

Régi kutatások veszélyben lévő adatainak megóvása

Meiszterics Enikő

Mi a KDK, hol és hogy érhető el?

- A KDK intézményi repozitórium, 2013 óta működik.
- A TK 4 intézetének kutatási anyagait gyűjti és teszi elérhetővé.
- Adatmenedzsment szolgáltatást is nyújt.
- Honlapja a kdk.tk.hu, repozitóriuma az openarchive.tk.mta.hu oldalon érhető el.

Milyen kutatási anyagok, milyen módon és kik számára érhetőek el a KDK-ban?

- A KDK repozitóriumban kvalitatív és kvantitatív módszerekkel készült társadalomtudományi kutatások anyagai egyaránt megtalálhatóak.
- Gyűjteményeiben sokféle kutatási adat (interjúk felvételei, leiratai, vezérfonalai; kérdőíves felmérések kérdőívei, módszertani leírásai, adatbázisai; terepnaplók, megfigyelések jegyzőkönyvei stb.) érhető el, különféle (szöveg, kép, videó, archiv stb.) formátumokban.
- A repozitóriumban az adatokat leíró **metaadatokhoz** minden esetben bárki szabadon hozzáférhet.

- A ***letétbe helyezett kutatási adatokhoz*** a letétbe helyezővel kötött, a Letéti Nyilatkozatban rögzített megállapodásnak megfelelően
 - bárki szabadon hozzáférhet, vagy
 - a TK kutatói szabadon, külsős kutatók egy email-es regisztráció után férhetnek hozzá, vagy
 - az adminisztrátorokon kívül mindenki más csak a letétbe helyező jóváhagyásával férhet hozzá.
- A gyűjtemények mindegyikének regisztrálunk digitális objektum azonosítót, DOI-t.

Mi a 20. Század Hangja Archívum és Kutatóműhely, és hol és hogy érhető el?

- A 20. Század Hangja Archívum és Kutatóműhely (Voices) a Magyar Tudományos Akadémia Szociológiai Kutatóintézetében alakult meg 2009 tavaszán.
- Egy évtizeden át OTKA támogatással működött, 2019 tavasza óta a KDK-val közös szervezeti egységben dolgozik.
- Az archívum a 20szazadhangja.tk.hu oldalon érhető el.



20. Század Hangja
Archívum és Kutatóműhely

Milyen kutatási anyagok, milyen módon és kik számára érhetőek el a Voices-ban?

- A 20. Század Hangja Archívum az elmúlt 60 év kvalitatív módszerekkel készült, szociológiatörténetileg is meghatározó kutatásainak anyagait tartalmazza.
- Folyamatosan bővülő gyűjteményei között megtalálhatóak romákkal kapcsolatos kutatási anyagok, a Holokauszt történetét feldolgozó interjúk, a 60-as évek munkáséletmódját bemutató nagyszabású interjúk, iskolakutatások és gyermekrajzok a 70-es évekből, szociálpszichológiai kutatások stb.
- A ***metaadatok*** a Voices esetében is minden érdeklődő számára szabadon elérhetőek.
- Az archívumban lévő ***írott, hangzó és képes dokumentumokhoz*** email-ben történő regisztráció és Kutatói Nyilatkozat kitöltése után lehet hozzáférni.

A KDK-s HRDA pilot projekt célkitűzése

- A pilot projekt célkitűzése két – a múlt század 80-as, 90-es éveiben végzett – kutatás audio kazettán lévő interjúk hanganyagainak
 - digitalizálása, leiratozása, metaadatokkal történő ellátása, majd
 - repozitóriumba helyezése és kutathatóvá és újrafelhasználhatóvá tétele volt.

Kovách Imre a helyi társadalom és a mezőgazdaság átalakulását vizsgáló kutatásának interjúi

- A vizsgálat témája a helyi társadalom és a mezőgazdaság átalakulása, a 90-es években a termelőszövetkezetek felbomlása, az érintett magánemberek, háztájizók kialakult helyzetre adott megoldásai voltak.
- Az adatfelvételek két hullámban, a 80-as évek közepén Hajdú-Bihar megyében (Hajdúnánás, Hajdúböszörmény) és a 90-es évek elején a Dunántúlon (Bajna, Epöl), történtek.

Kovács Évanak a „Shifting Identities in Mixed Marriages” című, Fejős Zoltán által vezetett kutatás keretében készített interjúi

- A projekt során kiegészítették és elemezték a vegyes házasságok adatbázisát Kassán és egész Szlovákiában.
- A demográfiai elemzéseket házasságokra, családtörténetekre fókuszáló életút interjúkkal tették teljessé, ezek hanganyaga volt az átadott kazettákon.

A projekt indítása

- A TK-ban minden induló projektben ki kell tölteni a kutatási adatkezelési tervet.
- Mivel ebben a kutatásban személyes adatokat is kezeltünk, kapott egy naplót is, ez egy - a TK felhő szolgáltatásában tárolt - táblázatos fájl a személyes adatokhoz hozzáférő résztvevők nyilvántartására.
- Ezeknek a dokumentumoknak a sablonjai a <https://kdk.tk.hu/hirek/2018/07/gdpr-es-kutatasi-adatok> oldalról bárki számára elérhetőek.

A hanganyagok átvétele

- Az átvételkor a Letéti Nyilatkozatban minden esetben rögzítjük, hogy mit kaptunk a letétbe helyezőtől.
- A Kovách Imre kutatás esetében a kazetták mellett kaptunk még papíralapú, írógéppel készített interjú leiratokat is. Az átvételkor nem áll rendelkezésre információ arról, hogy van-e, és ha van, milyen mértékű az átfedés a hanganyagok és a leiratok között.
- Kovács Éva interjúinak hanganyaga mellett papír alapon kaptunk még statisztikai táblákat és egy rövid elemzést, illetve digitálisan egy kutatási jelentést, interjúleiratokat valamint biográfiákat. Az esetleges átfedésekről ebben az esetben sem állt rendelkezésre információ.

A hanganyagok digitális formába történő konvertálása az Audacity szoftver használatával

- A hanganyagokat az Audacity szoftver használatával digitalizáltuk, zajcsökkentettük, meghallgatás után beazonosítottuk, az egy interjúhoz tartozó felvételeket összefűztük.
- A kapott digitális hangfájlokat metaadatokkal (cím, időpont, helyszín, kérdező neve, válaszadó neve) láttuk el és a TK felhőjében létrehozott mappákban helyeztük el.
- Ekkor már lehetővé vált a kapott hanganyagok és leiratok összevetése is. A vidék átalakulását vizsgáló kutatás esetében voltak átfedések, így ahol volt leirat és hanganyag is, összekötöttük azokat egymással.

Hanganyagok leiratozása a Régens Zrt. Alrite programjával

- A hanganyagok leiratozásához a Régens Zrt. Alrite programját használtuk, amely egy mesterséges intelligencia alapú beszédfelismerő szoftver a magyar nyelvre optimalizálva.
- A programot előzetesen teszteltünk, és már tesztelésnél világos volt, hogy a leiratozás mai hanganyagok esetében jellemző nagy - 90% feletti – pontosságától, amivel egy órányi hanganyag leiratának átnézett, javított változata 70 perc alatt készült el, jelentősen el fogunk maradni.
- Ugyanakkor kipróbáltuk azt is, hogy a régi szövegek hagyományos módon, csak hallás után történő gépelése milyen eredménnyel jár, és arra a megállapításra jutottunk, hogy a nehézségekkel együtt is jobb megoldás az Alrite használata, mert az általa leírt szöveg támpontot, kiindulást nyújt a problémás szövegrészek megértéséhez.

Míg a napjainkban készített hangfelvételek esetében a zavaró háttérzajokat a felvevő eszközök szinte teljesen kiszűrik, addig a projekt keretében digitalizált régi hangfelvételek minősége eleve nem volt tökéletes és egyenletes, ráadásul romlott is a rögzítés óta eltelt több mint 30 év során.

A felvételeknél a következő problémákat tapasztaltuk:

- rossz minőségű, nehezen érthető a hanganyag, különösebb zavaró tényező nélkül,
- halk a beszéd, maximális hangerő mellett is alig lehet hallani,
- zavaró zajok – kutyaugatás, ajtócsapkodás, madárcsicsergés - hallatszanak,
- sok esetben - feltehetően a magnó/mikrofon elhelyezkedésének függvényében- vagy csak a kérdező vagy csak az interjúalany hangját lehet jól hallani.

Nehezítette a leiratozást az is, hogy több esetben idős, tájszólással beszélő, tájnyelvi vagy ma már kevésbé elterjedt szavakat használó emberek voltak az interjúalanyok.

A felmerülő problémák miatt 1 órás régi hanganyag leiratának kiegészítése, javítása átlagosan 4 órát vett igénybe.

Gévelt leiratok digitalizása OCR technológiával, Abbey FineReader programmal

- Az, hogy a vidék kutatásban a hanganyagok egy részéhez megvolt a gévelt leirat is, lehetőséget adott azok optikai szövegfelismeréssel (OCR technológiával), Abbey FineReader programmal történő digitalizálására is.
- Az érintett gévelt leiratok viszonylag jó minőségűek és „tiszták” (nem voltak bennük kézzel írt bejegyzések) voltak, ennek köszönhetően az így készített digitális változat esetében a kapott szöveg pontossága – az Alrite mai technológiával felvett interjú leiratozásának pontosságához hasonlóan - 90 % feletti volt, de a szövegeket még újra kellett tördelni, formázni.
- Úgy döntöttünk, hogy a hanganyagtól független leiratok teljeskörű feldolgozásával a közel jövőben még kiegészítjük a gyűjteményt, de ezek között vannak rosszabb minőségű leiratok is, amelyek javítása, kiegészítése már tovább fog tartani.

Az interjú átiratok FAIR alapelveknek megfelelő archiválása, kutathatóvá tétele

- A leiratozást/OCR-t követően a KDK munkatársai még egyszer átnézték, egységes formába szerkesztették, anonimizálták és metaadatokkal látták el az interjúkat.
- A vegyes házasság kutatásnak korábbról megvolt a gyűjteménye a KDK repozitóriumában, a vidék átalakulását feltáró interjúk számára új gyűjteményt hoztunk létre.
- Mindkét gyűjteménynek DOI-t regisztráltunk az MTA Könyvtár DOI irodáján keresztül.

- A KDK gyakorlatának megfelelően a projektben keletkezett, a kutatásokat és kutatási adatokat leíró para- és metaadatokat, az archiválás módjának leírását, az adatok dokumentációját, a kapcsolódó publikáció listáját a repozitóriumban minden érdeklődő számára korlátozások nélkül hozzáférhetővé tettük.
- Az interjúk leiratait csak a teljes nevek szintjén anonimizáltuk, mivel meglátásunk szerint a települések, városrészek, foglalkozások anonimizálása olyan információvesztést okozott volna, amely ennyi idővel az adatfelvételek után már nem indokolt. Részben emiatt, részben pedig az interjúkban felmerülő szenzitív témák (pl. a magyar-szlovák viszony vagy a településeken élő zsidó közösség és a hozzájuk való viszony leírása) miatt az interjúkat nem tettük mindenki számára szabadon hozzáférhetővé, azok kutatói szinten elérhetőek, azaz a TK kutatói szabadon, más kutatók, egyetemi hallgatók email-ben történő regisztráció után férhetnek hozzájuk.

Konkluzió

- A bemutatott nehézségek mellett is érdemes volt belevágni a projektbe, érdekes és valószínűleg pótolhatatlan interjúkat tettünk így elérhetővé, újra felhasználhatóvá.
- Fontos lenne a még nálunk – vagy bárki másnál – lévő hasonló anyagokat menteni, de nem egyszerű feladat.

Köszönöm a figyelmet!