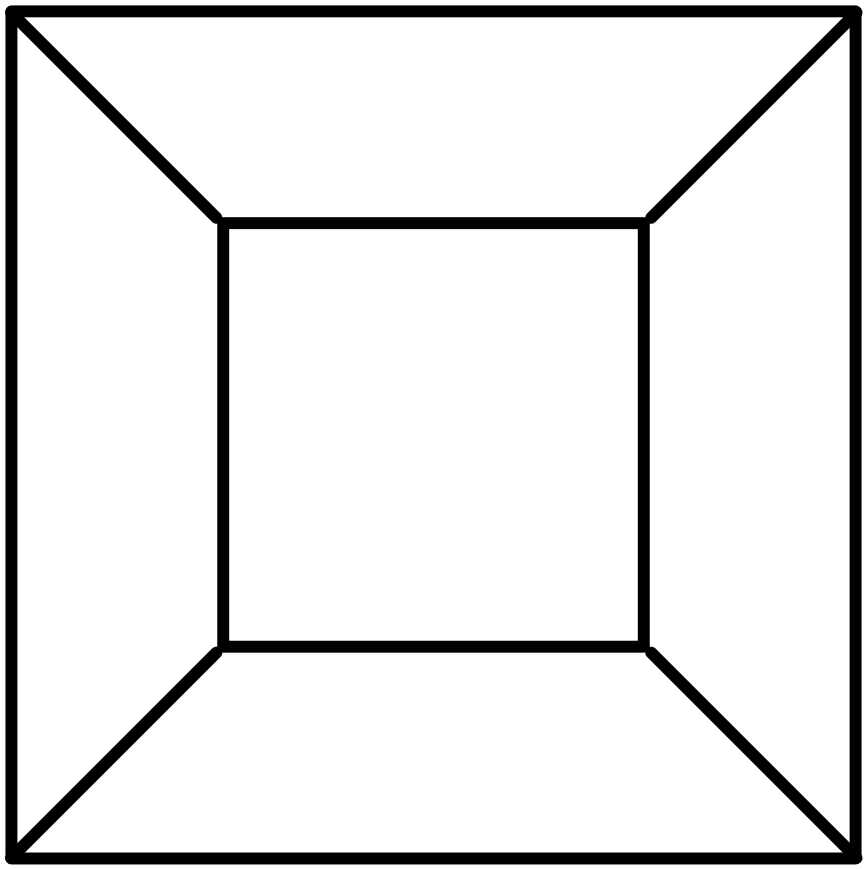


ORSZÁGOS MESTERSÉGES
INTELLIGENCIA KUTATÁS
FEJLESZTÉSI
SZUPRASTRUKTÚRA

Tartalom

OMIKI
Mesterséges intelligencia kutatás-fejlesztés támogató országos szuprastruktúra[1] létesítése és adatkezelési funkciói[2]

Dr. Kovács László (laszlo.kovacs@sztaki.hu)
SZTAKI, Számítástechnikai és Automatizálási Kutató Intézet, Budapest

Pallinger Péter (peter.pallinger@sztaki.hu)
SZTAKI, Számítástechnikai és Automatizálási Kutató Intézet, Budapest

Vers. 1.4 (2019.11.14)

Összefoglaló

E dokumentumban elsősorban a mesterséges intelligencia (MI) kutatás-fejlesztésekhez szükséges adatkezelés, adatfeldolgozás kérdéseit elemezzük, meghatározzuk egy MI kutatás-fejlesztés célú specializált és célorientált adatkezelés általános funkcióit.

Javaslatot teszünk a magyar MI kutatás-fejlesztést támogató, országos lefedettséget biztosító digitális hálózati szuprastruktúra (OMIKI) létesítésére. Ezen szuprastruktúra létrehozását és folyamatos üzemeltetését egy elkülönített szervezet végzi, melyet mint egy új, létrehozandó intézmény képzelünk el.

Tárgyaljuk az új szuprastruktúra, a fenntartó szervezet és egyéb szervezetek viszonylatait az adatáramlás és adatfelhasználás tekintetében, az szuprastruktúra vázlatos belső működésének terveit, főbb komponenseit és azok kapcsolatait, valamint az szuprastruktúrán belüli adatkezelési részszuprastruktúra által nyújtott adatfeldolgozási szolgáltatásrendszert nagy vonalakban.

Nem tárgyaljuk részletesen viszont itt, hogy az e helyen felvázolt szuprastruktúrát fenntartó szervezet hogyan illeszkedik bele az MI K+F teljes hazai és nemzetközi

[1] A “szuprastruktúra” olyan felettes, az adatok, szoftverek, logikai entitások stb. kezelésére szolgáló adat- és tartalom-infrastruktúráját, interoperábilisan szervezett adat- és tartalomkezelő szolgáltatások rendszerét jelöli, mely központilag segíti ezen entitások hatékony menedzselését. A “szuprastruktúra” fogalom megkülönböztetését az infrastruktúra fogalmától az teszi szükségessé, hogy a köznapi nyelv infrastruktúrán leginkább az energiaellátást, a közlekedést, a kommunikációs stb. igényeket kiszolgáló, leginkább hardver jellegű hálózati rendszereket ért. Az infrastuktúra, mint “alant” elhelyezkedő rendszertől eltérően a “szuprastruktúra” a “felett”, a köznapi értelmű infrastruktúra felett elhelyezkedő, a saját szintjén infrastrukturális jelentőségű hálózatosan szervezett rendszert jelöli.

[2] E dokumentum a Mesterséges Intelligencia Koalíció: Sandbox kialakítása, adatmegosztási és felhasználási kísérleti környezet megteremtése munkacsoport keretében, egyben a HRDA - Hungarian National Node of RDA, the Research Data Alliance of Europe támogatásával jött létre.

ökoszisztémájába, hanem tételezzük, hogy a teljes ökoszisztéma tervei illetve áttekintése már más MIK dokumentum(ok)ban vázolt illetve részletesen elemzett, a létrehozás, a működtetés, a fenntarthatóság stb. szempontjaiból.

E dokumentumhoz csatoljuk, jó példaként a nemzetközi EGI által kezelt pán-európai (cloud) infrastruktúra áttekintését, mely bár nem specifikusan MI K+F célokra épült ki, azonban jó példákat találhatunk benne arra, hogy milyen módon lehetséges nagy nemzetközi (adatkezelő) infrastruktúrákat létrehozni és működtetni, valamint azt, hogy melyek a fontosabb funkciók.

Jelen dokumentum viszonya más MIK létrehozta dokumentumokhoz

MIK dokumentumok viszonylata

Jelen dokumentum készítésekor rendelkezésünkre állt az MI Koalíció (MIK)

- "*Adatpolitikai stratégiai javaslat az MI-alapú innováció beindítására Magyarországon*" v. 2.0 (2019.04.15) című, az adatipar, adatvagyon munkacsoport által létrehozott dokumentum, valamint a

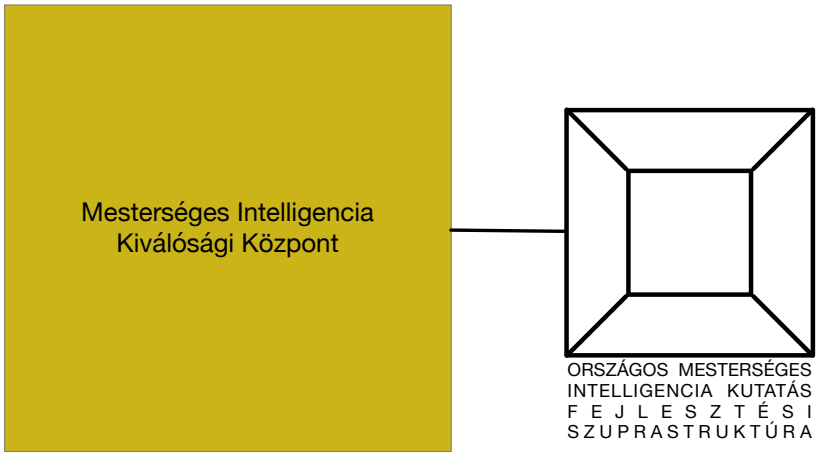
- "*Mesterséges intelligencia kutatási felhő"* v 1.0 (2019.09.09) záródokumentum, melyet a technológia és biztonság munkacsoport hozott létre.

Mindkét dokumentum tartalmát ismertnek tételezzük és a bennük foglaltakat nem ismételjük meg, legfeljebb csak hivatkozunk rá.

A MIK technológia és biztonság munkacsoport által jelenleg még csak prezentáció (kézirat) formájában létező "Adatpiac és felhőközpont" koncepcióban felvetett funkciók (leginkább az adatpiac funkció) az OMIKI szuprastruktúra néhány fontos részfunkciójáról szóló gondolkodásnak tekinthető és mint ilyen, harmonikusan simul bele az OMIKI szuprastruktúra magasabb szintű komplex szolgáltatásrendszerébe.

Intézményi viszonylatok

Jelen dokumentumban javasolt OMIKI szuprastruktúra intézményi viszonylatait a következő ábra jellemzi:



A továbbiakban tételezzük, hogy a tervbe vett MI Kiválósági Központ maga, avagy egy neki alárendelt fenntartó szervezet hozza létre, működteti és fejleszti tovább jelen dokumentumban javasolt OMIKI országos MI K+F támogató digitális szuprastruktúrát.

Az MI kutatások és fejlesztések (generikus) munkafolyamata

Az MI kutatás-fejlesztések speciális adatkezelési megoldásokat követelnek meg. Az MI kutatások, mind pedig az MI alapú (pl. applikáció) fejlesztések során egy jól körülhatárolható, de facto standardizálható munkafolyamat alakul ki. E munkafolyamatot ebben a dokumentumban kizárólag az adatkezelés illetve adatáramlás tekintetében írjuk le, más tekintetben (pl. humán és gépi erőforrás felhasználás, szervezés, financiális kérdések, stb.) az erre vonatkozó, más munkacsoportok által létrehozott specifikációs célú leírásokra hivatkozunk.

Az általános K+F célú generikus adatkezelés, mint adatkezelési ciklus főbb állomásai és főbb tevékenységei a következők:

1. Adat létrehozás
Az adatok létrehozásának sokfélesége (mérések, megfigyelések, kísérletek, szimulációk, stb. révén) mellett a meglévő adatok elérését szükséges biztosítani, mely funkciókat legtöbbször egy adatkezelési terv tartalmazza és benne a tevékenységhez szükséges adatok formátumának, mennyiségének, tárolásának és hozzáférésének mikéntje kerül meghatározásra. Az adatok mellé az adatok eredeti létrehozásának mikéntjét leíró illetve eredeti adatforrásokat meghatározó kísérő metaadatok (provenance metadata) társulnak.

2. Adatfeldolgozás
Az adatfeldolgozás tevékenységi körei felölelik a tárolás, szűrés, tisztítás, mozgatás, többszörözés, adatellenőrzés, validálás, adattranszformáció, (meta)adattársítás, adatleírások stb. tevékenységeinek igen széles palettáját.

3. Adatelemzés
Elemzés az adatok értelmezését, az adat - információ - tudás folyamat megvalósítását jelenti különféle célokból (pl. származtatott adat létrehozása, a kutatási feladat elvégzése, a publikálás avagy más adatkezelési célok – pl. hosszú távú tárolás – előkészítése érdekében).

4. (Hosszú távú) digitális adatmegőrzés
E tevékenység magában foglalja az adatok fizikai (változatlan formában történő) megőrzését, példányosítást, többszörözést, archiválást, az adatok logikai és szemantikai struktúráinak megőrzését (pl. formátum-migráció segítségével).

5. Adathozzáférés biztosítása
Az adatok megosztása és elosztása, a hozzáférés szabályozása, copyright meghatározása.

6. Adat (újra) felhasználás
Az adat felhasználását és újrafelhasználását elősegítendő tevékenységek mint például az adatok megtalálását lehetővé tévő/elősegítő metaadatok létrehozása és terjesztése/megosztása, az adattartalmak szemantikájának informális és formális leírása, az adatfelhasználás mikéntjének leírása és ezen leírások (pl. mint best practice) terjesztése, hirdetése.

Az MI célú K+F speciális igényei az adatok feldolgozásának mikéntje tekintetében

Az MI K+F célú tevékenységek speciális, ugyanakkor majdnem egységesíthető, kvázi standardizálható adatkezelési elvárásokat teremtenek. Ebben a fejezetben ezeket a célorientált, generikus MI adatkezelési tevékenységeket tekintjük át, továbbra is tartva magunkat az előzőleg felvázolt általános adatkezelési ciklus főbb állomásaihoz.

Az MI felhasználás célú adatállományok létrehozás tekintetében nem különböznek szignifikánsan az általános kutatási adatok létrehozásától, legfeljebb a big data területen nagy(obb) tömegű adatkezelés jellemzőbb (legalábbis a gépi tanulás MI megközelítés esetén).

MI esetében a **szenzoriális adatkeletkeztetés** egy jellemző adatlétrehozási módozat, mely esetében a jó adatminőség biztosítása (nagyobb felbontású adatok, megbízható real-time adatok, távoli adatkeletkeztetés, soha nem volt (mért) adatok létrehozása,

stb.) erősen befolyásolják az MI célú adatfelhasználást, a gépi tanulás alapú MI módszertanok esetében.

Az adatbizonytalanságok mérése, becslése, meghatározása már pl. a szenzoriális adatkeletkezés pillanatában, valamint a kialakult adatbizonytalansági paraméterek propagálása az adatfeldolgozási lánc teljes hosszában teszi majd lehetővé azt, hogy az MI megoldások és rendszerek megbízhatóságát mértékkel ellátva számszerűsíthetővé tegyük azokat, ami az MI felhasználás és az MI társadalmi elfogadásának egyik megalapozója, fő komponense lehet.

Gépi tanulás

Az adatfeldolgozás folyamatában már felismerhetők bizonyos feldolgozási mintázatok, melyek az MI gépi tanulás (machine learning) típusú megközelítésére vonatkozó adatfeldolgozásra jellemzőek. Ezek közül elsődleges az adatállományból kiemelt, a különféle tanuló algoritmusok betanítását célzó betanító adatállományok létrehozásának igénye.

A **betanító adatállomány**ok létrehozása az állományok teljes anyagának bizonyos, meghatározott szempontok szerinti leválogatását, szűrését, előfeldolgozását igényli, mely válogatási folyamatokat, különösen a nagy állományok esetében nem lehetséges kézi módszerekkel biztosítani. E feldolgozási lépésekben jelentős szerepet kapnak az elemi adatokhoz társított különféle metaadatok megléte, elérhetősége, a metaadatok sémájának, szemantikájának egyértelműsége, egyértelműsítése és az adattartalmak, valamint a metaadatok alapján történő adatleválogatási eljárások.

Ugyancsak jelentős szerepe van azoknak az adat előfeldolgozási eljárásoknak, melyek segítségével a teljes adatállományra valamilyen szempontból jellemző adatelemek illetve metaadatok kiválogatása lehetségessé válik. Ehhez legtöbbször az kell, hogy valamilyen áttekintésünk legyen a teljes feldolgozandó adatállományról, az állomány globális, átlagos, jellemző stb. paramétereiről, az adatok megbízhatóságáról. Az adatminőségi garanciák megléte illetve a megbízható adatminőséget biztosító feldolgozási eljárások használata az MI témakör egyik központi (kutatási) kérdése.

Mivel az MI algoritmusok működése, működésének hatékonysága jelentősen függ a betanító adatállományokba válogatott adatelemektől, ezért az MI fejlesztési gyakorlatban számos, egymással versengő betanító adathalmaz leválogatása tapasztalható. **A betanító halmazok kezelése** (leválogatás, tárolás, felhasználás, majd alkalmazhatóságuk? (amennyiben alkalmasnak bizonyulnak)/ esetében a betanító adatállományok hosszan tartó tárolásának és újrafelhasználásának igénye) **az MI célú adatfelhasználás specifikus jellemzője.**

A betanító halmazok kiválasztásánál azok méretének meghatározása, az adatok zömének jellemzése avagy a ritka adatok (long-tail) ábrázolása jelentős intellektuális

teljesítményt és néha intuíciót igényel, akár algoritmussal akár manuálisan történik a folyamat. **Terjed az MI alkalmazása is erre az adatelőfeldolgozási célra**.

A betanító adathalmazokon jellemző az esetleg jelentős humán kézimunkát igénylő adatfeldolgozási lépések sora (a humán intelligencia társítása céljából, címkézés, annotáció, stb.), mely sok esetben specializált human-in-the-loop interaktív gépi adatfeldolgozó infrastruktúra igénybevételét követeli meg. A humán interakcióban a résztvevők száma, szakmai háttere, stb. szempontjából jelentősen eltérő támogató környezetek használata jellemző, a szimpla adattársítást (pl. domain szakértő igénybevételével) lehetővé tévő környezettől egészen a (managed) crowdsourcing jellegű megközelítésekig.

A multimodális gépi tanulás, a tudásfeltárás a különféle eltérő modalitású adatállományokból (lásd. még szenzorfúzió) (diszkrét, folytonos, szöveges, képi, térben-időben rendezett, gráf stb. állományokból) a fenti adatelőfeldolgozási lépéseket jelentősen bonyolítja és jelenleg fontos terepe az új megközelítések kutatásának.

Az MI algoritmusok illetve az azokat implementáló szoftverrendszerek önmagukban is adatként viselkednek és olyan metaadatok társításával írhatók le, melyek egyrészről az algoritmusok (illetve az azokat implementáló szoftverrendszerek) belső működését magyarázzák el ha az lehetséges, illetve lehetővé teszik az algoritmusok klasszifikációját, valamint az algoritmusok és rendszerek működésére, hatékonyságára jellemző indikátorokat közlik többek között. A hatékonyság-indikátor betanító halmaz függő, ezért az indikátorok mellé a betanító halmaz megléte, mentése, a jellemző paramétereinek meghatározása, illetve a halmaz újrafelhasználhatósága fontos kritérium.

Az MI algoritmusok és a betanító halmazok hosszú távú megőrzése, valamint az újrafelhasználhatóságot lehetővé tévő illetve az azt elősegítő feldolgozási lépések tekintetében ma még nincsenek kialakult, megszilárdult, kvázi standardizálható módozatok, a témakör ma még kutatás-igényes és jelentős fejlődés előtt áll.

Nem gépi tanulás alapú MI megközelítések

Az MI nem gépi tanulás jellegű megközelítései esetében, melyek a gépi tanulással kombinálva azt kiegészítve avagy teljesen önállóan kerülnek felhasználásra (példák: problem solving, planning and scheduling, probabilistic, commonsense reasoning, combinatorial optimization, knowledge representation, natural language processing, (group) decision making, human-machine interaction) a szükséges adatfeldolgozási lépésekben eltérő, a módszertanokra egyedileg jellemző mintázatokat mutatnak. Jellemzővé válik a szemantikus, logikai, statisztikai, multimédia stb. **strukturált adatok** használata.

Ilyen esetekben kevéssé lesz hangsúlyos a nagytömegű adatfeldolgozás, sokkal inkább az erősen strukturált adatállomány kezelés válik jellemzővé és kívánatossá. A

bottom-up (gépi tanulás) helyett a top-down adatfeldolgozási, adatanalízis módszertanok és rendszerépítmények, és az ehhez elengedhetetlen know-how-k kerülnek használatba.

Gyorsan fejlődő MI kategóriát képviselnek az adatok nélküli illetve minimális adatfelhasználást megkívánó MI módszerek/módszertanok, melyek szándékoltan nem az emberi intelligencia által már eleddig létrehozott tudás reprodukálására, annak (újra)felhasználása alapján céloznak, hanem az adott peremfeltételek mellett globálisan avagy lokálisan optimalizált megoldásokat keresnek (lásd. generatív tervezés) a lehetséges megoldások sokaságának gyors digitális előállítása és a megcélzott paraméterek különféle optimalizálása révén. Ide tartoznak az evolúciós algoritmusok témakörei is.

Humán interfészek szerepe

Mint ahogy azt már a gépi tanulással foglalkozó előző részben is említettük, az emberi intelligencia és az emberi aktivitás fontos szerepet játszik a gépi tanulást alkalmazó MI módszerek K+F-e, az ehhez szükséges adatfeldolgozások esetében. Ennek egy lényegi folyománya az, hogy az adat-vizualizációknak valamint a felhasználói interfészeknek központi szerepe van az MI fejlesztési témakörben. Jobb vizualizációk, jobb felhasználó interfészek a nagytömegű adatok jobb áttekintését, használatát alapozzák meg, lehetővé teszik a komplex adatállományok mélyebb megértését, mely különösen a gyorsan változó környezetben elengedhetetlen.

Itt most nem részletezzük az ember - MI kollaborációban majdan szükséges vizualizációk és felhasználói interfészek problémakörét, mert ez a dokumentum az adatkezelés kérdéseire koncentrál csupán, ugyanakkor fontos észben tartani, hogy az MI felhasználás esetében a humán jelenlét illetve a humán MI használat ugyancsak felveti a humán - gép/MI interfészek súlyos kérdéseit.

Az MI K+F támogató szuprastruktúra főbb funkciói

Az elkülönülő, eltérő MI megközelítések gépi, infrastrukturális, szoftver stb. támogatása egymástól eltérő felépítésű, más-más hardver és szoftver komponensekből álló K+F infrastruktúrát igényelnek. Egy generikus, mindenfajta MI K+F-et támogató egyetlen rendszert ma még nem lehetséges összeállítani. Emiatt sokkal inkább célszerű meghatározni azokat a hardver-szoftver-rendszer-működés stb. komponenseket, melyek szerepet kaphatnak egy-egy adott MI kutatás-fejlesztés során, ugyanakkor az ezeket a kompenenseket egységes rendszerbe fűző elrendezések alapkoncepcióit szintén érdemes felvázolni.

Az adattárolás illetve adatmegosztás szükségessége

Az MI célú adatkezelés és különösen a nagy tömegű adatfeldolgozási igény esetében az adatkeletkezés és az adatfelhasználás legtöbbször térben és időben elkülönül, az adatok forrásai és az adatok felhasználói legtöbbször elkülönülő intézmények, cégek. Ekkor elengedhetetlen az (esetleg) átmeneti, a közép és/vagy hosszú távú adattárolásról gondoskodni. Az **adattárolási igények** ma már inkább terrabájtban mérhetők (1-50-500 TB) mintsem gigabájtban. A szenzoriális adatkeletkeztetés esetében a szenzorokból eredő legtöbbször real-time adatáramok (streams) feldolgozása olyan, ma még nem ismert eljárásokat igényel, melyek révén ésszerű határra csökkenthetők az adattárolás (illetve az adatkommunikációs, lásd később) igények.

A **nagy tömegű adatok** tér-időben történő **mozgatása nagy kommunikációs kapacitásokat** is felemészthet. Mindez jelentős igényeket támaszt a hálózati kommunikációs rendszerek és azok adott időben rendelkezésre álló átviteli sávszélességét illetően. Különösen a szenzoriális adatkeletkeztetés esetén tapasztalható nagy kommunikációs sávszélesség igény, ezért fontos olyan adatfolyam előfeldolgozási módszerek alkalmazása (illetve azok kutatása, fejlesztése) melyek segítségével csökkenthető, illetve optimalizálható az adatfolyamok mérete. Mindemellett a területen nem ritka a 10-100-1000 Gigabit adatátviteli igény felmerülése.

Az MI K+F célokra olyan adatállományokra van szükség, melyek egyrészről elegendő méretben, elegendő mennyiségben állnak rendelkezésre, másrészről az adatállományok a **valóság teljes komplexitását** tartalmazzák, vagyis nem egy egyszerűsített, valamilyen szempontból szimplifikált valóságot tükröznek. Ezek az adatállományok (függetlenül attól, hogy az állományokat a kormányzat avagy valamilyen más pl. ipari szereplő hozta létre) önmagukban számos, egymással is ellentétes érdekviszonylatokba ágyazva léteznek és az adatállományokban az egyéni, csoport, állami, kommerciális stb. érdekek párhuzamos teljességével kell számolnunk, illetve ezeket az érdekeket pl. jogok formájában, kezelni szükséges.

Az MI esetében (legalábbis amennyiben valóságosan is jól működő MI megoldásokra törekszünk a papírtudomány művelése helyett) csak tényleges, a valóságból származó adatok, adathalmazok használhatók fel és mint ilyen, az MI K+F területet ez mindenképpen megkülönbözteti a korábbi mérnöki tervezési-fejlesztési területeken működő módszertanoktól. Különösen fontos tehát a magas adatminőség és adat-megbízhatóság biztosítása.

Ebből az következik, hogy az MI célokra a jó minőségű (és ezért nagy társadalmi, kommerciális stb. értékű) adatállományok **többszörös felhasználását**, az adatállományok **megosztását** kell eleve tételezni illetve tervezni és mint ilyen, ezen adatokhoz való széleskörű hozzáférést kell biztosítani. E hozzáférést akár mint az osztott, nyilvánosan hozzáférhető adatállomány (public goods), akár mint a kommerciálisan kezelt adatállomány (adatpiac) biztosítja.

Az adatmegosztás esetében (még akkor is, ha ez csak kutatás céljából történik) az adatforrásoknak, adattulajdonosoknak számolniuk kell bizonyos rizikófaktorokkal, mely az adathozzáférés hiányos, pontatlan szabályozottságából, esetleg az adatok hibás (pl. az elégtelen szemantikai feltárás avagy az elégtelen származási információkból fakadó) felhasználásából ered.

Ugyancsak jelentős feladat a **személyes adatokat (is) tartalmazó adatállományok** kezelése és az azokhoz való széleskörű hozzáférés biztosítása. Ilyen esetekben a személytelenítés illetve az újraszemélyesítés (de-identification) adatelőfeldolgozási lépések létfontosságúak az MI felhasználás tekintetében. Ezen adatfeldogozási lépések megtételéhez az adatkezelési szuprastruktúránkban (lásd. később) e speciális feldolgozásokat lehetővé tévő komponenseket kell biztosítanunk.

A magas minőségű adatállományok egyik fontos jellemzője a pontos társított **származási (provenance) metaadatok** megléte. A származási adatok létrehozása (már az adatkeletkeztetés legelső fázisától lásd. pl. szenzoriális adatkeletkeztetés) az MI adatfelhasználás specifikusan kritikus igénye, mert csak így válik lehetővé az MI területen az algoritmusok és alkalmazások eredményességének mérése, az adattisztítás, adatszaporítás stb. elvégzése.

A jelenleg rendelkezésre álló **adatállományok minősége** csak ritkán üti meg az új MI terület igényeit, ezért nem csupán a majdan létrehozandó adatállományok minőségével kell foglalkozni, hanem a már meglévő adatok megfelelő feldolgozásával (az inkonzisztenciák és hiányok kiküszöbölésével, szűréssel, válogatással, megfelelő metaadatolással, stb.) szükséges fokozni az adatminőséget. Ez jelentős befektetést igényel, hiszen újra hozzá kell nyúlni a már meglévő adatállományokhoz és ez a komplett adatkezelési szuprastruktúra újraalkalmazását igényli.

Az MI K+F célokra alkalmazott szuprastruktúrában, az egyre nagyobb számú és méretében is növekvő tendenciákat mutató adatállományok kezelését, begyűjtését, tárolását, (specifikus) előfeldolgozását lehetővé tévő (preferáltan elosztott) repozitórium rendszerek számára csak a legújabb digitális könyvtári kutatások eredményezte legfejlettebb, az ún. **szemantikus repozitóriumok** felelnek meg. A klasszikus adattárolók, pontosan a szegényes metaadatkezelési funkcióik miatt ténylegesen alkalmatlanok ezekre a célokra. A szemantikus tárolókban gazdag, szemantikus szintű metaadatkezelés lehetséges, mely nélkül a jól célzott keresés, válogatás, előfeldolgozás, stb. MI által megkövetelt színvonala nem biztosítható.

Az MI kutatás-fejlesztés lényegében kikényszerít egy generációváltást az alkalmazott repozitórium technológiákban, továbbá a szemantikusan definiált metaadat-kezelés mellett a nagyon nagy adatállományok (big data) illetve az adatfolyamok (streams) kezelése együttesen válik szükségessé.

MI célú adatfeldolgozási algoritmusok és azok megosztása

Az MI célú adatfeldolgozás általában bonyolult adatfeldolgozási környezetbe integrálva működik (gépi tanulásnál bizonyosan, egyéb esetekben pedig jellemzően), melyet számos, egymáshoz adatkapcsolatokkal illesztett feldolgozási workflow együttműködése jellemez legjobban. Ezek a **feldolgozási workflow-k** végzik az adatok előfeldolgozását, az adatelőkészítéstől a skálázáson, az adattisztításon, az adatellenőrzésen, az adat kivonatoláson, tömörítésen, szűrésen stb. keresztül egészen a betanító halmazok, tesztelő halmazok, stb. előállításáig. A workflow-k működése során figyelemmel kell lenni a teljes integrált workflow-rendszer adatfolyam vezérlésére, a beiktatott adattárolók (mint adatpufferek) méretezésére, és a rendszervezérlés mikéntjére. A teljes adatfeldolgozási rendszer globális állapotának folyamatos monitorozása szintén követelmény, mely nélkül a teljes adatfeldolgozási rendszer helyes vezérlése nem lehetséges. A rendszer összeépítésénél nem csupán a feldolgozó szoftverkomponensek saját tulajdonságait kell figyelembe kell venni, hanem a rendszereken keresztül áramló adatelemek egymáshoz való illesztése szemantikus szintű függéseinek (dependenciák) rendszerét is.

Az adatfeldolgozási lépések egy hosszú sora határozza meg tehát a teljes feldolgozást. Az adatfeldolgozási sor bármely lépésénél keletkező bármely hiba erősen befolyásolhatja a feldolgozási lánc működését és annak áteresztő képességét, teljesítményét. Mivel a feldolgozási sorban működő szoftverkódok legtöbbször különböző helyről származnak, különböző nyelven, különböző szoftverkörnyezetben készültek más-más fejlesztők által, más-más minőségben, ezért a teljes integrált rendszer minősége, megbízhatósága könnyen kérdőjelessé válhat. A jelenség sokban hasonlít a **“spagetti kód”** néven a szoftveriparban ismert jelenséghez, mely valóságos **káoszkezelést** tesz szükségessé. Egy-egy feldolgozási elem cseréje (pl. verzióváltás során avagy alternatív megoldás beiktatásánál) a teljes feldolgozási hálózat és rendszer adat- és vezérlés-összefüggéseinek felülvizsgálatát foglalja magában.

Az adatfeldolgozási lépéseket megvalósító szoftverkódok és szoftver könyvtárak minőségét jó irányba befolyásolhatja, ha ezeket a kódokat és könyvtárakat a **nyílt forráskódú szoftver** mozgalom keretében hozzák létre, kezelik és a közösségi fejlesztés hosszú távon biztosít egyfajta minőségkontrollt.

Ez azt jelenti, hogy az adatfeldolgozási lépéseket végző szoftverek illetve szoftver-könyvtárak önmagukban is egy adatkezelési szuprastruktúrában kezelendő objektumként viselkednek és mint ilyenek, közösségi megosztást, közösségi fejlesztést, közösségi karbantartást igényelnek.

MI algoritmusok kollaboratív kutatása és fejlesztése

Az MI jelenlegi fejlettségi állapotában, mikor még az MI algoritmusok, megoldások elmélete nagyrészt kidolgozatlan, a gyakorlati MI kutatás-fejlesztés sokban hasonlít a

kipróbálás, mérnöki kísérletezés, tesztelés alapú, nagy számban történő fejlesztési ciklusok szinte végtelenbe nyúló végrehajtásához.

Ilyen esetekben, különösen az erőforrás ínséggel jellemezhető ökoszisztémákban (ilyen a jelenlegi magyar MI szféra is) a relatív kevés és alulfinanszírozott kutató-fejlesztői tevékenység szisztematikus összehangolása nem csupán hatékonysági kérdés de elemi **létfeltétel** is. Ilyen környezetben ugyanis nem lehetséges az erőforrások (még elfogadható szintű) pazarlása sem olyan módon, hogy az egyes kutatók illetve kutatócsoportok, fejlesztők újból és újból elvégezzék a mások által már elvégzett és negatív eredményekkel járó korábbi kísérleteket.

Mivel a műszaki tudomány jelenlegi (amúgy világszerte egyöntetűen tapasztalható) működése nem nyújt elegendő információt a sikertelen kísérletekről, azok kezdeti avagy peremfeltételeiről, beállítási paraméterekről, stb. ezért az országos kutató-fejlesztői tevékenységek szoros csatolására (hosszabb távú vízióként **országos szinten kvázi egyesített, egyintézményi működésre**) kell felkészülnünk, és mihamarabb megteremteni e működési mód teljes feltételrendszerét.

A szoros csatolás a különféle kutatók és fejlesztő csoportok adott elvégzett kísérleteinek pontos leírása, a felhasznált algoritmusok, adathalmazok, betanító halmazok, eredmények, és eredmények minőségi indikátorainak stb. országosan nyílt **közösségi megosztására** épülhet.

Ez a szituáció lényegében kikényszerítheti a nyílt kutatás (**open science**) mozgalom valamiféle (korai) megvalósítását az MI területen. Mivel ez egy jelenleg minden szempontból eléggé ingoványos terület (lásd. pl. az eltérő kutatói, intézményi, céges stb. érdekek összehangolásának várható nehézségeit), ezért elsődlegesen az országosan központilag kezelt, közösített és megosztott K+F támogató szuprastruktúra(ák) létesítésében és működtetésében látunk egy lehetséges kimenetet.

A nagyértékű, szolgáltatásgazdag (pl. adatfeldolgozási) szuprastruktúrák központi létesítése feleslegessé teszi az egyedi adatfeldolgozók saját, költséges feldolgozó kapacitásának kiépítését, egyben olyan szolgáltatások igénybevételét teszi lehetővé számukra, melyek központosítva lehetségések avagy célszerűek (pl. standardizált adat formátumok és azokra történő adattranszformációk mint központi szolgáltatások), megteremtve az egyéni érdekeket a csatlakozáshoz. (Jó jelenlegi példa a SZTAKI és Wigner kutatóintézetek által működtetett MTA Cloud, amely a kutatók számára hozott létre központi felhő infrastruktúrát és amelynek népszerűségét bizonyítja az eddig rajta indított 100+ különböző projekt.)

Az MI K+F szorosan összehangolt országos működtetése újfajta működési és/vagy üzleti modellek kidolgozását igényli, melyek egyike lehet pl. az adatfeldolgozó és adatforrás partnerek egymásra találását lehetővé tévő matchmaking működést támogató funkció (lásd pl. [2], [3]) megvalósítási javaslat.

MI K+F célhardverei és HPC alkalmazása

Az MI, és azon belül főleg a tanuló rendszerek robbanásszerű elterjedésével megnőtt az igény az MI jellegű számítási műveletek hatékonyabb és gyorsabb végrehajtására. Az általános célú processzorok (CPU - Central Processing Unit) nem alkalmasak az MI nagy párhuzamosságot, de viszonylag kevés fajta számítást igénylő feladatára. Ezért számos más létező és feltörekvőben lévő számítási architektúrát alkalmaznak, és próbálnak alkalmazni az MI területén, mind tanulási, mind végrehajtási oldalon.

Jelenleg hozzáférhető technológiák

- **GPU** (Graphical Processing Unit - Grafikus feldolgozóegység)
  A ma minden számítógépben és a legtöbb mobil eszközben is megtalálható grafikus feldolgozóegységek belső felépítése (sok kisméretű, számítási egység hosszú pipeline-okkal) alkalmasabb mind modellek betanítására, mind modellek végrehajtására, mint a hagyományos CPU-k, kb. egy nagyságrend gyorsulást lehetővé téve.

- **TPU** (Tensor PU (Google)), **IPU** (Intelligence PU), **Inference Chips** (Amazon Inferentia, Intel Nervana, Huawei Ascend, stb.)
  Ezek tipikusan GPU-khoz hasonló architektúrájú, de azok grafikus és egyéb szükségtelen részeit nem tartalmazó általános MI célú chip-ek. Általában a GPU-khoz képest még egy nagyságrenddel gyorsabbak mind betanításra, mind végrehajtásra.

- **ASIC** (Application-Specific Integrated Circuit - Felhasználás-specifikus integrált áramkör)
  Minden feladathoz készíthető egyedi tervezésű IC, ami ezáltal a specifikus feladatot sokkal gyorsabban és hatékonyabban látja el, mint egy általános számítóegység. Általánosságban véve elmondható, hogy minél specifikusabb egy ASIC, annál hatékonyabb, de ugyanakkor annál szűkebb körben használható fel. MI területén léteznek csak egyes tipikus MI-műveleteket gyorsító (itt tulajdonképpen nem egyértelmű a határ az ASIC-ok és a TPU/IPU-k között), specifikus modelleket megvalósító, valamint konkrét betanult modellt megvalósító ASIC-ok is.

- **FPGA** (Field-Programmable Gate Array - felhasználás helyén programozható logikai kapumátrix)
  Olyan chip-ek tartoznak ide, amelyek gyakorlatilag teljes felépítése újraprogramozható. Az FPGA-k rendkívül kisméretű (tipikusan egybites operációkat lehetővé tevő, és azt tárolni tudó) logikai blokkokból állnak, amik programozhatóan kapcsolhatók a szomszédos logikai blokkokhoz. A programozás percek alatt elvégezhető, és ennek eredményeképpen gyakorlatilag egyedi ASIC használható minden számítási feladat felgyorsításához. Természetesen a programozhatóságnak ára van, így egy azonos felépítésű ASIC-hoz képest egy

FPGA némileg kisebb teljesítményű és magasabb fogyasztású, valamint jelentősen drágább lesz. Az FPGA-kat gyakran használják ezért ASIC-ok fejlesztésekor prototipizálásra.

- **Neuromorphic Chips** (Neuromorf Chip-ek - IBM TrueNorth, Intel Loihi):
  Ezek az IC-k az emberi agy működését imitáló számítási architektúrát tartalmaznak, tulajdonképpen egy-egy neurális háló modell ASIC megvalósításának tekinthetőek, amikben az egyes neuronok viselkedése kötött, míg az átmeneti súlyok és a neuronok összekapcsolása programozható.

- **Hagyományos HPC technológiák** (nagysebességű hálózatok, tárolóegységek, processzorok)
  A gépi tanulás minőségének legfőbb feltétele, hogy minél több tanuló adat álljon rendelkezésre. Ezeket az adatokat kell tudni tárolni, valamint betanításkor megfelelő sebességgel rendelkezésre kell tudni bocsátani, a hagyományos HPC igényeknek megfelelő, vagy akár azokat meghaladó léptékben.

Most fejlesztett technológiák

- **Graphcore IPU**
  Rugalmasabb, menet közben tanulásra képes architektúrát ígér kisebb feldolgozóegységekkel, integrált memóriával, és nagysebességű összekötő buszokkal.

- **Dataflow Processors** (Adatfolyam-processzorok, pl. Wave computing)
  Az adatfolyam-processzorokat régóta használják jelfeldolgozás terén, és úgy tűnik, hogy az MI jellegű feladatok, azon belül is a végrehajtási oldal támogatására is kiválóan alkalmasak. Több cég is fejleszt ilyen jellegű IC-ket.

- **Analóg chip-ek** (IBM, Syntiant, Mythic)
  Az MI mélytanuló rendszerek általában digitális rendszerben végrehajtva is közelítő eredményt adnak, rengeteg tömbművelet egymás utáni alkalmazásával. Ezek a tömbműveletek azonban analóg módon is végrehajthatóak, ha nem pontos, hanem megközelítő eredményt várunk el. Az analóg végrehajtás elsődleges előnye a több nagyságrenddel kisebb energiafelhasználás mellett is egy nagyságrenddel gyorsabb műveletvégrehajtás. Több cég is rendelkezik már ilyen IC prototípusával.

- **Photonic Processors** (Fotonikus/Optikai Processzorok)
  A fotonikus megközelítés MI alkalmazások esetén is használható. Nagyobb sebességet, és jelentősen nagyobb hatékonyságot ígér, mint az elektronikus megoldások. Egyelőre nagyon kis méretű rendszerek léteznek csak, de több cég és egyetem is fejleszt ilyen rendszereket, pl. Intel, Lightmatter és az MIT.

Összefoglalásul, a fent említett technológiák mindegyikének helye van egy K+F szuprastruktúrában: a betanítást segítő technológiák az új modellek, algoritmusok és adathalmazok alapján gyorsabb betanítást engednek meg, míg a végrehajtást segítő

technológiák a betanított modellek tesztelését és teljesítményanalízisét teszik lehetővé. K+F esetében a különböző technológiákhoz gyakran korai (tehát piaci értékesítést megelőző) hozzáférés is lehetséges, így érdemes lehet a még csak fejlesztési fázisban lévő prototípusokat is beszerezni, és kutatók/fejlesztők számára hozzáférhetővé tenni. Mindezek mellett figyelni kell a hagyományos jellegű számítóközponti elemekből álló támogató infrastruktúra (megfelelően gyors és nagy adattároló egységek, előfeldolgozásra processzorok, valamint megfelelően gyors hálózati kapcsolatok) méretezésére is, mert enélkül nem lehet kihasználni a célhardverekben rejlő számítási kapacitást.

MI felhő mint rugalmas erőforráskezelő és osztott tároló

A jelenleg mainstream technológiának tekinthető felhő (cloud) használata MI K+F célokra nem csupán a felhő illetve (federált) felhők megfelelő és rugalmasan skálázható nagy számítási és/vagy adattárolási kapacitása miatt érdekes, hanem egyben azért, mert a felhő univerzálisan, mindenki által hálózaton hozzáférhető osztott erőforrásnak tekinthető.

Mint ilyen, a felhő technológia jelenleg a legalkalmasabb technológia az országot átfogó kutatás-fejlesztési támogató szuprastruktúra, így jelen javaslat, az OMIKI implementációjára is. A felhő szolgáltatások jelenlegi individuális sokszínűsége azonban önmagában nem elegendő céljainkra, hanem szükség van átfogó, egymással harmonikus működési rendszerbe szervezni e szolgáltatásokat és az MI specifikus új funkciókat integálni a meglévő funkciókhoz.

A felhő technológiák és a korábban bemutatott célhardverek és/vagy szuperszámítástechnika integrálása gyorsan fejlődő terület. Az OMIKI nagy hangsúlyt szándékozik fektetni arra, hogy az ország lehetőleg minden egyes MI célú célhardvere elérhetővé váljon felhőn keresztül bármely OMIKI felhasználó számára. E cél elérése jelentős anyagi ráfordítást igényel.

Szabványosítás és benchmarking

Az MI területen történő szabványosítás, mely esetünkben, itt és most persze elsődlegesen az adatok, metaadatok szabványosítását, az adatfeldolgozási lépések, adattranszformációk szabványosítását jelentik számunkra, nem korlátozódhat csupán az adatkezelés szférájára.

Az MI területen a szabványok tágabb értelemben egyben követelményeket, specifikációkat rögzítenek, segítenek abban, hogy az MI technológiák felhasználás tekintetében megfeleljenek az **össztársadalmi elvárásoknak** az MI funkciók, az együttműködési képességek (interoperability) (lásd pl. MI-humán együttműködési kontextus), a biztonságos és megbízható felhasználás stb. tekintetében.

Az MI algoritmusok valamint adatállományok egyértelmű jellemzésére, a megtalálhatóság és újrafelhasználhatóság, az összehasonlíthatóság stb.

elősegítésére legelőször is ki kell dolgozni azokat a **fogalmi rendszereket** (adatsémákat, taxonómiákat, ontológiákat), melyeket az ilyen adat és algoritmus objektumokhoz társítva a szemantikus kereső, tároló, tudáskezelő rendszerek működését teszik lehetővé. Egyben fogalmi készletet nyújtanak az MI tudás és alkalmazás-területről szóló szélesebb értelemben vett társadalmi kommunikáció számára is.

A szabványosítási erőfeszítések szükségesek a szoftverfejlesztési területeken a szoftver architektúrák, a bonyolult de biztonságosan fenntartható és üzemeltethető rendszerek megvalósítása céljából. Szabványok szükségesek a rendszerek megbízhatóságának, pontosságának, robosztusságának, skálázhatóságának stb. jellemzésére.

Az MI felhasználás még kritikus biztonsági megfontolások és használati szabályozások mellett is folyamatosan **rizikófaktorokat** tartalmaz, ezek ismerete, a rizikófaktorok kvalitatív és számszerűsítésükhöz szükséges **metrikák** révén kvantitatív jellemzése és elemzése nélkül nem lehetséges gyakorlati MI felhasználás.

A társadalmi méretű MI felhasználást a **használhatósági elemzés**, az MI rendszerek felhasználó interfészeinek pontos meghatározása és a lehetőség szerinti szabványosítása elősegíti.

Az adatfeldolgozási területen az adatformátumok, adatsémák szabványos volta nem csupán a potenciálisan egymással felcserélhető adatfeldolgozási entitások miatt érdekes, hanem az adatösszefüggések (mint dependenciák) számosságának mérséklése, egyben az adatfeldolgozási workflow-k komplexitásának csökkentése, az integrált adatrendszer megbízhatóságának növelése miatt is létfontosságú, nem beszélve az adatfeldolgozás erőforrásigényesség csökkentéséről.

Az MI K+F tevékenységek (pl. adatfeldolgozási lépések és eljárások) szabványosítása egyben támogatja a korábban már megindokolt tevékenység transzparencia megvalósulását, az egyes gyakorlatban működő MI megoldások más-más területeken történő **(újra)hasznosításának** lehetőségeit.

A szabványoknak való megfelelés vizsgálatához (compliance) illetve (pl. biztonsági stb.) kategóriákba való soroláshoz új műszaki vizsgálati módszerek és metrikák kidolgozása, szabványosítása és azok bevezetése válik szükségessé. Standardizált tesztelési módszerek alapján végzett standard tesztek (benchmarks) segítségével az MI K+F technológia fejlődése egyértelműen nyomonkövethető, mely nem csupán a résztvevőknek, hanem a szféra (állami) irányítóinak is megalapozott muníciót ad e tudományos és műszaki terület (pro)aktív fejlesztéséhez, irányításához illetve vezérléséhez. A szabványosítás az MI területen létkérdéssé válik és mint ilyen, újból állami, össztársadalmi cselekvést és korrekt **intézményi** (!) kezelést igényel.

A szabványos tesztek elvégzéséhez MI tesztelő központok, tesztkörnyezetek (**testbeds**) szükségesek, melyek kontrollált környezeteket nyújtanak standardizált

adatállományokon elvégzendő teszt vizsgálatok számára. Az MI teszkörnyezetek erősen technológia és adatállomány függőek.

A későbbiekben javasolt OMIKI adatkezelési szuprastruktúra biztosítja a zárt és kontrollált környezetben végzendő standard tesztek elvégzését is, tehát nem csupán **innovációs teret** alkot MI K+F célokra, hanem egyben benchmarking (standard teszt) funkció szolgáltatásokat is nyújt. Mint ilyen, nem elhagyható és nem helyettesíthető mással.

A szabványosított és zártan kezelt (lásd OMIKI sandbox modell később) tesztkörnyezetben azokat a valós működésből származó, tényleges nyers adatokat tartalmazó, eredeti állapotban rendelkezésre álló (pl. nem anonimizált) adatállományokat is használhatjuk benchmarking célokra, melyek széles és nyílt terjesztése az adott formátumban és állományban nem lehetséges. Ahogy ezt már korábban jeleztük, az MI K+F minőség a felhasznált tényleges és valós adatminőség függvénye. Csak ilyen zárt tesztkörnyezetben tudjuk mindezt biztosítani a kutatók és fejlesztők számára.

Hosszú távú digitális megőrzés

Az adat illetve adatvagyon szerepére, értékére az MI K+F tekintetében már korábban felhívták a figyelmet [1], itt és most csak arra szeretnénk emlékeztetni, hogy az adatok hosszú távra (akár 50-100 évre) történő megőrzése, mely az MI területen könnyedén válik kiemelt követelménnyé (pl. az MI széles, tömeges, nagy időállandójú társadalmi alkalmazásakor) speciális repozitórium és megőrzési technikákat, eljárásokat és intézményeket igényel. Ilyen időtávlatban nem csupán az adat fizikai megőrzéséről kell gondoskodni (pl. megfelelő adathordozók alkalmazásával illetve adathordozók között tervezett és ütemezett migrációs eljárásokkal), hanem az adatok hordozta szemantikai tartalom logikájának megőrzését is biztosítani kell a szoftver és adatformátum avulás miatt. Ehhez sokkal nagyobb informatikai felkészültségre van szükség.

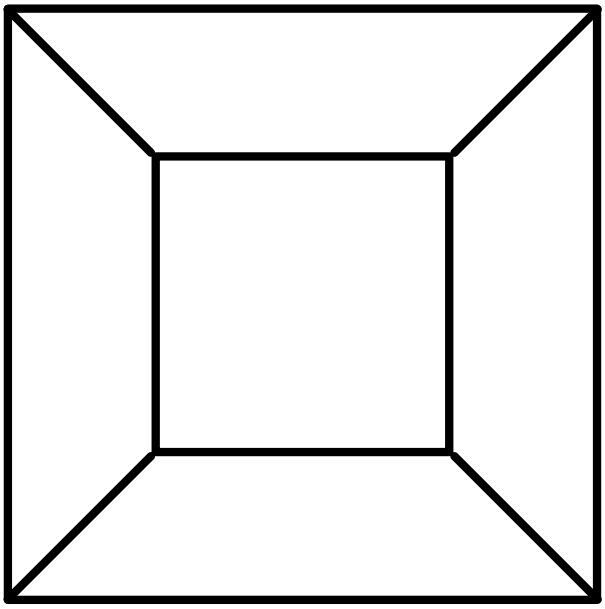
Az OAIS nemzetközi szabvány, a hosszú távú digitális tárolás jelenlegi alapkoncepcióit ugyan lefekteti, ugyanakkor az OAIS inkább a klasszikus digitális dokumentumok tárolását körvonalazza, mintsem azon entitásokét, melyek az MI területen leginkább szerepet kapnak. Ilyenek pl. a szoftver algoritmusok, hardver specifikációk, adatfeldolgozási workflow-k és az ezekhez társított pl. (betanító) adatállományok, minőség hatékonyság indikátorok stb. Ezek a digitális entitások nagyságrendekkel bonyolultabb megőrzési problémákat vetnek fel és mint ilyenek, jelenleg még **megoldatlan** feladatok elé állítják az ezen entitások hosszú távú megőrzését célzó projekteket, munkatársakat. E terület nem csupán fejlesztés, hanem (alap)kutatás igényes is.

Nemzetközi jó példák

Nemzetközi szinten találunk jó példákat, működő európai rendszereket az infrastrukturális megközelítésre, melyek a felépítés, a szolgáltatásrendszer tekintetében ötleteket adhatnak, kiinduló alapot képezhetnek számunkra. Erre a célra e dokumentumhoz csatoljuk az EGI és EOSC európai kutatástámogató infrastruktúrák rövid bemutatását tartalmazó kiváló papírt (Sipos Gergely, Kacsuk Péter: "*Releváns EGI és EOSC szolgáltatások, aktivitások*").

Az EGI és az EOSC nem specifikus MI kutatást támogató infrastruktúrák hanem általános, generikus infrastruktúráknak tekinthetők. Az MI indukálta újfajta infrastrukturális szolgáltatások kialakítása túlmutat az EGI, EOSC bemutatott meglévő szolgáltatásain és mind infrastruktúra **kutatásokat, fejlesztéseket, mind pedig (szolgáltatás) innovációkat igényel**. Ezt próbáljuk megtenni e dokumentum utolsó fejezetében a konkrét OMIKI javaslatunk kidolgozásakor.

Az OMIKI Javaslat



ORSZÁGOS MESTERSÉGES
INTELLIGENCIA KUTATÁS
F E J L E S Z T É S I
S Z U P R A S T R U K T Ú R A

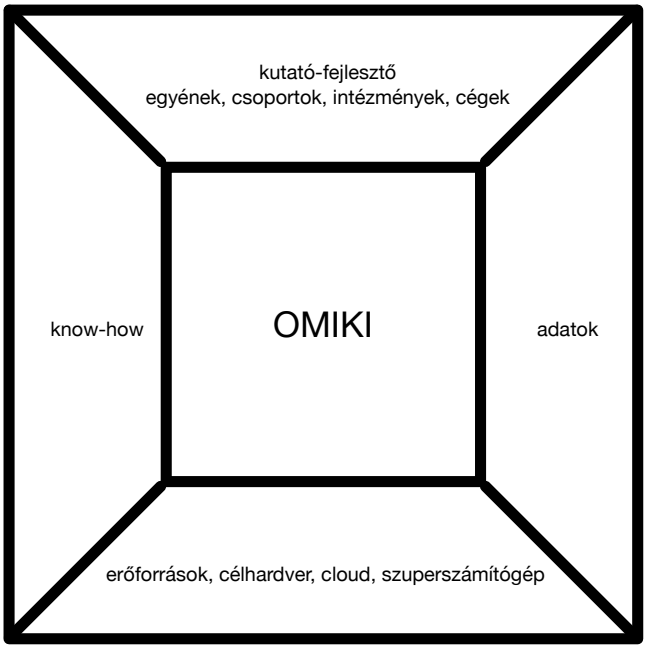
E fejezetben kidolgozzuk a javaslatunkat, mely az OMIKI, Országos Mesterséges Intelligencia Kutatás-fejlesztési Szuprastruktúra nevet viseli. A digitális **szuprastruktúra** létrehozását, folyamatos üzemeltetését, továbbfejlesztését, az szuprastruktúra hosszú távú (15-20 év) stabil fennmaradását reálisan csak intézményi keretekben lehet elképzelni, ezért ez a javaslat egyben egy OMIKI-t fenntartó szervezet létrehozását is tartalmazza.

Jelen fejezetben nem foglalkozunk részletesen azzal, hogy az OMIKI-t fenntartó szervezet és az MI Koalícióban megjelenő intézményi elképzelések (pl. Adatvagyon Ügynökség, MI Kiválósági Központ) között milyen reláció áll fenn valamint azzal, hogy milyen intézményi működési forma (pl. nonprofit vállalkozás) lenne a legmegfelelőbb erre a célra, hanem csak amellett érvelünk, hogy szükség van egy ilyen funkciókat lefedő, önálló szervezetre, mint a tervezett országos hatáskörű, digitális alapon működő funkcionalitásokat nyújtó kutatás-fejlesztési szuprastruktúra létrehozó, fenntartó és fejlesztő entitásra.

OMIKI, Országos Mesterséges Intelligencia Kutatás-fejlesztési Szuprastruktúra felépítése és főbb funkciói

Ebben a fejezetben bemutatjuk a tervezett kutatás-fejlesztés támogató szuprastruktúrát, a rendszer komponenseit, a tervezett funkciókat, kezdve az általános funkciókkal, majd bemutatva az adatkezelés specifikus funkciókat.

OMIKI entitások



ORSZÁGOS MESTERSÉGES INTELLIGENCIA KUTATÁS FEJLESZTÉSI SZUPRASTRUKTÚRA

Az **OMIKI, Országos Mesterséges Intelligencia Kutatás-fejlesztési Szuprastruktúra** egységes működéssel, központilag vezérelt módon kapcsolja egybe az országban az MI témakörben dolgozó, kutató, fejlesztő, oktató, stb. egyéni, csoport, intézményi, céges stb. **humán résztvevőket,** nem egyéni hanem intézményi granulációs szinten értelmezve működésüket.

Biztosítja számukra az MI K+F-hez szükséges **adatállományokhoz** (legyen az állomány bárhol és bármilyen formátumban elérhető) való direkt hálózaton keresztüli (a hozzáférés jogosultságokat, az esetleges használati díjazást, az esetleges védett módozatú elérést, stb. is figyelembe vevő) hozzáférést.

Biztosítja számukra az MI K+F-hez szükséges **know-how-hoz, szoftver rendszerek**hez, módszertanokhoz (pl. adatfeldolgozói vagy tudományos workflow, szabadalommal védett eljárás, stb.) való közvetlen (de a know-how konkrét és egyedi használati feltételeit is figyelembe vevő) hozzáférést, azok szuprastruktúrán belüli könnyed alkalmazását.

Az OMIKI egyben lehetővé teszi humán entitások számára a kutatás-fejlesztési tevékenység **gépi erőforrásai**hoz való direkt és közvetlen hozzáférést, legyen az a gépi erőforrás bármilyen, pl. egyedi célhardver, szuperszámítógép, felhő alapú szoftver- és/vagy szolgáltatásrendszer.

Virtualizált kutatás-fejlesztési tér

Az OMIKI szuprastruktúra teljesen digitalizált funkcionális keretet biztosít az MI területen a kollaboratív kutatás-fejlesztéseket kanalizáló virtuális kutató és fejlesztő környezetek létrehozására (**Virtual Research/Development Environment (VR/DE)**), az azokban létrejövő (kutatási) adatok, dokumentumok, módszertanok (tudományos workflow), technológiák, eredmények, KPI-k (teljesítménymérő indikátorok) stb. hosszú távú digitális (OAIS konform) megőrzésére. A VR/DE egyben a mesterséges intelligencia informatikai interoperábilis eszközrendszerének virtualizálását és minőségbiztosítását is jelenti (lényegében egy teljesen elektronikus működésű Workspace funkcionalitást nyújt, mely munkatér a benne létrejött entitások hosszú távú megmaradását (és a későbbi, esetleges, azonos avagy más területen történő újrahasznosítását) is biztosítja).

Virtuálisan integrált, elosztott adattárolás

Az szuprastruktúra vállalja a mesterséges intelligencia kutatás-fejlesztési projektek (kutatási) adatainak **közös, felhő federáció alapú adattárolásához**, a meglévő offline és online **adatbázisok, repozitóriumok szemantikus szintű virtuális integrálásához** és hosszú távú fenntartásához szükséges módszertanok és know-how-k biztosítását (*DataCloud*), egyben ezen szoftver entitások (preferáltan) közösségi karbantartását és minőségbiztosítását is irányítja/támogatja a releváns digitális támogató szolgáltatásrendszerén keresztül. Az szuprastruktúra tehát az MI célú adatmegosztás központi vezérlő és szükség esetén tároló szereplőjévé avanzsál.

Ez a gyakorlatban azt jelenti, hogy nem csupán a hálózaton teszi elérhetővé, kereshetővé, integrálhatóvá (adatfúzió) a tevékenységekhez szükséges **adatállományokat és adatbázisokat,** mintegy virtuálisan egyesítve azokat, de az adatbázisok hosszú távú fenntartását is elősegíti a megfelelő technológiák felajánlásával, vagy adatbázis megszűnés veszély esetén gondoskodik az szuprastruktúrán belüli **adatmigráció** segítségével történő folyamatos adat fenntartásról.

Adatállományok és adatgyűjtemények nyilvántartása

Az OMIKI szuprastruktúra működési tevékenységének részeként elvégzi a **nemzeti digitális adatállományok** (adatgyűjtemények, adat repozitóriumok, adat aggregátorok stb.) felmérését, központi regisztrálását, az MI kutatás-fejlesztés számára releváns **adatok** kutathatóvá-felhasználhatóvá tételét; a releváns (**meta)adatszabványok** alkalmazását, azok magyar adaptálását, a tudomány-fejlesztési területek tárgyszójegyzékeinek, ontológiáinak, névtereinek **szemantikus szintű összekapcsolását** (linked data), ezáltal azok kereshetővé, kutathatóvá, annotálhatóvá és megoszthatóvá tételét; megteremti, fejleszti, megosztja az adatállományokban az MI tekintetében létfontosságú **szemantikus feltárás** (entitás identifikálás) és **az adatok közötti logikai-szemantikai kapcsolati hálózat** felismerés módszertanait.

Entitás profilok

Az OMIKI-ban digitális létet élvező bármely entitás (legyen az humán, adat, know-how, avagy gépi erőforrás, stb.) saját, kiterjedt, részletes **szemantikus szintű profillal** rendelkezik, mely nem csupán az entitás képességeit leíró statikus (pl. név, elérhetőségi adatok adatállományok esetében) avagy lassan változó (pl. szakmai profil humán entitás esetében) metaadatokat tartalmazza, de tartalmaz az entitás konkrét és történeti működése során keletkező, az entitás (pl. adatállomány) felhasználásáról, a működés milyenségéről stb. árulkodó historikus (log) adatokat is. E profilok alapján lehet pl. kiválasztani egy adott MI témakörhöz értő humán entitásokat, avagy egy feladat elvégzésére alkalmas gépi erőforrásokat, avagy a konkrét adatfeldolgozáshoz szükséges workflow-kat, stb.

Általánosított partnerkeresés és kapcsolat létrehozás

A gazdag adattartalmú entitás profilok alapján **szemantikus szintű partnerkeresés** és kapcsolat létrehozás is lehetővé válik, entitás típusoktól függetlenül, pontosabban azokra teljesen transzparens módon (bármely entitás típus bármely más entitás típussal kapcsolatba hozható).

Ez a **matchmaking szolgáltatás** tehát általánosabb és generikusabb mint a [2], [3] dokumentumokban leírt adatforrás és adatfeldolgozó egymásra találását végző B2B és B2C brókerezési modell, ugyanis lehetővé teszi bármely kívánt profillal bíró entitás megtalálását és közöttük a kapcsolatépítést, legyen az a kapcsolat bármilyen (pl. közös munka, projekt, hasonló szakmai érdeklődés, stb.) Ez az általánosított modell lehetővé tesz, többek között, humán-humán, workflow-gépi erőforrás, know-how – humán, stb. párosításokat. Erre az általános matchmaking szolgáltatásra épül például a következő pontban leírt virtuális projektszervezési funkció is.

Virtuális projektek és virtuális téma központok

Az OMIKI digitális szolgáltatásrendszerének legfontosabb eleme a teljesen digitális, virtualizált térben történő **projekt** és/vagy **téma központ** szervezés támogatása. Az szuprastruktúrán belül globális matchmaking szolgáltatás segítségével partnereket találhatunk egy adott MI projekt végrehajtásához, összeválogatva a projekthez leginkább értő, szükséges kompetenciákkal bíró kutatókat, fejlesztőket, a feladathoz elengedhetetlen tesztadatokat, adatállományokat, a tervezett adatfeldolgozási work-flow-kat, azok végrehajtását végző szoftver stack-et és a projekt megtalálhatja a számára legmegfelelőbb (adott időben pl. rendelkezésre álló) végrehajtási (pl. szuperszámítógép avagy felhő) környezetet.

Míg a virtuális *projekt* adott célra, véges időtartamban, véges erőforrásokat mozgósító aktív folyamat, addig a *virtuális téma központ* időkorlát nélküli, egy-egy adott szakmai témakör vagy akár gyakorlati probléma köré csoportosítható humán entitások, know-how-k, publikációk, eredmények, stb. közösen megosztott munkaterületét hozza létre. Mindkét esetben az OMIKI szuprastruktúra teljes szolgáltatásrendszere a résztvevők rendelkezésére áll.

Eredmény újrahasznosítás

Befejezett projekt esetében egy projekt nyoma nem mosódik el teljesen, hanem az OMIKI szuprastruktúra által kikényszerítve illetve az szuprastruktúra kodifikált működési szabályai miatt legalább a projekt eredmények (pl. egy projektben kidolgozott új algoritmus avagy létrehozott tisztított adatállomány, stb.) permanensen megmaradnak, megteremtve az **újrahasznosítás, újrafelhasználás** reményét és lehetőségét. Természetesen az újrahasznosítás feltételrendszerét az eredménytermékekhez mint entitásokhoz társított profil kötelezően kell, hogy tartalmazza, mely feltételek betartását majd az szuprastruktúra, annak működése kényszeríti ki később.

Ez a mechanizmus neutrális a projekt eredményesség szempontjából, vagyis itt az OMIKI szuprastruktúra biztosítja a negatív eredmények “újrahasznosítását”, hogy a későbbi résztvevőknek, a későbbi projektekben ne kelljen erőforrás pazarlóan újból és újból megismételniük a korábban már elkövetett kudarcokat.

Lényegében ezzel az OMIKI egy intézményszintű, esetünkben tehát országosan összehangolt MI kutatás-fejlesztési **tevékenység optimalizálást** biztosít, az MI területre szánt (pl. pénzügyi) erőforrások hatékony felhasználását elősegítendő.

Sandbox szolgáltatás

Az OMIKI szuprastruktúra létrehozza a **generikus sandbox szolgáltatást** mint (erősen) védett környezetben végrehajtható virtuális projekt környezetet.

Amennyiben egy-egy adatállomány tartalma, eredete stb. nem teszi lehetővé még kutatási célokra sem az adatállomány direkt MI felhasználását, akkor az OMIKI sandbox szolgáltatását lehetséges igénybe venni erre a célra. A sandbox mintegy teljesen átlátszatlan fekete doboz, egy védett projekt végrehajtási környezetet biztosít, melybe helyezett entitások (bármely típusú entitás amit behelyeznek a sandboxba) a sandboxon kívülről nem láthatók és nem elérhetők. A sandboxok léte, száma, kiterjedése sem nyilvános információ.

Példák:

- Speciális szakértelmű kutatók-fejlesztők szeretnének valamilyen projektet együtt végrehajtani úgy, hogy sem a személyük, sem pedig a projekt léte, célja, tartalma, állapota nem nyilvános.

- Nem anonimizált, kényes személyes adatokat is tartalmazó nyers adatállományokhoz akarunk egy adott fejlesztési projekt keretében közvetlenül hozzáférni pl. valós tesztelés céljából.

- Speciális, pl. jelentős kommerciális értékekkel bíró új hardver eszközünket szeretnénk benchmark vizsgálatoknak kitenni még a szabadalmaztatás előtt, stb.

Ezekben az esetekben a védett sandbox szolgáltatás gondoskodik arról, hogy még az OMIKI, elosztott szoftver szuprastruktúra karbantartói (az OMIKI rendszergazdák) se tudjanak arról, hogy egy sandboxban kik, mivel, milyen tevékenységeket végeznek.

Az ilyen sandbox funkcionalitás megvalósítása egyrészről az adatállományoknál illetve más entitásoknál a titkosítás, homomorf titkosítás, de-anonimizálás, a szigorúan védett és szabályozott hozzáférés, stb. eljárások alkalmazását igényli és a végrehajtás fizikai környezetét tekintve akár megkövetelheti a konkrét intézményi keretekben értelmezett adatszobák avagy Trusted Execution Environment technológia kötelező igénybevételét is. Mindezt úgy, hogy a felhasználónak eközben nem kell kilépni az OMIKI szuprastruktúrájából.

Erőforrás használat és foglalás

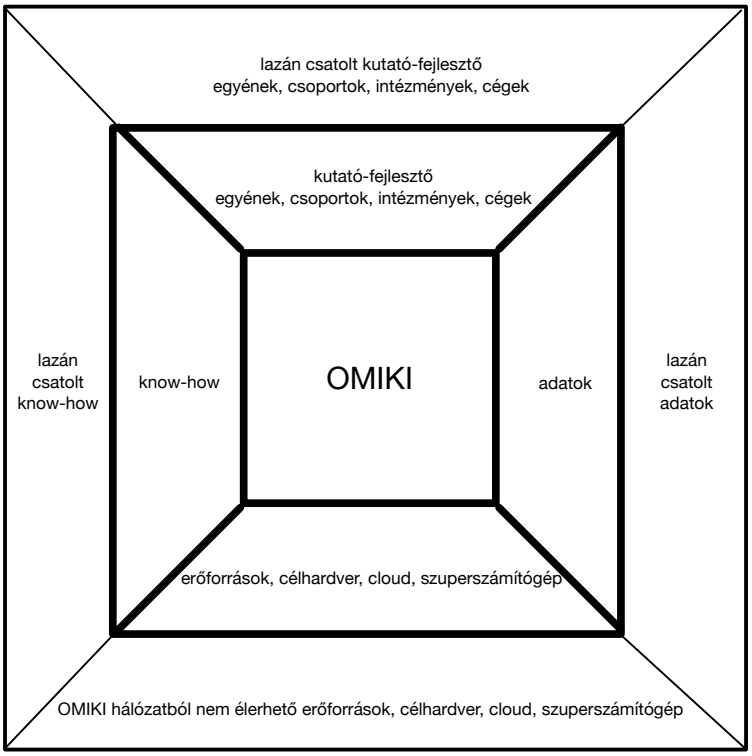
Az OMIKI rendszerbe kapcsolt erőforrások (távoli) igénybevétele preferáltan hálózaton keresztül történik. Az erőforrások foglalását és az erőforrás felhasználás ütemezését az OMIKI integrált, pontosabban **federált erőforrásfoglaló rendszer** segítségével az szuprastruktúra felhasználói számára transzparens módon kezeli. Ez azt jelenti, hogy az erőforrások (legyen az pl. egy szuperszámítógép avagy egyedi adatállomány) hozzáférését kezelő lokális rendszerek összekapcsoltak az OMIKI közös erőforráskezelő és -foglaló rendszerével így a felhasználóknak csak egy közös rendszerben kell a foglalást regisztrálni, kezelni.

Az OMIKI erőforrásnak tekint minden az szuprastruktúrában aktív/passzív entitást, tehát a humán entitásokat ugyanúgy, mint ahogy a (cél)hardver avagy véges számú

licensz lehetőséggel bíró szoftver entitások igénybevételét. Így lehetséges egy adott szakértő avagy akár szakmai csoport távoli (pl. hálózati konzultáció) avagy helyhez kötött (pl. f2f megbeszélés) igénybevételét kezdeményezni és azonos módon kezelni mint egy számítógépre történő gépidő-foglalást.

Az erőforrások felhasználása esetleg fizetéshez kötött (gépidő díj, szakértői óradíj, fejlesztői csoport munkadíj, stb.), melyre az szuprastruktúra az entitásokhoz kötött profilokban hívja fel a figyelmet és (esetleg a jövőben) akár automatikus fizetési szolgáltatásokat is nyújthat teljesen integrálva.

Szorosan és lazán csatolt entitások



ORSZÁGOS MESTERSÉGES INTELLIGENCIA KUTATÁS-FEJLESZTÉSI SZUPRASTRUKTÚRA

Az OMIKI-hez csatlakozó entitások a csatolás szempontjából lehetnek szorosan avagy lazán csatoltak. A **szorosan csatolt entitások** létét és működését a rendszer direkt szolgáltatásain keresztül támogatja.

Az entitások igénybe veszik a digitális szolgáltatásokat és a **virtualizált kutatás-fejlesztési térben** végzik mindennapos MI célzatú tevékenységeiket. A rendszer nyomon követi az entitások e tevékenységeit és a tevékenység fonalakat

(tevékenység log) bizonyos jól szabályozott hozzáférési rendszeren keresztül a OMIKI-n belül láthatóvá teszi.

Természetesen van/lehetséges K+F lét és tevékenység az OMIKI-n kívül is. Itt olyan erőforrásokra kell gondolni amelyek (még) nem integrálódtak be az OMIKI működésébe, avagy ismerten MI kutatást végző kutatók akik nem használják az OMIKI rendszereit, a rendszert nem használó de az MI kutatás-fejlesztést végző intézmények, cégek, olyan létező adatállományok, melyek nem érhetők el szabadon avagy jól szabályozott módokon az OMIKI szuprastruktúrájából közvetlenül hálózaton keresztül, de független létezésükről tudomásunk van. Általában a **lazán csatolt entitásokról** csak korlátozott ismeretek állnak rendelkezésre, részleges profil birtokában vagyunk. Ilyen esetekben az OMIKI minimálisan az entitások létét és a részleges profil információk regisztrálását, nyilvántartását végzi.

Az MI adatkezelés és adatfeldolgozáshoz szükséges OMIKI szuprastruktúra elemei és rendszerei

Itt és most, terjedelmi és egyéb okokból nincs lehetőségünk részleteiben kidolgozni az MI K+F komplex adatfeldolgozási szolgáltatás rendszerét, ezért csak azt vázoljuk, hogy milyen kategóriákban szükséges adatfeldolgozási és egyéb szolgáltatásokat fejleszteni és beintegálni az OMIKI szuprastruktúrába.

Az OMIKI saját integrált szolgáltatások mellett közvetíti (lásd pl. adatpiac szolgáltatás [3]) a rendszerhez kapcsolódó entitások (legyen az kommerciális cég, akadémiai kutatóintézet, egyetem, vagy akár egyén, kutatócsoport stb.) által nyújtott és a rendszerben **szemantikus profillal** ellátott szolgáltatásokat. E szolgáltatások lehetnek digitális, hálózaton keresztül elérhető számítógépes szolgáltatások vagy akár valós (fizikai) szolgáltatások is (pl. egy szakértő személyes munkavégzése avagy egy cég valamilyen szállítási szolgáltatása stb.). A szolgáltatások közül persze számunkra itt a digitális hálózati szolgáltatások és azokon belül is az adatfeldolgozás célú szolgáltatások érdekesek.

Entitás azonosítás

Az OMIKI infrastuktúrában minden kapcsolt entitás a rendszerben **egyedi azonosítóval** rendelkezik, mely azonosító birtokában az entitás minden pillanatban elérhető, mind statikus mind pedig dinamikus (pl. historikus) profilja megkereshető és használható. Az entitások aktuális állapota (ha az éppen nem védett) szintén felhasználható akár algoritmikusan is.

Osztott munkaterületek

Az entitások egymásrahatását az OMIKI szuprastruktúrában értelmezett **osztott munkaterületek** (projekt, sandbox, témaközpont, megbeszélés, kommunikációs tér,

workflow stb.) határolják és kanalizálják. Az osztott munkaterületek objektumai hosszú távon megőrződ(het)nek, a hozzáférés szabályozott és (esetleg) erősen védett.

Kommunikációs és csoportmunka szolgáltatások

Az OMIKI integrált **csoportmunka szolgáltatásrendszert** nyújt (osztott dokumentumkezelés és -szerkesztés, csoportos task és workflow kezelés, többkritériumú csoportos döntéstámogatás, elektronikus audio-videó konferencia szolgáltatás, szinkron és aszinkron üzenetközvetítés stb.) melyet a humán felhasználókon kívül szoftver applikációk is igénybe tudnak venni. Ez egy olyan új vonás, mely segítségével az MI adatfeldolgozói workflow-k humán interakciós igényei kezelhetővé de legfőképpen algoritmikusan programozható válnak.

Az emberi és gépi aktivitások mind homogénebb, teljesen integrált kezelése egy speciális MI specifikus cél, melynek megoldása új innovációkat és újfajta OMIKI hálózati szolgáltatásokat, újfajta, interaktívitást tartalmazó szoftvereket igényel.

Repozitórium (tárolási) és regisztrálási szolgáltatások

Az OMIKI hálózaton elérhetővé teszi az adatok tárolását végző digitális tárolókat, egyben olyan regiszter szolgáltatásokat is nyújt, melyek az OMIKI entitások globális nyilvántartását végzik. Nyilvántartják és elérhetővé teszik a humán, gépi, adat stb. erőforrások mellett pl. szoftvereket és know-how-kat, adatkezelési workflow leírásokat és vezérlőket, integrált szoftver stackeket stb. melyeket már valakik kipróbáltak és a hozzájuk társított használati tapasztalatok, vélemények, publikációk, megjegyzések, beszélgetések stb. alapján az OMIKI **globális K+F "géppé"**, működési mechanizmussá válik az MI területen.

Adatsémák, adatformátumok, szabványok, ajánlások kezelése

Az adatséma, adatformátum, ontológiák, (adat)szótárak szabványosítását, közös interoperábilis használatát elősegítendő, mind a séma stb. **megosztást**, mind pedig az ezekről az objektumokról való **csoportos tudásfelhalmozást** és tudáskezelést támogatja szolgáltatásaival az OMIKI.

Adatfeldolgozás, entitás kivonatolás

Az OMIKI nem csupán az egyszerűbb adatfelhasználást segíti elő a támogatott szabványosítás, egységesítés révén, de a szabványos és/vagy elterjedt (multimedia) adatok, adattípusok esetén a konkrét adatfeldolgozó eljárások, adatfeldolgozási workflow-k terjesztésével, megosztásával, központi üzemeltetésével, azok központi karbantartásával veszi le a terhet a rendszerbe kapcsolt adatfeldolgozók válláról. Az entitás kivonatolásra (entity extraction), az adatformátum transzformációra, az adattömörítésre, automatikus annotációra, a személytelenítésre és más típikus MI

adatfeldolgozási feladatokra az OMIKI standard, azonnal, akár nagy tömegű adatfeldolgozásra képes, **működő megoldásokat** kínál közvetlenül.

Publikációs szolgáltatások

Az OMIKI támogatja a szféra szakmai és-vagy laikus (ismeretterjesztés célú) publikációs tevékenységeit, a publikáció készítés különféle fázisaiban, a publikációk létrehozása, szerkesztése, nyomdakésszé tétele, terjesztése stb. területeken. Az OMIKI nyílt portált üzemeltet az eredmények, közlések, híradások stb. **nyílt disszeminálása** érdekében.

Felügyeleti és monitoring szolgáltatások

Az OMIKI létrehozása és működtetése esetében egy országos hatáskörű és kihatású, **"mission critical" rendszerré** válik. Mint ilyen, a rendszer folyamatos rendelkezésre állása, rendszerszerű működése a teljes MI K+F szféra létfeltétele lesz. Ezért a rendszert úgy, olyan szolgáltatás minőségben kell üzemeltetni, mint bármely más országos (pl. energetikai) ellátó rendszert.

Nemzetközi kapcsolatok

Az OMIKI szuprastruktúra különféle (fizikai, logikai, szemantikus, szervezeti stb.) szinteken kapcsolható össze (digitálisan és/vagy szervezetileg is) más nemzetközi kutatás-fejlesztést támogató infrastruktúrákkal, tehát nem csupán adatátadás (adat import-export) lehetőségeket nyújt meg, de megteremti a potenciális **nemzetközi szervezeti kapcsolódásokat**, kapcsolatokat a magyar MI K+F szféra teljes egésze számára.

Az OMIKI szuprastruktúra megvalósítása, működése és finanszírozása

Ebben a koncepcionális dokumentumban itt és most nem lehet célunk annak kidolgozása, hogy milyen szervezeti és pénzügyi, stb. feltételek mentén lehet megvalósítani, majd pedig üzemeltetni ezt az országos MI K+F támogató szuprastruktúrát, ezért csak néhány alapkövet szeretnénk letenni ezen a téren.

- Az OMIKI digitális hálózati rendszer és mint ilyen, manapság a **felhő technológia** kínálkozik a megvalósítás alapjául.

- Az OMIKI egyben egy országos szervezet ezért a klasszikus szervezetépítés, szervezetfejlesztés bevált módszertanai szükségesek a megvalósításhoz, de nem elégségesek, mivel a klasszikus szervezetfejlesztési módszertanok nem készültek fel az elsődlegesen hálózati működésmódra. Új **szervezetépítési innovációkra** és megoldásokra, szabályozásokra, esetleg törvényekre van szükség.

- Az OMIKI nem egy egyszer létrehozandó hálózati rendszer, hanem egy **folyamatosan fejlődő, fejlesztendő rendszer**, mely folyamatos beruházást is igényel, a terület technológiai stb. fejlődésével lépést tartandó.

- Mint az MI K+F központi támogató szuprastruktúrája, a magyar MI-vel foglalkozó **munkatársak 10-15%-ra** bizton szükség lesz az szuprastruktúra fenntartására és folyamatos fejlesztésére. Kezdetben ez egy **kb. 60-70 főt számláló, informatikában magasan képzett szakember** garnitúrával rendelkező szervezet lehet.

- Beruházási összeg tekintetében, alsó becslésként **az OMIKI szuprastruktúrára az évi 8-10 milliárd forint** befektetés szükségessége habkönnyen alátámasztható, ugyanakkor a járulékos (MI-n kívüli), de az MI feltételrendszerét szignifikánsan befolyásoló rendszerek megteremtése, fejlesztésigénye, a leginkább az adattárolás, adatkezelésben elmaradott magyar technikai, technológiai színvonal fejlesztés, adattartalom minőség javítás extra igénye itt és most nem becsülhető meg.

Célzott kutatási és oktatási tevékenységek

Néhány olyan kutatási, oktatási tevékenységet sorolunk fel példaképpen, a teljesség igénye nélkül, mely elengedhetetlen az OMIKI megvalósításához. A számos, már korábban regisztrált MI alapkutatási cél mellett az alábbiak a szűkebb értelmű OMIKI megvalósítását teszik elérhetővé.

- Az MI K+F szférájában, de leginkább az MI megoldások használatának elterjedésével **vadonatúj munkakörök** és szakmák jelennek meg, melyek közép- és felsőfokú képzése jelenleg még megoldatlan. Álljon itt egyetlen példa az adatkurátor szakmájára, mely se nem informatikus, se nem könyvtáros, de mindkét szakma eljárásait/tudását/praktikus gyakorlatát használja egy vagy több konkrét tématerületi (domain) szakértelem mellett.

- Az OMIKI szuprastruktúra létesítése szervezeti innovációkat, támogató szuprastruktúrák és szolgáltatások kutatását, fejlesztését követeli meg. Ezeket a kutatásokat, fejlesztéseket, melyek leginkább az **IT különféle más, klasszikus területein** jelentkeznek, szintén támogatni szükséges annak ellenére, hogy nem célzottan, vagy nem feltétlenül MI tartalmúak, hanem sokkal inkább pl. a csoportmunka, a csoportos döntéstámogatás, a humán kommunikáció stb. területeken jelennek meg. Ezek nélkül azonban nem lehetséges korszerű OMIKI.

- A tervezett és lényegi OMIKI sandbox szolgáltatás megvalósításához a **titkosítási** és biztonsági területeken új (alap)kutatások szükségesek.

- Az MI társadalmi befogadásának egyik alapfeltétele az **interpretálható** (megmagyarázható, megérthető működésű) **MI algoritmusok** és alkalmazások kifejlesztése. Ez azonban nem csupán MI kutatási terület, de **új fogalmi**

rendszerek, megnevezések, nevek stb. széles társadalmi bevezetését is igényli. Az OMIKI szuprastruktúrának ez is egyik fontos missziója kell, hogy legyen.

Utószó

Az **OMIKI, Országos Mesterséges Intelligencia Kutatás-fejlesztési Szuprastruktúra** létrehozásának **jelentősége** túlmutat egy digitális kutatás-fejlesztést támogató mechanizmus létrehozásán. Nem csupán egy adott szakmai terület (esetünkben az MI) kutatás-fejlesztésének digitális platformra ültetése válik lehetségessé (annak minden, elvben jótékony következményével), hanem abból a szándékból, hogy a kutatás-fejlesztéshez szükséges minden alapvető tényezőt (emberi, gépi, adat, alkalmas szoftver) egyszerrre kínálja fel e digitális hálózati szuprastruktúra, a **kutatás-fejlesztési tevékenységek hatékonyságát** ma még nem belátható módon növelheti meg. Egy azonban bizonyos, csak egy ilyen országosan egységes szuprastruktúra teszi lehetővé az erőforrás felhasználások globális optimumának potenciális elérését. Minden más megközelítés, jó esetben csak lokálisan optimális erőforrás felhasználást eredményez.

Egyetlen egy alapvető tényezőt azonban e fenti tervezetben nem kezeltünk, nem vettük számításba, sem entitás sem pedig működési szinteken. Ez pedig a kutatás-fejlesztési tevékenységekhez szükséges **pénzügyi források kezelése**.

Ennek az az oka, hogy e dokumentum a műszaki, technológiai, és azon belül is, az adatkezeléssel kapcsolatos kérdésekre és tervezett mechanizmusokra koncentrált. A kutatás-fejlesztés jelen javaslatból hiányzó lényegi tényezője a tevékenységek finaszírozásának kérdése.

Nem célunk a finanszírozás új modelljének kidolgozása, de itt a javaslat utószavában vázlatosan megpróbálhatjuk felvetni azokat az elképzeléseket, melyek révén, egy megvalósult OMIKI szuprastruktúra szolgáltatásainak keretében, lehetőséget látunk arra, hogy az MI területen **újfajta K+F finanszírozási modell-kísérletek** legyenek beindíthatók.

Az országos átfogású OMIKI szuprastruktúra pl. lehetővé teszi azt, hogy fokozatosan szakítani lehessen a klasszikus, ma már több évszázad hagyományára épülő általánosított értelmű akadémiai kutatás, tudáskezelés és gyarapítás módozataival és a **poszt-akadémikus tudáslétrehozás** irányába tegyünk határozott lépéseket.

Az akadémiai jellegű tudáslétrehozás, pontosan az egyéni (pl. professzor központú), kiscsoportos (pl. PI (principal investigator) vezette kutató csoport) vagy akár a több tucat kutatót-fejlesztőt tartalmazó intézményi (pl. kutatórészleg, department) tudástermelés kis mérete és szerény, lokálisan szervezett szervezettsége miatt csak rendkívül lassan, többszörös áttéten keresztül tud (siker esetében) társadalmi méretű változásokat indukálni. Az így létrehozott új tudás bármilyen formájú (pl. innovációs) hasznosulása megfelelő mérettel és országos, társadalmi avagy multinacionális

szinteken is észlelhető nyomatékkal rendelkező szervezetektől várható csupán. Ezek leginkább a nagy (multinacionális) cégek, országos avagy nemzetközi intézmények, állam stb. Miközben tehát kvázi “háziipari”, “kisipari” módszerekkel hozzuk létre az új tudást az akadémiai szférában, aközben, az elvárt társadalmi szintű hasznosulás csak ezekkel az akadémiai intézményekkel nem egy súlycsoportban lévő, nagy szervezetektől várható.

Az OMIKI országos szuprastruktúra létesítése azzal a jövőképpel kecsegtet, hogy (legalább az MI területen) megvalósulhat egy jobb, hatékonyabb, országos kontextusban szervezett, és **racionálisan megtervezett “nagyipari” tudástermelés,** a tudástermelés egy poszt-akadémikus formája.

A finanszírozás egy lehetséges folyamatos modellje

A manapság tapasztalható finanszírozási modellek (intézményi finanszírozás, pályázat-projekt alapú finaszírozás) egyik közös jellemzője, hogy nem tudnak rugalmasan alkalmazkodni a K+F szférában tapasztalható gyors változásokhoz, a gazdaság igényeihez és jelentős, a piacon nem tolerált késéseket okoznak.

Az OMIKI létrehozta digitális környezetben (elvileg) egy finanszírozó (szervezet) folyamatosan nyomon követheti a digitális térbe terelt teljes K+F újratermelési ciklus minden lépését, aktuális állapotát, eredményét, eredménytelenségét, sőt akár jó előre prognosztizálhatóvá teheti a futó tevékenységek lehetséges kimenetelét.

Az OMIKI virtuális térben minden szükséges (emberi, intézményi stb.) tényező profilja, a tényezők részletes tevékenységi története rendelkezésre áll. Így pl. egy új projekt kezdeménynek nem kell várni a következő pályázati kiírás (call) megjelenésére, majd a pályázat megírása után az esetleg sok hónapot igénybe vevő kiértékelésre, hanem a projekt ötlet megjelenésekor azonnal fordulhatnak a finanszírozó intézményhez, mely szinte azonnal képes a finanszírozási döntéseket meghozni, kvázi kialakítva egy üresjáratok és késések nélküli **folyamatos finanszírozási modellt** (finanszírozási pipeline) vagy akár proaktív finanszírozást (azonnali finanszírozási modell a finanszírozó kezdeményezésével). A pályázóknak csak az új projekt elképzelés megfogalmazására kell időt és energiát szánni, szakmai előéletük, korábbi projektjeik, eredményeik, értékelésük stb. az OMIKI térben a lehetséges és folyamatos tevékenység evaluáció miatt már azonnal rendelkezésre állhatnak.

A folyamatos finanszírozás e modellje mellett az OMIKI virtuális környezet az akadémiai szférában jelenleg járatos jutalmazási és elismerési rendszert is drasztikusan változtathatja meg, a szokásos egyéni és intézményi performancia indikátorok mellett sokkal részletesebb, **átfogóbb, mélyreható értékelési metrikák** lehetséges bevezetésével és azok teljesen digitális kezelésével.

Hivatkozott MI Koalíciós dokumentumok

[1] “Adatpolitikai stratégiai javaslat az MI-alapú innováció beindítására Magyarországon” v. 2.0 (2019.04.15)
[2] “Mesterséges intelligencia kutatási felhő” v. 1.0 (2019.09.09)
[3] “Adatpiac és felhőközpont” prezentáció (kézirat)

Releváns EGI és EOSC szolgáltatások, aktivitások

Dr. Sipos Gergely (gergely.sipos@egi.eu)
EGI Foundation

Dr. Kacsuk Péter (kacsuk@sztaki.hu) (tanácsadó, konzultáns)
SZTAKI, Számítástechnikai és Automatizálási Kutató Intézet

Bevezetés

Az EGI az egyik páneurópai e-infrastruktúra. Az EGI-t az Amszterdami székhelyű “EGI Foundation” koordinálja. Az intézetnek éves tagdíjat fizető tagjai vannak, az ún. nemzeti e-infrastruktúrák. (Magyarország 2010-2014 között volt tag, azóta finanszírozási problémák miatt már nem veszünk részt.). Az EGI egy Európán túlnyúló, federált infrastruktúrát üzemeltet. Az infrastruktúra illetve annak üzemeltetői számítási, adatkezelési, biztonságtechnikai és oktatási szolgáltatásokat nyújtanak kutatóknak, oktatóknak és innovációban résztvevő személyeknek/szervezeteknek. 2018 januárja óta az EGI Foundation koordinálja a 3 éves, 33 millió euró költségvetésű EOSC-hub H2020 projektet is. Az EOSC-hub fektette/fekteti le az European Open Science Cloud (EOSC) központi elemeit.

Mind az EGI, mind az EOSC és azok szolgáltatásai relevánsak a Magyar MI koalíció számára: segíthetnek az MI koalíciónak a tipikus minták és feladatkörök kidolgozásában, továbbá technológiákat, szolgáltatásokat adhatnak a magyar MI piactér létrehozásához.

EGI szolgáltatások

A következőben áttekintjük azokat az EGI szolgáltatásokat[3], amelyek relevánsak lehetnek a magyar MI adatpiac számára:

- Az **EGI Cloud** szolgáltatás [GS1] [GS2] egy több telephelyet összekötő ún. „multi-cloud” rendszer. Az EGI Cloud-ban részt vevő IaaS szolgáltatók

    - egy közös, OpenID alapú azonosítást használnak (az EGI Check-in Authentikációs-Authorizációs rendszerre, EduGAIN és szociális IdP-kre alapozva);
    - azonos forrásból telepítenek megbízható VM-eket (EGI AppDB Cloud Marketplace-ből)
    - közös információs rendszerben vannak beregisztrálva, amely alapján lehetővé válik a felhők rendelkezésre állásának monitorozása, valamint a felhasználás monitorozása (user accounting)
    - harmonizált user interface-eket nyújtanak (OpenStack Nova API and CMD line; AppDB VMOps Dashboard GUI; különböző orchestration rendszerek)
      Az EGI multi-cloud rendszerében jelenleg 25 IaaS felhő szolgáltató vesz részt és a fent felsorolt funkciók lehetővé teszik a közös alkalmazásfejlesztést, közösség vagy projektek által jóváhagyott VM-ek több felhőre való telepítését, a felhasználói VM-ek migrálását, stb. Az EGI Cloud-ban a kapacitás ún. Virtuális organizációkon keresztül kerül elosztásra. Létezik bárki által 2x6 hónapig ingyenesen használható, általában tesztelésre és alkalmazás fejlesztésére szolgáló virtuális organizáció, továbbá a hosszú távon üzemelő tudományos kísérletek/projektek saját virtuális organizációi.

- Az **EGI DataHub** szolgáltatás a felhőkben tárolt fájlokból egységes, könyvtárstruktúrájú adattároló réteget készít. A felhasználóknak, illetve az általuk a felhőkben indított VM-eknek nem kell a fájlok fizikai helyével foglalkozni, a DataHub szolgáltatás tárolja és ha szükséges, replikálja a fájlokat több felhőbe. A DataHub szolgáltatással nyílt adathalmazokat (Open Dataset) is ki lehet ajánlani, így azokat más felhasználók és az ő VM-jeik el tudják érni. A DataHub a CYFRONET intézet (Lengyelország) által fejlesztett és karbantartott OneData technológiára épül.

- Az **EGI Data Transfer** szolgáltatás képes FTP, GridFTP, WebDAV és más tároló szerverek között adatokat továbbítani. A szolgáltatást elsősorban nagy méretű, tipikusan több ezer fájlból álló adathalmazok másolására használják. A szolgáltatás felelős a fájlok megbízható továbbításáért, esetleges hiba esetén az átvitel megismétléséért. A szolgáltatást storage-storage közötti, és storage-compute szerver közötti másolásra is használható. Utóbbi esetben a felhőben futó VM-ek számára kerülnek az adattároló helyről a felhőbe az adatok. A szolgáltatás jelenleg két példányban fut a CERN-ben és az STFC (UK) intézetben.

---

[3] Az EGI szolgáltatás-portfóliója a https://www.egi.eu/services/ címen érhető el.

- [GS3] Az **EGI Notebooks** szolgáltatás az EGI Cloud-ban hosztolt JupyterHub, amely képes a felhasználók számára on-demand módon notebook-okat létrehozni és menedzselni. A szolgáltatás az EGI Check-in-t használva képes az EduGAIN vagy szociális azonosítóval rendelkező felhasználókat beléptetni. A szolgáltatás egy példánya bárki által elérhető a https://notebooks.egi.eu címen felhasználónként 1 CPU, 1GB RAM és 10 GB tárolóhely kapacitás-megkötéssel. Tudományos projektek és közösségek számára az EGI dedikált szolgáltatást nyújt egyedi kapacitás és egyéb paraméterekkel (melyeket a felek SLA-ban fektetnek le).

A leírt EGI szolgáltatások MI piactérben való használati feltételei további megbeszéléseket igényelnek (elsősorban a magyar EGI tagság hiánya miatt).

Digital Innovation Hub

Az EOSC-hub H2020 projekt keretében az EGI egy “EOSC Digital Innovation Hub”-ot (DIH) üzemeltet. Többek között a DIH támogatja kis és középvállalatok (angolul SME) felhő alkalmazásainak és szolgáltatásainak kidolgozását ún. Business Pilot-ok keretében. A pilotok hozzáférést kapnak az EOSC-hub-ban részt vevő IaaS felhőkhöz, felhasználói technikai támogatásban részesülnek, és utazási pénzt kapnak a létrehozott alkalmazásaik, szolgáltatásaik promotálásához. A felhő használat kuponok segítségével történik, amely behatárolja az egy-egy SME által lehívható erőforrás kapacitást.

Releváns EOSC szolgáltatások

EOSC Portal[4] a European Open Science Cloud (EOSC) interfésze. A portál listázza az EOSC-ban elérhető szolgáltatásokat, egy form-ot biztosít amin keresztül új szolgáltatók csatlakozhatnak, illetve háttérinformációt ad az EOSC-ról. Az EOSC portálhoz csatlakozik egy EOSC Marketplace, amely egy webshop-hoz hasonló felületet nyújt a felhasználóknak hogy a hozzáférés szabályozással működő szolgáltatásokhoz hozzáférjenek[5]. Az EOSC portál „mögött” több kiszolgálócsoport is működik, ezek közül a számunkra legfontosabbak:

1. Új szolgáltatások validálását végző csoport: Előre megadott kritériumok (az ún. „rules of participation”) alapján elvégzik a csatlakozni kívánó szolgáltatók „minőségellenőrzését”, melynek legfontosabb eleme a szolgáltatás, annak dokumentációjának és user support helpdesk-nek az ellenőrzése. Ezen felül begyűjtik a szolgáltatásról mindazt az információt amely az EOSC portálban, illetve

[4] https://eosc-portal.eu

[5] Megjegyzés: Az EOSC-ban vannak teljesen nyílt szolgáltatások amelyek authentikációt sem kívánnak. Vannak authentikációt használó de szabadon hozzáférhető szolgáltatások. És végül vannak athentikációt és hozzáfárás szabályozást (authorizációt) is kívánó szolgáltatások. Ez utóbbiaknak van szüksége az EOSC Marketplace-re, melyen keresztul hozzáférési igény küldhető a szolgáltatónak aki elbírálhatja azt.

a hozzáférés igényléseket kezelő EOSC Marketplace-ben való megjelenítéshez szükséges.

2. Hozzáférési igények elbírálását végző csoport: Az authorizációt is használó szolgáltatások hozzáférési kérései ehhez a csoporthoz futnak be. Ők ellenőrzik, hogy a kérés reális-e (nem spam, nincs benne félreértés vagy félreértelmezés). Ha szükséges a csoport javítást vagy kiegészítést kér az igénylőtől. A helyes kéréseket vagy közvetlenül a szolgáltatóknak, vagy a felhasználói támogató csoportnak továbbítja (ld. Következő pont)

3. Felhasználói támogató csoport: Feladata a komplex, vagy nem eléggé pontosítható felhasználói kérések támogatása. A komplex kérések több szolgáltatást is igényelnek, és az adott szolgáltatásokat ismerő technikai segítség nélkül nem vagy nehezen implementálhatók. A nem eléggé pontosítható kérések pedig olyan felhasználóktól származnak akik nincsennek tisztában (általában nem is lehetnek eléggé tisztában) azzal, hogy az ő use case-e pontosan milyen EOSC szolgáltatásokat igényelne, melyek lennének a legmegfelelőbb szolgáltatók akikhez fordulniuk kell. Ez a csoport segíti az ilyen felhasználókat a use case-ek strukturált analízisével, a legalkalmasabb szolgáltatások és szolgáltatók kiválasztásával.

Diszkusszió

Az EOSC (és egyébként egyéb hasonló csoportok, pl. a European Earth Observation terület vagy az AI4EU kezdeményezés) lényegében a következő résztvevőkből[6] épül fel:

1. Adatszolgáltatók (nálunk adatgazda, angolban data providers): Alapvető feladatuk a nyers, digitális adatok előállítása/begyűjtése, tisztítása és azok „data product”-ként való közzététele. Ide sorolhatók a Kutatási Infrastruktúrák (pl. részecskegyorsítók, teleszkópok, stb.), a hozzájuk hasonló pán-Európai társulások (pl. Sentinel szatelitek), a nemzeti adatszolgáltatók (pl. KSH), illetve az intézeti-közösségi adatszolgáltatók (pl. EMBL-EBI).

2. Tematikus szolgáltatók (nálunk adatfeldolgozók, angol nyelvhasználattal thematic service providers, providers of exploitation platforms): Előre megszabott formátumú vagy területhez kapcsolódó adatok feldolgozására, vizualizálására vagy általában véve kezelésére használható szolgáltatást nyújtanak. Tulajdonképpen SaaS szolgáltatóként működnek kifejezetten valamilyen tudományos célcsoportra specializálva. Általában egy vagy néhány előre megszabott általános szolgáltatást használnak (ld. következő pont) és mint „one-stop-shop” üzemelnek a kutatók számára, azaz a skálázást, felhasználói azonosítást, adatkezelést mind elvégzik

[6] A háttérben vannak további résztvevők is, például a finanszírozók, de ők az EOSC portálon lezajló tranzakciókban közvetlenül nem vesznek részt.

3. Általános szolgáltatók (angolban generic service vagy common services): informatikai szolgáltatásokat nyújtó szolgáltatók, amelyek hasznosak lehetnek elsősorban a tematikus szolgáltatóknak, másodsorban a nagy adatmennyiséggel közvetlenül dolgozni kívánó egyéni felhasználóknak. Tipikusan ide tartoznak a cloud (IaaS-PaaS), az adatkezelés, a security témakör szolgáltatásai (Pl. az EGI szolgáltatásai is).

Egy ilyen rendszerben az általunk elképzelthez (“Adatpiac és felhőközpont”) hasonló B2B és B2C brókerezési modell működik:

- Az EOSC egy B2B piactérként funkcionál a tematikus szolgáltatók számára, akik ott a legmegfelelőbb általános szolgáltatókat tudják megtalálni a szolgáltatásuk létrehozásához és/vagy üzemeltetéséhez. Általában H2020 projektek és kutatási infrastruktúrák keresik az EOSC-ot ilyen célból, tipikusan cloud szolgáltatást, tárolóhely szolgáltatást, adatkezelő szolgáltatót (mozgató/replikáló) és authentikációs-authorizációs szolgáltatókat keresve. Fontos megjegyezni, hogy egy-egy tematikus szolgáltató egyszer (vagy mindenképp kevés alkalommal) fordul így az EOSC-hoz akkor, amikor a saját szolgáltatását létrehozza. Egy-egy tematikus szolgáltató létrehozása, hozzáillesztése cloud, storage és hasonló általános szolgáltatásokhoz, a teljes rendszer tesztelése és validálása nem triviális munka, ezért nem gyakori a tematikus szolgáltatók részéről a „szolgáltatóváltás”.

- Az EOSC egy B2C piactérként funkcionál a tudományos felhasználók számára. Ők az EOSC-ban elsősorban a tematikus szolgáltatásokat keresik, amelyeknél vagy a fixen „bedrótozott” vagy a saját adataikat elemezhetik, analizálhatják, összevethetik, vizualizálhatják. A tudományos felhasználók aránylag kis része képzett és nyitott eléggé ahhoz, hogy az EOSC-on keresztül általános szolgáltatásokat keressen és használjon. Amíg a tematikus szolgáltatások tipikusan magas szintű grafikus interfészekkel rendelkeznek, az általános szolgáltatások programozási vagy parancssori interfészt nyújtanak és legalább közepes szintű informatikai tudást feltételeznek a felhasználótól.