató intézet stb. a magasan képzett személyi állomány munkaideje a legnagyobb költségfaktor, a bemutatott modell jó közelítéssel ad egy alsó határt a megtérülésre.

ÚJ TUDOMÁNYOS EREDMÉNYEK

1. Újraértelmeztem a nyílt forrású keresés fogalmát. Megállapítottam, hogy a nyílt forrású keresésnek nincs definíciója a magyar jogrendszerben. Javaslatot tettem ennek pótlására.
2. Feltártam azokat a módszereket, amelyek akadályozzák a közadatok kereshetőségét a ide vonatkozó EU-s irányelvek és a magyar jogszabályok szerint. Ajánlást dolgoztam ki a változtatásokra.
3. Újraértelmeztem a gazdasági hírszerzés fogalmát, és ebbe a kontextusba elhelyeztem az üzleti hírszerzés és ipari kémkedés fogalmait.
4. Feltártam az információkeresés magyar jogi környezetét az általam kidolgozott értékelési szempontok alapján, a magyar és nemzetközi szakirodalom figyelembevételével, és megállapítottam, hogy a magyar jogi környezet nem követi a technológiai változások adta követelményeket, valamint kidolgoztam ajánlásokat a változtatásokra.
5. Megvizsgáltam a nemzetbiztonsági szolgálatok és rendvédelmi szervek számonkérhetőségét a nemzetközi és hazai szakirodalom alapján, és rendszereztem azokat a módszereket, amelyek növelik egyrészt az állampolgári bizalmat, másrészt a szervezetek munkájának hatékonyságát.
6. Megvizsgáltam az információkeresés – mint a tudásmenedzsment – egy módszerének abszorpciós akadályait egy szervezetben, és javaslatokat tettem ezen akadályok elhárítására.
7. Kidolgoztam, és kísérleti vizsgálattal ellenőriztem egy jogi témájú keresés hatékonyságának és gazdaságosságának modelljét interjúk és konkrét mérési eredmények alapján. Ezek felhasználásával következtetéseket vontam le, és ajánlásokat tettem a magyar bírói rendszer technológiai megújítására.

AJÁNLÁSOK

Az értekezésben megfogalmazottak alapján a következő hasznosításokat ajánlom.

- Az értekezés szövege vagy annak részei alkalmasak egyetemi jegyzetként oktatási segédanyagnak.
- A kidolgozott gazdaságossági modell vagy annak szükség szerinti kibővített változata beruházásoknál, fejlesztéseknél alkalmazható megtérülési számításokhoz.
- Az információkeresés humán és biztonsági keretei rész ipari tapasztalataim szerint napi szinten használható rendszerszervezőknek, értékelő-elemzőnek és projektvezetőknek.
- Az információkeresés alapjai fejezet alkalmas felsővezetői továbbképzés során informatikában kevésbé gyakorlott, ezért túlzottan is óvatos és visszafogott felsővezetők látókörének szélesítésére.
- Javaslom, hogy a 2015. évi XCVI. törvényt módosítsák két irányban. Egyrészt legyen kötelező a közérdekű adatok rövid határidőn belüli, gépi úton olvasható és kereshető formában történő közzététele. Másrészt szankcionálják ennek a kiegészítésnek a be nem tartását.

A TÉMAKÖRBŐL KÉSZÜLT PUBLIKÁCIÓIM

Lektorált folyóiratban megjelent cikkek

Vadász Pál, Séllei Márton

Az információkeresés magyar jogi környezete

HADTUDOMÁNY: A MAGYAR HADTUDOMÁNYI TÁRSASÁG FOLYÓIRATA 27:(1-2) pp. 178-191. (2017)

Vadász Pál

Információkeresés a nyílt forrású hírszerzésben

FELDERÍTŐ SZEMLE XIV:(1) pp. 81–100. (2015)

Vadász Pál

Semantic technologies in sentiment analysis

BOLYAI SZEMLE XIV:(4) pp. 42–51. (2015)

Vadász Pál

INFORMÁCIÓKERESÉS A GAZDASÁGI HÍRSZERZÉSBEN

HADMÉRNÖK IX.:(2.) pp. 343–357. (2014)

Vadász Pál

A metakeresés alkalmazása a bűnüldözés és felderítés világában

NEMZETBIZTONSÁGI SZEMLE (ONLINE) II:(2) pp. 58–71. (2014)

Vadász Pál

Case study for measuring the feasibility of a semantic search system

HADMÉRNÖK VII:(2) pp. 405–415. (2012)

Vadász Pál

Egy nyílt forrásokra épített szemantikus keresőrendszer bemutatása

HADMÉRNÖK VII:(2) pp. 351–359. (2012)

Idegen nyelvű kiadványban megjelent cikkek

Vadász P, Benczúr A, Füzesi G, Munk S: Identifying Illegal Cartel Activities from Open Sources In: Akhgar B, Bayerl P S, Sampson F szerk.: Open Source Intelligence Investigation:

From Strategy to Implementation. 304 p. Cham (Svájc): Springer, 2016. pp. 251–273., (Advanced Sciences and Technologies for Security Applications) ISBN: 978-3-319-47670-4)

Konferenciakiadványban megjelent előadás

Vadász Pál

A Case Study on Finding Fraudulent Practices in the Public Procurement Process Using Text-Mining Methods from Open Internet Sources in: Alexander Balthasar, Blaž Golob, Hendrik Hansen, Robert Müller-Török, András Nemeslaki, Johannes Pichler, Alexander Prosser szerk.: Central and Eastern European eIDem and eIGov Days 2016: Multi-Level (e)Governance: Is ICT a means to enhance transparency and democracy?, 607 p., Konferencia helye, ideje: Budapest, Magyarország, 2016.05.12-2016.05.13. Wien: Austrian Computer Society, 2016. pp. 471–480. ISBN 978-3-903035-11-9

KÖSZÖNETNYILVÁNÍTÁS

Mindenekelőtt szeretném megköszönni Munk Sándor professzor úrnak az elmúlt hat év során nyújtott szakmai irányvezetésért, és a töretlen emberi támogatásért.

Támaszkodtam dr. Kovács Zoltán, dr. Vas Zsolt, dr. Péterfalvy Attila, dr. Lévay Szabolcs dr. Bálint László, Hering Ivetta és Gyuris András értékes útmutatásaira.

Köszönettel tartozom családomnak a megértésért, amiért olyan sokszor nem tudtam elég aktívan részt venni a közös életünkben.

Nem szeretnék megfeledkezni a kollégáimról (különösen Nagy Dánielről) sem, akik ötletekkel, magyarázatokkal segítettek fel a botladozásaimkor.

HIVATKOZOTT IRODALOM

[1] BOVERO, Silvia: *Big data analytics*, 2016, forrás: https://www.slideshare.net/silviabovero2/rulex-big-data-and-analytics, (letöltés ideje: 2017.10.30.)

[2] FIRESTONE, Joseph. M.: *Enterprise Information Portals and Knowledge Management*, Butterworth-Heinemann (Elsevier utánnyomás), Burlington, USA, 2003, ISBN 0 7506 7474 1

[3] FAJSZI Bulcsú et al.: *Üzleti haszon az adatok mélyén*, Alinea, Budapest, 2010, ISBN 978 963 9659 46 9

[4] MUNK Sándor: *Az információkeresés alapjai*, Hadmérnök, 2013. 1. sz., ISSN 1788-1919, forrás: www.hadmernok.hu, (letöltés ideje: 2014. 01.03.)

[5] *How many pages has Google indexed?*, 2017, forrás: https://www.quora.com/How-many-pages-has-Google-indexed, (letöltés ideje: 2018.01.28.)

[6] ROWLEY, Jennifer: *The wisdom hierarchy: representations of the DIKW hierarchy*, Journal of Information Science 2007; 33; 163 originally published online Feb 15, 2007; forrás: http://inls151f14.web.unc.edu/files/2014/08/rowleydikw.pdf, (letöltés ideje: 2018.04.20.)

[7] WEINBERGER, David: *The Problem with the Data-Information-Kowledge-Wisdom Hierarchy*, Harvard Business Review, 2010, Forrás: www.hbr.org/2010/02/data-is-to-info-as-info-is-not/, (letöltés ideje: 2018.05.08.)

[8] COPESTAKE, Ann: *Natural Language Processing*, *Lectures,* University of Cambridge, 2003-2004, forrás: https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf, (letöltés ideje: 2018.05.06.)

[9] SAUSSURE, Ferdinand de (1998): *Bevezetés az általános nyelvészetbe*, Corvina, Budapest, ISBN 963 13 4414 2, 91. oldal

[10] MUNK Sándor: *Szemantika az informatikában*, Hadmérnök, 2014 IX. 2., ISSN 1788-1919, 311-331. oldal

[11] Techopedia: *Semantic search*, forrás: https://www.techopedia.com/definition/23731/semantic-search, (letöltés ideje: 2018.05.06.)

[12] TIKK Domonkos et al.: *Szövegbányászat*, Typotex, Budapest, 2007, ISBN 978 963 9664 456

[13] *ISO/IEC 11179 Information Technology – Metadata Registries (MDR) – Part 1: Framework* – International Organization for Standardization, Geneva, 2004, 4. oldal, forrás:

https://www.ftb.ca.gov/aboutFTB/Projects/ITSP/Part_1_Framework.pdf (letöltés ideje: 2018.02.17.)

[14] *Dublin Core Metadata Intiative*, forrás: http://dublincore.org/specifications/, (letöltés ideje: 2018.02.17.

[15] BERNERS-LEE, T. – HENDLER, J. – LASSILA, O.: *The semantic Web***,** Scientific American 284 (5): 34., forrás: https://www.scientificamerican.com/magazine/sa/2001/05-01/#article-the-semantic-web, (letöltés ideje: 2013.04.27.)

[16] HORVÁTH Ádám: *Az ALIADA és az URI névkonvenció*, 2015, forrás: https://tmt.omikk.bme.hu/tmt/article/viewFile/563/528, (letöltés ideje: 2017-07-29)

[17] Wordnet, forrás: https://wordnet.princeton.edu

[18] MIHALTZ Mihály: *Magyar WordNet*, 2015, forrás: https://github.com/mmihaltz/huwn, (letöltés ideje: 2018.01.12.)

[19] LAVRENKO, Viktor: *A generative theory of relevance*, Springer, Berlin Heidelberg, 2009, ISBN 978 3 540 89363 9

[20] RIJNSBERGEN, C. J. van: *Information retrieval*, Butterworth-Heinemann, Newton, MA, USA, 1979, ISBN 0408709294

[21] KOWALSKI, Gerald: *Information retrieval Systems*, Kluwer Academic Publishers, Boston/Dordrecht/London, 1997, ISBN: 0 7923 9899 8

[22] MEADOW, Charles T. – BOYCE, Bert R. – KRAFT, Donald H. – BARRY, Carol: *Text Information Retrieval Systems*, Elsevier, Amsterdam – London – New York – Oxford – Paris – Shannon – Tokyo, 2007, ISBN 13: 978 0 12 369412 6

[23] LEXpert határozatkereső navigációs felülete, forrás: www.lexpert.hu, (letöltés ideje: 2018.01.28.)

[24] SPENCER, Stephan: *Google Power Search*, O'Reilly, Sebastopol, CA, USA, 2011, ISBN 978 1 449 31156 8

[25] MANNING, Christopher D. – SCHÜTZE, Hinrich: *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts – London, England, 1999, ISBN 978 0 262 13360 9

[26] MANNING, Christopher D. – RAGHAVAN, Prabhagar – SCHÜZTE, Hinrich: *Introduction to Information Retrieval*, Cambridge University Press, 2009, 4. oldal, forrás: www.informationretrieval.org/ (letöltés ideje: 2017.09.03.)

[27] SCHÜTZE, Hinrich – LIOMA, Christiana: *Introduction to Information Retrieval*, 2011, forrás: https://nlp.stanford.edu/IR-book/ppt/01intro.pptx, (letöltés ideje: 2018:01.12.)

[28] NAGY T. István – FARKAS Richárd: *Személynév-egyértelműsítés a magyar weben*, Szegedi Tudományegyetem, Szeged, 2010, forrás: http://www.textrend.org/publications/49.pdf, (letöltés ideje: 2018.05.11.)

[29] MUNDIE, D. A. – MCINTIRE, D. M.: *An Ontology for Malware Analysis*, International Conference on Availability, Reliability and Security, ARES, 2013, 556–558.oldal, forrás: http://doi.org/10.1109/ARES.2013.73, (letöltés ideje: 2017.11.03.)

[30] Az Infovadász keresőrendszer. Az Infovadász a HMEI Zrt. tulajdona 2015.07.07-től.

[31] HEDDEN, Heather: *The Accidental Taxonomist*, Information Today Inc., Medford, New Jersey, 2016, ISBN 978 15387 5288

[32] *DR. Search*: *What's the difference between Taxonomies and Ontologies?*, 2014, forrás: http://www.ideaeng.com/taxonomies-ontologies-0602 (letöltés ideje: 2013-04-28)

[33] GILCHRIST, Alan: *Thesauri, taxonomies and ontologies – an etymological note*, Journal of Documentation, Vol. 59, Issue 3., 2003, ISSN: 0022-0418, 7–18. oldal

[34] KŐ Andrea: *Az információtechnológia szerepe és lehetőségei a tudásmenedzsmentben*, PhD-értekezés, BCE, 2004, 17. oldal, http://phd.lib.uni-corvinus.hu/183/ (letöltés ideje: 2017.07.29.)

[35] *Gartner emerging tech curve*, forrás: https://blogs.gartner.com/smarterwithgartner/files/2015/10/EmergingTech_Graphic.pngGartnerGroup, (letöltés ideje: 2018.05.05.)

[36] SZABÓ, M.K. – VINCZE V.: Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai. In: Tanács A., Varga V., Vincze V. (szerk.) *XI. Magyar Számítógépes Nyelvészeti Konferencia* (MSZNY 2015). Szeged, Szegedi Tudományegyetem, ISBN: 978 963 306 359 0

[37] SZASZKÓ Sándor – SEBŐK Péter – KÓCZY László (2009): *Magyar szövegek véleményanalízise*, forrás: http://www.inf.u-

szeged.hu/projectdirs/mszny2009/MSZNY2009_press_b5_mod_opt.pdf, (letöltés ideje: 2015.09.29.)

[38] FELDMAN, Ronen (2013): *Techniques and Applications for Sentiment Analysis*, Communications of the ACM, Vol. 56, No. 4, 2013, ISSN 0001 0782, 82–89. oldal

[39] BING, Liu: *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012, Vol. 5. No.1., 161–167. oldal, ISBN 978 160 845 88 44

[40] RUPPENHOFER, Josef – REHBEIN, Ines: *Semantic frames as an anchor representation for sentiment analysis,* WASSA '12 Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, forrás: http://www.uni-hildesheim.de/ruppenhofer/pubs/longversion.pdf, (letöltés ideje: 2018 01.07.)

[41] PANG, Bo – LEE, Lillian: *Opinion Mining and Sentiment Analysis*, Proceedings of the 42nd Annual Meeting on Association for Computer Linguistics, Article No. 271, Foundations and Trends in Information Retrieval, Volume 2, Issue 1-2, 2008, 271–278. oldal, forrás: http://www.cs.cornell.edu/home/llee/omsa/omsa-published.pdf, (letöltés ideje: 2018.01.07.)

[42] VIRMANI, Deepali – MALHOTRA, Vikrant – TYAGI, Ridhi: *Aspect Based Sentiment Analysis to Extract Meticulous Opinion Value*, International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3262–3266. oldal

[43] EKMAN, Paul: *Emotion in the human Face*, Malor Books, Los Altos, California, 2013, ISBN 978 933779 82 9

[44] PLUTCHIK, Robert – KELLERMANN, Henry: *The theories of emotion*, Academic Press, New York – London – Toronto – Sydney – San Francisco, 1980, ISBN 0 12 558 701 5

[45] DHAWAN, Sanjeev – SINGH, Kulvinder – SEHRAWAT, Deepika: *Emotion Mining Techniques in Social Networking Sites*, International Journal of Information & Computation Technology, Volume 4, Number 12, 2014, ISSN 0974 2239

[46] A Plutchik-kerék. https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg (letöltés ideje: 2018.04.01.)

[47] OGNEVA, Maria: *How Companies Can Use Sentiment Analysis to Improve Their Business*, 2010, forrás: http://mashable.com/2010/04/19/sentiment-analysis/#Gyr3NMpX6iqk, (letöltés ideje: 2015.06.10.)

[48] JACSÓ Péter: *Internet Insights - Thoughts about Federated Searching,* forrás: http://www2.hawaii.edu/~jacso/extra/federated/federated.htm, (letöltés ideje: 2018.01.06.)

[49] LAWRENCE, S. - GILES. C.L.: *Inquirus, the NECI meta search engine*, Seventh International World Wide Web Conference, Brisbane, Australia, Elsevier Science, 1998, 95–105. oldal, forrás: http://www7.scu.edu.au/1906/com1906.htm (letöltés ideje: 2018.01.06.)

[50] WOLFF, Ch.: Effektivität von Recherchen im WWW, in KNORZ, Gerhard - KUHLEN, Rainer szerk.: *Informationskompetenz – Basiskompetenz in der Informationsgesellschaft*: proceedings des 7. Internationalen Symposiums für Informationswissenschaft, Darmstadt, 8. – 10. November 2000. Schriften zur Informationswissenschaft, 38. UVK, Konstanz, ISBN 3-87940-753-3, 31–48. oldal

[51] KERSCHBERG. L. – JEONG, H. – KIM, W.: *Emergent Semantics in Knowledge Sifter: An Evolutionary Search Agent Based on Semantic Web Services*, Journal on Data Semantics VI. Lecture Notes in Computer Science, vol 4090. Springer, Berlin, Heidelberg, 2006, ISBN 978 3 540 36712 3, 187–206. oldal

[52] JACSÓ Péter: *Trends in Professional and Academic Online Information Services*, INFORUM 2007, 13th Conference on Professional Information Resources, 2007, forrás: http://www.inforum.cz/pdf/2007/jacso-peter.pdf, (letöltés ideje: 2018.01.06.)

[53] KERSCHBERG, L.: *A semantic taxonomy-based personalizable meta-search agent*, Web Information Systems Engineering, Proceedings of the Second International Conference, Volume:1, 2001.12.3-6., ISBN 0-7695-1393-X, 41–50. oldal

[54] SADEH, T. (2006): *Google Scholar versus metasearch systems*, 2006, forrás: http://webzine.web.cern.ch/webzine/12/papers/1/index.html (letöltés ideje: 2018.01.07.)

[55] GLENN, Haya – NYGREN, Else – WIDMARK, Wilhelm: *Metalib and Google Scholar: a user study*, Online Information Review, Vol. 31 Issue: 3., 2007, 365–375. oldal, forrás: http://www.emeraldinsight.com/doi/abs/10.1108/14684520710764122, (letöltés ideje: 2018.01.07.)

[56] VADÁSZ Pál: *A metakeresés egy alkalmazása a bűnüldözés és felderítés világában*, Nemzetbiztonsági Szemle, II. évfolyam, 2. szám, 2014, HU ISSN 2064-3756, 58–71. oldal

[57] Google árlista, forrás: http://www.google.com/enterprise/search/products/gss.html#pricing_content, (letöltés ideje: 2013.02.16.)

[58] PHAM, Tien et al. (2008): *Intelligence, Surveillance, and Reconnaissance Fusion for Coalition Operations*, U.S. Army Research Laboratory, 2800 Powder Mill Road, Adelphi, MD, 20783, 2008, forrás: http://ieeexplore.ieee.org/document/4632340/, (letöltés ideje: 2017.09.03.)

[59] *Fusion Center Guidelines*, Department of Justice, Washington, USA, 2006, Forrás: https://it.ojp.gov/documents/fusion_center_guidelines_law_enforcement.pdf, letöltés ideje: 2018.05.05.]

[60] GRUSZCZAK, Artur: *Establishing an EU law enforcement fusion centre*, EUROPEAN JOURNAL OF POLICING STUDIES, 4(1), 2016, ISBN 978 90 466 0829 6]

[61] LÉVAY Gábor: *A nemzeti OSINT*, Felderítő Szemle, III. Évfolyam, 4. Szám, Budapest, 2004, ISSN 1588-242X, 27–42. oldal

[62] BÉRES János – KENEDLI Tamás: *Információs-fúziós központok és információ-megosztás a jövő útja*, Budapest, Szakmai Szemle, A KBH Tudományos Tanácsának kiadványa, 4. szám, 2009, ISSN 1785-1181

[63] VADÁSZ Pál: *Egy nyílt forrásokra épített szemantikus keresőrendszer bemutatása*, HADMÉRNÖK 2012 VII. 2., ISSN 1788-1919, 351–359. oldal

[64] TEKIR, Selma: *Open Source Intelligence Analysis, A methodological approach*, VDM Verlag Dr. Müller, Saarbrücken, Germany, 2009, ISBN 978 3 639 14036 1, 8. oldal

[65] KAHANER, Larry: *Competitive Intelligence*, Touchstone-Simon & Schuster, New York, 1997, ISBN 978 0 684 84404 6

[66] ROLINGTON, Alfred: Hírszerzés a 21. században, Antall József Tudásközpont, Budapest, 2015, ISBN 978 963 87486 3 8

[67] VIDA Csaba: Nyílt forrású adatszerzés (OSINT) in KOBOLKA István (szerk.): *Nemzetbiztonsági alapismeretek*, NKE, Nemzetbiztonsági Intézet, Budapest, ISBN 978-915-5344-32-9

[68] GROSE, Peter: *Gentleman Spy: The Life of Allen Dulles*, Boston, Houghton Mifflin, 1994, ISBN, 978-0-395-51607-2, 525–528. oldal

[69] GIBSON, Stevyn D.: *Open Source Intelligence: A contemporary intelligence lifeline*, Defence Colllege of Management and Technology, Cranfield University, PhD tézis, 2007, https://dspace.lib.cranfield.ac.uk/bitstream/1826/6524/1/PHD%20-%20Gibson,%20S.pdf (letöltés ideje: 2015.01.02.)

[70] PILCH Jenő (1936) szerk.: *A hírszerzés és kémkedés története*, Budapest, Franklin, reprint Kassák,1998, ISBN 963 9100 19 6, 79. oldal

[71] TORMA Béla: *Hírszerzés és felderítés az 1241. évi tatárjárás idején*, Felderítő Szemle, VI. évf. I. szám, 2007, ISSN 1588-242X, 152–163. oldal

[72] KRAUS, Hans P.: *Sir Francis Drake: A Pictoral Biography, The Cadiz Raid 1587*, 2010, forrás: http://www.loc.gov/rr/rarebook/catalog/drake/drake-7-cadizraid.html, (letöltés ideje: 2014.06.15.)

[73] LÉVAY Gábor: *OSINT (OPEN SOURCE INTELLIGENCE) – NYÍLT INFORMÁCIÓS HÍRSZERZÉS,* NKE, egyetemi jegyzet, kézirat, Zrínyi Miklós Nemzetvédelmi Egyetem, Hadtudományi Kar, Budapest, 2006, 7. oldal

[74] POLMAR, Norman - ALLEN, B. Thomas: *Spy Book, The encyclopedia of Espionage*, Greenhill Books (Random House), London, 1997 ISBN 0 679 42514 4, 414. oldal

[75] STEELE, Robert: Open Source Intelligence in JOHNSON, Loch K. szerk.: *Handbook of Intelligence Studies*, Routledge-Taylor & Francis, New York, 2009, ISBN 978 0 415 77783 4

[76] *NATO Open source Intelligence Handbook* (2001), forrás: http://www.au.af.mil/au/awc/awcgate/nato/osint_hdbk.pdf (letöltés ideje: 2017.11.03.)

[77] GDPR: 2016/6/679/EK, 5. cikk

[78] HOBBS, Christopher – MORAN, Matthew: Armchair safeguards: The Role of Open Source Intelligence in Nuclear Proliferation Analysis, in HOBBS, Christopher – MORAN, Matthew – SALISBURY, Daniel (szerk.): *Open Source Intelligence in the Twenty-First Century*, Palgrave Macmillan, New York, 2014, ISBN 978 1 137 35331 3, 65-80. oldal

[79] VADÁSZ Pál: A Case Study on Finding Fraudulent Practices in the Public Procurement Process Using Text-Mining Methods from Open Internet Sources in: Alexander Balthasar, Blaž Golob, Hendrik Hansen, Robert Müller-Török, András Nemeslaki, Johannes Pichler, Alexander Prosser szerk.: *Central and Eastern European eIDem and eIGov Days* 2016: Multi-Level (e)Governance: Is ICT a means to enhance transparency and democracy?, 607 p., Konferencia helye, ideje: Budapest, Magyarország, 2016.05.12-2016.05.13. Wien: Austrian Computer Society, 2016. ISBN 978-3-903035-11-9 , 471–480. oldal

[80] VADÁSZ P. – BENCZÚR A. – FÜZESI G. – MUNK S.: Identifying Illegal Cartel Activities from Open Sources In: Akhgar B, Bayerl P S, Sampson F (szerk.): *Open Source Intelligence Investigation: From Strategy to Implementation*. 304 p. Cham (Svájc): Springer, 2016.

(Advanced Sciences and Technologies for Security Applications), ISBN: 978-3-319-47670-4, 251–273. oldal

[81] GRITZALIS, Dimitris (2014): *Open-source Intelligence as a Means to reveal Insiders and protect Critical Infrastructures*, Athens, 2014 May 5th European Union – United States – Canada Experts Meeting on Critical Infrastructure Protection conference, forrás: http://www.cis.aueb.gr/Publications/EU-US-Canada-2014%20Social%20Media.pdf, (letöltés ideje: 2014.06.30.)

[82] BINDER, Clemens: *Happenings Foreseen: Social Media and the Predictive Policing of Riots*, S&F Sicherheit und Frieden, S+F, 34. Évfolyam (2016), 4, ISSN: 0175274X, 242–247. oldal

[83] WIBBERLEY, Simon – MILLER, Carl: Detecting Events from Twitter: Situational Awareness in the Age of Social Media in HOBBS, Christopher – MORAN, Matthew – SALISBURY, Daniel (szerk.): *Open Source Intelligence in the Twenty-First Century*, Palgrave Macmillan, New York, 2014, ISBN 978 1 137 35331 3, 147-167. oldal

[84] AVNER, Gabriel: *Can monitoring social media predict lone wolf attackers*, 2015, forrás: http://www.geektime.com/2015/10/13/can-monitoring-social-media-predict-lone-wolf-attackers/, (letöltés ideje: 2017.03.12.)

[85] VADÁSZ Pál: *Semantic technologies in sentiment analysis*, BOLYAI SZEMLE XIV:(4) pp. 42–51., 2015, ISSN 1416-1443

[86] SZABÓ Martina Katalin – VINCZE Veronika: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai In TANÁCS A.−VARGA V.− VINCZE V. (szerk.): *XI. Magyar Számítógépes Nyelvészeti Konferencia* (MSZNY 2015). Szegedi Tudományegyetem, Szeged. 219–226. oldal, ISBN: 978-963-306-359-0

[87] NÉGYESI Imre: *A befogadó nemzeti támogatás és a civil-katonai együttműködés (CIMIC) feladatrendszerének kapcsolata a feladat végrehajtáshoz szükséges információk tükrében*, Budapest, Nemzetvédelmi Egyetemi Közlemények, 2002. 1.sz., ISSN 1417-7323, 145–154. oldal

[88] GUPTA, Ravi – BROOKS (2013), Hugh: *Using social media for global security*, John Wiley & Sons, Indianapolis, USA, ISBN 978 1 118 44231 9, 177–326. oldal

[89] *Untagling the Web*: *An Introduction to internet Research*, Fort Meade, Maryland, 2007, https://www.nsa.gov/public_info/_files/untangling_the_web.pdf, (letöltés ideje: 2014.09.30.)

[90] SHERMAN, Chris – PRICE, Gary: *The invisible web: uncovering sources search engines can't see*, Information Today Inc. Medford, New Jersey, USA, 2007, ISBN-13 978 0910965514

[91] KALPAKIS, George et al.: OSINT and the Dark Web in: AKHGAR B. – BAYERL, P. S. – SAMPSON, F. szerk.: *Open Source Intelligence Investigation: From Strategy to Implementation*, Cham (Svájc): Springer, (Advanced Sciences and Technologies for Security Applications), 2016, ISBN: 978-3-319-47670-4, 111–132. oldal

[92] SUI, Daniel- CAVERLEE, James – RUDESILL, Dakota: *The Deep Web and the Darknet*, Wilson Center, Washington D.C., 2015. forrás: https://www.wilsoncenter.org/sites/default/files/deep_web_report_october_2015.pdf, (letöltés ideje: 2018.05.08.)

[93] *Shedding light ont he dark web*, 2016, forrás: https://www.economist.com/news/international/21702176-drug-trade-moving-street-online-cryptomarkets-forced-compete (letöltés ideje: 2017.11.12.)

[94] TORPROJECT: https://www.torproject.org

[95] BAZZELL, Michael (2016): *Open Source Intelligence Techniques*, ISBN 13 978 153 050 890 7

[96] OBH Elvi bírósági határozatok, forrás: http://www.kuria-birosag.hu/hu/elvi-birosagi-hatarozatok (letöltés ideje: 2018.04.15.)]

[97] Az Európai Parlament és a Tanács 2003. november 17-én kelt 2003/98/EK számú irányelve a közszféra információinak további felhasználásáról, forrás: https://eur-lex.europa.eu/legal-content/HU/ALL/?uri=CELEX%3A32013L0037, (letöltés ideje: 2018.05.20.)

[98] 2012. évi LXIII. törvény a közadatok újrahasznosításáról, forrás: https://net.jogtar.hu/getpdf?docid=a1200063.tv&targetdate=20160101&printTitle=2012.+évi+LXIII.+törvény, (letöltés ideje:2018.05.20.)

[99] Az Európai Parlament és a Tanács 2013. június 26-án kelt 2013/37/EK irányelve a közszféra információinak további felhasználásáról szóló 2003/98/EK irányelv módosításáról, forrás: https://eur-lex.europa.eu/legal-content/HU/ALL/?uri=CELEX%3A32013L0037, (letöltés ideje: 2018.05.20.)

[100] 2015. évi XCVI. törvény az információs önrendelkezési jogról és az információszabadságról szóló 2011. évi CXII törvény és a közadatok újrahasznosításáról szóló 2012. évi LXIII

törvény módosításáról, forrás: https://net.jogtar.hu/jogszabaly?docid=A1500096.TV&txtreferer=A1100112.TV, (lewtöltés ideje:2018.05.20.)

[101] *Thomson Reuters: Legal Taxonomy*, Forrás: http://2.sweetandmaxwell.co.uk/online/taxonomy/login.jsp (letöltés ideje: 2017.04.23.)

[102] Eurovoc taxonómia, forrás: http://eurovoc.europa.eu/ (letöltés ideje: 2017.04.23.)

[103] ALMÁSI Attila et al.: *Adó- és jövedéki jogi wordnet (TaXWN)*, SZTE, Informatikai Tanszékcsoport, 2009, forrás: http://maszeker.all.hu/download/publication/ado_es_jovedeki.pdf, (letöltés ideje: 2017.04.23.)

[104] VALENTE, A.: *Legal Knowledge Engineering*, IOS Press, Amsterdam, 1995, ISBN 90 5199 230 0, 47–81. oldal

[105] HAMP Gábor – MARKOVICH Réka: *Jogszabályok hivatkozásainak automatikus felismerése és a belső hivatkozások struktúrája*, XII. Magyar Számítógépes Nyelvi Konferencia, Szeged, 2016, forrás: http://syi.hu/pdf/MSZNY2016_linkanalysis.pdf, 220–229. oldal, (letöltés ideje: 2017.04.29.)

[106] ZŐDI Zsolt: *A korábbi esetekre történő hivatkozások mintázatai a magyar bíróságok ítéleteiben*, MTA Law Working Papers 2014/01, MTA, Budapest, 2014, ISSN 2064-4515

[107] *RIS: Das Rechtsinformationssystem des Bundes*, forrás: https://www.ris.bka.gv.at/ (letöltés ideje: 2017.05.07.)

[108] NABIZAI, Arzo – FILL, Hans-Georg: *Eine Modellierungsmethode zur Visualisierung und Analyse von Gesetztexten*, Universiaet Wien, Fakultaet Informatik, 2017, forrás: http://homepage.dke.univie.ac.at/fill/papers/Nabizai_Fill_IRIS_2017.pdf, (letöltés ideje: 2017.05.07.)

[109] STAUDEGGER, Elisabeth: *Recht ohne gratis. RIS/EUR-Lex*, Springer Wien, 2013, ISBN 978 3 211 00587 3

[110] LexPraxis, forrás: http://www.lex-praxis.hu/, (letöltés ideje: 2017.04.23.)

[111] Justeus, forrás: https://www.hmei.hu/hu/justeus.html, (letöltés ideje: 2017.05.01.)

[112] SUSSKIND, Richard – SUSSKIND, Daniel (2015): *The future of Professions*, Oxford University Press, Oxford, ISBN 978 0 19 871339 5

[113] INSALL, Jemma – BORTHAKUR, Ankur: *From Brawn to Brains, The impact of technology on jobs in the UK*, Deloitte Insight, 2016, forrás: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/Growth/deloitte-uk-insights-from-brawns-to-brain.pdf, (Letöltés ideje: 2017.05.07.)

[114] WOLFRAM, Stephen: *Computational Law, Symbolic Discourse and the AI Constitution,* 2016, forrás: http://blog.stephenwolfram.com/2016/10/computational-law-symbolic-discourse-and-the-ai-constitution, (letöltés ideje: 2016.10.12.)

[115] LOIS: forrás: *http://cordis.europa.eu/project/rcn/78314_en.html* (letöltés ideje: 2017.05.02.)

[116] CURTONI, Paolo et al.: *Semantic access to multilingual legal information*, 1997, forrás: http://www.di.uevora.pt/~pq/papers/eu-ws-lois.pdf, (letöltés ideje: 2017.05.07.)

[117] TISCORNIA, Daniela: *The LOIS Project: Lexical Ontologies for Legal Information Sharing*, Institute of Legal Information Theory and Techniques - Italian National Research Council, 2006, forrás: http://www.e-p-a-p.com/dlib/9788883980466/art14.pdf (letöltés ideje: 2017.05.07.)

[118] Estrella projekt, forrás: http://www.estrellaproject.org/ (letöltés ideje: 2017.05.07.)

[119] Lex Machina, forrás: https://lexmachina.com/, (letöltés ideje: 2018.02.11.)

[120] Ravel: forrás: http://ravellaw.com/, (letöltés ideje: 2017.0.5.07.)

[121] Ravn: forrás: https://imanage.com/product/ravn/, (letöltés ideje: 2018.02.11.)

[122] MADHUMITA, Murgia: SFO expected to promote Ravn's crime-solving AI robot, FT, 2017, forrás: https://www.ft.com/content/55f3daf4-ee1a-11e6-ba01-119a44939bb6 , 2017. 02.13. (letöltés ideje: 2017.05.07.)

[123] Luminance, forrás: https://www.luminance.com/, (letöltés ideje: 2017.05.07.)

[124] IBM Ross, forrás: https://rossintelligence.com/, (letöltés ideje: 2018.02.11.)

[125] DIETL, Wilhelm – MECK, Georg (1995): *Freund hört mit*. Forrás: http://www.focus.de/politik/ausland/industrie-spionage-freund-hoert-mit_aid_156611.html, (Letöltés ideje: 2014.05.04)

[126] *Canadian Spies*, 2013: http://www.cbc.ca/news/canadian-spies-targeted-brazil-s-mines-ministry-report-1.1927975, (letöltés ideje:2014.05.04.)

[127] Institue of Competitive Intelligence, http://www.institute-for-competitive-intelligence.com/, (letöltés ideje: 2016.10.05)

[128] SASVÁRI Rudolf: *Üzleti hírszerzés, avagy Az ügynöktartás ábécéje*, Budapest, Agave, 2006 ISBN 963-7118-26-8

[129] ORMOSY Gábor László, et al.: *Üzleti hírszerzés*, POSYS-BME Mérnöktovábbképző Intézet, Budapest, Belső tanfolyami anyag, 2002

[130] GILAD, Ben: *Early Warning*, AMACOM, New York, 2004, ISBN 0-8144-0786-2, 158. oldal

[131] PORTEOUS, Samuel (1994): Commentary No. 46: *Economic Espionage (II)*, Analysis and Production Branch of the Canadian Secrurity and Intelligence Service, Ottawa, Canada, 1994, forrás: http://www.datapacrat.com/True/INTEL/CSIS/COM46E.HTM, (letöltés ideje: 2017.11.02.)

[132] HAIG Zsolt – VÁRHEGYI István (2005): *Hadviselés az információs hadszíntéren*, Zrínyi Kiadó, Budapest, ISBN 963 327 391 9

[133] PORTER, Michael E.: *Competitive Strategy*, Free Press-Simon & Schuster, New York, 1998, ISBN 0 7432 6088 0

[134] FLEISHER, Craig. S. – BENSOUSSAN, Babette, E.: *Business and competitive analysis*, FT Press, Upper Saddle River, N.J., 2007, ISBN 0 13 187366 0

[135] FULD, M. Leonard (2006): *The secret language of competitive intelligence*, Crown Publishing-Random House, New York, ISBN 978 0 609 61089 3

[136] MICHAELI, Rainer: *Competitive Intelligence*, Springer, Berlin, 2006, ISBN 3 540 030816, 6. oldal

[137] *Börtönre ítélték ipari kémkedésért a Diagon-ügy vádlottjait,* forrás: http://www.origo.hu/itthon/20111010-bortont-es-harommilliardos-penzbuntetest-kaptak-elso-fokon-a-diagonugy-vadlottjai.html, (letöltés ideje: 2014.05.02.)

[138] DOBÁK Imre – KOVÁCS Zoltán: Új technológiák hatása a hírszerzésre, in DOBÁK, Imre szerk.: *A Nemzetbiztonság általános elmélete*, NKE, Nemzetbiztonsági Intézet, Egyetemi jegyzet, Budapest, 2014, ISBN 978-615-5305-49-8, 206–211. oldal

[139] *Bulk Collection of Signals Intelligence: Technical Options*, US National Research Council of the National Academies, Washington DC, 2015, ISBN 978 0 309 32520 2

[140] FLEINER Rita – MUNK Sándor: *Közigazgatási adatbázisok összekapcsolásának biztonsági kérdései*, Hadmérnök, VII. 4., 2012, ISSN 1788-1919, 124. oldal, forrás: http://hadmernok.hu/2012_4_fleiner_munk.pdf, (letöltés ideje: 2016. 03. 26.)

[141] 15/1991. (IV. 13.) Alkotmánybírósági határozat, forrás: http://public.mkab.hu/dev/dontesek.nsf/0/DD1D6974489E3975C1257ADA00529D49?OpenDocument (letöltés ideje: 2016.04.17.)

[142] *APEH elismerte* (2003), forrás: http://www.origo.hu/itthon/20030213nemerdeklodhetne.html, (letöltés ideje: 2017.01.24.)

[143] VADÁSZ Pál – SÉLLEI Márton: *Az információkeresés magyar jogi környezete,* HADTUDOMÁNY: A MAGYAR HADTUDOMÁNYI TÁRSASÁG FOLYÓIRATA 27:(1-2), 2017, ISSN 1215 4121, 178-191 oldal

[144] Magyarország Alaptörvénye, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A1100425.ATV (letöltés ideje: 2017.01.31.)

[145] Az emberi jogok és az alapvető szabadságok védelméről szóló, Rómában, 1950. november 4-én kelt Egyezmény és az ahhoz tartozó nyolc kiegészítő jegyzőkönyv kihirdetéséről szóló 1993. évi XXXI. törvény 8. cikk 2. pontja, forrás: http://www.lb.hu/hu/egyezmeny-az-emberi-jogok-es-alapveto-szabadsagok-vedelmerol (letöltés ideje: 2017.01.02.)

[146] 1976. évi 8. törvényerejű rendelet az Egyesült Nemzetek Közgyűlése XXI. ülésszakán, 1966. december 16-án elfogadott Polgári és Politikai Jogok Nemzetközi Egyezségokmánya kihirdetéséről https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=97600008.TVR, (letöltés ideje: 2017.01.31.)

[147] 2011. évi CXII. törvény az információs önrendelkezési jogról és az információszabadságról, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A1100112.TV (letöltés ideje: 2017.01.31.)

[148] HETESY Zsolt: *A titkos felderítés, doktori értekezés*, 2011, 87. oldal, forrás: http://ajk.pte.hu/files/file/doktori-iskola/hetesy-zsolt/hetesy-zsolt-vedes-ertekezes.pdf, (letöltés ideje: 2017.01.08)

[149] Szabó és Vissy kontra Magyarország ügy, forrás: https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-176354&filename=CASE%20OF%20SZAB%D3%20AND%20VISSY%20v.%20HUNGARY%20-%20%5BHungarian%20Translation%5D%20by%20the%20Hungarian%20Ministry%20of%20Justice.pdf, (letöltés ideje: 2018.05.08)

[150] 2/2007 (I.24.) Alkotmánybírósági határozat, forrás: http://public.mkab.hu/dev/dontesek.nsf/0/6BF683C5CD89975CC1257ADA00529A78?OpenDocument, (letöltés ideje: 2013.01.31.)

[151] Az Európai Parlament és a Tanács 1995. október 24-én kelt 95/46/EK irányelve a személyes adatok feldolgozása vonatkozásában az egyének védelméről és az ilyen adatok szabad áramlásáról, forrás: https://eur-lex.europa.eu/legal-content/HU/ALL/?uri=CELEX:31995L0046, (letöltés ideje:2018.05.20.)

[152] JÓRI András: *Az adatvédelmi jog generációi és egy második generációs szabályozás részletes elemzése*, PhD dolgozat, 2009, forrás: http://ajk.pte.hu/files/file/doktori-iskola/jori-andras/jori-andras-vedes-ertekezes.pdf, (letöltés ideje: 2016. 03.28.)

[153] HETESY Zsolt: *A büntetőeljárás szükségtelen eleme: a célhoz kötött bizonyíték elve*, PHD tanulmányok 4., Pécs, 2005., Pécsi Tudományegyetem Állam- és Jogtudományi Karának Doktori Iskolája, forrás: http://nbsz.gov.hu/docs/pub_a_be_szuksegtelen_eleme.pdf, (letöltés ideje: 2017.01.31.)

[154] 1995. évi CXXV. törvény a nemzetbiztonsági szolgálatokról, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=99500125.TV (letöltés ideje: 2017.02.05.)

[155] 2016/681/EU irányelv az utas-nyilvántartási adatállománynak (PNR) a terrorista bűncselekmények és súlyos bűncselekmények megelőzése, felderítése, nyomozása és a vádeljárás lefolytatása érdekében történő felhasználásáról, forrás: https://eur-lex.europa.eu/legal-content/HU/ALL/?uri=CELEX%3A32016L0681, (letöltés ideje: 2017.11.02.)

[156] 1994. évi XXXIV. törvény a rendőrségről, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=99400034.TV (letöltés ideje: 2017.02.05.)

[157] 2016. évi LXIX. törvény a terrorizmus elleni fellépéssel összefüggő egyes törvények módosításáról, forrás: http://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A1600069.TV&timeshift=fffffff4&txtreferer=00000001.TXT (letöltés ideje: 2017.02.05.)

[158] 2002. évi LIV. törvény a bűnüldöző szervek nemzetközi együttműködéséről, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=a0200054.tv (2017.02.05.)

[159] A Tanács 2008/615/IB határozata (2008. június 23.) a különösen a terrorizmus és a határokon átnyúló bűnözés elleni küzdelemre irányuló, határokon átnyúló együttműködés

megerősítéséről, forrás: http://eur-lex.europa.eu/legal-content/HU/TXT/?uri=URISERV%3Ajl0005 (letöltés ideje: 2017.02.05.)

[160] 2011. évi CLXIII. törvény az ügyészségről, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A1100163.TV (letöltés ideje: 2017.02.05.)

[161] 2016. évi CXVI. törvény az egyes belügyi tárgyú törvények módosításáról, forrás: https://net.jogtar.hu/jogszabaly?docid=A1600116.TV&timeshift=fffffff4&txtreferer=00000001.TXT, (letöltés ideje: 2018.05.20.)

[162] 2003. évi C. törvény az elektronikus hírközlésről, forrás: https://net.jogtar.hu/jogszabaly?docid=A0300100.TV, (letöltés ideje: 2018.0.5.20.)

[163] ZŐDI Zsolt: *Privacy és a Big Data,* Fundamentum, ELTE Társadalomtudományi Kar, Emberi Jogi Információs és Dokumentációs Központ Alapítvány, 2017, ISSN 1417-2844 , forrás*: http://fundamentum.hu/sites/default/files/teljes-szamok/fundamentum-17-1-2-beliv.pdf*, 18-30. oldal, (letöltés ideje: 2018.02.03.)

[164] 1998. évi XIX. törvény a büntetőeljárásról, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=99800019.TV (letöltés ideje: 2017.02.05.)

[165] 2001. évi CVIII. törvény az elektronikus kereskedelmi szolgáltatások, valamint az információs társadalommal összefüggő szolgáltatások egyes kérdéseiről, forrás: https://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=a0100108.tv (2017.02.05.)

[166] KOVÁCS Zoltán: *Az infokommunikációs rendszerek nemzetbiztonsági kihívásai, doktori értekezés*, NKE, HHK, KMDI, 2015, forrás: http://archiv.uni-nke.hu/feltoltes/uni-nke.hu/konyvtar/digitgy/phd/2015/kovacs_zoltan_2015.pdf, (letöltés ideje: 2018.02.10.)

[167] ANDERSON, David: *Report of the Bulk Powers Review*, Crown copyright, London, 2016, ISBN 978 1474136921

[168] JUVENALIS, Junius Decimus: *Junius Decimus Juvenalis Satírái*, ford. Barna Ignácz, Tettey Nándor és Tsa., Budapest, 1876

[169] SCHMID, Gerhard: *Report on the existence of a global system for the interception of private and commercial communications (ECHELON interception system)*, 2001, forrás: http://cryptome.org/echelon-ep-fin.htm (letöltés ideje: 2016.12.30.)

[170] MACASKILL, Ewen – DANCE, Gabriel *NSA Files: Decoded*, 2013: forrás: http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/1 (letöltés ideje: 2017.11.05.)

[171] BULL, Hans Peter: A szabadság pátosza és a biztonságpolitika. SZIKLAY Júlia szerk.: *Az információs jogok kihívásai a XXI, században*, Adatvédelmi Biztos Irodája, Budapest, 2009, ISBN 978 963 08 7155 6, 63–67. oldal

[172] PÉTERFALVI Attila: *Átláthatóság a védelmi igazgatásban*, PhD-értekezés, NKE, Budapest, 2014, forrás: http://uni-nke.hu/downloads/konyvtar/digitgy/phd/2014/peterfalvi_attila.pdf (letöltés ideje: 2017.02.26)

[173] *Standard operating procedure*, The Economist, 2016. november 12., Special report, 9. oldal, forrás: http://www.economist.com/news/special-report/21709776-how-war-terror-turned-fight-about-intelligence-standard-operating-procedure (letöltés ideje: 2017.01.24.)

[174] WEBER, Rolf H. – STAIGER, Dominic N.: Privacy versus Security in KULESZA, Joanna – BALLESTE, Roy szerk.: *Cybersecurity and Human Rights in the Age of Cybervelliance*, Rowman & Littlefield, Lanham, ML, USA, 2016, ISBN 978 144 2260 412

[175] MANOLOPOULOS, Constatinos: *Surveillace by intelligence services: fundamental rights safeguards and remedies in the EU*, European Union Agency for Fundamental Rights, Vienna, Austria, 2015, forrás: http://fra.europa.eu/en/publication/2015/surveillance-intelligence-services (letöltés ideje: 2017.02.26)

[176] BORN, Hans – WILLS, Aidan: *Overseeing Intelligence Services, A toolkit*, DCAF, Geneva, Switzerland, 2012, ISBN: 978-92-9222-222-2

[177] *Case of Öcalan v. Turkey*, European Court of Justice, Strasbourg, 2005, forrás: http://hudoc.echr.coe.int/eng?i=001-69022, (letöltés ideje: 2017.11.05.)

[178] DICKINSON, William B. – MERCER, Cross – POLSKY, Barry: *Watergate: chronology of a crisis*. 1. Washington D. C.: Congressional Quarterly Inc., 1973, ISBN 0-87187-059-2, 8, 133–140, 180–188. oldal

[179] *Életfogytiglant kaptak a Politovszkaja-gyilkosság vádlottai*, 2014, forrás: http://hvg.hu/vilag/20140609_Eletfogytiglant_kaptak_a_Politivszkajagy (2017.02.26.)

[180] LICHFIELD, John: *PRESIDENT SARKOZY OF FRANCE ACCUSED OF USING SECURITY SERVICES TO SPY ON JOURNALISTS…,* 2010*,* forrás:

https://blendz72.wordpress.com/2010/11/04/president-sarkozy-of-france-accused-of-using-security-services-to-spy-on-journalists/ (letöltés ideje: 2017.11.05.)

[181] AMIR, Oren: *Mordechai Vanunu Indicted*, 1986, forrás: http://www.haaretz.com/israel-news/1.530146, (letöltés ideje: 2017.02.26.)

[182] ILEY, Chrissy: *Valerie Plame Wilson: the housewife CIA spy who was 'fair game' for Bush*, 2011, forrás: http://www.telegraph.co.uk/culture/film/8318075/Valerie-Plame-Wilson-the-housewife-CIA-spy-who-was-fair-game-for-Bush.html (2017.02.26.)

[183] BÍRÓ János: *Jelentés a kémprogramok magyar nemzetbiztonsági célú alkalmazásáról*, NAIH, Budapest, 2014, Forrás: https://www.naih.hu/files/adatved-jelentes-1904-6-2014-T_kemprogram.pdf/, (letöltés ideje: 2018.05.08.)

[184] BALL, James: *US and UK struck secret deal to allow NSA to 'unmask' Britons' personal data*, 2013, forrás: http://www.theguardian.com/world/2013/nov/20/us-uk-secret-deal-surveillance-personal-data (letöltés ideje: 2016.03.28.)

[185] VOELZ, Glenn J.: *Contractors and Intelligence: The Private Sector in the Intelligence Community*, International Journal of Intelligence and Counteritelligence, Vol. 22., Iss. 4., 586 - 613. oldalak, 2009, ISSN: 0885 0607 (letöltés ideje: 2018.02.10.)

[186] SHORROCK, Tim: *Spies for hire*, Simon and Schuster Paperbacks, New York, 2008, ISBN 978 0 7432 8225 3

[187] GILAD, Ben: *Business blindspots*, Calne, Wiltshire, Infonortics, 1966, ISBN 1 873699 33 6

[188] BLANCHARD, Kenneth: *Empowerment*, Edge 2000, Budapest, 2007, ISBN 978 963 869 278 8]

[189] TATE, Julie: *Judge sentences Bradley Manning to 35 years*, 2013, forrás: https://www.washingtonpost.com/world/national-security/judge-to-sentence-bradley-manning-today/2013/08/20/85bee184-09d0-11e3-b87c-476db8ac34cd_story.html?utm_term=.e4f70f05f360/ (letöltés ideje: 2013. 08. 21.]

[190] MACASKILL, Ewen: *Edward Snowden, NSA files source*, 2013, forrás: https://www.theguardian.com/world/2013/jun/09/nsa-whistleblower-edward-snowden-why/ (letöltés ideje: 2016.08.13.)]

[191] DERNONCOURT, Franck: *What are the good aplications for selecting statistics on keyboard use*, 2013, forrás: https://www.quora.com/What-are-good-applications-for-collecting-statistics-on-keyboard-use/, (letöltés ideje: 2016.08.14.)

[192] ROBIDEAU Rob: *Incognito toolkit*, under Creative Commons Attribution, 2014, ISBN 13 978 0 9850491 4 0

[193] *"Darkhotel" Cyberespionage Group Boosts Attacks with Exploit Leaked from Hacking Team* ,2015, forrás: http://newsroom.kaspersky.eu/en/texts/detail/article/darkhotel-cyberespionage-group-boosts-attacks-with-exploit-leaked-from-hacking-team/?no_cache=1&cHash=779a78f65c7ddbc79b1a9c36018449c6/ (letöltés ideje: 2016.08.14.)]

[194] MAHMOOD, Adam Mo – MANN, Gary J.: *Measuring the Organisational Impact of Information Technology Investment: An Exploratory Study*, Journal of Management Information Systems, Vol. 10, Issue 1, 1993, ISSN: 0742-1222, 97–122. oldal

[195] SCHMEIER, Sven - NEUMANN, Günter: Interactive Topic Graph Extraction and Exploration of Web Content, in POIBEAU, T. et al szerk.: *Multi-source, Multilingual Informaton Extraction and Summarization*, Chapter 7, Springer, 2012, ISBN 978-3-642-28568-4, 137–161. oldal

[196] BREALEY, Richard A. – MYERS, Stewart C.: *Modern vállalati pénzügyek I-II.,* Panem, Budapest, 1992, ISBN 963 7628 13 4

[197] PARKER, John: *Calculating ROI on Information Technology Projects*, Enfocus Solutions, 2012, forrás: http://enfocussolutions.com/calculating-roi-on-information-technology-projects/ (letöltés ideje: 2017.06.05.)

[198] FERRARI, Mascia: ROI in text mining projects, in ZANASI, Alessandro szerk.: *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, WIT Press, Southampton, UK, 2007, ISBN 978 1 84564 131 3

[199] VADÁSZ Pál: *Case study for measuring the feasibility of a semantic search system*, HADMÉRNÖK VII. 2., 2012, ISSN 1788-1919, 405–415. oldal

[200] FELDMAN, Sussan et al.: *The Hidden Costs of Information* Work, IDC, 2005, forrás: http://www.slideshare.net/PingElizabeth/the-hidden-costs-of-information-work-2005-idc-report, (letöltés ideje: 2014.05.13.)

[201] SVEIBY, Karl Erik: *Szervezetek új gazdasága: a menedzselt tudás*, KJK-Kerszöv, Budapest, 2001, ISBN 963 2245 997

[202] POLÁNYI Mihály: *Személyes tudás: úton egy posztkritikai filozófiához I-II.,* Budapest, Atlantisz, 1994, ISBN 963 7978 275

[203] POLANYI, Michael: *The Tacit Dimension*, Chicago, University of Chicago Press, 2009ISBN 978-0226672984 (utánnyomás)

[204] BODA György: *A tudástőke kialakulása és hatása a vállalati menedzsmentre*, Információs Társadalomért alapítvány, Budapest, 2008, ISBN 9789638778802

ÁBRÁK JEGYZÉKE

TÁBLÁZATOK JEGYZÉKE

GRAFIKONOK JEGYZÉKE

FOGALMAK ÉS RÖVIDÍTÉSEK JEGYZÉKE

Rövidítés	**Angol megnevezés**	**Magyar megnevezés**
AH		Alkotmányvédelmi Hivatal
AI	artificial intelligence	mesterséges intelligencia
API	Application Programming Interface	felhasználói program felület
API	application programming interface	felhasználói programozói felület
BH		Bírósági Határozatok
BTK		Büntetőtörvénykönyv
C4I	Command, Control, Communication, Computers, Intelligence	parancs, ellenőrzés, kommunikáció, számítógép, hírszerzés
CI	compettive intelligence	üzleti hírszerzés
CIA	Central Intelligence Agency	Központi Hírszerző Ügynökség
CMOS	Complementary Metal-Oxyd Semiconductor	komplementer fém-oxid félvezető
ECS	Enterprise Content Search	vállati tartalmat kereső rendszer
EJEB		Emberi Jogok Európai Bírósága
ELINT	Electronic Intelligence	elektronikus hírszerzés
EWS	early warning system	korai előrejelző rendszer

FININT	financial intelligence	pénzügyi hírszerzés
GCHQ	Government Communication Headquarters	Kormányzati Kommunikációs Központ (Főhadiszállás)
GVH		Gazdasági Versenyhivatal
HMEI		Honvédelmi Minisztérium Elektronikai, Logisztikai és Vagyonkezelő Zrt.
HTML	Hyper Text Markup Language	kiemelt szövegű jelölőnyelv
IM		Igazságügyi Minisztérium
IMINT	image intelligence	képi hírszerzés
IOSWG	International Open Source Work Group	Nemzetközi Nyílt Forrás Munkacsoport
IP	internet protokol	internet protokol
ISKO	International Society for Knowledge Organisation	Nemzetközi Szervezet a Tudásszervezésért
KNN	K-Nearest Neighbors	K-legközelebbi szomszéd
MAC	media access control	média elérés szabályzó
MI5	Military Intelligence 5 (osztály)	az Egyesült Királyság polgári elhárítása
MI6	Military Intelligence 6 (osztály)	az Egyesült Királyság polgári hírszerzése
MKLK		Magyar Közlöny Lap- és Könyvkiadó Kft.
NAIH		Nemzeti Adatvédelmi és Információszabadság Hatóság
NEAK		Nemzeti Egészségügyi Alapkezelő

NIBEK		Nemzeti Információs és Bűnügyi Elemző Központ
NISZ		Nemzeti Infokommunikációs Zrt.
NLP	Natural Language Processing	természetes nyelvű feldolgozás
NSA	National Security Agency	Nemzetbiztonsági Ügynökség
OCR	Online Character Recognition	aktív karakter felismerés
OEP		Országos Egészségügyi Pénztár
OSINT	open source intelligence	nyílt forrású keresés
OWL	Web Ontology Language	Hálózati Ontológiai Nyelv
PDF	Portable Document Format	hordozható dokumentum formátum
PSI	public sectore information reuse	közösségi információt újrafelhasználó (bróker)
RDF	Resource Description Framework	Forrás Leíró Keretrendszer
RDFS	RDF Schema	RDF séma
S2T	Speech-to-text	hanganyag írott szöveggé történő átalakítása
SCIP	Strategic and Competitive Intelligence Professionals	stratégiai és üzleti hírszerző szakemberek (egyesülete)
SIGINT	Signal Intelligence	elektromágneses spektrumot izsgáló hírszerzés
SocMed	Social Media	Közösségi média
SPARQL	Simple Protocol and RDF Query Language	egyszerű protokol és RDF lekérdező nyelv

SVM	Support Vector Machine	támogató vektor gép
TED	Tenders Electronic Daily	napi elektronikus tender(-figyelő),
TIBEK		Terrorelhárítási Információs és Bűnügyi Elemző Központ
UI	user interface	felhasználói felület
UX	user experience	felhasználói élmény
W3C	World Wide Web Consortium	Világszerte Működő Konzorcium
XML	Extensible Markup Language	kiterjeszthető jelölőnyelv

MELLÉKLET

Fontosabb külföldi jogi adatbázisok

EUR-Lex

Az EUR-Lex a legszélesebb és legelterjedtebb alkalmazás, mely az Európai Unió jogi hivatalos adatbázisa, amely közvetlen és ingyenes hozzáférést biztosít az európai uniós jog mindhárom forrására és az EU Hivatalos Lapjának hiteles elektronikus kiadásához mind a 24 hivatalos nyelven. A napi rendszerességgel frissített adatbázis az EU joggal kapcsolatos szakmai anyagokkal és speciális nyilvántartásokkal is segíti a jogalkalmazók munkáját. Az elsődleges, másodlagos és a kiegészítő jogi források, valamint a legfontosabb szakosodott uniós adatbázisok és honlapok elérését aktív linkkel biztosítja. Az EUR-Lex tartalmazza az Európai Bíróság három igazságszolgáltatási fórumának (Bíróság, Törvényszék, Közszolgálati Törvényszék) ítélkezési gyakorlatával kapcsolatos döntéseket, kapcsolódó anyagokat.

CURIA

Az Európai Unió Bírósága az Európai Unió és az Európai Atomenergia-közösség (Euratom) igazságszolgáltatási szerve. Feladata az Unió jogi aktusai jogszerűségének vizsgálata, aminek keretében felülvizsgálja az Európai Unió intézményei jogi aktusainak jogszerűségét, gondoskodik arról, hogy a tagállamok teljesítsék a szerződésekből eredő kötelezettségeiket, és a nemzeti bíróságok kérelmére értelmezi az uniós jogot. Ez utóbbi keretében a tagállami bíróságokkal együttműködve gondoskodik az uniós jogegység értelmezéséről és alkalmazásáról. Az Európai Bíróságnak három igazságszolgáltatási fóruma van: az 1952-ben létrehozott Bíróság, az 1988-ban létrehozott Törvényszék, valamint a 2004-ben létrehozott Közszolgálati Törvényszék. A három igazságszolgáltatási fórum eddig hozzávetőleg 28 000 ítéletet hozott. Az Európai Unió Bíróságának hivatalos adatbázisa, a Curia hozzáférést biztosít a három igazságszolgáltatási fórum működésével kapcsolatos szakmai információkhoz, közérdekű adatokhoz, továbbá az Európai Bíróság elé terjesztett ügyek nyilvános adataihoz és dokumentumaihoz, az 1997. június 17. utáni ítélkezési gyakorlatához az EU minden hivatalos nyelvén. Ezen túlmenően az uniós joghoz kapcsolódó nemzeti és nemzetközi ítélkezési gyakorlat megismerését is segíti.

A honlap az anyagokat kollekciókba és alkollekcióba csoportosítva teszi közzé.

Külön önálló kollekciója van: az Európai Bíróság egészének mint intézménynek, az egyes bírói fórumoknak (Bíróság, Törvényszék és Közszolgálati Törvényszék), az ítélkezési gyakorlatnak és a sajtóval és médiával kapcsolatos tájékoztatóknak, anyagoknak.

A Könyvtár és dokumentáció cím alatt szakirodalom, uniós vonatkozású jogi információk, az uniós jog története, valamint szélesebb átfogóbb szakmai ismeretek érhetők el, így pl. az EUR-Lex, a nemzeti és nemzetközi joggyakorlat, továbbá a hasznos linkek. Az adatbázis eleget tesz az Európai Unió Bíróságával szembeni többnyelvűség kapcsán megfogalmazott szigorú követelményeknek is, melyek szerint az Unió bármely hivatalos nyelve lehet az eljárás nyelve, és biztosítania kell ítélkezési gyakorlatának valamennyi tagállamban való közzétételét is.

Az Európai Bíróság döntéseit és az eljárásával kapcsolatos nyilvános adatokat és dokumentumokat **„***A Bíróság, a Törvényszék és a Közszolgálati Törvényszék Határozatainak Tára”***,** az Európai Unió hivatalos lapja, valamint a Curia nevű hivatalos honlapja teszi közzé. A Határozatok Tárának van egy általános része, amely a Bíróság és a Törvényszék ítélkezési gyakorlatát tartalmazza, és van egy közszolgálati része, amely a Törvényszék és a Közszolgálati Törvényszék közszolgálati ügyekkel kapcsolatos ítélkezési gyakorlatát tartalmazza. Az általános rész 2011-ig, a közszolgálati rész pedig 2009-ig jelent meg papíralapú formában. Ezen időszak vonatkozásában kizárólag a papíralapú változat minősül hivatalosnak. Ezt követően az EUR-Lex-en és Curian közétett elektronikus változat is hivatalosnak minősül. *A Bíróság, a Törvényszék és a Közszolgálati Törvényszék Határozatainak Tára az EUBookshop honlapon elektronikusan az EU 22 nyelvén ingyenesen PDF formátumban elérhető.*

Az Európai Bíróság közzétételi szabályai értelmében meghatározott döntéseit a Határozatok Tárában már nem teszi közzé. E határozatok azonban a rendelkezésre álló nyelveken - azaz az eljárás nyelvén és a tanácskozás nyelvén – a Bíróság internetes honlapján, a Curián elérhetők.

JURE

Az angol és francia nyelvű ingyenes, szakosított jogi adatbázis a polgári és kereskedelmi ügyekben a joghatóságra, valamint a határozatoknak a meghozataluk államától eltérő államban való elismerésére és végrehajtására vonatkozó ítélkezési gyakorlatot tartalmazza. Ide tartozik az alkalmazandó nemzetközi egyezményekkel (vagyis az 1968. évi Brüsszeli Egyezmény és az 1988. évi Luganói Egyezmény) kapcsolatos, valamint az uniós és a tagállami ítélkezési gyakorlat is.

CaseLex

A szintén angol és francia nyelvű fizetős szakosított jogi adatbázis az Európai Bíróság ítélkezési gyakorlatát és a tagállamok nemzeti legfelsőbb bíróságainak vonatkozó határozatait tartalmazza 11 jogterületre vonatkozóan. Minden dokumentumhoz van angolul és az eredeti szöveg nyelvén lévő összefoglaló.

JURIFAST

Az angol és francia nyelvű ingyenes szakosított jogi adatbázis az Európai Bíróság által meghozott előzetes döntéseket és a tagállami bíróságok által előzetes döntéshozatalra előterjesztett megfelelő kérdéseket és egyéb nemzeti döntések értelmezése az uniós joggal összefüggésben.

DEC.NAT

Az angol és francia nyelvű ingyenes szakosított jogi adatbázis a tagállamok uniós jog alkalmazásával kapcsolatos ítélkezési gyakorlatát tartalmazza. A nemzeti határozatok az eredeti nyelven szerepelnek, angol és francia összefoglalóval, továbbá az Európai Bíróság kutatási és dokumentációs szolgálat által biztosított hivatkozások és elemzések. Az adatbázis „nemzeti döntései" tartalmazzák az uniós joggal kapcsolatos nemzeti joggyakorlatot is. Az adatbázis tartalmazza még kulcsszavas módon a döntések elemzését, összefoglalót a határozatok céljairól.

Külön kört képviselnek még a nem hivatalos jogi alkalmazások, amelyek lehetnek általánosak vagy szakosodottak vagy konkrét nemzeti joghoz, esetlegesen nyelvcsoporthoz kapcsolódóan hozták létre. Ezek egy része szintén ingyenes, de vannak fizetős alkalmazások, amelyek már számos speciális szolgáltatást is nyújtanak.

WestLaw

Naponta frissülő, fizetős angolszász jogi adatbázis. Két fő részből áll: hírek és jogi adatbázis szekcióból. A hírek (Globál News) általános és üzleti híreket tartalmaz, a világ minden tájáról a vezető napi- és hetilapokból. A jogi adatbázisban (Westlaw International) jogforrások, kommentárok, magyarázatok, jogszabályok háttéranyagai és jogi bibliográfiák találhatóak.

Legifrance

Alapvetően francia jogszabályokat tartalmazó alkalmazás, továbbá kiegészül egyes Európai Unió-s és nemzetközi jogi dokumentumokkal.

Juricaf

Egy frankofón adatbázis, melyben 43 országgal összefüggésben szerepelnek a Legfelsőbb Bíróság döntései. Ennek az oldalnak az a célja, hogy szabad hozzáférést biztosítson a jogi szakembereknek, a bíróságok határozataihoz és segítséget nyújtson a szakembereknek és civil embereknek a határozatok megértésében. Az adatbázis a frankofón tagállamok, részvevő államok és megfigyelő államokon belüli határozatokra terjed ki. Naponta frissülő tartalommal rendelkezik.

Dalloz

Egy 2006-ban indított francia jogi adatbázis, mely törvényeken belül hét területet dolgoz fel.

Leggiitaliane.it

Ebben az olasz alkalmazásban törvényekre, rendeletekre, irányelvekre és folyamatban lévő ügyekre és számos ítéletre lehet keresni.

RegioneUmbria

Egy olasz nemzeti adatbázis, melyen törvényekre és rendeletekre lehet keresni.

RicercaGiuridica

Olasz nemzeti adatbázis, ahol bírósági határozatok teljes szövegeit is meg lehet tekinteni. Továbbá ítéletek, törvények és rendeletek is fellelhetőek az oldalon.

Gesetze

A német lakosság számára a Belügyminisztérium hozta létre a portált, egy jogi céggel együttesen, hogy tájékoztatásul szolgáljon a polgárok fogyasztói jogvédelemre. Aktuális fogyasztóvédelmi nemzeti jogszabályokat tartalmazza, fogyasztói panaszok kezelése.

Beck-Online

A következő modulokat tartalmazza jogterület szerint: Polgári jog, Kereskedelmi és gazdaságjog, Polgári peres eljárás jog, Munka és szociálisjog, Közigazgatási jog, Büntetőjog és Közlekedési szabályokra vonatkozó jog, Adójog és könyvelési jog; Könyvek; Folyóiratok; Igazságszolgáltatás; Normák, szabályok/irányelvek (törvénygyűjtemények), továbbá vannak segédanyagok, formanyomtatványok. A szolgáltatások nagyon szerteágazóak, széleskörűek. Állásportálon túl könyvkereskedésként is működik, tehát nem kizárólag adatbázisra funkcionált az oldal.

JURIS

Ezen a német jogi portálon rendelkezésre bocsátják az aktuális jogi változtatásokat. Naponta frissül, tehát naprakész dokumentumok fogadják a felhasználót. Beírhatjuk a keresést bizonyítékok, normák, kommentárok, újságcikkek, mint pl. keresett szavak, kifejezések, ügyiratszám, jogszabály rövidítés vagy lelőhely szerint. A találati listák és törvények (jogszabályok) és rendeletek (rendelkezések) széles választéka ingyenesen is elérhetőek.

HEINONLINE

A világ legnagyobb kép-alapú kereső adatbázisa jogi cikkekre. Több mint 1800 törvényt és joggal foglalkozó periodikát tartalmaz. Az Egyesült Államok Kongresszusának döntéseit, törvényeit és a tagállamok jogi normáit tartalmazza. Több mint 9 századnyi jogi anyag. Régebbi dokumentumokat, kéziratokat is tartalmaz. A kép-alapú adatbázis lehetőséget nyújt arra, hogy a törvényszövegekhez írt lábjegyzetek, megjegyzések is digitalizálva legyenek. Lehet keresni a címben vagy a szerzőt, illetve kulcsszavas keresés is van. Speciális szolgáltatásként régebbi dokumentumokat, kéziratokat is tartalmaz. A kép-alapú adatbázis lehetőséget nyújt arra, hogy a törvényszövegekhez írt lábjegyzetek, megjegyzések is digitalizálva legyenek.

Ju§line

Az alkalmazás kínálja a legfontosabb osztrák; német törvények teljes szövegét. A kulcsszavas, témakörös és abc-rendű keresésen túl különböző nyilvántartásokat is tartalmaz (ingatlan nyilvántartás, cégjegyzék, bíróság, ügyvéd, szakértő, jogalkotási) és szakmai fórumot is fenntart. A törvényekhez – a törvényszövegek érthetősége kedvéért - lehet kommentárokat írni, amit szakértők és civilek is csinálhatják. A keresési eredményeknek a kommentárok is részei.

Justis, the law Online

A Justis egy teljesen szöveges online könyvtára az Egyesült Királyság, Írország, az Európai Unió és a nemzetközi jogszabályoknak és esetjogoknak. A portál sokoldalú, könnyen használható és egy megbízható jogi keresőként is szolgál rengeteg jogász számára a világon. A jogi adatbázisban az alábbi jogi tartalomra lehet keresni: Egyesült Királyság, Írország, Európai Unió és egyéb nemzetközi jog. A több mint 20 éves múlttal rendelkező portál széleskörű szolgáltatást nyújt. Legújabb kiadását 2012. júniusában indítottak el. A Justis könyvtára naprakészen gondoskodik a jogalkotás követéséről, az esetjogokról, és az európai uniós információkról. A dokumentumok egységesen egyszerűen letölthetők PDF formátumban. Az alkalmazás felhasználóbarát, jó kereső funkcióval rendelkezik.

Jogi tájékoztatási adatbázis /World Legal Information

Átfogó, független és publikus információt biztosít a világ országainak a jogi és állami apparátus felépítéséről, összegyűjti az egyes nemzetek jogszabályait, egyezményeit, tanulmányozhatók a nemzetközi jogi normák, illetve a különböző jogi reformokra vonatkozó anyagok.

Lauterpacht Nemzetközi Jogi Központ adatbázisa

A Cambridge-i Egyetemen létrehozott Lauterpacht Nemzetközi Jogi Központ adatbázisa, mely a Cambridge-i Egyetem jogi karához tartozik. Rendszeresen tematikus vitafórumokat szervez a nemzetközi közösség aktuális jogi kérdései kapcsán, illetve összegyűjti a publikációkat és a jogi beszámolókat.