

# Az elektronikus adatállományok közép- és hosszú távú archiválása

Székely Iván

**Tézis:** *Az elektronikus adatállományoknak a felhasznált hardver és szoftver élettartamát meghaladó időtávú archiválására már megszülettek a koncepciók és követelményrendszerek, növekszik az adattároló eszközök megbízhatósága és várható élettartama, azonban az alkalmazott technológiai és szervezési módszerek még nem kiforrottak; ezek egységesedése és szabványosítása várható néhány fő koncepció alapján, a gyakorlati tapasztalatok intenzívebb visszacsatolása mellett.*

## 1. Témakör

Már a XX. század utolsó évtizedére nyilvánvalóvá vált, hogy az információs technológiák egyre gyorsuló fejlődése sokkal hamarabb elavulttá teszi a mindenkori hardvert és szoftvert, mint az azokkal előállított és tárolt információt, vagy annak megőrzésre méltó elemeit. A néhány nagyobb nyilvánosságot kapott adatvesztés<sup>1</sup> csak a jelenség felszínét mutatta, a háttérben az adatokat előállító és felhasználó szervezetek és személyek egyre inkább szembesültek a néhány évvel korábbi formátumban és adathordozón archivált adataik felhasználhatóságának problémáival.

Mind a tudományos célú, mind az államigazgatási és üzleti célú, mind pedig a magáncélú számítógépes adathasználat mértéke robbanásszerű növekedésnek indult az ezredforduló körüli évektől kezdve, s növekedett az adathasználók függősége adataik – beleértve félkurrens vagy nem kurrens adataikat is – elérhetőségétől és felhasználhatóságától. Emellett fokozódik az igény az elsődleges aktualitásukat már elvesztett üzleti célú adatállományok újbóli felhasználására, utóelemzésére, adattárházak építésére, adatbányászati módszerek alkalmazására; a nagy internetes szolgáltatók pedig „minden” információ „örökre” történő megőrzésének illúzióját vetítik felhasználóik elé.

A tömeges felhasználást jelentő üzleti, államigazgatási és magáncélú adatkezelés archiválási és visszakereshetőségi igényeit jelenleg olyan archiváló és dokumentumkezelő rendszerek próbálják kielégíteni, amelyek csak néhány éves távlatban tudják garantálni az archivált adatállományok felhasználhatóságát. A középtávú (10–15 éves) és hosszú távú (több évtizedes, egyes esetekben elvileg korlátlan időtávú) archiválás követelményeinek és elvi megoldási lehetőségeinek kidolgozása megtörtént az elmúlt évtizedben, emellett néhány ambiciózus nemzetközi projekt is indult a digitális információtömeg archiválására. Az elkövetkező évek feladata a tartós archiválási technológiák és szabványok kidolgozása, valamint alkalmazói szintű elterjesztése.

## 2. Jelenlegi helyzet

Az adatgyűjtés, -feldolgozás és -elemzés technológiáinak fejlődésével, a felhasználói szintű alkalmazások terjedésével egyre több adatállomány jön létre elektronikus formában, ezek közül egyeseket eredetileg is számítógépen készített előállítója, mások papíron vagy más analóg hordozón születtek, és később digitalizálták őket, ismét másokat emberi beavatkozás nélkül automatizált rendszerek állítanak elő. Korlátozott körben már megvalósult a kizárólag elektronikus iratkezelést alkalmazó „papír nélküli iroda”, általánosságban a papír nélküli adatkezelés, emellett átfogó digitalizálási projektek születtek Magyarországon és nemzetközi

---

<sup>1</sup> Talán legismertebbjük a NASA adatvesztése, amely 2007-ben került nyilvánosságra: az előző évben bezárták azt a laboratóriumot, amely még képes volt az 1969-es holdsétáról készített és analóg szalagon tárolt mozgóképfelvételek és telemetriai adatok olvasására, és maguk a szalagok is eltűntek.

szinten is. (Az archiválás szempontjából közömbös, hogy az elektronikus adatállomány hogyan jött létre, ezért a digitalizálás problémakörét csupán megemlítjük, mint a megőrzendő elektronikus adatállományok létrejöttének egyik forrását.)

Ugyanakkor már ma sem tudjuk olvasni az egy-két évtizeddel ezelőtt készült adatállományok egy részét, részben az adathordozók öregedése miatt, részben a ma használatos adathordozók formátumának megváltozása miatt, részben pedig az állomány visszakereshetőségét, olvashatóságát biztosító szoftverek változása miatt. Ehhez járul a felhasználók által használt informatikai eszközök egyre rövidülő életciklusa, a gazdasági modell, amely a fogyasztói társadalom értékrendjén alapul, s az információs társadalom kívánatos technológiai fejlődését összeköti az állandó *innováció* kényszerével, miközben az *optimalizáció*, a rendszerek hosszabb távú működőképességének biztosítása háttérbe szorul. Vannak ugyan biztonságkritikus szektorok (például honvédelem, stratégiai fejlesztések, bankszektor), ahol a fokozott megbízhatóság igénye az adatállományok hosszabb távú felhasználhatóságát is részben magába foglalja, ezek azonban többnyire nem foglalkoznak a közép- és hosszú távú archiválás általános problémáival.

## 2.1 Adattároló eszközök

Az adatállományok fizikai tárolását végző eszközök fejlődésének általános áttekintése e helyütt nem célunk, azonban néhány, az archiválás szempontjából fontos fejleményt, illetve problémát kiemelünk. Archiválási szempontból elsősorban az adattároló eszközök élettartama és megbízhatósága (és e tényezők ismerete és tervezhetősége) bír jelentőséggel; emellett nem elhanyagolható a fajlagos tárolókapacitás sem, mivel az archiválandó adatok mennyisége időegységre vetítve is és kumulatív módon is egyre növekszik; és amennyiben az archivált adatok speciális eljárás nélküli hozzáférhetőségének biztosítása is fontos szempont, akkor a rendelkezésre állás, illetve az adatok elérési ideje is szerepet játszik az eszköz alkalmasságának értékelésében. A második, de különösen a harmadik jellemző nem csupán a fizikai adathordozó minőségén múlik, hanem az azzal egybeépített elektromos és mechanikus meghajtó elemek és firmware minőségén is. Ha pedig az adattároló eszközt a működtetéséhez szükséges host számítógép e célt szolgáló részegységeivel együtt tekintjük működőképes egésznek, akkor komplex, többretegű rendszert kapunk, amelynek minden rétege és eleme befolyásolja az „eszköz” minőségét és archiválási célra való alkalmasságát. Ilyen, „többretegű” adattároló eszköz például a ma általánosan elterjedt mágneses merevlemez (winchester). Az adathordozók öregedése – amennyiben meghibásodásuk nem katasztrófaszerűen következik be – viszonylag jól tervezhető (annak ellenére, hogy a hosszú távú öregedésükre vonatkozó tapasztalati adatok értelemszerűen nem állnak rendelkezésre, csupán becslések és a „gyorsított öregítés” kísérleti adatai). Ennélfogva az öregedés problémája jól kezelhető az adathordozók rendszeres cseréjével, vagyis az adatok más, ugyanolyan (vagy akár új típusú) hordozóra való átírásával. Természetesen az ilyen átírás nem oldja meg az adatformátumok, illetve a működtető szoftverek elavulásának problémáját, és önmagában nem nyújt védelmet az eszközök katasztrófaszerű meghibásodása ellen.

A winchesterek várható meghibásodásának időbeli alakulását egy jellemző „fürdőkád-diagram” illusztrálja (*1. ábra*). Amennyiben a kezdeti időszak mortalitását fokozott ellenőrzéssel és adat-helyreállítási módszerekkel ellensúlyozzák, a lemez hosszú időn át egyenletes, jól tervezhető meghibásodási valószínűséget mutat. A meghibásodási valószínűséget általában a Mean Time to Failure (MTTF) mutatóval jelzik a gyártók, de archiválási szempontból ennél fontosabb az úgynevezett Unrecoverable Error Rate (UER), amely helyreállíthatatlan adatvesztést eredményezhet. A jelenleg elterjedt merevlemezek ugyanis eleve tartalmaznak egy alapszintű helyreállító mechanizmust, amely az elemi adatok szektorokba szervezéséhez kapcsolódik: egy szektor jellemzően 512 bájt adatot tartalmaz, valamint néhány kiegészítő bájtot, amely a bithibák helyreállításához szükséges redundáns

bitsorozatot rejti (ezt általában az úgynevezett Reed-Solomon kódolással állítják elő). Ha több ilyen redundáns bájt szerepel szektoronként, ez javítja a helyreállítás esélyeit, viszont csökkenti a merevlemez hasznos fajlagos adattárolási kapacitását, ezért a gyártók igyekeznek arányát a lehető legkisebb értékre beállítani. (Egyes kutatók javasolják a szektorok méretének megnövelését, amely mind a fajlagos adattárolási kapacitás, mind a helyreállítás esélyeinek növelését eredményezné, de ennek a belátható jövőben kevés az esélye a széles körben elterjedt szoftverek miatt, amelyek 512 bájtos szektorokat tételeznek fel a merevlemez tárolók esetében.) A szektor-szintű hibajavítás azonban nem tud minden bithibát korigálni: az UER értéke ATA/IDE típusú merevlemezeknél 1 bithiba  $10^{13}$ – $10^{14}$  olvasott bitre, SCSI merevlemezeknél 1 bithiba  $10^{13}$ – $10^{15}$  olvasott bitre. Ezeket a hibákat már nem célszerű hardver szinten kezelni, egyfelől azért, mert a gyártók nem érdekeltek a fajlagos adattárolási kapacitás csökkentésében, másfelől pedig azért, mert e hibák csak egyetlen részterületét jelentik az adatintegritás sérülésének. Ezért az UER típusú hibákat magasabb, szemantikus elemeket is magukban foglaló szinteken kezelik az archiváló rendszerek. Ennek a gyakorlatban megvalósuló formája a folyamatos adatintegritás-ellenőrzés (auditálás), amely azonban hozzájárul a merevlemezek elhasználásához, ezért korlátai vannak, s leginkább más ellenőrzési folyamatokkal (pl. vírusellenőrzéssel) párosítva történik alkalmazásuk, a lemezek elhasználódását csökkentendő.



1. ábra. A winchesterek élettartamának megoszlása

A névleges élettartam tehát ismert, az öregedésből és elhasználódásból következő meghibásodások megelőzhetők rendszeres átírással, azonban a „gyermekhalandóság” típusú meghibásodásokra ez nem ad megoldást. (Ez utóbbi esetben alkalmazhatók a különféle adatmentési eljárások, amelyek kifejlesztésében és alkalmazásában jelentős magyar tapasztalatok állnak rendelkezésre – mégis ez a megoldás csak rendkívüli események esetén alkalmazható és nem képezheti egy archiválási stratégia betervezett részét.)

Az ilyen meghibásodások és adatvesztések elleni védekezés meghatározó jelentőségű csoportját a *redundáns adattároláson* alapuló megoldások képviselik. E megoldások ma is széles körben elterjedtek a nem-archiválási célú megbízható adattárolás területén, de mivel egyes hosszú távú archiválási elképzelésekben és megvalósított rendszerekben is szerepet játszanak, ezért itt is megemlítjük néhányukat. Az adathordozó szintjén végzett egyszerű *adattükörözés* csupán megkétszerezi vagy megtöbbszörözi az adathordozó eszközöket azonos adattartalommal, jellemzően ugyanabban a számítógépben. Hasonló elven alapul az adatok *replikálása*, amely a teljes adattartalmú tartalék adathordozó helyileg elkülönített (jellemzően

távoli) biztonságos tárolását jelenti. (A replikálás az elosztott tárolási-archiválási koncepciók egy részének alapját is képezi, lásd később.) Általánosan elterjedtek a számítógépekben az adatvesztés esélyét jelentősen csökkentő RAID (Redundant Array of Independent Disks) merevlemez-csoportok, amelyek alapszinten csak tükröznek, fejlettebb változataik viszont különféle adatintegritás-növelő és ellenőrző módszereket is alkalmaznak, jellemzően a paritás-bitek elosztott tárolásán vagy ellenőrző kivonatok (hash digest) alkalmazásán alapulókat (pl. CRC). Ez utóbbi módszerek alkalmasak bizonyos szintű adatvesztések automatikus detektálására és helyreállítására, ezzel új dimenzióval bővítve az adattároló eszköz megbízhatóságának eszköztárát.

Mind a kurrens és fél-kurrens, mind az archív adatok és dokumentumok esetében egyre növekvő elvárás a könnyű – és egyben távoli – hozzáférhetőség biztosítása. Amennyiben például a tárolás merevlemezeken történik, e lemezek tartalmának távolról is elérhetőnek kell lennie. Ahol pedig az elérhetőség sebessége döntő szempont a szolgáltatás megítélése szempontjából, különösen sokfelhasználós környezetben (tipikusan ilyenek a kurrens adatok tekintetében a keresőgépes szolgáltatások, vagy az Internet Archive), ott ezeknek a merevlemezeknek állandóan készenléti állapotban kell lenniük, vagyis forogniuk kell. Az ilyen hatalmas tárolókapacitást nyújtó szolgáltatások energiaigényének (és egyéb üzemeltetési költségeinek) csökkentésére fejlesztették ki az „alvó” winchesterek rendszerét, a MAID-et (Massive Array of Idle Disks), amelyben az egyedi lemezek csak akkor pörögnek fel, ha a rajtuk lévő adatok kiolvasására van szükség. Ez a megoldás természetesen megnöveli az átlagos elérési időt, ami a nem-kurrens adatok esetében elfogadható, azonban növeli a winchesterek meghibásodásának esélyét is (amelyek gyakran felpörgéskor, illetve leálláskor hibásodnak meg) és a rövid aktív periódusok miatt csökkenti az adatvesztés-helyreállító módszerek hatékonyságát. Emiatt nagyobb a hibák látenciájának időtartama is, mivel a rendszeres ellenőrzések (auditok) a lemezek felpörgetését, élettartamuk csökkenését okoznak, s ezért csak az adathasználat eseteire korlátozódnak. Tekintettel azonban a MAID rendszerekben alkalmazott olcsó, rövid élettartamú egyedi merevlemezekre, az ilyen rendszerek költségei versenyképesek a mágnesszalagos tárolóegységekből felépített rendszerekével.

A mágnesszalagos tárolás a kisebb alkalmazók körében kevésbé népszerű, mint a különféle lemezes megoldások, azonban továbbra is elterjedt a professzionális adattárolás terén. Az újabb rendszerek közül kiemelendő a HP, az IBM és a Quantum cégek által közösen kifejlesztett Linear Tape-Open (LTO) technológia, amely a Digital Linear Tape (DLT) alternatívájaként jelent meg, és amely a létrehozóinak állítása szerint nyílt formátumú, abban az értelemben, hogy az LTO alapon készülő későbbi termékek kompatibilisek lesznek a jelenlegiekkel. Az LTO jelenlegi implementációja az Ultrium, amelynek eddig hat generációja született meg és további generációi állnak tervezés alatt; a tervek szerint az LTO Ultrium 6. generációs mágnesszalagos kazetta 6,4 terabyte kapacitású lesz és a rendszer 540 MB/s adatátviteli sebességre lesz képes.<sup>2</sup>

A kis alkalmazók, olcsóságuk és egyszerűségük miatt előszeretettel alkalmaznak CD-ket és DVD-ket archiválási célra. Ezek élettartamát és megbízhatóságát azonban a gyártók nem garantálják, ezért alkalmazásuk hosszú távú archiválásra – a formátum-elavulás problémáitól eltekintve is – kockázatos. Az úgynevezett WORM (Write Once, Read Many, vagyis egyszer írható és sokszor olvasható) optikai lemezek élettartamát viszont a gyártó garantálja; legújabb generációjuknál ez az időtartam eléri az 50 évet. Előnyük a szalagos tárolókkal szemben a rövid elérési idő.

Több gyártó WORM elnevezéssel hagyományos winchesterekből felépített rendszert kínál, s ezzel csupán azt garantálják, hogy a lemez adattartalma sem szándékosan, sem véletlenül nem

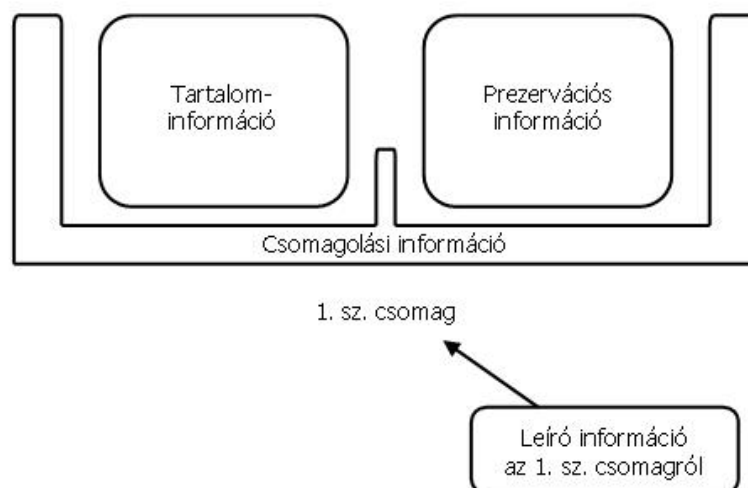
---

<sup>2</sup> <http://www.lto-technology.com/technology/default.php?section=0>

törölhető, illetve nem módosítható. Ilyen rendszer például a svájci FAST LTA (LTA = Long Term Archiving), amely három adatközpontban elosztva, hagyományos merevlemezekon tárolva, kilenc rétegű redundáns tárolási eljárással és a meghibásodó lemezek automatikus ellenőrzésével, cseréjével és újraírásával a felhasználóknak 30 éves adatmegőrzési időtartamot garantál.<sup>3</sup>

## 2.2 Archiválási modellek

A felsorolt és példaképpen megemlített elemek és megoldások csak részeit, esetenként szükséges, de önmagában nem elégséges feltételeit képezik a hosszú távú archiválásnak. Az archiválási és kapcsolódó (hozzáférhetőségi, integritási, hitelességi, gazdaságossági, megvalósíthatósági stb.) problémák megoldása rendszerszerű koncepciót igényel. Néhány ilyen, magas szintű, egységes rendszert alkotó koncepció létrejött az elmúlt évtizedben, közülük is a leginkább elfogadott, *de facto* szabványként kezelt, és a magyar stratégiákban, szakanyagokban is tükröződő ajánlás a Nyílt Archivumi Információs Rendszer (Open Archival Information System, OAIS). Az OAIS-t eredetileg az űrkutatási szervezetek dolgozták ki a digitális adatok hosszú távú megőrzési modelljének magas szintű leírására (részletes leírását lásd CCSDS 2002), később ISO 14721:2003 néven szabvánnyá vált. Az OAIS komplex információcsomagokat értelmez, amelyek általános felépítését a 2. ábra illusztrálja.



2. ábra. Az OAIS információcsomag-típusai és kapcsolatuk

A tartalomra vonatkozó információnak tartalmaznia kell a reprezentációjára vonatkozó információt is, más szóval azt, hogy a megőrzött bitsorozatot hogyan kell értelmezni, például képként vagy szöveggként. A prezervációs információ maga is négy alapvető elemből áll, ezek: az archivisztikából jól ismert *proveniencia*, amely a megőrzendő tartalom forrásáról, keletkeztetőjéről őriz információt, a *kontextus*, amely meghatározza a tartalom más tartalmakhoz való kapcsolódását,<sup>4</sup> a *referencia-információ*, amely egyedileg azonosíthatóvá teszi a tartalmat valamely nyilvántartási rendszerben,<sup>5</sup> valamint a *rögzítő információ*, amely megakadályozza, de legalább is detektálja a tartalom megváltozását. A csomagolási információ a fenti két fő információ típust tartalmazza: például ha a megőrzendő

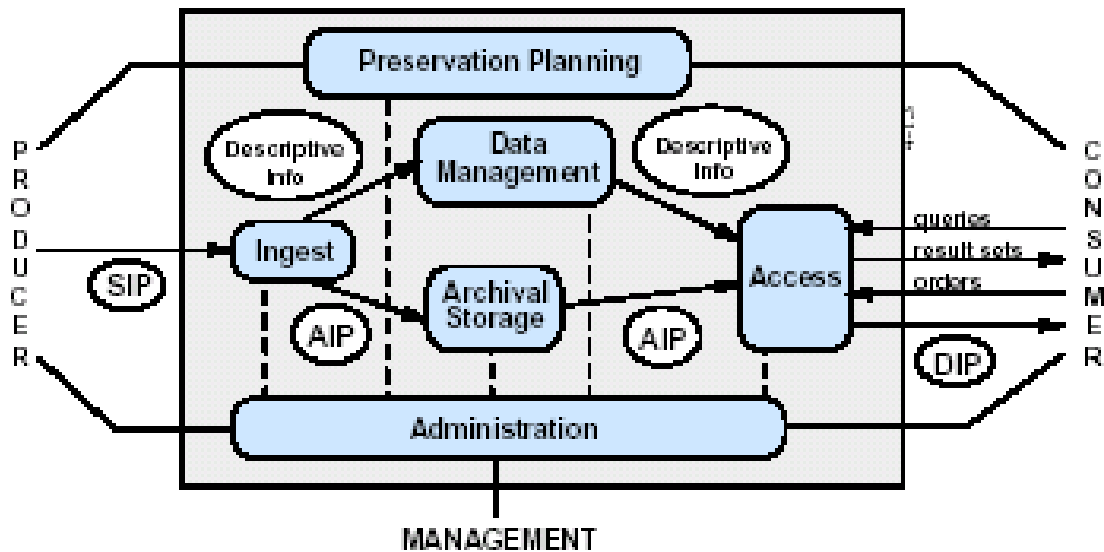
<sup>3</sup> <http://www.fast-lta.com>

<sup>4</sup> Az egyedi dokumentum vagy dokumentum-rész értelmezéséhez ismernünk kell azok kapcsolódását, viszonyát ugyanazon keletkeztető más dokumentumaihoz, más keletkeztető hasonló dokumentumaihoz, időben korábban vagy később készült dokumentumokhoz stb.

<sup>5</sup> Ilyen rendszer például könyvek esetében az ISBN szám.

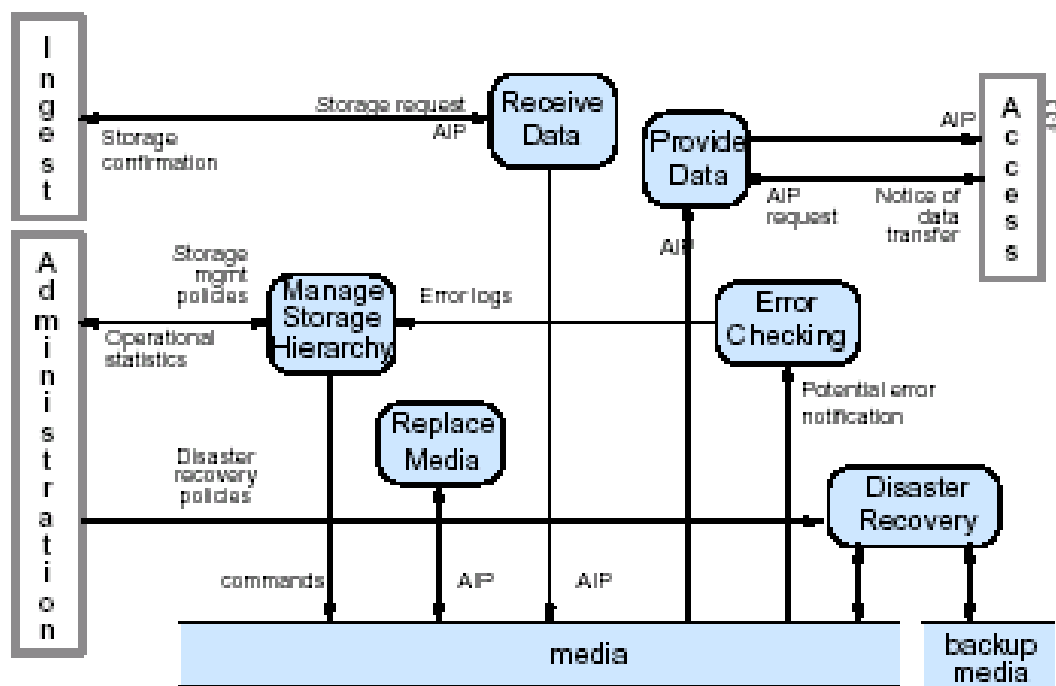
tartalmi és prezervációs információt CD-n tároljuk, akkor a csomagolási információ tartalmazza a fájlstruktúrát, a könyvtárak és fájlok neveit és összefüggéseit. A leíró információ pedig lehet egy egyszerű cím vagy elérési út, de tartalmazhat a katalógus-szerű kereséshez szükséges attribútumokat is. Mindezen elemeket és összefüggéseiket meg kell őriznünk ahhoz, hogy az eredetileg megőrzendő tartalom hosszú távon elérhető és értelmezhető maradjon.

Az OAIS funkcionális entitásait és azok főbb kapcsolatait a 3. ábra illusztrálja (CCSDS 2002 alapján). E csoportosításban az archiválásra vonatkozó információcsomagokat *AIP*, az archiválandó anyagot szolgáltató fél által adott információkat *SIP*, a jövőbeli felhasználónak adott, az értelmezéshez szükséges információkat *DIP* jelöli.



3. ábra. Az OAIS funkcionális entitásai (CCSDS 2002)

Ha ebből a rendszerből egy alapvető elemet, a tárolási entitást (Archival Storage) közelebbről megnézzük, az alábbi elemeket és főbb összefüggéseket ábrázolhatjuk:



4. ábra. A tárolási entitás (Archival Storage) funkciói (CCSDS 2002)

Látható, hogy az adathordozó (media), illetve a fentebb említett hibaellenőrző és javító megoldások csak részét képezik a tárolási entitásnak (a tárolási entitás pedig a teljes rendszernek), ezért azok megbízhatósága és élettartama csak szükséges, de nem elégséges feltétele az elektronikus adatállományok hosszú távú megőrzésének.

A megőrzendő adatállományok levéltári szemlélet szerinti leírását szolgáló metaadatok terén ugyancsak *de facto* szabvánnyá emelkedett az úgynevezett *Dublin Core* adatkészlet, amelynek 15 legfontosabb adateleméből ISO szabvány készült.<sup>6</sup> A Dublin Core-ban az adatelemek HTML-ben és XML-ben egyaránt címkézhetők, az elemek szabadon választhatók és ismételhetők, sorrendjük nem meghatározott, de az általános értelmezhetőség biztosítása céljából kötött szókészlet alkalmazása szükséges. A Dublin Core adatelemeit és magyar megnevezésüket az 1. táblázat tartalmazza.

Elemnév	Magyar megnevezése
Title	cím
Creator	létrehozó
Subject	tárgy- és kulcsszavak, jelzetek
Description	leírás
Publisher	kiadó
Contributor	közreműködő
Date	dátum
Type	típus
Format	formátum
Identifier	forrásazonosító
Source	eredeti információforrás

<sup>6</sup> A szabvány Magyarországon 2004-ben jelent meg „Információ és dokumentáció. A Dublin Core metaadat elemkészlete” címmel (MSZ ISO 15836).

Language	nyelv
Relation	kapcsolat
Coverage	tér-idő vonatkozás
Rights	jogok

1. táblázat: A Dublin Core elemkészlete

### 2.3 Az elvek és a gyakorlat viszonya

Hiába jelennek meg azonban a fenti modellek és követelményrendszerek egyes elemei a kereskedelmi termékekben és szolgáltatásokban, megtévesztő lehet a felhasználók szempontjából, hogy a piacon mind az üzleti alkalmazók, mind a közigazgatás szervezetei számára többségükben olyan dokumentumkezelő és archiváló rendszerek állnak rendelkezésre, amelyek csak rövid távon biztosítják a digitális adatok és dokumentumok tárolását és felhasználhatóságát, a hosszú távú archiválás követelményeinek kielégítése nem céljuk.

A nagy iratkezelők számára kínált úgynevezett dokumentumkezelő rendszerek olyan hibrid rendszerek, amelyek megfelelő hardvertámogatással szkennelik a papír-alapú dokumentumokat, emberi közreműködéssel metaadatokat, illetve leíró adatokat rendelnek hozzájuk, majd az eredetileg is elektronikus formátumban rendelkezésre álló dokumentumokkal (szövegfájlokkal, e-mailekkel stb.) együtt kulcsszavas keresést tesznek lehetővé bennük, és megfelelő mutatókkal segítik a megőrzött papír-alapú dokumentumok előkereshetőségét. Ügyviteli hasznuk jelentős, azonban a hardver és szoftver avulása, főként pedig a rendszer szállítójától való függés csak néhány éves távlatban nyújt megoldást az elektronikus dokumentumok archiválására. Ugyanezen okok miatt a szervezeti informatikai alkalmazásoknál szokásos, a biztonsági mentések mellett végzett ún. archív mentések szintén csak néhány éves távlatban tekinthetők megbízhatónak.

A digitális kép- és mozgóképrögzítés, a tömörítési algoritmusok, a zenefájlok és multimédia-állományok terjedése néhány széles körben elterjedt, *de facto* szabvány jellegű formátumot eredményezett (például JPEG, TIFF, MPEG stb.); használatuk tömeges. Kérdéses azonban e formátumok élettartama, különös tekintettel a szórakoztató elektronikai alkalmazások generációváltásaihoz fűződő üzleti érdekekre.

Általános szinten megállapítható, hogy noha egyes alkalmazók évtizedekkel ezelőtt felismerték az elektronikus adatállományok archiválásának megoldandó problémáit, a kutatók kidolgozták az alapelveket, egyes szabványokat és megőrzési megoldásokat; léteznek néhány évtized időtartamra garantált élettartamú adathordozó eszközök illetve rendszerek, és léteznek egyes, a probléma megoldását ösztönző jogszabályok is, mind a nagy gyártók, mind a nagy alkalmazók, mind pedig az állami szervek eddig elodázták a döntést a hosszú távú átfogó, komplex megoldásokról, tekintettel azok infrastrukturális jellegére, forrásigényére, szabványosítási követelményeire.

### 3. Folyamatban lévő kutatások, fejlesztések

Az *InterPARES* (International Research on Permanent Authentic Records in Electronic Systems) nemzetközi kutatócsoport alap kutatás jellegű projektjei (InterPARES 1 és InterPARES 2) 1999 óta az elektronikus dokumentumok hosszú távú megőrzésének alapvető követelményeire, módszertanának kidolgozására, majd a komplex digitális környezetben előállított művészeti, tudományos és elektronikus kormányzati dokumentumok megőrizhetőségére irányultak. A projekt jelenlegi, 2012-ig terjedő fázisa a kis és közepes méretű archívumokra és irattárakra koncentrál, s ennek részeként az eddigi elméleti kutatások

eredményeinek gyakorlatba ültetését és a szervezeten belüli képzések tananyagának kidolgozását célozza.<sup>7</sup>

Az elektronikus iratkezelés és archiválás projektjeinek összeurópai fóruma a *DLM Forum* (Document Lifecycle Management Forum), amely az Európai Unió 1994-ben született, az archívumi együttműködést szorgalmazó határozata alapján jött létre.<sup>8</sup> A fórumban régióink is aktívan képviselteti magát; 2005. októberében Budapesten, legutóbb pedig – 2008. áprilisában – Ljubljanában rendeztek DLM konferenciát. A DLM Forum megbízásából készítette el egy szakértői csoport a 2001-ben a *MoReq* követelményrendszert (Model Requirements Specification for the Management of Electronic Records – Mintakövetelmények az Elektronikus Iratok Kezeléséhez),<sup>9</sup> amely jelenleg az egyetlen olyan EU szintű ajánlás, amelynek alapján egységes elektronikus iratkezelési rendszerek alakulhatnak ki.<sup>10</sup> 2008. februárjára készült el és vált nyilvánossá a *MoReq2*, amely nem kevesebbet tűz ki célul, mint hogy az elektronikus iratkezelés *de facto* világszabványává váljon.<sup>11</sup>

Noha egy komplett, hosszú távú archiválási rendszernek magának kell tárolnia a megőrzött adatok értelmezéséhez, felhasználhatóságához szükséges információkat, az alkalmazott formátumok számának redukálása és csereszabatoságuk biztosítása szükségessé teszi a formátumok szabványosítását, de legalább is ismertségét és hozzáférhetőségét. Ez utóbbi célt kívánja megvalósítani a Harvard Egyetem és az MIT által indított projekt, a *Globális Digitális Formátum-nyilvántartás* (Global Digital Format Registry), amelyet elsősorban könyvtárak igényeinek kielégítésére fejlesztnek. Az Online Computer Library Center (OCLC) támogatásával folyó projekt felhasználói köréhez – formátum-szállítóként és lekérdezőként egyaránt – szabadon csatlakozhatnak a digitális archiválást végző könyvtárak, levéltárak és más szervezetek.<sup>12</sup>

A svájci és német kezdeményezésre indult *ArchiSafe* projekt elsősorban az elektronikus dokumentumok megőrzésére vonatkozó jogi előírások teljesítésére és számonkérhetőségére alkalmas termékek és rendszer fejlesztését tűzte ki célul. A fejlesztés alatt álló, több technológiát integráló rendszer támaszkodik a korábbi, az elektronikus aláírások érvényességének hosszú távú fenntartására (átírására) vonatkozó *ArchiSig* projekt eredményeire, az ISO szabvánnyá vált, kifejezetten a hosszú távú archiválás céljaira kifejlesztett PDF/A dokumentumformátum alkalmazására, és mindehhez egy web alapú munkafolyamat-kezelő rendszert társít.<sup>13</sup>

Az elektronikus adatállományok archiválásának egyik alapvető követelményét, az integritás folyamatos ellenőrzését és verifikálását költséghatékony módon biztosító új technikát javasol Song és JaJa (2007), amely bármilyen centralizált, elosztott vagy peer-to-peer archiválási architektúrában alkalmazható. Módszerük lényege egy háromlépcsős objektum-regisztráció, amely az időegység alatt regisztrált objektumok számához alkalmazkodóan dinamikus, egy másodperctől egy óráig terjedő időbélyegzési felbontást (granularitást) alkalmaz, a regisztrált objektumok hash kivonatait egymáshoz láncolja, majd egy bizonyos időegység eltelte után (a prototípusban egy hét után) az időegység alatt regisztrált objektumokból egy összesített hash kivonatot készít, amit „tanú”-nak (witness) nevez és nyilvánosan publikál. A „tanú” segítségével ellenőrizhető az archívum integritása a külső szemlélők számára.<sup>14</sup>

---

<sup>7</sup> <http://www.interpares.org>

<sup>8</sup> <http://dlmforum.typepad.com>

<sup>9</sup> Magyarul lásd: [http://www.inform-consult.com/download/moreq/MoReq\\_Hungarian.pdf](http://www.inform-consult.com/download/moreq/MoReq_Hungarian.pdf)

<sup>10</sup> A MoReq követelményrendszer tükröződik „A magyar levéltárak középtávú informatikai stratégiája és feladatterve (2006–2010)” c. dokumentumban és a vonatkozó rendeletekben is.

<sup>11</sup> <http://www.moreq2.eu>

<sup>12</sup> <https://collaborate.oclc.org/wiki/gdfr/news.html>

<sup>13</sup> [http://download.openlimit.com/website/case\\_studies/WEB\\_Case\\_ArchiSafe\\_EN\\_03.pdf](http://download.openlimit.com/website/case_studies/WEB_Case_ArchiSafe_EN_03.pdf)

<sup>14</sup> <http://www.umiacs.umd.edu/~joseph/dgo2007-ace.pdf>

Az adattárolási iparág nemzetközi szakmai szövetségének<sup>15</sup> hosszú távú archiválással foglalkozó programjából (Long-Term Archive and Compliance Storage Initiative, LTACSI) két munkacsoport és tevékenysége kíván említést: a „100 Year Archive Task Force”, amely – ambiciózus elnevezése ellenére – egyelőre csupán a létező legjobb gyakorlatok összegyűjtését tűzte ki célul, valamint az önleíró adatformátumokkal foglalkozó „Self-Describing Data Format (SDDF) Task Force”, amely a nyílt szabványok elterjesztésének mint a hosszú távú logikai olvashatóság zálogának trendjével szemben olyan megoldásokat kíván kifejleszteni, amelyek segítségével a szoftvercégek birtokában lévő formátumok hosszú távú olvashatóságát formátum-leíró metaadatok használata biztosítaná.<sup>16</sup>

### 3.1 Elosztott tárolás

A folyamatban lévő kutatások és fejlesztések külön csoportja foglalkozik az elektronikus adatállományok elosztott tárolásával. A Grid technológiák nemzetközi fóruma, a Global Grid Forum (jelenlegi nevén Open Grid Forum) számára dolgozták ki a NASA kutatói a grid alapú, redundáns tárolási koncepciójú, az adatelemek szabványos meghatározását és rendelkezésre állásuk automatikus ellenőrzését megvalósító hosszú távú archiválás részletes követelményrendszerét (Barkstrom, 2005), amely a kereső interfészek automatikus generálásától kezdve a költséghatékony működtetés szempontjain és a rendszerösszeomlások kezelésén át egészen az oktatásig és az archiválási rendszer egészének rendszeres, független auditálásáig 35 követelményt és azok teljesítésének megfelelő és nem megfelelő módjait írja le, és amely ezzel a jelenlegi fejlesztések egyik viszonyítási alapjává vált. Több kutatás folyik a Bázeli Egyetemen, köztük a DISTARNET (Distributed Archival Network), amely a fentebb tárgyalt Nyílt Archívumi Információs Rendszer (OAIS) alapján definiál egy XML alapú protokollt és szabályrendszert digitális objektumok elosztott tárolására és tartós megőrzésére.<sup>17</sup> A gyakorlatban digitális folyóiratok peer-to-peer archiválására fejlesztették ki a LOCKSS<sup>18</sup> rendszert, tartós adattárolásra és biztonságos adatmegosztásra az OceanStore<sup>19</sup> rendszert; blokkokra osztott fájlok elosztott tárolásán alapul az InterMemory<sup>20</sup> rendszer, fájlok teljes replikálásán a PAST<sup>21</sup>; az anonim hozzáférhetőség és a cenzúrázatlanság biztosítását célozza a FreeHaven<sup>22</sup> rendszere. Közgazdasági fogalmak és menedzsment modellek alkalmazásával javasol aukciós eljárást a szabad tárhelykapacitások archiválási célú kihasználására Cooper és Garcia-Molina (2005)

Az elmúlt években számos javaslat született konkrét rendszerek megvalósítására és ezek közül néhány elérte a gyakorlati alkalmazhatóság szintjét. Az elosztott rendszerű digitális adattárolási rendszereknek azonban nem mindegyike célozza a hosszú távú archiválás biztosítását, és nem foglalkoznak a formátumok elavulásának kérdéseivel. Lu és Chiueh (2006) nyomán az alábbi táblázatban foglaljuk össze a replikációt alkalmazó újabb elosztott digitális adattároló rendszerek néhány jellemző sajátosságát.

---

<sup>15</sup> Storage Networking Industry Association (SNIA)

<sup>16</sup> <http://www.snia.org/forums/dmf/programs/ltacsi>

<sup>17</sup> [http://www.distarnet.ch/papers/ist\\_distarnet\\_2006-pdf](http://www.distarnet.ch/papers/ist_distarnet_2006-pdf)

<sup>18</sup> <http://www.eecs.harvard.edu/~mema/publications/TOCS2005.pdf>; <http://lockss.org>

<sup>19</sup> <http://oceanstore.cs.berkeley.edu/publications/papers/pdf/asplos00.pdf>

<sup>20</sup> <http://pnylab.com/pny/intermemory/intermemory.odf>

<sup>21</sup> <http://research.microsoft.com/~antr/PAST/past-sosp.pdf>

<sup>22</sup> <http://www.freehaven.net>

Rendszer	Archiválási célra tervezett?	Replikáció típusa	Rejtjelezett?	Auditálás?	Adatok élettartama	Skálázható?
Cooperative File System	nem	teljes replikáció	nem	nincs	bérelt időtartam	
PAST	igen	teljes replikáció	opcionális	nincs	bérelt időtartam	
FreeNet	nem	teljes replikáció	igen	nincs	nem garantált	
FreeHaven	nem	Erasure Coding	igen	nincs	bérelt időtartam	
FarSite	nem	teljes replikáció	igen	nincs	nem garantált	
Eternity Service	igen	teljes replikáció	nem	nincs	meghatározott időtartamú	igen
InterMemory	igen	Erasure Coding	nem	nincs	korlátlan	igen, ha a rendszer mérete növekszik
OceanStore	igen	Erasure Coding	igen	van	meghatározott időtartamú	igen
LOCKSS	igen	teljes replikáció	nem	van	korlátlan	nem

2. táblázat. Elosztott tárolási rendszerek összehasonlítása

### 3.2 Adathordozók

Az adathordozók élettartamának és megbízhatóságának növelésére, illetve ilyen hordozók alkalmazására irányuló kutatások között megemlítendő a Bázeli Egyetemen folyó ARCHE<sup>23</sup> és Peviar<sup>24</sup> projekt. A folyamatban lévő kutatások zömével ellentétben e projektek nem digitális adathordozók fejlesztésével és alkalmazásával foglalkoznak, hanem a hagyományos, bizonyítottan hosszú élettartamú adathordozók – esetünkben a mikrofilm , illetve a microfiche – új célra való alkalmazásával. A Peviar digitális információt tartalmazó kétdimenziós vonalkódot ír mikrofilmre, az ARCHE pedig a freiburgi Fraunhofer Intézet által kifejlesztett speciális lézert használja a digitális adatok színes mikrofilmre (microfiche-re) írására. Az így készített adathordozók élettartama elérheti az 500 évet, egy microfiche lap tárolókapacitása pedig a 700 MB-t.

Az adathordozók tárolókapacitásának növelését célzó fejlesztések kurrens eredményei közül figyelemre méltók a holografikus optikai lemez megvalósított formátumai. A Holographic Versatile Disc (HVD) elméletileg 3,9 terabájt információt tárolhat és 1 Gbit/s átviteli sebességre képes; szabványosításának első eredményei megszülettek, ISO szabványként való elfogadtatása folyamatban van. Felhasználási területét az extrém adattárolási igényű szervezetekre szabták; a média élettartama azonban közelebből nem meghatározott. Riválisai közül a legígéretesebb az InPhase Technologies által kifejlesztett és 2007. végén termékként bejelentett Tapestry holografikus lemez, amely 300 GB tárolókapacitású és általános használatra tervezett. Élettartamát a gyártó 50 évre becsüli.

### 3.3 Speciális adat- és dokumentumformátumok

A megőrzendő adat- illetve dokumentumformátumok némelyike sajátos archiválási problémákat vet fel. Vannak olyan dokumentumok, amelyek eleve elektronikus formában születtek, de nem abból a célból, hogy kinyomtaszák őket és így papír-alapúvá váljanak,

<sup>23</sup> Alkalmazásáról lásd például: <http://www.newsfox.com/pte.mc?pte=070312031>

<sup>24</sup> [http://www.peviar.ch/peviar\\_abstract.pdf](http://www.peviar.ch/peviar_abstract.pdf)

hanem hogy mindvégig elektronikus formában használják őket. Jellemzően ebbe a kategóriába tartoznak azok a nem lineáris szövegből álló (vagy nem szöveges) dokumentumok, amelyek „iratként”, sőt „dokumentumként” való megítélése amúgy sem egyértelmű a keletkeztetőnél és a felhasználóknál egyaránt. Három dokumentum-típust kell kiemelnünk ebből a körből:

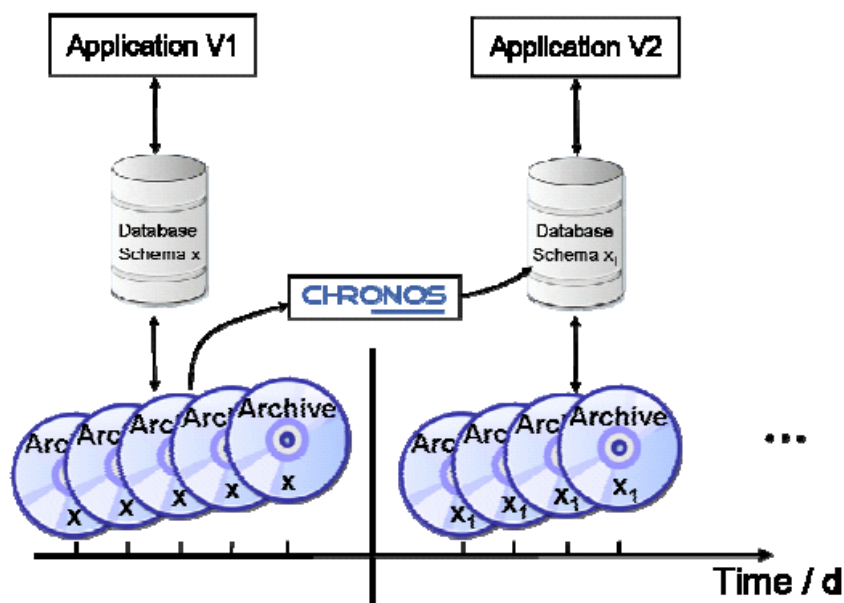
- (a) a működtető logikát is magukban foglaló, több állapotú adatbázisokat,
- (b) a csoportmunkában született, verziókat, állapotokat megőrző dokumentumokat, valamint
- (c) a digitális formátumú kép, hang és videofájlokat.

A relációs adatbázisok esetében a bennük kezelt adathalmaznak felmérhetetlenül sok rendezettségi állapota képzelhető el; nyilvánvalóan nem elég ezen állapotoknak csupán egyikét megőrzendőnek tekinteni. Ugyancsak nem kielégítő az a megoldás, amely egy korábbi adatbáziskezelő programmal előállított és használt adatállományt egy későbbi, fejlettebb (és így több lekérdezési, csoportosítási, adatelemzési lehetőséggel rendelkező) programmal használható formában őriz meg, hiszen például egy szervezet tevékenységének megítéléséhez hozzátartozik annak figyelembevétele, hogy az adott körülmények között, az adott döntések előkészítésénél milyen adatelemzési lehetőségek álltak a rendelkezésére. A csoportmunkában készült dokumentumok egyszerűbb esetben lehetnek akár lineáris szöveget tartalmazó iratok is, már ami a munka végeredményét vagy egyes részeredményeit illeti, azonban ezek előállításának folyamata nem egyetlen személy irat-előzményeinek és változatainak sorát tartalmazza csupán, hanem egy munkacsoport több tagjáéit is. A változatok értékének, irat-voltának, megőrzendőségének kérdései tehát itt is hatványozottan merülnek fel, ráadásul a megőrzendő „dokumentum” itt a csoportmunkát dokumentáló történeti adatállomány is lehet, amelynek archiválása éppen azt célozza, hogy megőrkítse az elektronikus eszközök útján végzett közös tevékenységet.

A Svájci Szövetségi Levéltár SIARD (Software Invariant Archiving of Relational Databases) projektjében a relációs adatbázisok archiválásának egyik lehetséges módját követi, az adatbázisok konvertálását egy alkalmazásfüggetlen generikus formátumba.<sup>25</sup> A SIARD a logikai adatbázis-struktúrát SQL-3 nyelven írja le, tekintettel annak nyílt szabvány voltára és részletes dokumentációjára. Egy másik lehetséges megoldást követ a kereskedelmi termékként és szolgáltatásként is megvalósított *Chronos* rendszer, amely nem konvertálja az eredeti adatbázisokat, hanem szemantikai és szintaktikai leírásukat mintegy kivonatolja és egyszerű szövegformátumban (a metaadatokat XML formátumban) tárolja; inkrementális – vagyis csak a bővülést rögzítő – archiválást tesz lehetővé, és a ma használatos adatbáziskezelő rendszerekről beépített – és feltehetően a jövőben bővülő – tudásbázist tartalmaz (Brandl és Keller-Marxer, 2007). Közös mindkét megoldásban, hogy a formátumok jövőbeli rendszeres konvertálásán, azaz migráltatásán alapulnak; a tervezett migrációs periódus a SIARD esetében 10–20 év közötti, ami a hosszú távú archiválás céljait tekintve nem tekinthető megnyugtatónak. A *Chronos* rendszer az adatbázisok inkrementális archiválása során az adatbázis sémákban várhatóan fellépő változások kezelésére is speciális migráltatást alkalmaz (5. ábra), ami szükségtelenné teszi az egyes séma-változatokhoz tartozó mindenkor teljes adatbázis újbóli archiválását. Ugyanakkor a rendelkezésre álló leírásokból nem világos, hogy ezek az adatbázis-archiváló rendszerek miként őrzik meg az adatbázisok *eredeti* (a későbbiekhez képest korlátozottabb) funkcionalitását, ami a történeti adatok kontextusának értékeléséhez szükséges lenne.<sup>26</sup>

<sup>25</sup> <http://arxiv.org/pdf/cs.DL/0408054>

<sup>26</sup> A *Chronos* például egységesen az eredeti SQL 92 szabványon alapuló lekérdezéseket végez.



5. ábra. Változó sémák és alkalmazások kezelése adatbázisok inkrementális archiválása során (Brandl és Keller-Marxer, 2007)

Ugyan a jelenkorban készülő audiovizuális felvételek és más megőrzendő dokumentumok egyre nagyobb hányada jön létre eleve digitális formában, az emberiség eddig felhalmozott tudásanyagának túlnyomó része hagyományos adathordozókon és formátumokban található. Ahhoz, hogy ezt a tudásanyagot a digitális archiválás tárgyává tegyük, először digitalizálni kell az egyes elemeit.

Az ambiciózus digitalizálási és archiválási projektek közül kiemelendő a két nagy rivális a Google és a Yahoo törekvése. Mindkét cég könyvek digitalizálását, elektronikus formában való archiválását és interneten keresztüli szabad hozzáférhetőségét tűzte ki célul. A legambiciózusabb projekt azonban az Internet Archive,<sup>27</sup> amely az interneten valaha elérhető összes weboldal archiválását tűzte ki célul. Időgépes szolgáltatásával e kézirat lezártakor több mint 85 milliárd weboldal érhető el 1996-tól napjainkig terjedő állapotában. Egy kapcsolódó projektben pedig az ókori Alexandriai Könyvtár teljesítményéhez hasonlóan minden valaha kinyomtatott könyv összegyűjtését, digitalizálását és teljes szöveges kereshetőségű online elérhetőségét próbálják megvalósítani, ami az így létrejövő adatállományok hosszú távú archiválásának problémáit is felveti.<sup>28</sup> Jelenleg 13 szkennelő központban összesen napi 1000 – egyébként is szabad felhasználhatóságú – könyvet digitalizálnak, a digitalizált állomány meghaladja a 300.000 könyvet – ez még természetesen csak töredéke a világtörténelem eddigi könyvtermésének. Az Internet Archive jelenlegi tárolókapacitása körülbelül 3 petabyte (3 millió gigabájt).

### 3.4 Az univerzális virtuális számítógép víziója

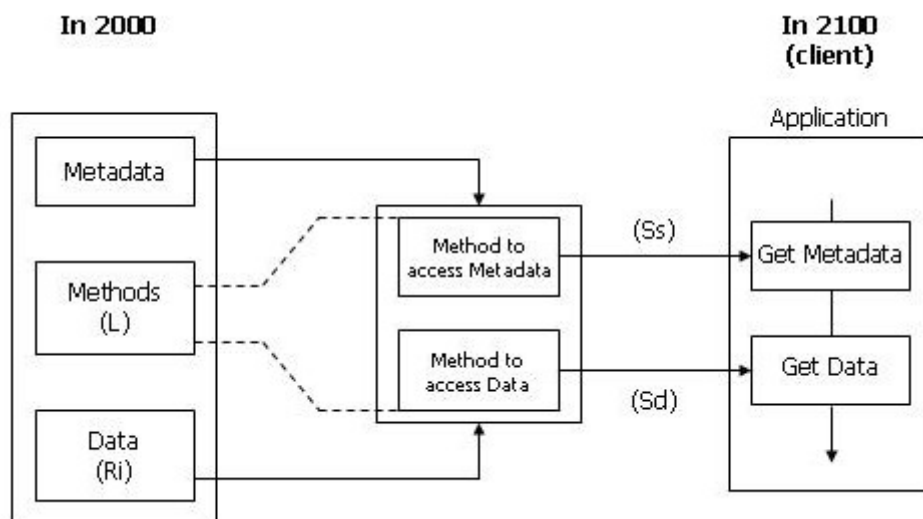
Végül, a jövőben születendő elektronikus állományok archiválásának és tartós rendelkezésre állásának elvi megoldására született az IBM univerzális virtuális számítógép koncepciója és működőképes megvalósítása. Eszerint minden jövőbeli digitális állomány létrehozásánál ugyanazt a szabványos kódolási eljárást kellene alkalmazni, ami lehetővé tenné, hogy azok az

<sup>27</sup> <http://www.archive.org>

<sup>28</sup> Az Új Alexandriai Könyvtár – némiképp szimbolikusan is – tükrözi az Internet Archive időgépes szolgáltatását, s ezzel nemcsak az online hozzáférhetőséget javítja, hanem biztonsági másolatként is szolgál.

aktuális környezettől függetlenül egységesen visszafejthetők és használhatók legyenek. Ehhez azonban egy világszabvány kidolgozása és elterjesztése lenne szükséges.

Az univerzális virtuális számítógépet (Universal Virtual Computer, UVC) nem kell a valóságban megépíteni, csupán követni működési elveit és használni számítógépes nyelvét. Célja a digitális objektumok eredeti formájában való helyreállítása egy tetszőleges jövőbeli időpontban; nem célja ugyanakkor biztosítani azt, hogy az így helyreállított objektumokkal közvetlenül dolgozni is lehessen. Az UVC két objektumtípust különböztet meg: adatot és programot.<sup>29</sup> Adatok archiválásánál a megőrzendő objektum tartalmazza a szöveg tárolásához felhasznált ábécé leírását, a kapcsolódó metaadatok alapvetően szöveges leírását, az adatokat bitfolyam formájában, valamint azt a kódot, amely az UVC instrukciókat tartalmazza, és amelyet a jövőben az UVC interpreter segítségével a felhasználó számítógépe értelmezni tud és ezáltal eredeti formájában helyreállítani a megőrzött tartalmat. A folyamat főbb lépéseit a 6. ábra illusztrálja, ahol  $L$  az algoritmus specifikálásához használt nyelvet,  $R_i$  a belső reprezentáció információit,  $S_d$  az adatmodellhez tartozó sémákat,  $S_s$  a sémák olvasásához szükséges sémát jelöli.



6. ábra. Adatok archiválása és felhasználása univerzális virtuális számítógéppel (Lorie, 2000)

Programok archiválásánál a megőrzendő objektum ugyancsak tartalmazza a szöveg tárolásához felhasznált ábécé leírását, a kapcsolódó metaadatok alapvetően szöveges leírását, a végrehajtható programot bitfolyam formájában, valamint az UVC instrukciókat tartalmazó, az UVC interpreter segítségével visszafejthető kódot, amely emulálja az eredeti környezet funkcionalitását és képes futtatni a megőrzendő programot e környezetben.

#### 4. A várható fejlődés

Az előző alfejezetekben bemutatott kutatási-fejlesztési irányokból, az adat-keletkeztetők és adatőrök gyakorlatából és stratégiai dokumentumaiból megállapíthatjuk, hogy az elektronikus dokumentumok közép-, illetve hosszú távú megőrzésére elvileg három, eltérő megközelítésen alapuló és eltérő technológiai követelményeket és következményeket involváló koncepció alakult ki: a megőrző, a migráltatáson alapuló és az „információ az információról” koncepció.

A *megőrző koncepció* megvalósítása tiszta formájában azt jelenti, hogy az elektronikus adatállományok megőrzéséért és rendelkezésre állásuk biztosításáért felelős (egyedi vagy

<sup>29</sup> <http://www.freepatentsonline.com/6691309.html>

központi) intézményeknek voltaképpen egy folyamatosan bővülő műszaki múzeumot kellene fenntartaniuk, amelyben megőriznek legalább egy működőképes példányt minden olyan hardverből és szoftverből, amelynek segítségével az archivált dokumentumokat előállították, illetve amelyek a dokumentumok használatához szükségesek voltak az adott kor technikai körülményei között. Ezek az intézmények a megőrzött hardverek és szoftverek segítségével teszik hozzáférhetővé és értelmezhetővé a megőrzött elektronikus dokumentumokat. Ezenkívül biztosítják a rendszerek hosszú távú működőképességét, beleértve a szervizelést, az alkatrész-utánpótlást (ami például a chip-gyártás esetében igen magas költségű kis sorozatok előállításánál), illetve olyan számítástechnikai szakemberek – afféle „paleo-informatikusok” – képzését és alkalmazását, akik értenek a régi adatkezelő eszközök működtetéséhez, használatához, javításához.

A *migráltatáson alapuló koncepció* szerint a korábbi formátumban készült, illetve tárolt elektronikus adatállományokat időről időre konvertálni (migráltatni) kell a mindenkori aktuális formátumokra. A migráltatást meg kell különböztetni az adathordozók időszakos átírásától, bár alapvetően más elven működő hordozókra való átírásnál a két folyamat történhet egymással összefüggésben is. A migráltatás alapvetően kumulatív jellegű feladat: ma a tegnapi állományt kell migráltatnunk, holnap a tegnapi és a mai is, és így tovább. A kumulatív terhek növekedése mellett a migráltatás két kritikus problémája az állományok azonosságának és eredeti funkcionalitásának biztosítása. Az azonosság követelménye elsősorban azon dokumentumoknál jelentkezik, amelyek használatához joghatás fűződhet, de hasonló fontosságú a tudományos célt szolgáló adatállományoknál, dokumentumoknál is. Az eredeti funkcionalitás megőrzése jellemzően a nem-lineáris olvasásra, használatra szánt állományok esetén bír jelentőséggel, például hypertext, táblázatkezelők, adatbázisok esetében, de ide sorolhatók a szövegszerkesztővel készült állományok jegyzetei, kereszthivatkozásai is. Az „*információ az információról*” típusú koncepció két megvalósítási formája az *emuláció* és a „becsomagolás” (bundling). Az emulációt a legtöbb felhasználói alkalmazásra szánt szoftvercsomag jelenleg is nyújtja: e funkció segítségével a szoftvercsomag elődjeivel (pl. 16 bites környezetben futó változataival) készített dokumentumokat a későbbi változatokkal is lehet olvasni, sőt a korábbi formátumban is lehet módosítani, illetve új dokumentumokat is lehet régebbi formátumokban menteni. Mivel a szoftvercsomagok gyártóinak üzleti érdeke, hogy megtartsák a korábbi változatokat használó felhasználóikat és áttereljék őket az újabb verziók használatára, a saját korábbi és későbbi formátumok átjárhatósága egy néhány éves időszámban biztosítottnak tekinthető.<sup>30</sup> A hardver-emulációs megoldások szoftveres úton hozzák létre azt a hardverkörnyezetet, amelyben az eredeti adatállományt létrehozták és használták, illetve amelyben az archivált adatállomány használatához szükséges szoftverek eredetileg futottak.<sup>31</sup>

Amíg az emuláció rövid- és középtávon használható, a „becsomagolás” kifejezetten hosszú távú alkalmazásra szánt elképzelés. Ennek is a lényege az eredeti szoftver- és hardverkörnyezet emulációja, azonban az ehhez szükséges dokumentációt és pontos specifikációt egy független információs burok tartalmazza, amelynek segítségével előállítható az a környezet, amelyben a megőrzött adatállomány újból használhatóvá válik. Ezt az információt az elképzelések szerint emberi nyelven olvasható, papír vagy mikrofilm hordozójú leírás tartalmazná, amely az elektronikus állomány adathordozóján vagy annak csomagolásán lenne elhelyezve.

---

<sup>30</sup> A más gyártók (korábbi vagy jelenlegi) formátumaival való átjárhatóság csak egyik irányban érdeke a szoftvercégeknek: a más formátumból saját formátumba alakítás előnyös számukra, a visszafelé irány nem; ez tükröződik a szoftvercsomagok szolgáltatásaiban.

<sup>31</sup> A hardver-emulációk közismert kommerciális alkalmazása egyes "klasszikus" számítógépes játékok "retro-környezetben" való futtatása.



7. ábra. Várható fejlődés 2008–2018

A Gartner Group ezredforduló körüli ajánlása (Logan *et al.*, 2001), amely szerint minden tíz évnél hosszabb távra megőrzendő dokumentumot ember által közvetlenül olvasható formában, például mikrofilmen célszerű tárolni, a kérdés megoldatlanságát, inkább megkerülését tükrözi. Az elektronikus adatállományok elektronikus formában történő archiválását nemcsak az elektronikus formában létrejövő állományok egyre növekvő mennyisége teszi szükségessé, hanem visszakereshetőségük, felhasználásuk, elemezhetőségük biztosítása is. Az elektronikus üzenetek tárhely alapú archiválására szakosodott piaci szegmens átalakulóban van<sup>32</sup> és egyre inkább alkalmassá válik többféle formátum és irattípus integrált kezelésére – azonban ezek a szolgáltatások alapvetően csak a kurrens és félkurrens állományok kezelését célozzák, tehát csak a rövid-, esetleg középtávú archiválás problémáira nyújtanak jelenleg megoldást.

A megőrző koncepciót csupán néhány kutatóintézet követi, korlátozott körben, de nyilvánvaló, hogy a koncepció széleskörű, hosszú távú alkalmazása nem járható. Az emuláció, mint a felhasználói szoftvercsomagok áthidaló szolgáltatása várhatóan fennmarad, sőt a szabad szoftverek alkalmazói szintű terjedésével szerepe növekedhet. A „becsomagolás”, illetve az univerzális virtuális számítógép alkalmazása kísérleti fázisban van, a vizsgált időszakban alkalmazásukban áttörés nem várható. A dokumentumkezelő rendszerek választéka és alkalmazói köre várhatóan bővül, azonban önmagukban csak néhány éves időszámban adnak kielégítő megoldást. A megőrzendő adatállományok tárolása terén várhatóan tovább terjed az elosztott és peer-to-peer megoldások alkalmazása, azonban jelentős elvárásbeli és teljesítésbeli különbségek alakulnak ki a „best effort” és a „quality of service” típusú szolgáltatások között. A hosszabb távú archiválási szükségletek kielégítésére leginkább olyan törekvések folytatása várható, amelyek meghosszabbítják az elektronikus dokumentum-

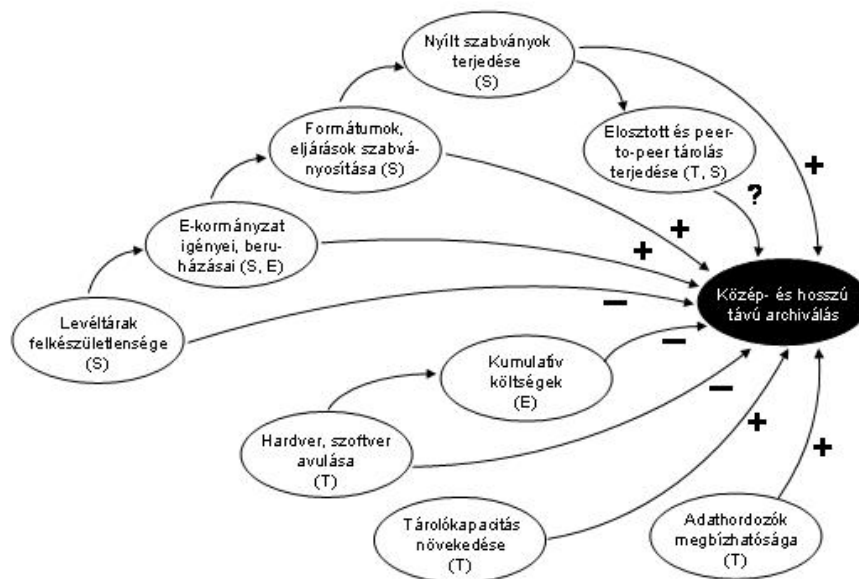
<sup>32</sup> A Forrester Research 2004. szeptemberében azt jósolta, hogy a piaci szegmens üzleti forgalma három év alatt megötszöröződik és 2006-ra elérheti az egymilliárd dollárt; ehhez képest 2006-ban elemzői arról írtak, hogy a tárhelyszolgáltatók feltámadtak hamvaikból (<http://www.forrester.com/Research/Document/Excerpt/0,7211,38034,00.html>). – A piaci viszonyok gyors változásai nem kedveznek a hosszú távú megoldások elterjedésének.

formátumok felhasználhatóságának időtartamát, és elvileg biztosítják egy későbbi migráció lehetőségét.

## 5. Befolyásoló tényezők

A közép- és hosszú távú archiválás technológiai előfeltételei közül a tárolóeszközök fejlesztése viszonylag belátható pályán halad. Az optikai tárolók fejlesztésében a vizsgált időszakban a kéklézeres eszközök elterjedésétől, illetve a holografikus lemezek megjelenésétől várható a fajlagos kapacitás növekedése. Fontos feltétel a tárolóeszközök megbízhatósága, ellenőrizhetősége, hosszú távú üzembiztossága és tartóssága; ez elsősorban a RAM típusú eszközök esetében szorul bizonyításra és fejlesztésre. A technológiai előfeltételek között megemlítendő a szükséges migrálási hardver- és szoftver-eszközök rendelkezésre állása, valamint a szükséges migrálási kapacitás biztosítása.

Tekintettel arra, hogy a közép- és hosszú távú archiválás tárgyát olyan információk képezik, amelyek élettartama hosszabb a létrehozásukhoz, illetve tárolásukhoz eredetileg használt hardver és szoftver élettartamánál, ez utóbbiak technológiai és erkölcsi avulása akadályát képezi a tartós megőrzésnek, illetve speciális eljárások alkalmazását teszi szükségessé. Mind az adathordozók rendszeres cseréje, mind a redundáns tárolás esetén a meghibásodó adattároló egységek működésének időszakos helyreállítása, mind pedig a rendszeres migráltatás kumulatív költségeket jelent az adatállományok megőrzési időtartama alatt. Ezeket a költségeket az archiválási rendszer létrehozásának és működtetésének költségei között legalább is figyelembe kell venni, ha nem is lehet pontosan számítani.



8. ábra. Befolyásoló tényezők

A hagyományos levéltárak többsége még nincs felkészülve az elektronikus adatállományok tömeges fogadására és hosszú távú megőrzésére, hozzáférhetővé tételére. Ez a tény is az egyik ösztönzője azoknak az elektronikus kormányzati (e-Government) fejlesztéseknek és beruházásoknak, amelyek nemcsak a közigazgatási eljárások elektronizálását, hanem az ennek során keletkező adatállományok tartós megőrzését is célozzák.

Az adat- és állományformátumok, szoftverek egységesítése, csereszabotossága, illetve egységes archiválási szabványok kidolgozása és követése alapvető fontosságú. Jelentősen növeli a hosszú távú archiválás esélyeit a nyílt szabványon alapuló eszközök, ezen belül a nyílt forráskódú szoftverek alkalmazása. Hasonlóan fontos az archivált dokumentumok

hozzáférhetőségének, kereshetőségének biztosítása, amelynek alapvető feltétele a dokumentumok egységes metaadat-szabványok alapján történő indexálása.

Az elosztott, illetve peer-to-peer tárolási megoldások hatása egyelőre kérdéses: egyfelől segítik a tárolókapacitás és hozzáférés korlátainak lebontását, másfelől viszont nem bizonyított a „best effort” alapon működő rendszerek hosszú távú megbízhatósága és rendelkezésre állása.

## **6. Várható hatások**

Az elektronikus dokumentumok közép- és hosszú távú archiválásának igénye ösztönözheti az adatbiztonság, illetve az informatikai biztonság fenntartására és fejlesztésére irányuló törekvéseket, elsősorban az állományok integritásának és rendelkezésre állásának területén. Ösztönözheti továbbá olyan, széleskörűen használható alkalmazások fejlesztését, amelyek az archivált dokumentumok hitelességét közép- és hosszú távon is biztosítják – ez a ma használatos elektronikus aláírási rendszerek időszakos „felülbélyegzését” igényelheti. Az archiválási igények felvetik a szabványok egységesítésének igényét, és ösztönzik az adattároló eszközök teljesítményének, megbízhatóságának növelését. További igény a tartalom- és kontextus-orientált kereshetőség biztosítása az archív állományokban is, ez pedig – az univerzális hozzáférhetőség igényével együtt – ösztönzi a nyílt szabványok és szoftverek használatát. Az archivált állományok számának és terjedelmének növekedésével megoldandó feladat lesz a könnyű átjárhatóság biztosítása az archív és az operatív/kurrens rendszerek között.

Társadalmi szinten az elektronikus dokumentumok archiválásához ugyanolyan alapvető érdekek fűződnek, mint a papír-alapú vagy más analóg hordozójú dokumentumokéhoz: a jogbiztosítás, igazgatási és üzleti elemzések visszamenőleges végezhetősége, tudományos kutatások (hosszabb idősoros elemzések, történeti jellegű kutatások) végezhetősége, végső soron a kollektív (csoportszintű, szervezeti szintű, nemzeti, regionális, európai, sőt globális) emlékezet megőrzése, a kulturális identitás fenntartása. Az archivált információk egy része ugyan "elavul", ezeket a szó szoros és átvitt értelmében felülírjuk, de a változások történetisége még a gyorsan avuló információk terén is külön értéket képvisel, a humán szférában pedig a felülírás egyenesen ellentétes lenne alapvető értékrendünkkel.

A közép- és hosszú távú archiválás körülményeinek biztosítása infrastrukturális jellegű beruházásokat és üzemeltetést igényel, ezért piaci alapon csak egyes üzletileg érdekelt szektorokban lehet megvalósítani, bár ott is fennáll a szelektív megőrzés, illetve konvertálás veszélye. A költségek hasznosulása csak áttételesen értékelhető; a jelenlegi becslések szerint a migráltatás költsége legalább kétszerese az előállítás költségének, ráadásul kumulatív jellegű. Ezért a közép-, illetve hosszú távú megőrzésre szánt adatállományok előállításakor figyelembe kell venni azok tárolásának jövőbeli járulékos költségeit is.

Az üzleti életben feltehetően a közép- (és hosszú) távú adatsorok elemzése csak olyan szektorokban jelent versenylőnyt, ahol a fogyasztói mintázatok viszonylag állandóak. Egyébként az archiválás szervezeti szinten is beruházásigényes, és a fenntartás is forrásokat igényel (állandóakat, pl. őrzés, az adathordozó védelme, és időszakosakat, pl. a hordozó átmásolása, adatok migráltatása). A papír (mikrofilm) alapú tárolás megszüntetése ezért csak rövid távon nyújt tiszta költségmegtakarítást, ugyanakkor a későbbi hozzáférhetőség szempontjából számottevő kockázatot is jelent. Az adattárolási költségek csökkentésének igénye tovább ösztönözheti az outsourcing jelleggel nyújtott adattárolási és archiválási szolgáltatások terjedését, ezek azonban nem tévesztendőek össze a közép- és hosszú távú archiválás követelményeit „quality of service” alapon kielégítő rendszerekkel.

Az archivált elektronikus adatállományok hozzáférhetőségének szerzői jogi kérdéseit e helyütt csak megemlíthetjük: az információs társadalom hálózatos, digitális világában a szerzői

jog megújítási igénye, sokak szerint válsága a közép- illetve hosszú távra archivált adatállományok esetében is jelentkezik.

A csak elektronikus formátumban létező dokumentumok számának növekedése, a „papír nélküli iroda” gyakorlatának terjedése, illetve az elektronikus dokumentumok archiválási problémáinak tartós megoldása közötti „olló” egyelőre még nyílik; bezárulása csak a vizsgált időszakot követően várható. Addig viszont bizonyos értelemben „lyukak” keletkezhetnek a történelemben; a jövő generációk kutatói vagy érdeklődő utódaink egyes, számukra fontos elektronikus adatállományokat már nem találnak meg, vagy ha meg is találnak, nem tudják felhasználni.<sup>33</sup>

Másfelől viszont megkérdőjelezhető az a törekvés, hogy a technológia segítségével a jövőben „minden” információt „örökre” meg kell őrizni; az archiválás több ezer éves történelme mindig magában foglalta az értékelés, szelektálás mozzanatát. Az egyre növekvő információtömeg nemcsak annak kezelését, hanem az értelmes tájékozódás lehetőségét is nehezíti. A probléma enyhítésére egyfelől az intelligens ágensek, másfelől a hálózati keresés általános intelligenciájának mint beépített szolgáltatásnak a javítását célzó újítások – például a szemantikus web megalkotására vonatkozó elgondolások – kínálnak jövőbeli megoldásokat. A kollektív emlékezet megőrzése mellett az archiválási projektekben megjelenik a globális hozzáférhetőség, mint távlati cél megvalósításának igénye is.<sup>34</sup> Emellett várható az emlékezetörző intézmények (múzeumok, könyvtárak, levéltárak) konvergenciájának növekedése, amit a digitalizálás és az adatállományok archiválása, közös kereshetősége tovább erősíthet.

## 7. Hazai helyzet

Magyarországon az elmúlt években megszülettek azok a jogszabályok, amelyek a kurrens elektronikus dokumentumok kezelésének egyes szabályait tartalmazzák, azonban a közép- és hosszú távú archiválás szakmai követelményrendszerét a jogalkotó még nem alkotta meg. Megszületett a magyar levéltárak középtávú (2006–2010) informatikai stratégiája és feladatterve, a hagyományos levéltári intézményrendszer azonban, néhány kezdeti kísérlettől eltekintve, alapvetően felkészületlen az elektronikus dokumentumok tömeges levéltári kezelésére, aminek nemcsak az elektronikus dokumentumok átvételére, tárolására és visszakereshetőségük biztosítására kellene kiterjednie, hanem a levéltárak felügyeleti szerepének gyakorlására, például a dokumentumkezelés ellenőrzésére, a selejtezésre is. A stratégia szerint nem indokolt valamennyi közlevéltár felkészítése az elektronikus iratok átvételére és megőrzésére; első lépésként a Magyar Országos Levéltár szervezeti keretei között felállítandó stratégiai központ felállítását tartja sürgető feladatnak, az e-levéltár működésének beindulását pedig – szintén a MOL keretei között – 2010-re ütemezi. Néhány működő intézmény már végez digitális archiválási feladatokat Magyarországon. A Nemzeti Audiovizuális Archívum (NAVA) magyarországi műsorszolgáltatók által sugárzott műsorszámokat gyűjt és tárol törvényi felhatalmazás alapján, és gyűjteményét az ún. NAVA-pontokon keresztül oktatási és kutatási célra hozzáférhetővé teszi; a saját tevékenységéről nyilvánosságra hozott dokumentumokból azonban nem derül ki, hogy milyen archiválási koncepciót követ hosszú távon.<sup>35</sup> Egy elektronikus archiválási szolgáltató nyilvántartásba vétele már megtörtént meg a Nemzeti Hírközlési Hatóságnál – mégpedig minősített szolgáltatóként – és ez a vállalkozás látja el a közjegyzői elektronikus levéltár üzemeltetését is. A szolgáltató elektronikus aláírással és időbélyegzéssel ellátott dokumentumokat archivál,

<sup>33</sup> E problémát is említi: „A történelemben lesz egy lyuk” (Talyigás, 2003)

<sup>34</sup> Erről bővebben lásd Székely (2007).

<sup>35</sup> <http://www.nava.hu>

migráltatást nem végez, hanem az eredeti formátum olvashatóságához szükséges szoftver- és hardverkörnyezetet biztosítja, ennek időtávja azonban nem meghatározott.

A hosszabb távú archiválás problémáinak megoldására leginkább a központi államigazgatási szerveknél van esély, az elektronikus kormányzati infrastruktúra kiépítéséhez kapcsolódva; az önkormányzatok jelentős támogatásra szorulnának e téren.

## **8. Összegzés**

Az elektronikus dokumentumok közép- és hosszú távú archiválása terén megtörtént a problémák tudatosulása, az archiválás koncepcióinak és elvi követelményrendszerének kidolgozása, vannak egyes részterületeket lefedő K+F projektek, elindultak az ambiciózus digitalizálási programok és az elosztott tárolást nyújtó szolgáltatások; a közép- és hosszú távú megoldások azonban még nem egységesedtek és nem terjedtek el. Az EU aktívan ösztönzi az egységesítést és szabványosítást; Magyarország eddig elsősorban a kurrens dokumentumok kezelése terén tett érdemi lépéseket. Tekintettel egyfelől a beruházási és üzemeltetési költségekre, másfelől a szabványosítás igényére, a problémák megoldásához állami szerepvállalás szükséges, ami infrastrukturális jellegű szemléletet, szakmai megalapozottságot, valamint megfelelő oktatási, továbbképzési háttérrel feltételez.

