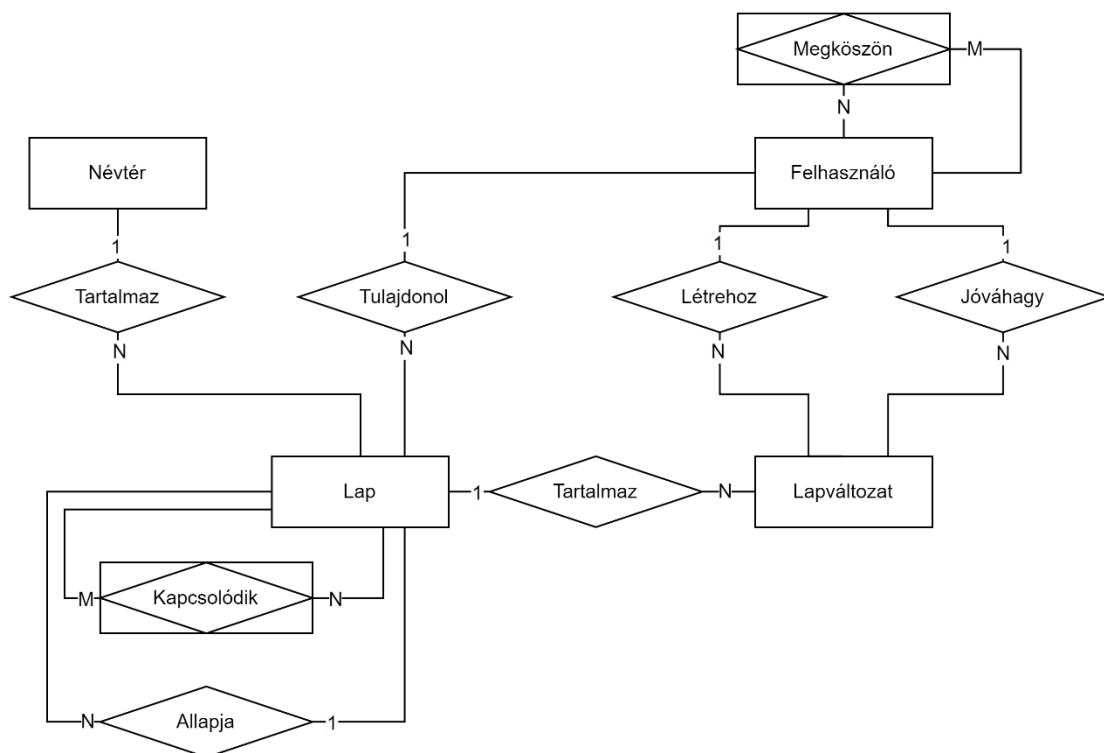




26. ábra: Wikipedia fogalmi adatmodell (Saját szerk.)

A Wikipedia alapvető célja, hogy a **Felhasználók** új **Lapváltozatokat** hozzanak létre, amelyek folyamatosan frissítik és bővítik az enciklopédia tartalmát. A felhasználók egymás között nem nyilvánítják ki a kapcsolataikat, azok tevékenységükből vezethetők le. A direkt felhasználó-felhasználó interakcióra példa, ahogy a felhasználók megköszönhetnek egymásnak szerkesztéseket. Az, hogy pontosan melyik szerkesztés megköszönése történt, nem nyilvános, így ez a kapcsolat felhasználó-felhasználó viszonylatban modellezhető.

A felhasználók más felhasználók lapváltozatait jóváhagyhatják (megfelelő jogosultság esetén), ezt követően a lapváltozat nyilvánossá válik és az olvasók számára is alapértelmezett tartalomává válik (nem jóváhagyott változat is megtekinthető, de nem elsődleges keresési találatokban). E funkció bevezetettsége platformonként eltér, például a magyar Wikipédián minden szerkesztésnek jóváhagyást kell nyernie valamilyen formában (megfelelő jogosultsági szint esetén saját szerkesztés is jóváhagyható), míg az angol nyelvű Wikipédián csak néhány oldal van ilyen értelemben korlátozva.

A lapváltozatokat **Lapok** fogják össze, és amikor egy Wikipedia oldalt megtekintünk, az utolsó vagy az utolsó úgynevezett ellenőrzött változatot látjuk (ez a lapváltozat egy attribútuma). Minden lapváltozat teljes értékű, a teljes szövegtartalmat hordozza. Lapok

önállóan nem hozhatóak létre, az első lapváltozat az, ami a lapot, mint önálló, azonosítóval rendelkező entitást létrehozza.

A lapok **Névterekbe** szerveződnek, amelyek meghatározzák a lapok célját és funkcióját. A névterek lehetnek például szócikkek, kategóriák, felhasználói lapok vagy viták. A névterek segítenek rendszerezni a tartalmat és elkülöníteni a különböző típusú információkat. Például, a „Szócikk” névtérbe tartozó lapok enciklopédikus tartalmakat tartalmaznak, míg a „Kategória” névtér lapjai a szócikkek logikai csoportosítását segítik. Ezen kívül a névterek meghatározzák a lapokon megjelenő szövegtartalom természetét, például, hogy az hozzászólásokból álló párbeszéd vagy folyószöveg.

A lapok egymással két módon állhatnak kapcsolatban: beágyazással vagy alásorolással. A beágyazás annyit tesz, hogy az adott oldalra mutató hivatkozás az oldal valamely lapváltozatában beágyazásra kerül. Egy lap technikailag bármely más lapra hivatkozhat és egy lap is bármely helyen hivatkozható. A rendszer rendkívül sok, az 5.3.1 fejezetben ismertetett rendszerekben önállóan létező funkciót ilyen módon old meg, például így lehet oldalakat kategorizálni. Az alásorolás azt jelenti, hogy a lap címe egy másik lap címének kiegészítését tartalmazza. Egyik tipikus használati módja az archívum: a lap tartalmának eltávolítása és egy alárendelt oldalra helyezése⁷.

Fontos még kiemelni, hogy a Felhasználók önálló, a szokott fogalomnak megfelelő entitások a Wikipédián, akik saját Lapokkal rendelkezhetnek (erre a célra két külön névtér szolgál). Ezek a felhasználói lapok lehetőséget biztosítanak a felhasználók számára, hogy információkat tegyenek közzé magukról, például a szerkesztési tevékenységükről vagy érdeklődési köreikről. Az úgynevezett felhasználói vitalapok a felhasználóval való kommunikáció színterei, ahol más felhasználók üzeneteket hagyhatnak, kérdéseket tehetnek fel vagy vitákat folytathatnak. Felhasználó olyan lappal nem áll birtokosi viszonyban, amely nem felhasználói lap vagy felhasználói vitalap, ezt a névtér határozza meg. Az ilyen lapokat a felhasználókkal a lap neve köti össze, például, ha a használó neve Teszt, akkor a Tesztre vonatkozó felhasználói információk a User:Teszt oldalon férhetők

⁷ Például a https://simple.wikipedia.org/wiki/Talk:Wikipedia/Archive_1 címen található oldal a <https://simple.wikipedia.org/wiki/Talk:Wikipedia> archívuma.

hozzá, míg a Teszt számára üzenetet írni szándékozó társak számára a User vita:Teszt lap áll rendelkezésre.

5.4.2 ETL Esettanulmány

A demonstrációhoz egy specifikus nyelvi változatot választottam, de az alkalmazott módszerek, eszközök bármely azonos szerkezetben kezelt MediaWikire alkalmasak a közös platformból adódóan, módosítás nélkül átültethető. Az esettanulmány adatkészletének kiválasztásánál a következő megfontolások vezettek:

- Az adott MediaWiki oldal adatbáziskivonata legyen hozzáférhető.
- A tartalmak angol nyelvűek legyenek, a szövegbányászati eszközök minél szélesebb körű alkalmazhatósága érdekében. A Wikipedia nyelvi változatok között angol nyelven készül az en.wikipedia.org és a simple.wikipedia.org. A legnépszerűbb Fandom oldalak jellemzően angol nyelvűek.
- Legyen általános célú, azaz változatos érdeklődési körű felhasználókat vonzzon.

A Simple English Wikipedia (a továbbiakban: simplewiki) a fentiek alapján alkalmas választás. Az adatkészlet hozzáférhető, a tartalmak angol nyelvűek, a tartalom általános. A simplewiki 2001. szeptember 18-án indult (mindössze nyolc hónappal az első Wikipedia, az általános angol nyelvű változat után), azaz 23 évnyi tartalom feldolgozásának lehetőségét biztosítja. A több mint két évtized alatt közel 260 ezer szócikk és közel 600 ezer egyéb oldal készült. Később bemutatom, hogy a 600 ezer egyéb oldal között megtalálhatóak a közösségi kommunikáció területei is. A simplewiki másfélmillió regisztrált felhasználóval rendelkezik, amelyek közül a lekérdezés időpontjában 1600 aktív volt, azaz szerkesztette a tartalmakat a statisztika összeállítását megelőző 30 napban (Simple English Wikipedia, 2024a).Az alábbi fejezetekben bemutatom a Simple English Wikipedia adatkészletén lefuttatott ETL folyamatot.

5.4.2.1 Esettanulmány tervezése

Az esettanulmány célkitűzése az 5.3.4 fejezetben lefektetett elvek mentén feldolgozni a simplewiki teljes adatkészletét. Az alábbiakban az esettanulmány főbb feladatainak összefoglalása:

- **Elérhetőség felmérése:** bemutatom, hol és milyen formában férhetők hozzá a simplewiki adatok.

- **Releváns adatforrások ismertetése:** a közzétett adatokból felhasználásra tervezett struktúrák ismertetése.
- **Felhasználói adatok lehetséges forrásainak felmérése:** a felhasználásra tervezett adatok közül a felhasználófüggő adatok lehetséges adatforrásainak válogatása.
- **Tartalom felmérése: kontextusok és tartalmak azonosítása:** a felhasználásra tervezett adatok közül a tartalomfüggő adatok lehetséges adatforrásainak válogatása, kontextusok és tartalmak azonosítása.
- **Felhasználói, kontextus és tartalmi adatok megfeleltetése:** szisztematikus összerendelés a fogalmi adatmodellel.
- **Konverziós folyamat és algoritmusainak tervezése:** szükséges adatgyűjtési és transzformációs folyamatok áttekintése és tervezése.
- **Konverziós folyamat futtatása:** implementáció Microsoft SQL Server platformon.
- **Ellenőrzés:** helyreállítási tesztek, illetve leíró statisztikák segítségével.

5.4.2.2 Adatgyűjtési eljárások tervezése

A simplewiki adatainak gyűjtéséhez az üzemeltető által elérhetővé tett adatforrások: a MediaWiki API és a simplewiki adatbázis-kivonat.

MediaWiki API: A MediaWiki alapú rendszerek lehetővé teszik az adatok lekérdezését egy alkalmazásprogramozási interfészen keresztül. A MediaWiki API egy olyan interfész, amely lehetővé teszi a fejlesztők számára, hogy webes hívások segítségével hozzáférjenek a MediaWiki-alapú weboldalak által tárolt tartalomhoz. Az API többek között támogatja az adatok lekérdezését, oldalak szerkesztését, felhasználói fiókok kezelését, képek feltöltését, valamint a változások nyomon követését. REST-alapú hívásokon keresztül működik (MediaWiki, 2024a). Az API használat ugyanakkor jelen esettanulmány szempontjából komoly korlátozásokkal rendelkezik: a hitelesítés nélküli használat esetén óránként 500, minden más esetben 5000 hívás valósítható meg (Wikimedia API portal, 2023).

A MediaWiki adatbázis-kivonatok a MediaWiki platformot használó weboldalak, például a Wikipedia tartalmának teljes vagy részleges adatbázis-másolatait tartalmazzák. Nagyrészt XML, részben SQL formátumban kerülnek közzétételre. Ezek a kivonatok

rendszeresen, általában havonta készülnek, és az üzemeltető döntésétől függően nyilvánosan hozzáférhetők. Egy kivonat közel 40 féle állományt tartalmaz, amelyek a legkülönbözőbb célokat szolgálják, például külön állományban érhető el az aktuális állapot, a teljes történet, vagy éppen a különféle keresőmotorok számára kibontott absztraktok (MediaWiki, 2024b; Wikimedia, 2024a).

Mivel a cél a teljes adatállomány feldolgozása, amely a simplewiki indulása óta közel 10 millió közreműködést jelent, az API használat nem célravezető: azon túl, hogy több ezer órát venne igénybe mind a két híváskorlátozási szinten, az API hívások paraméterezéséhez sem állna minden esetben a kellő mennyiségű információ, például egy felhasználó adatainak lekéréséhez szükséges annak azonosítója vagy neve, amely a folyamat kezdetén még nem ismert. Ezért célként az adatbázis-kivonatok feldolgozását jelöltem ki.

A korábban leírtak szerint a felhasználók és a tartalom kölcsönhatásait a felhasználók, lapok és lapváltozatok hármasa írja le. A felhasználók önmagukban csak kulcs-érték párként (azonosító és egy egyedi név) vannak jelen a rendszerben, a velük való kommunikáció és saját identitásuk meghatározása szintén a lapokon történik. Ennek elősegítésére lapok rendelkeznek egy felső, kötött kategorizálással, amelyeket névtereknek neveznek: a lap névtérbeli hovatartozása határozza meg a használat célját. Ezek felett is meghatározhatunk egy felettes kategorizálást: egyes névterek célja a tartalom, míg másoké a tartalmakkal kapcsolatos felhasználói interakciók tárolása.

A felhasználók tartalommal történő műveleteit a **laptörténet** (revisions) írja le⁸. A laptörténet minden szerkesztés után menti a lap teljes tartalmát. A törölt lapok szerkesztései az állományból eltűnnek és semmilyen módon nem nyerhető ki róluk információ. A simplewiki 2024. augusztus elsején közzétett kivonatában a laptörténet egy 63 Gb méretű xml állomány, amely közel 9 millió lapváltozatot tartalmaz.

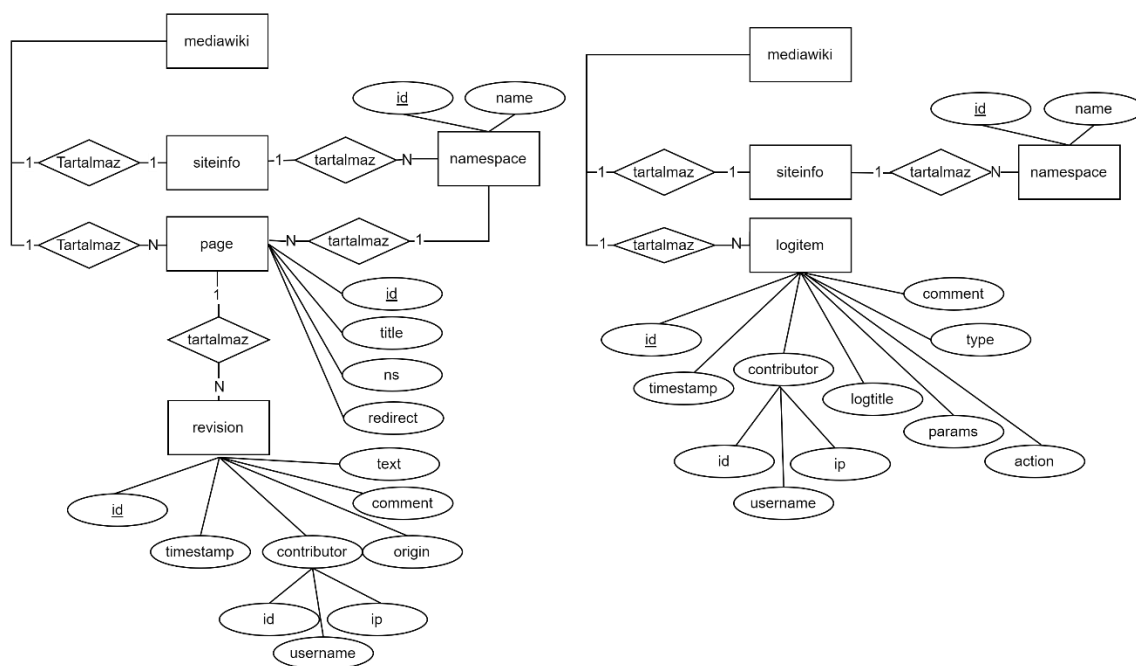
A felhasználók és a tartalmak közötti, szerkesztések túli interakciókat a **rendszernapló**⁹, írja le. Például kép feltöltése, tartalom jóváhagyása, felhasználói fiók regisztrációja, lap törlése, vagy egy felhasználó kitiltása. A rendszernaplók esetében a MediaWiki rendszer

⁸ Példa: <https://simple.wikipedia.org/w/index.php?title=Hungary&action=history>

⁹ Példa: <https://simple.wikipedia.org/w/index.php?title=Special:Log>

fejlesztése jól látható a különféle évek adatain: a művelet célja (a törölt lap címe, vagy a kitiltott felhasználó neve) megjelenhet a comment, a logtitle és a params mezőben is, utóbbi mezőben serializált tömb és egyszerű szöveges adatok formájában is megjelenhetnek leíró adatok, naplótípustól függően. A simplewiki 2024. augusztus elsején közzétett kivonatában a rendszernapló egy 1 Gb méretű xml állomány, amely mintegy 2,5 millió logbejegyzést tartalmaz.

A lapváltozatokat, illetve a rendszernaplót leíró adatállomány szerkezetét a 27. ábra szemlélteti. Az összevont laptörténetek és napló alapján az általános adatmodell alapján készített adatbázis feltölthető. Az ábra attributált Chen-féle notációval készült, melynek leírását a 8.1 számú melléklet tartalmazza.



27. ábra: A laptörténet és a napló adatbázis-kivonat szerkezete (Saját szerk.)

5.4.2.3 Összegyűjtött adatok elemzése

A kivonatból kinyerhető lapváltozatok és naplóállományok összefoglalását adja a 16. táblázat.

Megnevezés	Tartalom	Számosság
revisions ¹⁰	Nem törölt lapok lapváltozatai teljes szöveges tartalommal	8 752 431
namespaces	A lapok kategorizálásához	22
log/newusers	Új felhasználói regisztrációk	1 462 332
log/create	Lapok létrehozása	439 014
log/delete	Lapok törlése	291 791
log/patrol	Újonnan létrehozott lapok jóváhagyása	113 447
log/move	Lapok átnevezése	71 963
log/block	Felhasználó kitiltása és visszaengedése	56 366
log/renameuser	Felhasználók átnevezése	36 550
log/import	Más nyelvi változatokból történő importálás	34 753
log/thanks	Felhasználók köszönetnyilvánítása más felhasználók felé	17 732
log/protect	Lapvédelem	10 887
log/growthexperiments	Új szerkesztők bevonását célzó program logja, amely olyan apró feladatokat kínál a szerkesztők számára, ami segíti a beilleszkedésüket, például linkajánlás.	7 215
log/rights	Jogosultságok kiosztása és elvonása.	4 799
log/upload	Audiovizuális állományok feltöltése	777
log/tag	Lapváltozások megjelölése, címkézése	590
log/massmessage	Tömeges (rendszer)üzenetek naplója	41
log/merge	Lapváltozatok összefűsülése egyetlen oldal alá	17
log/contentmodel	A Wikipédián az oldalak többsége wikitext jelölőnyelvet használ, ez a napló gyűjti az egyes oldalak tartalmi modelljének megváltoztatását	10
log/gblblock	Globális, az adott nyelvi változaton túlmutató kitiltás	10
log/managetags	Az elérhető címkék kezelése	2
log/usermerge	Felhasználók összefűsülése	2
log/globalauth	Globális, az adott nyelvi változaton túlmutató felhasználói fiók menedzsment	2
log/review	Jelölt lapváltozatok funkció naplója	-
Összesen		11 300 753

16. táblázat: Adatbáziskivonat (Saját szerk.)

Az esettanulmány tervezésénél ismertetett hármás bontás szerint az alábbiakban áttekintem az összegyűjtött felhasználói, kontextus és tartalmi adatkörök jellemzőit.

Felhasználói adatok

A felhasználói adatok forrásait mutatja a 17. táblázat.

¹⁰ 833 813 egyedi lap

Attribútum/Entitás	Forrás	Megjegyzés
Azonosító	Newusers napló Bármely naplóállomány contributor mezői	A naplóállomány hibájából adódó hiányosságai részben pótolhatók az egyéb naplóállományokból nyert tényadatokkal.
Felhasználónév	Newusers napló Bármely naplóállomány contributor mezői Korábbi felhasználónevek időbélyeggel: renameusers napló	A naplóállomány hibájából adódó hiányosságai részben pótolhatók az egyéb naplóállományokból nyers tényadatokkal. A renameusers naplóból felhasználók korábbi nevei állíthatók helyre idősorosan. A felhasználónév egy adott időpillanatra nézve egyedi érvényességű, felhasználónként egy nyitott végű név engedhető meg.
Létrehozás időpontja	Newusers napló	A newusers naplóban szereplő felhasználókról ismert a regisztráció időpontja. Az ott nem szereplő felhasználók részére ilyen adatot nem tárolunk.
Jogosultságok	Rights napló, időbélyeggel	A Wikipedia jogosultságok kiosztását és elvételét idősorosan a rights napló tartalmazza.
Kitiltások	Block napló, időbélyeggel	Egy felhasználót örökre, vagy hosszabb-rövidebb időre kitilthatnak a rendszerből.
Egyéb attribútumok	Revisions	A felhasználók magukról leíró adatokat tehetnek közzé tulajdonolt oldalaikon. Ezek a revisions állományból állíthatók helyre.
Relációk	Thanks	A köszönetnapló tárolja a köszönetet küldő felhasználó azonosítóját és a köszönetet fogadó felhasználó aktuális nevét. Nem csak regisztrált felhasználóknak küldhető köszönet.
Relációk	Block/Unblock	Megfelelő szerepkörű felhasználó más felhasználók jogait korlátozhatja és a korlátozást megszüntetheti.

17. táblázat: A felhasználó adatok forrásai (Saját szerk.)

Megjegyzések:

Regisztrációs rekord: Bármely naplóállomány, különösen a korai rekordok esetében, mutathat hiányosságokat, nem mentes ettől a felhasználói regisztrációk naplója sem (newusers). Mivel az adatbázis felépítésének alapköve a felhasználói bázis minél pontosabb kialakítása, így a hiányos adatokat egyéb naplóállományokban felelhető adatokból vezetem le. Az ilyen felhasználók esetében a regisztráció időpontja nem áll rendelkezésre, egyedi döntés kérdése, hogy az első tevékenység dátumának rögzítéséből adódó pontatlanságok akceptálhatók-e.

Időfüggő attribútumértékek: A modell lehetőséget biztosít attribútumértékek időfüggő tárolására. Döntést kell hozni arról, hogy mely attribútumok esetében engedünk meg többszörös érték-hozzárendelést. A létrehozás időpontja esetében kikötés, hogy csak 0 vagy 1 érték tartozhat a felhasználóhoz, míg egyes jogosultságcsoportokból és felhasználónévből egyetlen érvényes engedhető meg.

Névtelen és automatizált felhasználók: Tekintve, hogy a Wikipedián regisztráció nélkül is lehetséges szerkeszteni, minden naplóban megtalálható regisztrálatlan felhasználóra mutató bejegyzés, ami nem tartalmaz userid-t. Az ilyen felhasználóknak az IP címe érhető el az állományokból. Az ilyen felhasználókat nem töltjük be a users táblába, mivel az IP cím nem alkalmas hosszabb időn keresztül való stabil azonosításra. Az automatizált felhasználók tömeges szerkesztéseket végeznek a rendszerben, az automatizáltságot egy jogcsoport („bot”) jelzi, ez egyben azt is jelenti, hogy egy felhasználó lehet bizonyos ideig „bot”, majd visszatérhet aktuális szerepéhez. Az ilyen felhasználókat tárolni szükséges.

Szerepkörök: MediaWiki megkülönböztet egymástól szerepköröket, amelyek kezelése hasonló módon történik, mint a „bot” besorolás megadása. azaz egy felhasználónak több jogosultsági szintje is lehet egy időben. A szerepkörökhöz „role” attribútum hozandó létre, és a felhasználó attribútumok között tárolható maga a jogosultság.

Egyéb tulajdonságok: A felhasználóknak a nevükön túl semmilyen adatot nem szükséges magukról megadni. Adatgyűjtési lehetőség a sablonok és kategóriák használatával adódik, ezek viszont nem a felhasználói entitáshoz, hanem a felhasználói entitás által tulajdonolt laphoz kapcsolódnak, így ez a kontextus entitások és a tartalom entitások feldolgozásakor kerül ismertetésre.

Felhasználók közötti relációk: A felhasználók közötti közvetlen (azaz nem üzenetküldéssel járó és tartalomhoz nem kötődő) explicit interakciók a Wikipediában csekély számúak. Három ilyen interakció deríthető fel: köszönet, tiltás és tiltás feloldása. A köszönet két, bármilyen szerepkörű felhasználó közötti esemény, a napló „thanks” részéből vonhatók ki. A tiltás és a tiltás feloldása magas szintű szerepkörrel rendelkező felhasználók által elvégezhető művelet más felhasználókkal kapcsolatban. A log tábla „block” és „unblock” eseményeiből vonható ki. Megjegyezhető, hogy például a felhasználóátnevezési vagy jogosultsági napló esetében is értelmezhető a kezdeményező és érintett felhasználó viszonya, de az adott közösségben ezek a műveletek nem a kezdeményező felhasználó saját döntéséből történnek, így kutatásfüggő, hogy extrakciójuk a UserRelations táblába megtörténik-e.

Kontextus

A kontextusadatok forrásait mutatja a 18. táblázat.

Attribútum	Forrás	Megjegyzés
Azonosító	Revisions napló page_id Revisions napló namespace_id	Ismétlődések nélkül
Cím	Revisions napló title Revisions napló namespace Szövegváltozatok alapján szekciócímek	Ismétlődések nélkül
Létrehozás időpontja	Revisions napló időbélyeg	Ismétlődések nélkül; Törölt lapok esetében a Create napló
Jelleg	Névtér alapján	Közösségi/Felhasználói/Tartalom
Létrehozó felhasználó	Revisions napló első contributor	Ismétlődések nélkül; Törölt lapok esetében a Create napló
Tulajdonos	Korábban összegyűjtött felhasználónevek	User és User talk névtérben; elnevezés alapján
Kontextusok kapcsolata	Revision szövegváltozatok összehasonlítása	Beágyazás alapján
Szerkesztés korlátozása	Protect napló	Felhasználó-kontextus interakció Kontextusattribútum
Dinamikus felhasználói attribútumok	Revision szövegváltozatok összehasonlítása	Beágyazás alapján

18. táblázat: A kontextusadatok forrásai (Saját szerk.)

Az ETL folyamat ajánlása szerint a rendelkezésre álló adatok közül a kontextusok elhatárolása a következő lépés. A Wikipédiában a legmagasabb szinten álló kontextustípus megválasztásánál két lehetőség kínálkozik: a névtér és a lap. Maga a lapváltozat már konkrét felhasználóhoz kötődő közreműködés.

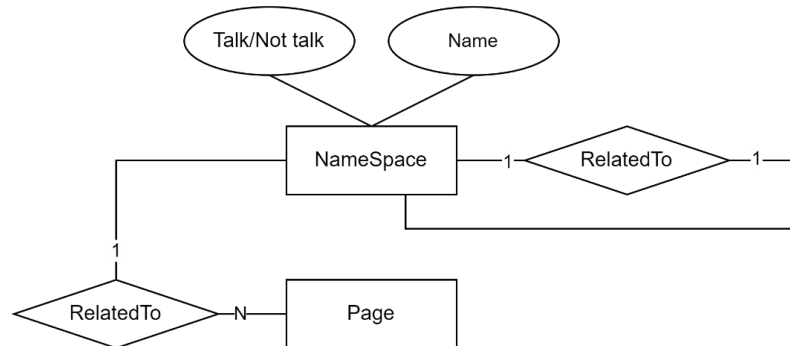
A „névtér” egy, a benne foglalt lapok felhasználási és értelmezési kontextusát meghatározó tulajdonságot jelent a Wikipédiában. Példával illusztrálva a 0 azonosítóval rendelkező, a simplewikin üres elnevezéssel szereplő főnévtér szolgál a cikk típusú oldalak tárolására. Az 1-es névtérű „Talk” névtér a cikk típusú oldalakkal kapcsolatos diskurzusokra. A simplewikin alkalmazott névtéreket a 19. táblázat foglalja össze.

Kód	Név	Típus	Példa
0		Cikk típusú oldalak	https://en.wikipedia.org/wiki/Hungary
1	Talk	Cikk típusú oldalakhoz kapcsolódó viták	https://en.wikipedia.org/wiki/Talk:Hungary
2	User	Felhasználók saját oldalai	https://en.wikipedia.org/wiki/User:Angela
3	User talk	Felhasználók megszólítására alkalmas oldalak	https://en.wikipedia.org/wiki/User talk:Angela
4	Wikipedia	Közösségi névtér, amelyen egyaránt folyhatnak beszélgetések vagy szerepelhetnek útmutatók	https://simple.wikipedia.org/wiki/Wikipedia:Simple_talk
5	Wikipedia talk	Wikipedia névtérbeli lapokhoz kapcsolódó diskussziók	https://simple.wikipedia.org/wiki/Wikipedia_talk:Simple_talk
6	File	Helyben felöltött állományok, pl. képek	https://simple.wikipedia.org/wiki/File:Wiki.png
7	File talk	File névtérbeli lapokhoz kapcsolódó diskussziók	https://simple.wikipedia.org/wiki/File_talk:Wiki.png
8	MediaWiki	Technikai névtér, stílusok, javascriptek tárolásához	https://simple.wikipedia.org/wiki/MediaWiki:Monobook.css
9	MediaWiki talk	MediaWiki névtérbeli lapokhoz kapcsolódó diskussziók	https://simple.wikipedia.org/wiki/MediaWiki_talk:Sidebar
10	Template	Oldalakra beilleszthető sablonok	https://simple.wikipedia.org/wiki/Template:Wikipedia_policies_and_guidelines
11	Template talk	Template névtérbeli lapokhoz kapcsolódó diskussziók	https://simple.wikipedia.org/wiki/Template_talk:Wikipedia_policies_and_guidelines
12	Help	Wikipedia szerkesztésről szóló útmutatók	https://simple.wikipedia.org/wiki/Help:Contents
13	Help talk	Help névtérbeli lapokhoz kapcsolódó diskussziók	https://simple.wikipedia.org/wiki/Help_talk:Contents
14	Category	Más oldalak kategorizálására szolgáló oldalak, beillesztéssel érvényesülnek.	https://simple.wikipedia.org/wiki/Category:Names
15	Category talk	Category névtérbeli lapokhoz kapcsolódó diskussziók	https://simple.wikipedia.org/wiki/Category_talk:Names
710	TimedText	Videók feliratozására szolgáló modul, a simplewikin üres	-
711	TimedText talk	TimedText névtérbeli lapokhoz kapcsolódó diskussziók, a simplewikin üres	-
828	Module	Olyan speciális sablonok (Template), amelyekbe programkód is helyezhető	https://simple.wikipedia.org/wiki/Module:Storm_categories
829	Module talk	Module névtérbeli lapokhoz kapcsolódó diskussziók	https://simple.wikipedia.org/wiki/Module_talk:Storm_categories

19. táblázat: Wikipedia névterek (Saját szerk.)

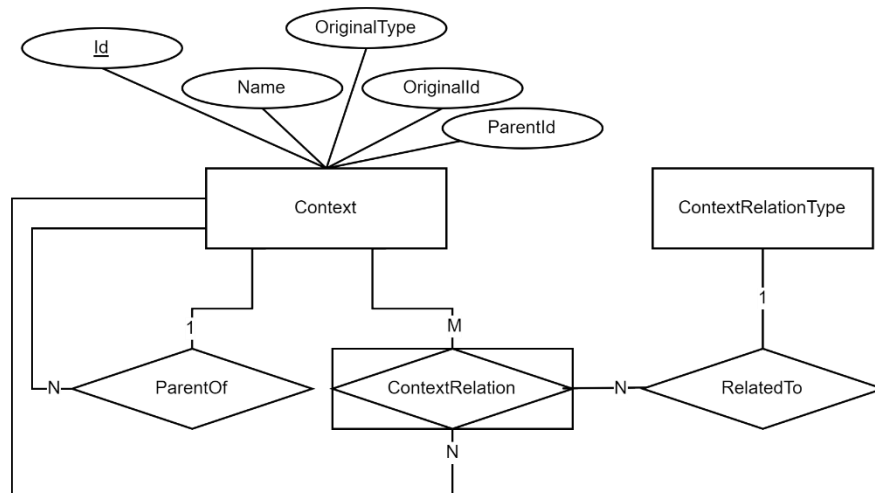
A névterek meghatározzák, hogy beszélgetés vagy tartalomszerkesztés zajlik-e rajtuk, ennek alapján a névterek kategorizálhatók a kutatás aktuális célja szerint (például felhasználói, közösségi, enciklopédikus névtér).

Minden oldalra jellemző pontosan egy névtér, és minden névtér párba állítható pontosan egy másik névtérrel: minden névtérhez kapcsolódik egy „Talk” névtér. A viszonyokat a 28. ábra mutatja.



28. ábra: Wikipedia névterek és lapok viszonya (Saját szerk.)

A közös adatmodell a kontextusokat hierarchikus formába ajánlja szervezni. A 29. ábra egy részletet tartalmaz a modellből:



29. ábra: Közös adatmodell kontextus entitások (Saját szerk.)

Ennek alapján a célravezető elrendezés a „talk” névterek besorolása a kapcsolódó névtér alá a kontextusok között fennálló ParentOf reláció elrendezés segítségével, majd az oldalakhoz a tényleges névtér beállítása. Ez a hierarchia lehetővé teszi a vonatkozó tartalmak rekurzív vizsgálatát.

Az oldal típusú kontextusok a laptörténet állományokból kibonthatók, szülőként a megfelelő névtér állítandó be, tulajdonságuk a cím.

A következő kérdés, hogy vajon a lap áll a hierarchia alsó szintjén, vagy van-e más, alacsonyabb szintű befoglaló blokk? Ehhez célszerű megvizsgálni egy-egy tipikus Wikipedián előforduló beszélgetést, illetve kollaboratív tevékenységet. Előbbire legyen példa Magyarország simplewiki vitalapja¹¹, utóbbira maga a szócikk¹². A 30. ábrán látható, hogy a Wikipedia oldalak szekciókra bomlanak, amelyek külön cím alá rendeződnek. Tartalmi oldalaknál ezek a szócikkek alcímeit jelentik, míg vitaoldalakon a beszélgetések elkülönítésére szolgálnak. A legalsó szintű tartalmi blokk tehát a szekció.



30. ábra: Wikipedia vitalap (Simple English Wikipedia, 2024b)

A szekciók az eredeti adatállományról kizárólag a szövegtartalom összehasonlításával bonthatók ki, önálló azonosítót a forrásállományok nem rendelnek hozzájuk. Egy szerkesztéssel: történhet tartalom hozzáadása egy vagy több már meglévő szekcióhoz¹³, létrehozható egy vagy több új szekció tartalommal és anélkül¹⁴. A szekció megadása nem kötelező, és egy-egy szerkesztésben több szekció is megadható. Így elmondható, hogy a szekciók azonosítása egyrészt a laptörténetek összevetését, másrészt a beillesztett

¹¹ <https://simple.wikipedia.org/wiki/Talk:Hungary>

¹² <https://simple.wikipedia.org/wiki/Hungary>

¹³ <https://simple.wikipedia.org/w/index.php?title=Talk:Europe&diff=next&oldid=7874984>

¹⁴ <https://simple.wikipedia.org/w/index.php?title=Talk%3AEurope&diff=6724244&oldid=5258691>

szövegrészek pozíciójának elemzését teszi szükségessé. A szekciókhoz azonosító generálandó, ami a létrehozó laptörténet azonosítóját és az új szekció pozícióját tartalmazza, mivel a szekciók neve nem egyedi. Kezelní szükséges a szekciók eltávolítását is az oldalról.

Mivel az adatbázis építése közösségvizsgálati célokból történik, kollaboratív munkavégzés esetén a „page” szintű helymeghatározás a tartalmi oldalak esetén elegendő annak meghatározásához, hogy két felhasználóról feltételezhető-e a közös érdeklődési kör. Vitaoldalak esetén viszont az oldal túlzottan elnagyolt szemcsézettséget jelentene, hiszen, ha két felhasználó azonos beszélgetésben jelenik meg, az sokkal jelentőségtejjesebb információ, mint az, hogy mindketten szerkesztették a beszélgetések folytatására szolgáló oldalt (egy Facebook analógiával élve, tagja lenni ugyanannak a Facebook csoportnak kevesebbet jelent, mint ugyanabban a csoportban azonos poszthoz hozzászólni).

Kontextusok egymás közötti kapcsolatai: két oldal között az eddig ismertetteken túl beágyazással létesülhet kapcsolat. Két oldal között kapcsolat linkek (kategóriák, más oldalak) vagy beágyazás (sablonok, modulok) segítségével építhető ki. Ezek az elemek időfüggő módon a revízió szövegéből nyerhetők ki és tárolható a ContextRelations táblában. A beágyazás módot nyit újabb attribútumok meghatározására: a felhasználói lapokon beágyazott sablonok és kategóriák olyan strukturált információk, amelyekből felhasználói attribútumok határozhatók meg a UserAttributes táblában¹⁵.

Kontextusok és Felhasználók kapcsolatai: Minden nevesített felhasználó rendelkezhet saját lapokkal. A kapcsolat a felhasználónév alapján teremthető meg, „Test” nevű felhasználó felhasználói lapja a User:Test, vitalapja a User talk:Test néven található meg az oldalak között. Mind a két laptípus rendelkezhet allappal, például: User:Test/Subpage, User talk:Test/Subpage. Ezek a fő- és aloldalak egyértelműen a felhasználóhoz kötődnek, az adott felhasználó a lapok tulajdonosa, a User talk oldalon az adott felhasználó közvetlenül megszólítható. Ennek megfelelően beállítandó az OwnerId mező.

¹⁵Példa: <https://simple.wikipedia.org/w/index.php?title=User:Barliner&oldid=480313>. A felhasználó vállalt egészségügyi információkat ad meg magáról ebben a lapváltozatban.

Tartalmak

A tartalomra vonatkozó adatok összerendelését adja a 20. táblázat.

Attribútum	Forrás	Megjegyzés
Azonosító	Revisions napló	Generálás szükséges
Létrehozás időpontja	Revisions napló időbélyeg	
Létrehozó felhasználó	Revisions napló első contributor	Ismétlődések nélkül; Törölt lapok esetében a revision szövege nem áll rendelkezésre
Tartalom	Revision szövegváltozatok összehasonlítása	
Felhasználók és tartalmak kapcsolata	Patrol napló	Jóváhagyás
Dinamikus attribútumok	Revisions napló szövegváltozatok összehasonlítása	A kutatás célja szerint: nyers szöveg tárolása helyett/mellett NLP is végezhető.

20. táblázat: A tartalmi adatok forrásai (Saját szerk.)

Tartalmi blokkok: A tartalom az egyetlen felhasználó egyetlen közreműködésével létrehozott egység. Minden lapváltozattal új tartalom (vagy több szekció esetén több tartalom) jön létre egy adott kontextuson belül, amely egyértelműen egy adott felhasználóhoz kötődik. A közzétett kivonat a lapváltozatok teljes szövegét tárolja, így a tartalom kinyeréséhez a változatot az előző változattal össze kell hasonlítani, és csak a szövegdiffereenciákat (ún. deltákat) tárolni.

Szekciók: A kontextusok kapcsán kifejtésre került a szekciók, mint kontextus extrakciós folyamata. A tartalmak szekcióhoz rendelése szövegpozíció vizsgálat alapján történik: minden hozzáadott tartalom a pozícióban legközelebb elhelyezkedő szekcióhoz kötődik.

Attribútumok: A kutatás céljától függ, hogy a konkrét szövegtartalmat attribútumként tároljuk-e. Amennyiben az adatbázis célfüggetlen, ez a kézenfekvő választás, de ha méretkorlátozásokkal kell számolni a tároló adatbázis kapcsán, felmerül például az, hogy csak a kulcsszavakat és/vagy érzelmi indexeket tároljuk.

Felhasználó-Tartalom interakciók: „patroller” jogosultságú felhasználók az újonnan létrehozott lapok első lapváltozatát ellenőrzöttnek jelölhetik (olyan wikiken, ahol a lapváltozat ellenőrző rendszer aktív, ez a jogosultság minden olyan lapváltozatra kiterjedhet, amelyet magasabb szintű jogosultság nélküli felhasználó hozott létre).

5.4.2.4 Bizonyítékok gyűjtése

E fejezet célja a fejlesztési folyamat tömör összefoglalása, a konverziós folyamat algoritmusainak tervezése. Az 5.3.3 fejezetben ismertetett adatmodellt Microsoft SQL

Server 2022 Express adatbázis platformom implementáltam, SQL és TSQL parancsállományok készítésével. Az adatgyűjtés megvalósításához Python nyelven előfeldolgozó készült. Az alábbiakban tömören összefoglalom a megvalósított lépéseket és az alkalmazott algoritmusokat.

Az előfeldolgozás során a nagyméretű szöveges állományok tisztítása és tömörítése, valamint a szövegdiffereciák vizsgálata történt meg. A bemeneti (xml) adatállomány a lapváltozatok kapcsán 63 GibiByte (9 millió lapváltozat), a naplóállományok kapcsán 1 GibiByte (2,5 millió logbejegyzés). Ezek az állományméret, illetve tételek számok a szöveges állomány szekvenciális feldolgozását követelik meg, a hagyományos xml deszerializáció nem képes hatékonyan működni. A szekvenciális feldolgozás során is tekintettel kell lenni a visszakeresés optimalizálására. A lépéssorozat a következő.

1. **Állományok beszerzése:** a kivonatok közzétételi helyére a napló és a revisions állomány beszerzése.
2. **Adattisztítás:** a bemeneti formátumok célfüggő konszolidációja

2.1 Lapváltozatok: szövegtartalom nélküli lapváltozat állomány előállítás.

2.2 B-indexek előállítása: A szövegtartalom visszakereséséhez kétszintű indexállományt állítottam elő, amely tartalmazza a lapok-lapváltozatok kapcsolatát és azt, hogy egy adott lapváltozat szövegtartalma a kiinduló állományban hol érhető el.

2.3 Szövegdiffereciák keresése: az időben egymás után következő lapváltozatok szövegének összehasonlítása. A feldolgozáshoz a difflib csomag Differ osztályát használtam, amely Myers algoritmusát (Myers, 1986) implementálja. Az előfeldolgozó a beszúrák és törlések sorozatával, az úgynevezett deltákkal tér vissza, amelyből a szekciócímek, a hozzáadott és törölt tartalom azonosítható.

2.4 Logállományok: az előző fejezetben feldolgozásra kijelölt naplóállományok feldolgozása. Ez magában foglalja a napló felbontását típusonként kisebb, kezelhetőbb állományokra, valamint a params mezőben tárolt összetett adattípusok dimenziócsökkentését (a params tömbben php szerIALIZED tömbök fordulhatnak elő).

3. **Gyorsítótár építése:** a feldolgozás eredményeit köztes állományokba mentem. Erre a célra olyan bináris állományokat célszerű használni, amelyek önmagukban biztosítják a tömörítést, és gyors elérést biztosítanak. Az elérhető alternatívák közül a feather állományok használatát választottam (Apache Arrow, 2020). Ezzel a

módszerrel a naplóállományok bemeneti 1Gb-os mérete 80%-kal csökkenthető, a redundáns attribútumok kibontása dacára.

- 4. Relációs táblák előállítása:** A feldolgozó eljárások eredményeképpen pandas (McKinney, 2010) dataframe-ek állnak elő, amelyeket az sqlalchemy csomag (Bayer, 2012) segítségével relációs adattáblákká konvertálok.

Ezt követően az összegyűjtött adatok elemzése során rögzített következtetéseim alapján az adatbázistáblák feltölthetők. A feltöltés során figyelemmel kell lenni az ott megfogalmazott szabályokra, valamint az unicode nyelvi környezetre. A szöveges mezők visszakeresését megkövetelő eljárások optimális futása érdekében az attribútumtáblákat hashmap indexekkel láttam el. A futtatás sikeresen lezajlott, a vonatkozó parancsfájlok a disszertációhoz készült GitHub repozitóriumban érhetők el (Magyar, 2025). .

5.4.2.5 Következtetések

Az összegyűjtött és normalizált állományok betöltése eredményeképpen előáll egy célfüggetlen relációs adatbázis, amely lehetővé teszi a modell képességeinek alkalmazását tudományos kutatásokban. Erre mutat példát a következő alfejezet.

5.4.3 Exploratív adatelemzés

A Wikipedia egyszerűsített angol nyelvű változata 2001-ben indult, és az eltelt több mint 20 év alatt közel 264 ezer szócikket hozott létre a mögötte szerveződő önkéntes szerkesztői közösség. A vizsgálat célja a létrehozott tartalom elemzése annak tükrében, hogy a szerkesztői közösségbe való beilleszkedés milyen hatással van az enciklopédia termelékenységére.

5.4.2.1 Kutatási kérdés

Milyen összefüggés figyelhető meg a szerkesztői aktivitás koncentrálódása és a közösségi beágyazottság között?

A Wikipedia elsődleges célja a kollaboratív tartalom előállítása önkéntes közösségek munkája által. Az enciklopédia minden szempontból ingyenes, az alkalmazott licenz a tartalmak szó szerinti átvételét is lehetővé teszi (forrásmegjelöléssel). Az egyes nyelvi változatok felületén nincs reklám, minden szempontból önkéntes közreműködésre épít, azaz az egyes nyelvi változatokon belül nem tevékenykednek a fenntartó által fizetett alkalmazottak. Így a megfelelően fejlődő és önfenntartó, valamint produktív közösség

kialakítása az adott nyelvi változat szempontjából létfontosságú. Könnyen levonható a következtetés, hogy a tartalomelőállítást célzó kontextus mellett e közösség kialakítása és fenntartása, a szerkesztők motiválása az egyéb oldalakon valósul meg. Annak érdekében, hogy összefüggéseket derítsék fel a tartalom előállítását célzó és a szerkesztőségi kommunikációt szolgáló két szintér között, a vizsgálatba egy további adatsort vonok be: a kontextusok és a tartalmak felhasználásával megjeleníthető a kontextustípusonként elért közreműködések darabszáma, valamint az újonnan létrehozott oldalak száma. Ez az adatsor tehát a kontextus típusát, az időpontot, valamint a közreműködés időpontját tartalmazza.

5.4.2.2 Adatok meghatározása

A vizsgálathoz szükséges annak világos elhatárolása, hogy mely közreműködések tekintjük a kollaboratív munkavégzés eredményének és melyeket a közösségi lét megélésének terepének. Ehhez az adatmodellben hierarchikusan kialakított kontextusrendszer kiváló alapot nyújt. A hierarchia csúcsán álló kontextusok alkalmasak a felhasználói együttműködések elhatárolására.

- A **tartalmi tér** minden olyan felettes kontextus, amely az enciklopédia passzív látogatói számára megjelenő tartalmat szolgáltatják. Összesen 7 kontextus és az alájuk sorolt kontextusok tartoznak ide, úgymint cikk, kategória, médiaállományok stb.
- A felhasználók **személyes terét** kontextustípus jeleníti meg, az ezen való szerkesztés az adott felhasználó személyes megszólítását jelenti, de olyan módon is kialakulhat itt kapcsolat, hogy kettőnél több felhasználó folytat diskurzust.
- A **közösségi tér** a fentiekén kívül eső kontextustípusokat jelenti. Ebben a kontextusban a felhasználók megvitathatják a tartalmakat vagy más közösségi ügyeket, segítséget nyújtanak más felhasználóknak. Összesen 11 kontextusból áll, úgymint a főként közösségi beszélgetések teréül szolgáló Wikipedia kontextus, vagy a cikkek mentén kialakuló eszmecseréket tároló Article talk.

A Wikipedián belül számos automatizált szerkesztői tevékenység zajlik, amelyeket „bot” típusú jogkörrel rendelkező felhasználók kezdeményeznek. Az ilyen, valamint a nem azonosítható, anonim felhasználók a vizsgálatból kizárásra kerülnek.

5.4.2.3 Adatok kinyerése

Az adatok kinyerésének első lépéseként a felesleges kontextusokat szükséges megcímkezni a meghatározott besorolás szerint, ez a kontextusattribútumok között elhelyezhető. Ezt követően kontextus és a tartalom táblák relációja megadja a kívánt kategóriák szerinti közreműködések összességét.

A fentiek szerint szükséges az automatizált szerkesztők kizárása. Az ETL folyamat ajánlása szerint az automatizált műveleteket a felhasználó felhasználócsoportha határozza meg, amely időponttól-időpontig tárolható a felhasználók attribútumai között. Ezek alapján az automatizált szerkesztések (bot, illetve flood felhasználócsoportha) kizárhatók az eredményekből.

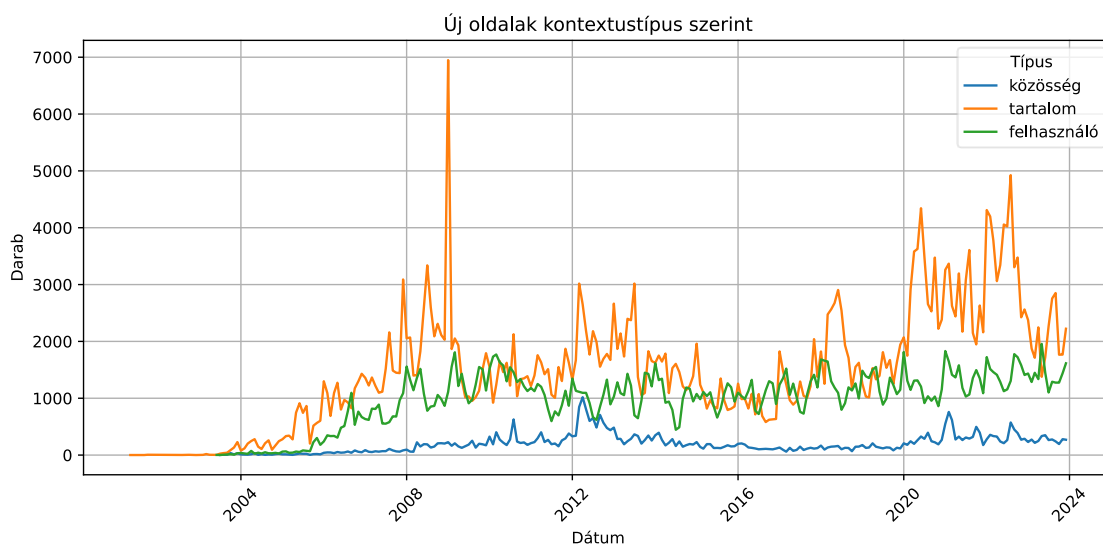
Az előálló adatkészlet a felhasználó azonosítóját, a közreműködés időpontját a kontextustípust tartalmazza.

5.4.2.4 Elemzés

A közösségben a három fő kontextus számossága egyértelműen mutatja a közösség kollaboratív munkavégzésének jellegét: az adatbázis felépítésének időpontjában a 420 ezer tartalmi lapra 51 ezer közösségi és 258 ezer felhasználói lap jut. Tehát a létrehozott lapok 58%-a tartalmi, 35% felhasználói, 7% pedig közösségi kontextust jelent. Érdekes megjegyezni, hogy a közösségi oldalak számának növekedési potenciálja alacsonyabb, mint a felhasználói vagy tartalmi lapokénak.

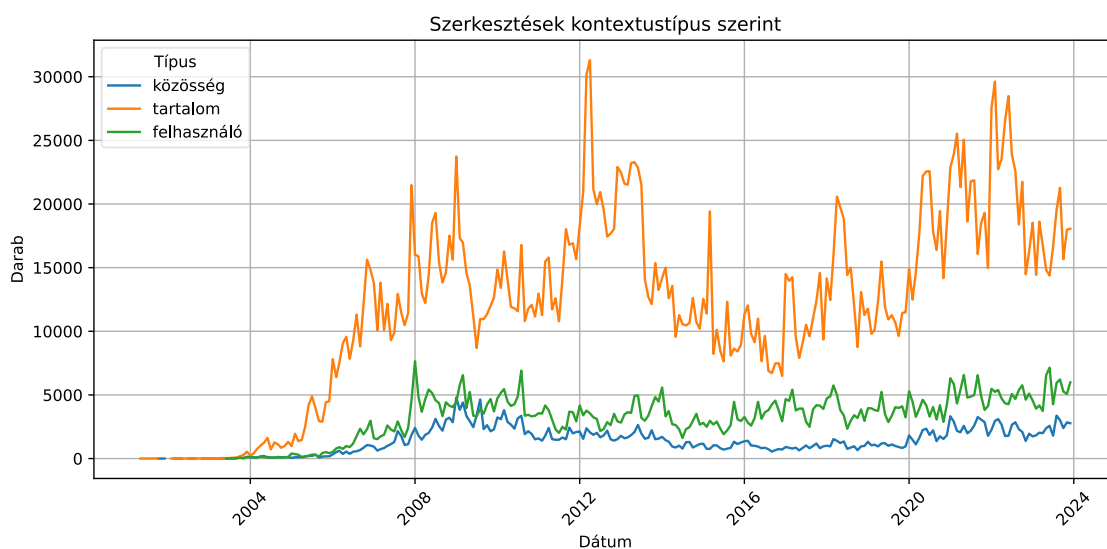
A 31. ábrán a különféle kontextustípusba tartozó oldalak létrejöttének historikus adatai tekinthetők meg. A közösség 2004-es indulása jól látható. A gyors felfutás első ízben 2009 körül akadt el, amely után egészen 2012-ig nem éri el a korábbi termelékenységét, amely a következő években ismét visszaesik, és 2020 környékén kezd ismét emelkedni (nagyjából a Covid járvány időszakával egy időben).

Érdekes megfigyelni, hogy bár a közösségi és felhasználói kontextusban végzett szerkesztések adatsora is mutatja ezt a trendet, de a visszaesés mértéke jóval szolidabb.



31. ábra: Simplewiki új oldalak (Saját szerk.)

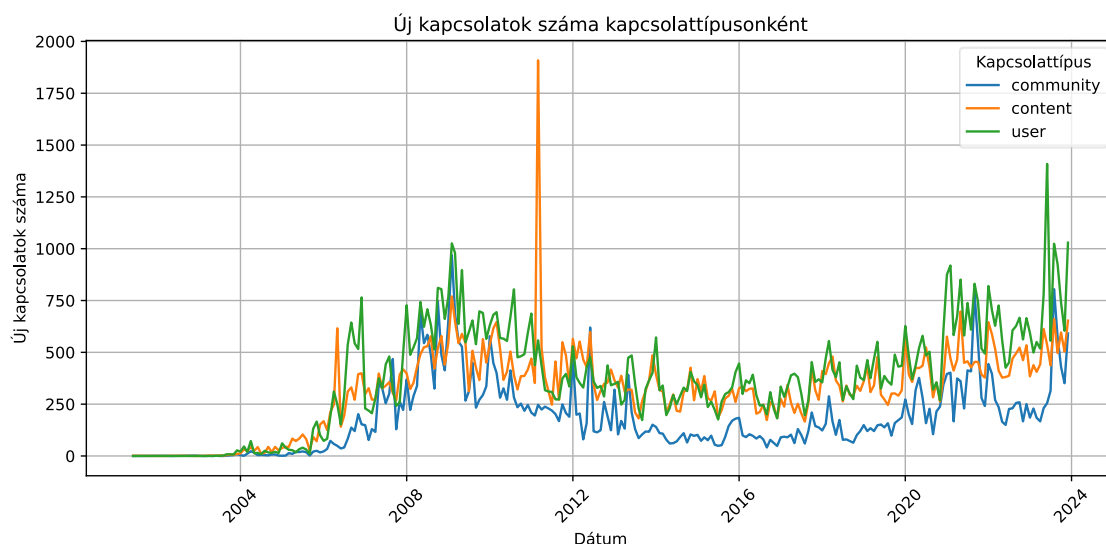
Az adatbázisban mintegy 7.2 millió szerkesztés található az adatbázis építésének időpontjában. Ennek egy jelentős része automatizált szerkesztés, ezeket kiszűrve mintegy 4,5 millió közreműködés származik valódi felhasználótól. Ezek 73%-a tartalmi, mintegy 20%-a felhasználói, és 8%-a közösségi kontextust érint, időbeli alakulásukat a 32. ábra szemlélteti. A trendek az új oldalak szerkesztéséhez hasonlóan alakulnak.



32. ábra: Simplewiki szerkesztések (Saját szerk.)

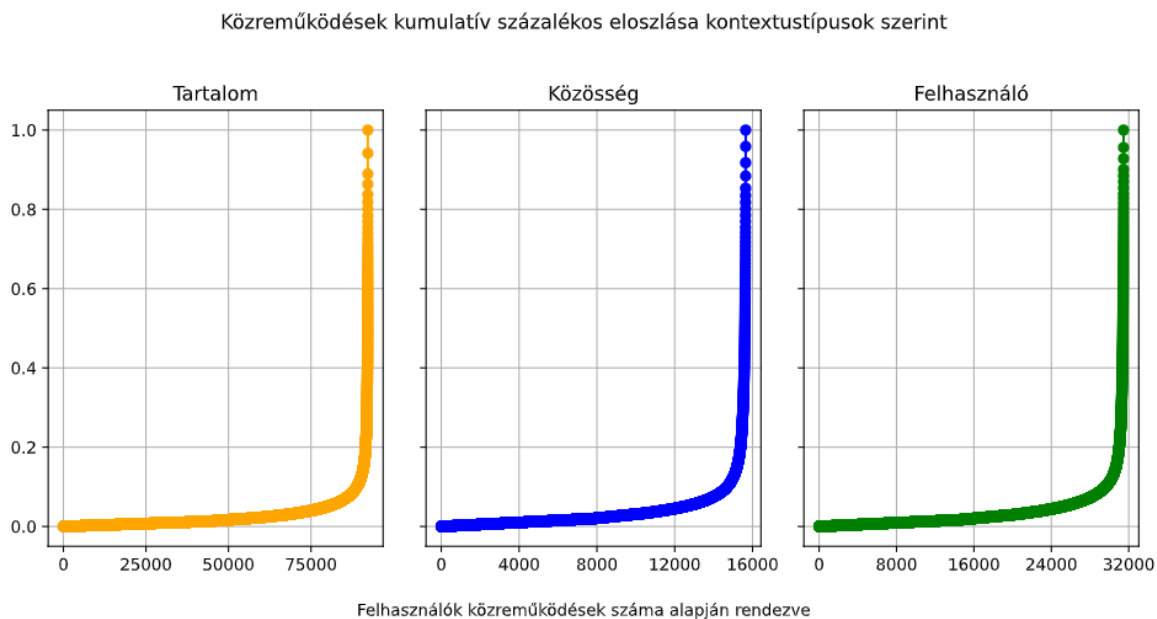
A 33. ábra a különféle kontextustípusokban kialakuló új felhasználói kapcsolatok alakulását mutatja. A korábbi arányoktól eltérően ezek az adatok a különféle kontextustípusokban már kiegyensúlyozottabb összetételt mutatnak. A kialakult mintegy negyed millió kapcsolat 43%-a személyes megszólítás, 36%-a közös szerkesztés, 21%-a

közösségi vita nyomán alakul ki. Érdeemes megjegyezni, hogy a tartalmi területen látható csúcspont az adott Wikipédiában kiemelt figyelemmel övezett fukusimai nukleáris katasztrófához köthető, például a tóhokui földrengés laptörténetében megfigyelhető a számos új felhasználói érkezése, akik így közös szerkesztés okán kapcsolatba kerültek egymással.



33. ábra: Simplewiki felhasználó kapcsolatok fejlődése (Saját szerk.)

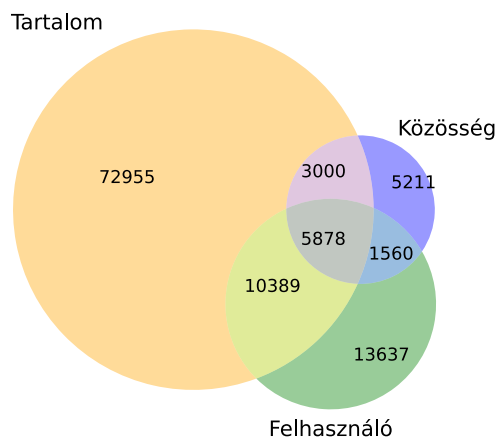
A 34. ábra még inkább rámutat a közösségen belüli viselkedés szélsőségeire. A grafikon a felhasználók szerkesztéseinek kumulatív eloszlását ábrázolja, ahol a felhasználók növekvő szerkesztésszám szerint vannak rendezve. Az eredmények alapján látható, hogy a görbe nagyon laposan indul mindhárom ábrázolt kontextustípus esetén, ami azt jelenti, hogy ezek a felhasználók csak nagyon kis százalékban járulnak hozzá az összes szerkesztéshez. A grafikon mindhárom kontextustípus esetén hirtelen meredek emelkedést mutat az utolsó pár ezer felhasználónál, ami azt jelzi, hogy ez a viszonylag kis csoport adja az összes szerkesztés nagy részét. Ez az úgynevezett „hosszú farok” jelenségre utal, ahol a nagy számú alacsony aktivitású felhasználó jelentős része csak marginális hozzájárulást nyújt, míg a közösség egy kis százaléka dominálja a közösségi aktivitást.



34. ábra: A „hosszú farok” jelenség (Saját szerk.)

Látható, hogy a kollaboratív térben sokkal nagyobb számú felhasználó fordul meg, amely a Wikipedia természetéből fakad: az enciklopédia az olvasók számára is szabadon szerkeszthető, így az olvasás során egy-egy kisebb javítás szinte természetes. A 35. ábra a három különféle kontextustípusban megforduló felhasználók metszetét és számosságát adja.

Felhasználói metszetek a kontextustípusok szerint



35. ábra: Felhasználók kontextustípus szerinti metszete (Saját szerk.)

Azok a felhasználók, akik csak a felhasználói, vagy csak a közösségi kontextusban szerkesztettek, nem adnak hozzá az enciklopédia fő célkitűzéséhez. A sejtés az, hogy a termelékenység egyik legfőbb jelzése, ha egy felhasználó mindhárom kontextustípusban

aktív, azaz betagozódtak a szerkesztői közösségbe. E sejtést megerősíti a 21. táblázatban összefoglalt vizsgálat eredménye.

Felhasználócsoporth	Felhasználók száma	Tartalmi közreműködés	Átlagos közreműködés	Hozzájárulás mértéke
Csak tartalom	72 955	304 462	4	9%
Tartalom és közösségi	3 000	55 058	18	2%
Tartalom és felhasználói	10 389	215 584	21	6%
Mindhárom kontextus	5 878	2 818 082	479	83%
Összesen	92 222	3 393 186	37	100%

21. táblázat: Felhasználói közreműködések beágyazottság szerint (Saját szerk.)

Ennek alapján elmondható, hogy az enciklopédia messze letermékenyebb szerzői azok, akik sikeresen bevonásra kerültek a kollaboratív munkavégzés mellett mind az interperszonális, mind a közösségi kontextusba.

5.4.2.5 Kiértékelés

A Wikipedia szerkesztői közösségének vizsgálatakor jól megfigyelhető a „hosszú farok” jelenség, amely - hasonlóan a korábban említett 90+9+1 megfigyeléshez -, amely azt mutatja, hogy a közösségi aktivitás, és így a kollaboratív munkavégzés egy korlátozott számú, magas szintű tapasztalattal rendelkező felhasználóhoz kapcsolható. Ez a jelenség a tartalom minőségének fenntartásában kívánatos, hiszen amennyiben a tartalom fejlesztését tapasztalt szerkesztők végzik, úgy a végeredmény minősége várhatóan magasabb lesz, és ez a felhasználói kör a közösségi normáknak és szokásoknak is jobban tudatában van. Aggasztó jelenség, hogy a közösségbe újonnan belépő, kevésbé tapasztalt szerkesztők részvétele alacsony. Az évek során koncentrálódó tapasztalt felhasználói körből való lemorzsolódás különösen érzékeny veszteséget jelent, egyszersmind megmegnehezíti az új felhasználók csatlakozását is.

5.4.2.6 Javaslatok

A kiértékelés alapján származtatott következtetések tükrében javaslatok fogalmazhatók meg a közösségi dinamika és a szerkesztői aktivitás ösztönzése céljából. A belépési pont a legtöbb felhasználó számára a tartalmi tér. Az egyetlen szerkesztéssel rendelkező felhasználók között bizonyára magas számban vannak jelen olyan ad-hoc szerkesztők, akiknek későbbi bevonódására a közösség nem feltétlenül számíthat, de kiaknázatlan lehetőségeket is nyújthat ez a felhasználói kör. Különösen érdemes lenne az olyan felhasználók kiemelt kezelése, akik több különböző napon vagy témában szerkesztenek, alacsony, de 1-et meghaladó számban. Célszerű ösztönözni a különféle

kontextustípusokban való szerkesztést, másképpen fogalmazva: megmutatni a Wikipedia közösségi oldalát.

A kiértékelésben megfogalmazott másik cél a kulcsfontosságú felhasználók megtartásának ösztönzése. E kutatás továbbfejlesztéseként érdemes az ebbe a körbe tartozó felhasználók fluktuációnak és motivációs faktorainak feltárása, valamint a lemorzsolódás impaktfaktorának elemzése. Ennek kézenfekvő eszköze a hálózatkutatás, amely választ adhat arra a kérdésre is, hogy mennyire valós jelenség a megfogalmazódott sejtés, ami az aktív szerkesztői közösség klikkesedésére utalhat, amely megnehezíti új felhasználók bevonódását.

6. Konklúzió

A kutatás fő célja az volt, hogy javaslatokat dolgozzon ki virtuális közösségek egységes adatrepresentációjára, ezáltal az adatmegosztási gyakorlatok fejlesztésére. Az alábbiakban összefoglalom a kutatási célokat, kérdéseket és eredményeket.

Kutatási célok

A 3.1.1 számmal jelölt kutatási cél középpontjában **virtuális közösségek viselkedési mintáinak és adatszerkezetének feltárása** állt, amelyet az 1. és a 2. fejezetben tekintettem át szakirodalomkutatás segítségével. Bármely közösséget érint egy tanulmány a társadalomtudományi háttér és összefüggésrendszer meghatározó, így az onnan megismerhető struktúrák és jelenségek hasznos kiegészítést jelentenek a téma kutatásában.

Ezen területek rövid áttekintése után tértem rá a 3.1.1.1 számmal jelölt részletcélra, amely a virtuális közösségek kategorizálására irányul. Kutatásom rávilágított, hogy a virtuális közösségeket a szakirodalomban elsősorban a használat célja szerint tipizálják, és az e téren a múlt században született Rheingold-féle tipológia (Rheingold, 1993) alapjaiban időtállóan bizonyult, ahhoz kiegészítések érkeztek. Egy eltérő tipológia a szabályozottság mértéke, a csatlakozás szabadsága és a nyereségorientáltság alapján csoportosítja a közösségeket (Plant, 2004). A használat céljához lazán kötődő saját csoportosításom a platformtípusok szerint osztályozza a közösségeket, a levelezőlistáktól a közösségi hálózatokig.

A 3.1.1.2 részlete cél fókuszában a felhasználói viselkedés és az azt leíró metrikák álltak. Az 1.1.3 fejezet a virtuális közösségekben betöltött szerep alapján közelíti a kérdést. Ennek kapcsán kutatásom azt mutatja, hogy a legtöbb kutatás a felhasználói interakció és a hozzáadott tartalom alapján alakít ki szerepeket. A fejezetben ismertetett Golder & Donath féle taxonómia (Golder & Donath, 2004) a legtöbb hasonló célú kutatásban megjelenik. A 1.1.4 fejezet a felhasználói életciklus felől közelíti a kérdést, amelyet nézetem szerint tekinthetünk úgy, mint a felhasználói szerepek időbeni vizsgálatát. E téren fontos hozzájárulás a Reader-to-Leader Framework (Preece & Shneiderman, 2011), amelynek középpontjában a felhasználó fejlődése áll. A kapcsolódó munkák fejezetben elemzett mintegy húsz publikációból kigyűjtésre kerültek az alkalmazott metrikák, amelyek kapcsán a közösségkutatásnál gyakran előforduló idő-, tartalmi és hálózati metrikák hármas tagolás mentén világosan elkülöníthetők.

A 3.1.2 számmal jelzett kutatási cél (**virtuális közösségek egységes fogalmi, logikai és fizikai adatmodellezése**) az adatmodellezésre fókuszál. Ennek előkészítéseképpen a szakirodalmi áttekintés részletesen foglalkozik a virtuális közösségek adatgyűjtési és tárolási kérdéseivel. A virtuális közösségek napjainkban hatalmas mennyiségben állítanak elő felhasználó által generált tartalmakat, amelyek ma már egyértelműen az üzleti haszon forrását is jelentik, így nem csak méretük és mennyiségük, hanem hozzáférhetőségük is mind bonyolultabbá válik. E témához fontos adaléknak tartom Bruns ismertetett cikkét, aki arra ösztönzi a kutatókat, hogy a lehető legtöbbet hozzák ki a még hozzáférhető adatokból, fejlesszenek adatgyűjtési és adatmegosztási gyakorlataikon (Bruns, 2019).

Az adatmegosztási gyakorlatok egyik fontos eszköze a generalizált adatmodellek használata, amelyek egyfajta közös nevezőre hozzák az azonos nemű, de eltérő struktúrájú adatbázisokat, ezáltal lehetővé teszik azok standardizált kezelését és így a kutatások megismételhetőségét. Az ilyen adatmodellek és adatbázisok fejlesztéséhez alkalmazott ISO szabványt áttekintettem és a megfelelő egyszerűsítésekkel élve beépítettem azt a kutatómódszertanomba (International Organization for Standardization, 2015).

A tárolandó adatokat a kiszolgáló igények alapjaiban meghatározzák, így 2. fejezetben áttekintettem számos, virtuális közösségeket célzó kutatást, annak érdekében, hogy a tárolandó adatokról, felhasznált adattudományi eszköztárról világos képet kapjak.

Az áttekintett elemzésekből kitűnik, hogy a területen fontos a direkt és indirekt kapcsolati hálók kialakítási lehetősége, a felhasználók és tartalmak és a köztük lévő interakciók minél pontosabb jellemezhetősége, és az időfüggő adattárolás.

Az áttekintett háttér alapján a modellek gyakorlati kivitelezésével az 5.3 fejezet foglalkozik, melynek keretében áttekintettem számos ma elterjedt közösségeket kiszolgáló platform mellett többcélú (azonos architektúrával több helyen alkalmazott) rendszereket, ezekhez fogalmi adatmodellt alkottam, meghatároztam és megfeleltettem az entitásokat, kialakítva így egy közös adatmodell alapjait, amely alkalmas a közösségi adatok tárolására. A háromszintű adatmodellezésben a fogalmi adatmodellezést logikai, majd fizikai modell követi, ezeket relációs platformot alkalmazva alkottam meg. A fogalmi, logikai és fizikai adatmodellek kielégítik a 3.2.1-3.2.3 részletcélakat.

Az elkészített adatbázis próbája állt a 3.1.3. kutatási cél, a **prototípusfejlesztés** középpontjában. A modell prototípusos megvalósítása megközelítésem szerint három elemből áll: ETL folyamat kialakításából, amely egy adott közösségre alkalmazható; annak implementációjából; és a kutathatóság igazolásából. Ehhez a modell kiegészítéseként általánosan alkalmazható iránymutatásokat állítottam össze, majd végrehajtottam egy ETL folyamatot egy MediaWiki adatkészleten. Végül az adathalmazon végeztem egy kutatást, amely megmutatja, hogy az adott kollaboratív közösség nagymértékben koncentrált közösségi struktúrát mutat, amely intő jel lehet az új felhasználók vonzási képessége és a meglévők megtartásának kritikus mivolta okán egyaránt.

Kutatási kérdések

Az alábbiakban röviden összefoglalom a kutatási célokból származtatott kutatási kérdésekre adott válaszokat.

Q1 Milyen adatszerkezettel, funkcionális jellegzetességekkel rendelkeznek a ma nagy elterjedtséggel rendelkező virtuális közösségei?

A ma elterjedt virtuális közösségek esetén szembeűnő, hogy funkcionalitásuk két központ körül csoportosul: a felhasználók hálózata és a tartalom. Az adott közösség típusától, céljától függően ezek a funkciók kisebb vagy nagyobb súllyal jelennek meg az egyes implementációkban. A skála egyik végén a felhasználók vállalt tulajdonságainak, kapcsolati hálójának kinyilvánítására alkalmas közösségi háló típusú funkciókörök

állnak, míg a másik végén a fórumjellegű funkciókör: a tartalom kezelése, visszakereshetősége nagy hangsúlyt kap. A kutatási kérdésre adott részletes választ az 5.3.1 fejezet tárgyalja, amelyhez a kontextust az 1. és a 2. fejezet biztosítja.

Q2 Milyen közös fogalmi adatmodell alkotható virtuális közösségekhez?

Kutatásom szerint a virtuális közösségek legkézenfekvőbb ábrázolási módja a felhasználók, az általuk létrehozott tartalom és a tartalom kontextusa közötti interakciók feltérképezése útján áll elő, amely alkalmas mind az entitások mind az asszociációk terén az azok időfüggő attributált tárolásra. Ez teszi lehetővé a kutatásfüggő adatkinyerést. A kutatási kérdésre adott részletes választ az 5.3.2 fejezet tárgyalja.

Q3 Hogyan lehet a fogalmi adatmodellből származtatott struktúrákat relációs platformon logikai és fizikai adatmodellekké átalakítani?

A kialakított fogalmi modell alkalmas relációs, RDF és property graph alapú adattárolásra egyaránt. A disszertáció a relációs modell alkalmazását tűzte ki maga elé, az implementáció relációs platformon történt. A generalizált modellek esetében a logikai és fizikai leképezés a relációs normalizálás általános szabályai szerint könnyen megtehető, egyetlen megjegyzéssel: míg a gráfadatbázisok szerkezete kifejezetten támogatja a dinamikus attribútumkezelést, addig a relációs modell magas normalizáltsági szinten kötött. Erre a problémára kínál megoldást az általam is alkalmazott EAV generikus modell, amellyel ez a probléma kezelhető. Mivel valódi adatmegosztási gyakorlat is cél, a betöltéshez irányelveket fogalmaztam meg, amely segít a fő entitások és asszociációk azonosításában. A kutatási kérdésre adott részletes választ a 5.3.3 és az 5.3.4 fejezet tárgyalja.

Q4 Hogyan lehet a kialakított adatstruktúrákat és modelleket alkalmazni a felhasználói életutak felismerésére és elemzésére a virtuális közösségekben?

A modell alkalmazásának első lépése az ETL folyamat kivitelezése egy valós közösségen. Ennek demonstrációjaként a disszertáció 5.4.1 fejezete implementál egy ETL folyamatot egy MediaWiki adatbázis-kivonaton, amely újrafuttatható. Mivel a Wikipédiát üzemeltető alapítvány adatbázis-kivonatok útján a teljes enciklopédiát közzéteszi havi rendszerességgel, ez azt jelenti, hogy több száz lehetséges szerkesztői közösség válik így feldolgozhatóvá.

Az adatbázist felhasználva a leggyakoribb elemzési célokat kiszolgáló tárolt lekérdezések készíthetők, melyek képesek adatokkal ellátni az aktuális kutatást. Egyedi attribútumigények és lekérdezési igények esetén a modell rugalmas struktúrája lehetővé teszi annak kiegészítését. A kutatás kivitelezhetőségének demonstrációjaként az 5.4.2 fejezetben ismertetek egy rövid kutatást.

A kutatás hozzáadott értéke

A kiválasztott platformok (Facebook, VBulletin, MantisBT, Reddit, GitHub) szerkezeti összehasonlítása azt mutatta, hogy fő entitásai és kapcsolataik mind visszavezethetők a kialakított közös adatmodell fő komponenseire: a felhasználóra, a kontextusra és a tartalomra. A maximális variációra törekvő minta nagymértékben különböző szerkezetű rendszerekből állt, a kialakított modellt egy, a mintán kívül eső platformra (MediaWiki) is sikerrel alkalmaztam. A feltárt eredmények alapján a felhasználó-közösség-tartalom entitások és kapcsolataik különféle jelentőséggel, de minden közösségi kommunikációra alkalmas rendszerben jelen vannak, és a vizsgált közösségek minden aspektusát lefedik. Így megalapozottan feltételezhető, hogy az általánosított modell sikerrel alkalmazható virtuális közösségek modellezésére.

Így kutatásom legfontosabb hozzáadott értéke, hogy a kifejlesztett adatmodell segítségével a különböző virtuális közösségek adatstruktúráit lehetséges egységesen kezelni, ami lehetővé teszi a standardizált adatfeldolgozást, a kutatások megismételhetőségét. Az ETL prototípusfejlesztés igazolta, hogy kutatásom eredménye alkalmazható valós környezetben. A disszertációban bemutatott megoldás lehetőséget biztosít a virtuális közösségek adatgyűjtési és adatmegosztási gyakorlatának fejlesztéséhez, elősegítve ezzel a tudományos közösség munkáját ezen a területen.

7. Irodalomjegyzék

- Abdi, F., & Abolmakarem, S. (2019). Customer Behavior Mining Framework (CBMF) using clustering and classification techniques. *Journal of Industrial Engineering International*, 15(Suppl 1), 1–18. <https://doi.org/10.1007/s40092-018-0285-3>
- Aggarwal, C. C. (Ed.). (2011). *Social Network Data Analytics*. Springer New York. <https://doi.org/10.1007/978-1-4419-8462-3>
- Aggarwal, C. C., & Wang, H. (2010). *Managing and Mining Graph Data* (C. C. Aggarwal & H. Wang (Eds.)). Springer US. <https://doi.org/10.1007/978-1-4419-6045-0>
- Al-Badayne, D. (2013). Human behaviour - When and where virtual society meets physical society? *European Journal of Science and Theology*, 9(1), 105–110.
- Angles, R. (2018). The property graph database model. *Alberto Mendelzon Workshop on Foundations of Data Management*, 1–10.
- Angles, R., Arenas, M., Barceló, P., Hogan, A., Reutter, J., & Vrgoč, D. (2018). Foundations of Modern Query Languages for Graph Databases. *ACM Computing Surveys*, 50(5), 1–40. <https://doi.org/10.1145/3104031>
- Apache Arrow. (2020). *Feather: A fast, lightweight, and easy-to-use binary file format for storing data frames [Computer software]*. Apache Software Foundation. <https://arrow.apache.org/docs/python/feather.html>
- Atlassian. (n.d.). *JIRA - Issue and Project Tracking System*. Retrieved June 20, 2025, from <https://www.atlassian.com/software/jira>
- Bayer, M. (2012). SQLAlchemy. In A. Brown & G. Wilson (Eds.), *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. <http://aosabook.org>. <http://aosabook.org/en/sqlalchemy.html>
- Baym, N. K. (2015). *Personal Connections in the Digital Age* (2nd ed.). Polity Press.
- Bégin, D., Devillers, R., & Roche, S. (2018). The life cycle of contributors in collaborative online communities - the case of OpenStreetMap. *International Journal of Geographical Information Science*, 32(8), 1611–1630. <https://doi.org/10.1080/13658816.2018.1458312>
- Bergstrom, K., & Poor, N. (2022). Signaling the Intent to Change Online Communities: A Case From a Reddit Gaming Community. *Social Media + Society*, 8(2), 1–10. <https://doi.org/10.1177/20563051221096817>
- Bernardi, M. L., Canfora, G., Di Lucca, G. A., Di Penta, M., & Distant, D. (2018). The relation between developers' communication and fix-Inducing changes: An empirical study. *Journal of Systems and Software*, 140, 111–125. <https://doi.org/10.1016/j.jss.2018.02.065>
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 34–43. <https://doi.org/10.1038/scientificamerican0501-34>
- Beuving, J. J. (2020). Ethnography's future in the big data era. *Information, Communication & Society*, 23(11), 1625–1639. <https://doi.org/10.1080/1369118X.2019.1602664>
- Biggerstaff, T. J. (1989). Design Recovery for Maintenance and Reuse. *Computer*, 22(7), 36–49. <https://doi.org/10.1109/2.30731>
- Blanchard, A. (2008). Sense of Virtual Community. In S. Kelsey & K. St. Amant (Eds.), *Handbook of Research on Computer Mediated Communication* (pp. 325–338). IGI Global. <https://doi.org/10.4018/978-1-59904-863-5.ch025>
- Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230. <https://doi.org/10.1111/j.1083-6101.2007.00393.x>

- Bruns, A. (2019). After the ‘APIcalypse’: social media platforms and their fight against critical scholarly research. *Information Communication and Society*, 22(11), 1544–1566. <https://doi.org/10.1080/1369118X.2019.1637447>
- BuiltWith. (2024). *Forum Software Usage Distribution on the Entire Internet*. Builtwith.Com. <https://trends.builtwith.com/cms/forum-software/traffic/Entire-Internet>
- Burger, A., Oz, T., Kennedy, W. G., & Crooks, A. T. (2019). Computational Social Science of Disasters : Opportunities and Challenges. *Future Internet*, 11(5), 1–31.
- Burrows, R., & Savage, M. (2009). Some further reflections on the Coming Crisis of Empirical Sociology. *Sociology*, 43(4), 762–772. <https://doi.org/10.1177/0038038509105420>
- Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, 1(1), 1–6. <https://doi.org/10.1177/2053951714540280>
- Burt, R. S. (1982). *Toward a Structural Theory of Action*. Elsevier. <https://doi.org/10.1016/C2009-0-21953-2>
- Chan, J., Hayes, C., & Daly, E. (2010). Decomposing Discussion Forums and Boards Using User Roles. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1), 215–218. <https://doi.org/10.1609/icwsm.v4i1.14063>
- Chen, P. P. S. (1976). The Entity-Relationship Model—toward a Unified View of Data. *ACM Transactions on Database Systems (TODS)*, 1(1), 9–36. <https://doi.org/10.1145/320434.320440>
- Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6), 377–387. <https://doi.org/10.1145/362384.362685>
- ConceptDraw. (n.d.). *Entity Relationship Diagram Symbols*. ConceptDraw. Retrieved June 30, 2025, from <https://www.conceptdraw.com/How-To-Guide/erd-entity-relationship-diagram-symbols>
- Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). SAGE Publications.
- Cui, J., Wang, Z., Ho, S., & Cambria, E. (2023). Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, 56(8), 8469–8510. <https://doi.org/10.1007/s10462-022-10386-z>
- Danese, M. D., Halperin, M., Duryea, J., & Duryea, R. (2019). The Generalized Data Model for clinical research. *BMC Medical Informatics and Decision Making*, 19(1), 1–13. <https://doi.org/10.1186/s12911-019-0837-5>
- Davidson, B. I., Jones, S. L., Joinson, A. N., & Hinds, J. (2019). The evolution of online ideological communities. *PLOS ONE*, 14(5), 1–25. <https://doi.org/10.1371/journal.pone.0216932>
- Desai, R. J., Marsolo, K., Smith, J., Carrell, D., Penfold, R., Pillai, H. S., Lii, J., Ngan, K., Winter, R., Adgent, M., Ramaprasan, A., Driscoll, M. R., Scarnecchia, D., Kiernan, D., Draper, C., Lyons, J. G., Khurshid, A., Maro, J. C., Zimmerman, R., ... Schneeweiss, S. (2024). The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). *Pharmacoepidemiology and Drug Safety*, 33(10), 1–12. <https://doi.org/10.1002/pds.70028>
- Dessewffy, T., & Láng, L. (2015). Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. *Replika*, 93(3–4), 157–170. http://www.replika.hu/system/files/archivum/92-93_11_dessewffy_lang.pdf
- Durkheim, É. (1895). A szociológia módszertani szabályai. In P. Léderer (Ed.), *A*

- társadalmi tények magyarázatához – válogatott tanulmányok (pp. 21–159). Közgazdasági és Jogi Könyvkiadó.
- Dybka, P. (n.d.). *Chen Notation*. Vertabelo. Retrieved June 30, 2025, from <https://vertabelo.com/blog/chen-erd-notation/>
- Edge Delta Team. (2024). *Breaking Down The Numbers: How Much Data Does The World Create Daily in 2024?* Edge Delta. <https://edgedelta.com/company/blog/how-much-data-is-created-per-day>
- Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of Management Review*, 14(4), 532–550.
- Elmasri, R., & Navathe, S. (2016). *Fundamentals of Database Systems* (7th ed.). Pearson.
- Farkas, Z. (2010). *Társadalomelmélet*. Bíbor Kiadó.
- Fergusson, K. (2008). *Entity Relationship Diagrams (ERDs) with draw.io*. Draw.Io. <https://drawio-app.com/blog/entity-relationship-diagrams-with-draw-io/>
- Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, 24(4), 230–235. <https://doi.org/10.1179/2047480615z.0000000000329>
- Fournier-Viger, P., He, G., Cheng, C., Li, J., Zhou, M., Lin, J. C., & Yun, U. (2020). A survey of pattern mining in dynamic graphs. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6), 1–30. <https://doi.org/10.1002/widm.1372>
- Fu, X., Chen, X., Shi, Y. T., Bose, I., & Cai, S. (2017). User segmentation for retention management in online social games. *Decision Support Systems*, 101, 51–68. <https://doi.org/10.1016/j.dss.2017.05.015>
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>
- Garza, M., Del Fiol, G., Tenenbaum, J., Walden, A., & Zozus, M. N. (2016). Evaluating common data models for use with a longitudinal community registry. *Journal of Biomedical Informatics*, 64, 333–341. <https://doi.org/10.1016/j.jbi.2016.10.016>
- Ghani, N. A., Hamid, S., Targio Hashem, I. A., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior*, 101, 417–428. <https://doi.org/10.1016/j.chb.2018.08.039>
- Gillespie, T. (2010). The politics of “platforms.” *New Media and Society*, 12(3), 347–364. <https://doi.org/10.1177/1461444809342738>
- GitHub. (n.d.). *GitHub Docs*. Retrieved July 23, 2025, from <https://docs.github.com/en>
- Golbeck, J. (2007). The dynamics of Web-based social networks: Membership, relationships, and change. *First Monday*, 12(11), 1–18. <https://doi.org/10.5210/fm.v12i11.2023>
- Golbeck, J., Robles, C., & Turner, K. (2011). Predicting Personality with Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 253–262. <http://mendeley.csuc.cat/fitxers/2d2e1ca56485e2ca0da32511da6ea353>
- Golder, S. A., & Donath, J. (2004). Social Roles in Electronic Communities. *Aoir, January 2004*, 1–25. papers2://publication/uuid/A44A65B0-3B94-4809-A728-DAB3D4674856
- Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26(2), 91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>
- Hagel, J. (1999). Net gain: Expanding markets through virtual communities. *Journal of Interactive Marketing*, 13(1), 55–65. [https://doi.org/10.1002/\(SICI\)1520-6653\(199924\)13:1<55::AID-DIR5>3.0.CO;2-C](https://doi.org/10.1002/(SICI)1520-6653(199924)13:1<55::AID-DIR5>3.0.CO;2-C)
- Halford, S., & Pope, C. (2014). *Big Data : Methodological Challenges and Approaches*

- for Sociological Analysis. *Sociology*, 48(4), 663–681.
<https://doi.org/10.1177/0038038513511561>
- Halford, S., & Savage, M. (2017). Speaking Sociologically with Big Data : Symphonic Social Science and the Future for Big Data Research. *Sociology*, 51(6), 1132–1148.
<https://doi.org/10.1177/0038038517698639>
- Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19(2), 87–92.
- Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems* (Vol. 22). Springer US. <https://doi.org/10.1007/978-1-4419-5653-8>
- Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly: Management Information Systems*, 28(1), 75–105.
<https://doi.org/10.2307/25148625>
- Hohpe, G., & Woolf, B. (2003). *Enterprise integration patterns: Designing, building, and deploying messaging solutions*. Addison-Wesley Educational.
- Hollweck, T., & Yin, R. K. (2016). Case Study Research Design and Methods. *The Canadian Journal of Program Evaluation*, 30(1), 108–110.
<https://doi.org/10.3138/CJPE.BR-240>
- Horváth, C. (2020). A közösség értelmezése, jellegzetességei és területi vonatkozásai. *Köztes-Európa*, 2(28), 103–125.
- Horváth, I. (2007). Comparison of Three Methodological Approaches of Design Research. *International Conference on Engineering Design*, 1–11.
- Ianni, M., Masciari, E., & Sperlí, G. (2021). A survey of Big Data dimensions vs Social Networks analysis. *Journal of Intelligent Information Systems*, 57(1), 73–100.
<https://doi.org/10.1007/s10844-020-00629-2>
- Iivari, J., & Venable, J. (2009). Action research and design science research - Seemingly similar but decisively dissimilar. *European Conference on Information Systems*, 1–12.
- International Organization for Standardization. (2015). *Information technology – Metadata registries (MDR)* (ISO Standard No. 11179:2015).
<https://web.archive.org/web/20160406042038/http://metadata-standards.org/11179/>
- Iriberry, A., & Leroy, G. (2009). A life-cycle perspective on online community success. *ACM Computing Surveys*, 41(2), 1–29. <https://doi.org/10.1145/1459352.1459356>
- Ishikawa, H. (2015). Social Big Data Mining. In *Social Big Data Mining*. CRC Press.
<https://doi.org/10.1201/b18223>
- Kalinichenko, L. A. (2004). Canonical model development Techniques aimed at semantic interoperability in the heterogeneous world of information modeling. *Knowledge and Model Driven Information Systems Engineering for Networked Organizations: Proc. of the CAiSE INTEROP Workshop*, 101–116.
- Kazienko, P., & Chawla, N. (Eds.). (2015). *Applications of Social Media and Social Network Analysis*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-19003-7>
- KhakAbi, S., Gholamian, M. R., & Namvar, M. (2010). Data Mining Applications in Customer Churn Management. *2010 International Conference on Intelligent Systems, Modelling and Simulation*, 220–225.
<https://doi.org/10.1109/ISMS.2010.49>
- Kimball, R., & Ross, M. (2015). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). John Wiley & Sons.
- Knoke, D., & Kuklinski, J. H. (1982). *Network analysis*. Sage Publications.

- Lee, F. S. L., Vogel, D., & Limayem, M. (2003). Virtual community informatics: A review and research agenda. *Journal of Information Technology Theory and Application*, 5(1), 47–61.
- Lucidchart. (n.d.). *Entity-Relationship Diagram Symbols and Notation*. Retrieved June 30, 2025, from <https://www.lucidchart.com/pages/ER-diagram-symbols-and-meaning>
- Magyar, M. (2020). *Perzisztencia és lemorzsolódás - Online motivációs tényezők vizsgálata adatbányászati eszközökkel a Wikipédia szerkesztői közösségében* [Corvinus University of Budapest]. http://szd.lib.uni-corvinus.hu/13512/1/Magyar_Melinda.pdf
- Magyar, M. (2025). *Kódmelléklet [Kísérő anyag a "Felhasználói interakciók és közösségi dinamikák elemzése és modellezése a virtuális térben" című doktori disszertációhoz]*. GitHub. <https://github.com/melindamagyarcorvinus/community>
- Magyar, M., & Burka, D. (2025). Using issue tracking as a groupwork facilitator in education. *Infocommunications Journal*, 17(Special Issue), 42–52. <https://doi.org/10.36244/ICJ.2025.5.6>
- Magyar, M., Kovács, L., & Burka, D. (2021). Forecasting the Spread of the COVID-19 Pandemic Based on the Communication of Coronavirus Sceptics. *The 7th International Conference on Time Series and Forecasting*, 1–10. <https://doi.org/10.3390/engproc2021005035>
- MantisBT. (n.d.). *Mantis Bug Tracker*. Retrieved June 20, 2025, from <https://mantisbt.org/>
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.
- McAuley, J., & Leskovec, J. (2013). From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, 897–907. <https://doi.org/10.48550/arXiv.1303.4402>
- McFarland, D. A., Lewis, K., & Goldberg, A. (2016). Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *The American Sociologist*, 47(1), 12–35. <https://doi.org/10.1007/s12108-015-9291-8>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. <https://doi.org/10.25080/Majora-92bf1922-00a>
- McNeill, L., & Zuern, J. D. (2015). Online Lives 2.0: Introduction. *Biography*, 38(2), v–xlvi. <https://doi.org/10.1353/bio.2015.0012>
- MediaWiki. (2012). *Manual:What is MediaWiki?* Mediawiki.Org. https://www.mediawiki.org/wiki/Manual:What_is_MediaWiki
- MediaWiki. (2024a). *API*. Mediawiki.Org. https://www.mediawiki.org/wiki/API:Main_page
- MediaWiki. (2024b). *Extension:DataDump*. Mediawiki.Org. <https://www.mediawiki.org/wiki/Extension:DataDump>
- Microsoft. (2022). *Common Data Model*. Microsoft Learn. <https://learn.microsoft.com/en-us/common-data-model/>
- Miller, T., Paloque-Bergès, C., & Dame-Griff, A. (2023). Remembering Netizens: an interview with Ronda Hauben, co-author of Netizens: on the history and impact of Usenet and the Internet (1997). *Internet Histories*, 7(1), 76–98. <https://doi.org/10.1080/24701475.2022.2123120>

- Moertini, V. S., & Adithia, M. T. (2021). Uncovering Active Communities from Directed Graphs on Distributed Spark Frameworks, Case Study: Twitter Data. *Big Data and Cognitive Computing*, 5(4), 46. <https://doi.org/10.3390/bdcc5040046>
- Moudene, Y. El, Idrais, J., El Abassi, R., & Sabour, A. (2024). Gender-Based Analysis of User Reactions to Facebook Posts. *Big Data Mining and Analytics*, 7(1), 75–86. <https://doi.org/10.26599/BDMA.2023.9020005>
- Mützel, S. (2015). Facing Big Data: Making sociology relevant. *Big Data & Society*, 2(2), 1–4. <https://doi.org/10.1177/2053951715599179>
- Myers, E. W. (1986). An O(ND) difference algorithm and its variations. *Algorithmica*, 1(1–4), 251–266. <https://doi.org/10.1007/BF01840446>
- Nadkarni, P. M., & Brandt, C. (1998). Data extraction and ad hoc query of an entity-attribute-value database. *Journal of the American Medical Informatics Association*, 5(6), 511–527. <https://doi.org/10.1136/jamia.1998.0050511>
- Narula, U. (2006). *Handbook of communication: Models, perspectives, strategies*. Atlantic Publishers & Distributors.
- Newcomb, T. M. (1953). An approach to the study of communicative acts. *Psychological Review*, 60(6), 393–404. <https://doi.org/10.1037/h0063098>
- Nielsen, J. (2006). *The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities*. Nielsen Norman Group. <https://www.nngroup.com/articles/participation-inequality/>
- Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009). Outline of a design science research process. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 1–11. <https://doi.org/10.1145/1555619.1555629>
- Ogden, J. (2022). “Everything on the internet can be saved”: Archive Team, Tumblr and the cultural significance of web archiving. *Internet Histories*, 6(1–2), 113–132. <https://doi.org/10.1080/24701475.2021.1985835>
- OHDSI CDM Working Group. (2024). *OMOP Common Data Model*. OMOP Common Data Model. <https://ohdsi.github.io/CommonDataModel/>
- Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data*, 4(1), 1–19. <https://doi.org/10.1186/s40537-017-0063-x>
- Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Community dynamics in social networks. *Noise and Stochastics in Complex Systems and Finance*, 6601(September), 660106. <https://doi.org/10.1117/12.724517>
- Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods*. Sage Publications.
- Pfeifer, P. E. (2005). The optimal ratio of acquisition and retention costs. *Journal of Targeting, Measurement and Analysis for Marketing*, 13(2), 179–188. <https://doi.org/10.1057/palgrave.jt.5740142>
- Phillips, S. (2007). A brief history of Facebook. *The Guardian*, 553–592. <https://www.theguardian.com/technology/2007/jul/25/media.newmedia>
- Plant, R. (2004). Online communities. *Technology in Society*, 26(1), 51–65. <https://doi.org/10.1016/j.techsoc.2003.10.005>
- Preece, J. (2001). Sociability and usability in online communities: Determining and measuring success. *Behaviour and Information Technology*, 20(5), 347–356. <https://doi.org/10.1080/01449290110084683>
- Preece, J., Maloney-Krichmar, D., & Abras, C. (2003). History of online communities. In K. Christensen & D. Levinson (Eds.), *Encyclopedia of Community: From Village to Virtual World*. (pp. 1023–1027). SAGE Publications, Inc.

- <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.3155>
- Preece, J., & Shneiderman, B. (2011). The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction*, 3(1), 1–25.
- Qin, X., Cunningham, P., & Salter-Townshend, M. (2016). Online trans-dimensional von mises-fisher mixture models for user profiles. *Journal of Machine Learning Research*, 17(200), 1–51.
- Qin, X., Greene, D., & Cunningham, P. (2016). A latent space analysis of editor lifecycles in wikipedia. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9546, 46–69. https://doi.org/10.1007/978-3-319-29009-6_3
- Rafaeli, S., & Jones, Q. (2000). Time to Split, Virtually: “Discourse Architecture” and “Community Building” Create Vibrant Virtual Publics. *Electronic Markets*, 10(4), 214–223. <https://doi.org/10.1080/101967800750050326>
- Reddit. (n.d.). *Reddit - The Heart of the Internet*. Retrieved June 20, 2025, from <https://www.reddit.com/>
- Rheingold, H. (1993). *The Virtual Community. Homesteading in the Electronic Frontier*. Addison-Wesley Publishing Company. <https://archive.org/details/virtualcommunity00rhei>
- Rowe, M. (2013). Changing with time: Modelling and detecting user lifecycle periods in online community platforms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8238 LNCS, 30–39. https://doi.org/10.1007/978-3-319-03260-3_3
- Rowe, M. (2014). Predicting Online Community Churners Using Gaussian Sequences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8851, pp. 66–79). https://doi.org/10.1007/978-3-319-13734-6_5
- Rowe, M. (2016). Mining User Development Signals for Online Community Churner Detection. *ACM Transactions on Knowledge Discovery from Data*, 10(3), 1–28. <https://doi.org/10.1145/2798730>
- Rowe, M. (2009). Interlinking distributed social graphs. *Proceedings of the Linked Data on the Web Workshop (LDOW2009)*, 1–8.
- Rowe, M., & Alani, H. (2012). What Makes Communities Tick? Community Health Analysis Using Role Compositions. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, 267–276. <https://doi.org/10.1109/SocialCom-PASSAT.2012.18>
- Rowe, M., Fernandez, M., Angeletou, S., & Alani, H. (2013). Community analysis through semantic rules and role composition derivation. *Journal of Web Semantics*, 18(1), 31–47. <https://doi.org/10.1016/j.websem.2012.05.002>
- Runeson, P., Höst, M., Rainer, A., & Regnell, B. (2012). *Case Study Research in Software Engineering*. Wiley. <https://doi.org/10.1002/9781118181034>
- Sadeque, F., & Bethard, S. (2019). *Predicting engagement in online social networks: Challenges and opportunities*. 85721, 1–13. <http://arxiv.org/abs/1907.05442>
- Safari, R. M., Rahmani, A. M., & Alizadeh, S. H. (2019). User behavior mining on social media: a systematic literature review. *Multimedia Tools and Applications*, 78(23), 33747–33804. <https://doi.org/10.1007/s11042-019-08046-6>
- Savage, M., & Burrows, R. (2007). The Coming Crisis of Empirical Sociology. *Sociology*, 41(5), 885–899. <https://doi.org/10.1177/0038038507080443>
- Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying

- CRISP-DM process model. *Procedia Computer Science*, 181(2019), 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>
- Schroer, J., & Hertel, G. (2009). Voluntary Engagement in an Open Web-Based Encyclopedia: Wikipedians and Why They Do It. *Media Psychology*, 12(1), 96–120. <https://doi.org/10.1080/15213260802669466>
- Shayo, C., Olfman, L., Iriberry, A., & Igbaria, M. (2007). The Virtual Society. In J. Gackenbah (Ed.), *Psychology and the Internet* (pp. 187–219). Elsevier. <https://doi.org/10.1016/B978-012369425-6/50027-8>
- Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 5(4), 13–22.
- Sheng, Z., Zhang, T., & Zhang, Y. (2021). Neurocomputing HTDA : Hierarchical time-based directional attention network for sequential user behavior modeling. *Neurocomputing*, 441, 323–334. <https://doi.org/10.1016/j.neucom.2021.02.006>
- Similarweb. (2024). *Top Websites*. <https://www.similarweb.com/top-websites/>
- Simple English Wikipedia. (2024a). *Statistics*. <https://simple.wikipedia.org/wiki/Special:Statistics>
- Simple English Wikipedia. (2024b). *Talk:Hungary*. Simple English Wikipedia. <https://simple.wikipedia.org/wiki/Talk:Hungary>
- Song, I.-Y., Evans, M., & Park, U. E. K. (1995). A Comparative Analysis of Entity-Relationship Diagrams. *Journal of Computer and Software Engineering*, 3(4), 427–459.
- Stephens-Davidowitz, S. (2019). *Mindenki hazudik – Az vagy, amire kattintasz*. Athenaeum.
- Sun, L. H., & Fichman, P. (2020). The collective trolling lifecycle. *Journal of the Association for Information Science and Technology*, 71(7), 770–783. <https://doi.org/10.1002/asi.24296>
- Takács, K. (2011). *Társadalmi kapcsolathálózatok elemzése*. Budapesti Corvinus Egyetem. https://regi.tankonyvtar.hu/hu/tartalom/tamop425/0010_2A_08_Kapcsolathalo_elemez_szerk_Takacs_Karoly/ch01.html#link_1
- Tarde, G. (1899). *Social Laws. An Outline of Sociology* (H. C. Warren, trans.). The Macmillan Company.
- Tilson, L. V., Excell, P. S., & Green, R. J. (1988). A generalisation of the Fuzzy c-Means clustering algorithm. *Remote Sensing. Proc. IGARSS '88 Symposium, Edinburgh, 1988. Vol. 3, 10(2)*, 1783–1784. <https://doi.org/10.1109/igarss.1988.569600>
- Tönnies, F. (1983). *Közösség és Társadalom* (B. Gábor (Ed.)). Gondolat Könyvkiadó.
- Turner, T. C., Smith, M. A., Fisher, D., & Welser, H. T. (2005). Picturing Usenet: Mapping Computer-Mediated Collective Action. *Journal of Computer-Mediated Communication*, 10(4), 1–37. <https://doi.org/10.1111/j.1083-6101.2005.tb00270.x>
- Van Aken, J. E. (2004). Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules. *Journal of Management Studies*, 41(2), 219–246. <https://doi.org/10.1111/j.1467-6486.2004.00430.x>
- Van Mierlo, T. (2014). The 1% Rule in Four Digital Health Social Networks: An Observational Study. *Journal of Medical Internet Research*, 16(2), 1–9. <https://doi.org/10.2196/jmir.2966>
- Velasquez, A., Wash, R., Lampe, C., & Bjornrud, T. (2014). Latent users in an online user-generated content community. *Computer Supported Cooperative Work: CSCW: An International Journal*, 23(1), 21–50. <https://doi.org/10.1007/s10606->

- Venditti, B. (2024). *Ranked: The Most Visited Websites in 2024*. Visual Capitalist. <https://www.visualcapitalist.com/ranked-the-most-visited-websites-in-2024/>
- Vercseg, I. (2018). *Közösségelmélet*. Közösségfejlesztők Egyesülete Választmánya.
- Wasko, M. M. L., Teigland, R., & Faraj, S. (2009). The provision of online public goods: Examining social structure in an electronic network of practice. *Decision Support Systems*, 47(3), 254–265. <https://doi.org/10.1016/j.dss.2009.02.012>
- Weeks, J., & Pardee, R. (2019). Learning to Share Health Care Data: A Brief Timeline of Influential Common Data Models and Distributed Health Data Networks in U.S. Health Care Research. *EGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, 7(1), 4. <https://doi.org/10.5334/egems.279>
- Wellman, B. (1997). An Electronic Group Is Virtually a Social Network. In S. Kiesler (Ed.), *Culture of the Internet* (pp. 179–205). Hillsdale. <https://doi.org/10.4324/9781315806389-17>
- Wellman, B. (2005). Physical place and cyberplace: The rise of networked individualism. In D. Eagle, B. Hague, L. Keeble, & B. D. Loader (Eds.), *Community Informatics: Shaping Computer-Mediated Social Networks* (pp. 17–42). Taylor & Francis. <https://doi.org/10.4324/9780203996683-7>
- Wellman, B., & Gulia, M. (1999). Virtual communities as communities: Net surfers don't ride alone. In P. Kollock & M. Smith (Eds.), *Communities in Cyberspace* (pp. 167–194). Routledge.
- Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. *Proceedings of the 2011 IConference*, 122–129. <https://doi.org/10.1145/1940761.1940778>
- West, M. (2011). Developing High Quality Data Models. In *Developing High Quality Data Models*. Elsevier. <https://doi.org/10.1016/C2009-0-30508-5>
- Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-43839-8>
- Wikimedia. (2024a). *Wikimedia Downloads*. Dumps.Wikimedia.Org. <https://dumps.wikimedia.org/>
- Wikimedia. (2024b). *WikiStats*. Wikistats 2.2. <https://wikistats.wmcloud.org/>
- Wikimedia API portal. (2023). *Rate limits*. https://api.wikimedia.org/wiki/Rate_limits
- World Wide Web Consortium (W3C). (2014). *Resource Description Framework (RDF): Concepts and Abstract Syntax*. <https://www.w3.org/TR/rdf-concepts/>
- Yin, R. K. (2003). *Case Study Research Design and Methods* (3rd ed.). Sage Publications.
- Zablith, F., Fernandez, M., & Rowe, M. (2015). Production and consumption of university Linked Data. *Interactive Learning Environments*, 23(1), 55–78. <https://doi.org/10.1080/10494820.2012.745428>
- Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social Media Mining. In *Social Media Mining: An Introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139088510>

8. Függelék

8.1 Chen-notáció

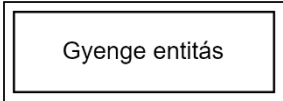
A Chen-féle notáció az egyik legismertebb és leggyakrabban alkalmazott módszer az adatmodellezés területén. A jelölésrendszert Peter Chen vezette be az 1970-es években, hogy az adatbázisok logikai szerkezetét egyértelműen és áttekinthetően lehessen ábrázolni (Chen, 1976). A notáció vizuálisan elkülöníti az entitásokat, azok attribútumait, valamint az entitások közötti kapcsolatokat, ezzel segítve a fejlesztőket és a tervezőket abban, hogy pontosan megértsék az információs rendszer adatstruktúráját, és könnyebben kommunikáljanak az üzleti felhasználókkal is: a jelölésrendszer egyszerű elemekkel dolgozik, amelyek megértéséhez nem szükséges mély informatikai jártasság, így különösen alkalmas fogalmi adatmodellek készítéséhez és értékeléséhez.

8.1.1 A notációs rendszer és kiegészítései

A notáció az entitásokat téglalappal, a köztük lévő kapcsolatokat rombusszal, míg az attribútumokat ellipszisekkel jelöli. Az elemi összetevőket vonalak kötik össze, a kapcsolatok kardinalitása az entitásokat és kapcsolatokat összekötő vonalakon jelölhető.

Amennyiben az aktuális modellezési cél ezt megköveteli, egyéb tulajdonságokat (pl. többértékű attribútumok, gyenge entitások, részleges vagy teljes részvétel, opcionális kapcsolatok stb.) módosított (pl. megkettőzött, szaggatott) körvonalakkal, illetve tipográfiai kiemeléssel jelezhetünk.

A 22. táblázat az entitások, a 23. táblázat a kapcsolatok jelöléséhez használatos ábrákat tartalmazza.

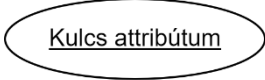
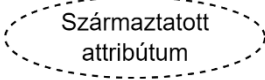
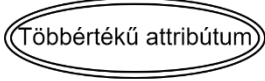
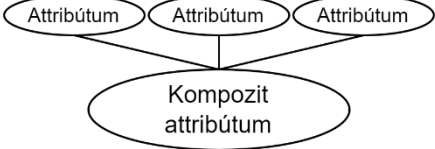
Ábra	Jelentés
	Entitás
	A gyenge entitás olyan entitás, amely pusztán attribútumai alapján nem azonosítható, ahhoz felettes (tulajdonos) entitás szükséges.
	Az eredeti Chen notációban nem szereplő, de széles körben elterjedt jelölésmód a több-a-többhöz típusú kapcsolatokat tároló táblák számára, különösen abban az esetben, ha ezek a kapcsolatok önálló attribútumokkal vagy relációkkal rendelkeznek (Chen, 1976; ConceptDraw, n.d.; Dybka, n.d.; Fergusson, 2008; Lucidchart, n.d.; Song et al., 1995).

22. táblázat: Chen notáció – entitások (Saját szerk.)

Ábra	Jelentés
	Két entitás között fennálló kapcsolat.
	Gyenge kapcsolat: gyenge entitások és felettes tulajdonos entitásuk között. Ez a kapcsolat valójában a gyenge entitás azonosítására szolgál, ezért azonosító kapcsolatnak is szokás nevezni.

23. táblázat: Chen notáció – kapcsolatok (Saját szerk.)

Amennyiben az aktuális modellezési cél ezt megköveteli, az entitásokat, illetve a kapcsolatokat attribútumokkal láthatjuk el a 24. táblázatban szereplő elemek felhasználásával.

Ábra	Jelentés
	Attribútum
	Kulcs attribútum
	Részleges kulcs attribútum: gyenge entitások részleges kulcsa.
	Származtatott attribútum: más attribútumok értékéből származó érték, amely a végleges logikai fizika modellben nem feltétlenül fog adatbázismezőként megjelenni.
	Többértékű attribútum: olyan attribútum, amely több értéket is felvehet, normalizáltsági szinttől függően önálló táblát is jelenthet.
	Kompozit attribútum: jelöléstechnikai eszköz, amely a tipikus attribútumszetteket csoportosítja.

24. táblázat: Chen-notáció – attribútumok (Saját szerk.)

8.1.2 A disszertációban alkalmazott jelölésrendszer

A disszertációban számos ábra alkalmaz Chen-féle notációra épülő jelölésrendszert. Az entitásokat téglalappal az 1:N, 1:1 kapcsolatokat rombusszal, míg az M:N kapcsolatokat asszociatív entitással jelölöm. A fogalmi adatmodelleket ábrázoló ábrákon a legtöbb esetben nem kerülnek feltüntetésre az attribútumok, egyrészt azok számos rendszer esetében azok dinamikus mivolta, másrészt az ábrák áttekinthetőségére való törekvés miatt. Az ábrákon a kapcsolatok opcionáltságát hasonló megfontolásból nem ábrázolom.

Az asszociatív entitások jelölésére alkalmazott elem felhasználási gyakorlata eltérő (Song et al., 1995), és, mivel az eredeti Chen-notációban ez az entitástípus nem szerepel, különböző útmutatók különböző ajánlásokat fogalmaznak meg:

- vagy minden M:N kapcsolat ilyen ábrával jelölését javasolják, mely gyakorlat elősegíti a relációs platformra történő logikai transzformáció során a táblák azonosítását (ConceptDraw, n.d.; Fergusson, 2008),

- vagy azt, hogy a saját attribútumokkal rendelkező kapcsolótáblák esetén alkalmazzuk (Lucidchart, n.d.) – habár a Chen-purista megközelítés sem tiltja azt, hogy a kapcsolatokhoz attribútumokat csatoljunk (Chen, 1976, p. 12),
- vagy akkor, ha az asszociatív entitás maga is részt vesz kapcsolatban. A Chen-notáció előírja, hogy az entitások rombusz beiktatásával kapcsolódjanak egymáshoz (Chen, 1976, p. 30), és az M:N kapcsolatokat is rombuszsal jelöli, így, ha az asszociatív entitás további kapcsolatban vesz részt, akkor az sok esetben szükségessé teszi a kapcsolatok felbontását akár 1:N szintre az egyértelmű ábrák alkotásához, amely azokat szükségképpen elbonyolítja.

A disszertációban minden M:N kapcsolatot asszociatív entitásként jelölök.