

Jasoa: 2025-05-19
Onartua: 2025-09-09

Online first argitalpen-data: 2026-02-13
DOI: <https://doi.org/10.1387/asju.27560>



FAIR Mitxelena: Aipamen-erlazioen erabilgarritasunerantz¹

FAIR Mitxelena: Towards the usability of citation relations

David Lindemann

Universidad del País Vasco/Euskal Herriko Unibertsitatea (EHU)

ABSTRACT: We present experimental steps in the computational processing of the contents of Mitxelena's publications, and of the relationships they contain. Regarding the written work of Koldo Mitxelena (1915-1987), a philologist from Erreterria, we extracted citations from two Basque scholarly journals and collected them as citation relations in a database, on the one hand, and worked with the mentions of persons in Mitxelena's own work, on the other. We recorded the mentions of persons in their context, along with relationships involving the entities describing these people in the Wikidata knowledge graph. We have opened up new access paths for students and researchers to discover the work of Mitxelena. The results have been stored in an online database called *Inguma Wikibase*, which is accessible and reusable by anyone.

KEYWORDS: Linked Data; Wikibase; Koldo Mitxelena; citation relations.

LABURPENA: *Mitxelenaren argitalpenetako edukien eta horiek dituzten erlazioen prozesamendu konputazionalerako zenbait urrats egin dugu. Alde batetik, Koldo Mitxelena*

¹ Hemen aurkezten ditugun emaitzetara heltzeko lanak «*Monumenta Linguae Vasconum* 6: avances en cronología de la historia y la prehistoria de la lengua vasca» (MINECO, PID2020-118445GB-I00) ikerketa-proiektuak eta «*Hizkuntzalaritza Diakronikoa, Tipologia eta Euskararen Historia / Diachronic Linguistics, Typology and the History of Basque (DLTB)*» (Eusko Jaurlaritza, IT1534-22) ikerketa taldeak sustatu dituzte.

* David Lindemann. Hizkuntzalaritza eta Euskal Ikasketak Saila - Letren Fakultatea (UPV/EHU). Unibertsitateen ibilbidea, 5 - 01006 Vitoria-Gasteiz - david.lindemann@ehu.eus -  <https://orcid.org/0000-0002-8261-6882>



(1915-1987) Errenteriako euskalariaren lan idatziek euskarazko bi aldizkari zientifikotan dituzten aipua erauzi eta aipamen-erlazio gisa datu-base batean jaso ditugu eta, bestetik, Mitxelenak berak bere lanetan aipatzen dituen pertsonak landu ditugu. Mitxelenaren laneko testuinguruarekin jaso ditugu pertsona horien aipamenak, Wikidata ezagutza-grafoan dituzten erlazioekin batera. Mitxelenaren lana ezagutzeko sarbide berriak ireki ditugu dagoeneko, ikasleentzat nahiz ikerlariarentzat. Emaizak Inguma Wikibase izeneko datu-basean jasota geratu dira, sarean, edonork ikusi eta berrerabiltzeko moduan.

HITZ GAKOAK: Datu Lotuak; Wikibase; Koldo Mitxelena; aipamen-erlazioak.

1. Sarrera

Euskal Filologiaren lanabesak euskarri digitaletan jasotzeko aukera berriak dakartza; lehen mailako iturriak (batik bat, euskarazko testu zaharrak) eta bigarren mailakoak (ikerketa-artikuluak) paperezko euskarritik ateratzeak edukiak metodo konputazionalarekin prozesatzea ahalbidetzen du, lehen eskuzko lana izaten zenari bidezidor eraginkorrak zabalduz. Bi alderditatik abiatu gara norabide horretan, eta honako hauek dira: ikerketa-artikuluek euren artean aipamen-erlazioak dituzte, artikulua bati erantsita datozen bibliografia-ataletan islaturik geratzen direnak; erlazio horiek jasotzea izan dugu helburu, euskarazko bi aldizkaritako artikulua guztietakoak, zehazki. Bestalde, Koldo Mitxelenaren idazlan guztietan aipatzen diren entitateak (pertsonak, tokiak, erakundeak, etab.) landu ditugu. Artikulu honetan, datu horiek lortzeko eta Datu Lotuen paradigmaren² arabera datu-base batean jasotzeko beharrezkoak izan diren lanak aurkeztu ditugu, zenbait emaitzarekin batera. Software libreaz baliatu gara beti ere; hau da, edonork errepika ditzake proposatzen ditugun urratsak, edo beste datu-multzo bati aplikatu. Aldi berean, lortutako emaitzek orokor bihurtzen ari den FAIR irizpide-sortari³ erantzuten diotela erakutsi nahi dugu. Euskal Filologiako datuak eskala handian digital eta FAIR bihurtzeko proiektua diseinatzeke ekarpena egitea da artikulua honen helburua.

2. FAIR irizpideak eta Euskal Filologiako datuak

FAIR (*findable, accessible, interoperable, reusable*) irizpide-sortak datu-multzoen ezaugarriak deskribatzen ditu (Wilkinson *et al.* 2016). Lehen eta bigarren mailako iturriak (testuak eta artikulua zientifikoak) eta ikerketa-emaitzak (hiztegiak eta abarrekoak)

2 Datu Lotuen paradigma, hau da, *Semantic Web* teknologiak, zitazio-erlazioetan aplikatzean, zitazio-grafo irekiak eraikitzeke tendentzia orokorretan oinarritzen gara (ikus, adibidez, Heibi *et al.* 2024). *Inguma* bildumako datuak modu honetan prozesatzeko proposamena Astigarraga *et al.* (2025) artikuluan aurkeztu dugu; hemen aurkeztu duguna egitasmo orokor horren parte da.

3 Ikus <https://www.go-fair.org/> ere.

argitaratzeko orduan, irizpide haiek betetzea baldintza bihurtzen ari da Horizon deialdietan adibidez,⁴ edo gomendatu egiten da, ANECaren ikerketa jardueraren ebaluazioan, esaterako.⁵ Irizpideok, labur esanda, paperezko argitalpenaren bitartez ezin dira bete, eta argitalpen digitalek ere ezaugarri jakin batzuk izan behar dituzte, edukiak datu-multzo (*dataset*) gisa modu eraginkor batean prozesatu ahal izateko.

FAIR irizpideak betetzeko bide bat eta argitalpen digitalen formatu batzuk kualitatiboki sailkatzeko balio duen eskema proposatzen du Berners-Lee autoreak (ikus Martin *et al.* 2013), *Five Star Open Data* eskema deiturikoa.⁶ Eskema hark lizentzia irekiko PDF argitalpen bati izar bakarra ematen dio, eta izarrek gehitzen dizkie prozesamendurako PDF argitalpena baino egokiagoak diren formatuei. Bost izarreko formatua *Linked Open Data* da, Datu Lotu Irekiak, non datu-multzoen edukiaren elementu batek makinak uler dezakeen helbide uniboko eta betikoa duen (URI bat, *uniform resource identifier*, hau da, betiko *http* helbide bat). URIaz, elementu hura beste elementu batzuekin erlazionatuko da, makinak ulertzeko modukoa den hizkuntza bat erabiliz: hirukote semantikoa, subjektu, predikatu eta objektua dituen. Hemen *elementu* deitzen dioguna edozer izan daiteke; hori irudikatzeko, saia gaitezen FAIR irizpideak Euskal Filologiaren lanabesekin erlazionatzen, Datu Lotuen paradigmaren arabera jokatz.

- *Findable*, aurkigarria. Datu-multzo bat (testu historikoen bertsio digitala, hiztegiak, artikulua, tesia, ikerketa-emaiza edo antzekoa) identifikatzaile uniboko batez deskribatuko da. Identifikatzaile haren bitartez datu-multzo hura deskribatzen dituzten metadatueta heldu behar da. Adibidez, argitalpen bati lotutako DOI (*Digital Object Identifier*)⁷ batean klik egiten duen pertsonari fitxa bibliografikoa eskaintzen zaio, argitaratzailearen webgunean; helbide horretara eskaera egiten duen makina bati, ordea, metadatu egituratuak eskaintzen zaizkio, makinak ezagutzen duen bokabulario estandar baten arabera. Metadatuaren parte honako hauek dira: ohiko argitalpen-metadatuak (aipamena egiteko modukoak) eta edukien formatuari eta erabiltzeko lizentziari buruzko xehetasunak, lizentzia irekia izanda beti ere.
- *Accessible*, atzigarria. Datu-multzoaren identifikatzaileak baliabidearen edukietara bertara eraman behar du, edukia deskargatzeko helbide batetik. Edukia web-interfaze batez ikus badaiteke, horren sarbideko helbidea ere baliabidearen metadatuaren parte

4 Ikus <https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm>.

5 Ikus 2024ko baremoari buruzko dokumentua, https://www.aneca.es/documents/20123/263604/BAREMO_091224.pdf.

6 Ikus <https://5stardata.info/en/>.

7 Ikus <https://www.doi.org/>.

izango da. Horretaz gain makina batek datu-base galdeketak egitera konektatzeko interfazea (APIa) badago, haren berri emango da.

- *Interoperable*, elkarreraginkorra. Edukiak mikro-mailan identifikatzaileak izan behar ditu, datu-multzoaren elementuak beste baliabide batzuetako elementuekin erlazionatu ahal izateko. Hiztegi-sarrera batek, adibidez, bere URIa izango du, kanpotik elementu horri berorri heldu ahal izateko, edo hirukote semantiko batean erabili ahal izateko. Artikulu baten egile bat, adibidez, bere ORCID zenbakiaz deskribatzen bada, egile beraren beste artikuluekin edo egile hori deskribatzen duten fitxa biografikoekin lot daiteke. Halaber, erreferentzia bibliografiko batek DOI bezalako identifikatzailea badu, pertsonak nahiz makinek argitalpenera zuzenean jotzeko modua dute, eskuzko bilaketa egin eta emaitzak desanbiguatu beharrik gabe. Beste adibide bat: Hiztegi-sarrera baten adiera bat beste hizkuntza bateko lema baten adierarekin bat badator, beste haren itzulpen-ordaina dela hirukote semantiko batez adieraz daiteke: $A_hiztegia_A1_adiera > itzulpena > B_hiztegia_B1_adiera$. Hirurok URIz adieraziko dira: bi hiztegi-tako elementuak, eta “itzulpena” erlazioa. Elementuen URIek elementura bertara bidaliko dute, eta erlazioaren URIak erlazio hura deskribatzen duten datu batzuetara bidaliko du. Bi elementuen arteko erlazioak, propietateak ere deituak, bokabulario estandarrekoak izatea komeni da, datuak ezagutza-grafo handietan integratzea helburua bada, behintzat.
- *Reusable*, berrerabilgarria. Gorago jada aipaturiko lizentzia irekiaz gain, ondorengo hau ere kontuan hartzekoa da irizpide honi dagokionez: datuen kalitatea, jatorriari (*provenance*) buruzko metadatu xeheetan bermaturik. Hau da, modu egituratuan emango dugu baliabidearen barruan egiten ditugun baieztapenen sorreraren berri: nork, noiz, non eta nola den sortua deskribatzen ari garen hiztegi-sarrera hori, corpus-eduki, artikulua, sarrera bibliografiko edo bestelako elementu hori.

3. Aurrekariak eta egungo egoera

Lehen mailako iturrien prozesamendu konputazionalari dagokionez, Estarrona *et al.* (2022) lana aurreratuen dela esan daiteke. Anotazio morfosintaktikoak dauzkan euskarazko lehen corpus historikoa da haien emaitza. Lema, kategoria gramatikala eta analisi morfologikoari dagozkien etiketez hornitu dituzte testu zaharretako hitzak. Web atariko erabiltzaile-interfaze batean⁸ bilaketa aurreratuak egin daitezke honezkero, eta emaitzak testuinguruan ikusi.⁹

8 Atariak <http://bim.ixl.es/> helbidea du.

9 Ikus proiektua aurkezteko posterreko irudiak: <https://www.clariah.eu/eu/node/12>.

Proiektuko lan-fluxua honako hau izan da: (1) *Armiarma* gordailutik hartutako testu zaharraren bertsioa berrikusi dute, testu historikoaren bertsio batera heltzeko non jatorrizko testu-hitzak *OEH*ko grafiari egokituta agertzen baitira (adib. *gucietan* > *guzietan*, Etxepare Hitzaurrea);¹⁰ (2) aurreko urratseko egokitzapena *Euskaltzaindiaren hiztegian* zerrendatutako lemaren formaren arabera egokitu dute (*guzietan* > *guztietan*); (3) Euskara Baturako erabilgarri dagoen tresneriaz baliatuz, etiketatzea burutu dute, testu-hitzentzat lema, kategoria gramatikala eta analisi morfologikoa lortuz.

Testu historikoen edizioa euskarri digitalean aurkezteko, ordea, Bilbao *et al.* (2011) lana da aipagarriena. Lazarragaren eskuizkribuko atalak eskaintzen ditu web-interfaze batek¹¹ bertsio batean baino gehiagotan, eta, kasu guztietan, jatorrizko grafia “erdipaleografikoan” eta “gaurko” grafian. Bertsio ezberdinak PDF dokumentu banatan eskaintzen dira, edizioa prestatu dutenen oharrak testuan bertan txertatuta agertzen direlarik, oin-ohar gisa. Bestetik, grafia erdipaleografikoa gordetzen duen bertsio digitalaren formategia erauzi dute, bakoitza jatorrizko testuinguruarekin. Datu horiekin, konkordantzia ekoitzi dute, eta edizio digitalari gehitu.¹² Zerrendak dira, PDF formatuan, eta, ondorioz, testu-bilaketa sinpleak baino ezin dira egin. Zerrendan ageri diren jatorrizko grafia duten formak (adib. *guztiaquin*) ez daude gaurko grafiako formekin (adib. *guztiakin*) erlazionatuta; hau da, gaurko grafiako formak ezin dira bilaketetan erabili. Halaber, dagokion lemara doan, lemaren arabeko bilaketa ahalbidetuko lukeen loturarik ez dute zerrendako formek, ezta deskribapen morfosintaktikorik ere. Konkordantzia-zerrenden sarreretatik ez dago ediziora berera salto egiteko modurik, zerrendan agertzen diren dokumentu-kodeak edizioa duten PDF dokumentuetan eskuz bilatzen ez bada.

Arlo honetan puntakoak diren proiektuen metodologiari jarraituz (Linking Latin; ikus Passarotti & Mambrini 2021),¹³ testu-hitzak Euskara Batuko formategi eta lemategi batera lotzeko datu-eredu bat proposatu dugu beste lan batean (Lindemann & Alonso 2025); lan horretan, testu batean ageri diren entitate izendunak Wikidatarekin lotzeko eta lotura hori Wikibase batean jasotzeko proposamena ere egiten dugu. Antzeko eredu bat azalduko dugu hemen, Mitxelenaren *Obras Completas* bildumak (Mitxelena 2011) dakarren izendegiko entitateak Wikibase batean errepresentatzeari dagokionez.

10 Ikus https://eu.wikisource.org/wiki/Linguae_vasconum_primitiae#2 (Wikitekako bertsioa, grafia egokiturik) eta <https://www.ixu.eus/wikisource-bim/index.php/Orrialde:Etxepare-LVP.djvu/4> (Estarrona *et al.* (2022) lanerako erabilitako bertsioa, faksimilearekin batera).

11 Ikus <https://www.ehu.eus/monumenta/lazarraga/>.

12 Ikus <https://www.ehu.eus/monumenta/lazarraga/hiztegia.html>.

13 Corpusetako testu-hitzak eta hiztegi-sarrerak elkarrekin lotzen ditu Linking Latin (LiLa) proiektuak; ikus <https://lila-erc.eu>.

Inguma Wikibase izeneko datu-basea aurreko proiektu txiki batean sortu genuen (Lindemann *et al.* 2023). Inguma¹⁴ datu-basearen kopia du, Datu Lotuen paradigmari egokitutakoa; hau da, deskribatzen diren entitate guztiek URI (web-helbide) uniboko eta betikoa dute.

4. Koldo Mitxelenaren bibliografia eta aipamen-erlazioak bi aldizkaritan¹⁵

Koldo Mitxelenaren lan idatzien multzo oparoa, lan honetan aztergai duguna, aski sakabanatuta zegoela esan daiteke: 1665 argitalpen eta berrargitalpen zenbatu ditugu,¹⁶ hainbat aldizkaritan eta liburutan agertzen direnak; horien artean gehien-gehienak behin edo behin baino gehiagotan berrargitaratu izan dira. 2011. urtean Mitxelenaren *Obras Completas* (OC; Mitxelena 2011) agertu arte, 745 testu garrantzitsuenak bateratzen dituen bilduma, ez zegoen Mitxelenaren lana batera ikusteko ituririk.

Gure lehenengo urratsa Koldo Mitxelenaren idazlanen argitaraldi bakoitzari identifikatzailea esleitzea izan da. Wikibasean, idazlan bakoitzaren argitalpen-metadatuak *entitate*¹⁷ bati lotu dizkiogu.¹⁸ Bigarrenik, lan haietako bat aipatzen duten artikulua bildu eta datu-base berean errepresentatu ditugu modu berean.¹⁹ Hirugarrenik, aipamen-erlazioak errepresentatu ditugu, aipatzen duen artikulua metadatuari Koldo Mitxelenaren lanera doan erlazio bat gehituz. Bi iturritako artikulua hartu ditugu aintzat urrats honetan: *ASJU*²⁰ aldizkaria eta *Uztaro*²¹ aldizkaria. Denera, 830 aipu-erlazio zehaztu ditugu, 725 aipu-erlazio *ASJU* aldizkariko 175 artikulutatik, eta 105 aipu-erlazio *Uztaro* aldizkariko 37 artikulutatik.²²

Gure ekimenez, OCko testuak aurreko argitaraldiekin loturik geratzen dira Inguma Wikibase datu-basean (ikus bisualizazioa 1. irudian).²³ Gure Wikibasetik kanpo, 2011 baino

14 <https://inguma.eus> helbidean Inguma web-ataria aurkitzen da.

15 Atal honetan aurkeztutako esperimendua UEU/Inguma talde gisa proposatu genion Errenteriako Udalak babestutako Humanitate Digitaleen Guneari, eta erakunde hark emandako bekaren laguntzaz burutu genuen; elkarrekin aurkeztu genuen poster gisa CLARIAH-EUS III. workshopean, 2024ko azaroan (Lindemann *et al.* 2024). Egitasmoaren nondik-norakoak Astigarraga *et al.* (2025) artikuluan zehaztu genituen.

16 Ikus https://wikibase.inguma.eus/wiki/Ziterauzi_KM#Mitxelenaren_argitalpen_guztiak. Elixabete Perez Gazteluren bibliografia, Worldcat, DialNet eta Inguma izan dira iturriak.

17 Datu basean, entitate bakoitzak identifikatzaile unibokoa izango du, eta entitatea deskribatzen dituen *propietate-balio* bikoteak eranstean zaizkio; adib., *A artikulua – noiz argitaratua – 1975*, edo *A artikulua – aipatzen du – B artikulua*, ingelesezko SVO ordenari egokitutako *hirukote semantikoak* osatuz. Hirukote horiek grafo itxurako sarea osatzen dute.

18 Hemen Mitxelenaren bibliografia: https://wikibase.inguma.eus/wiki/Ziterauzi_KM#KM_Bibliografia.

19 Ingumako edukiak dituen, *Uztaro*ko artikulua guztien eta *ASJU*ko euskarazko artikulua guztien sarrerak bazeuden aurretik Wikibasean.

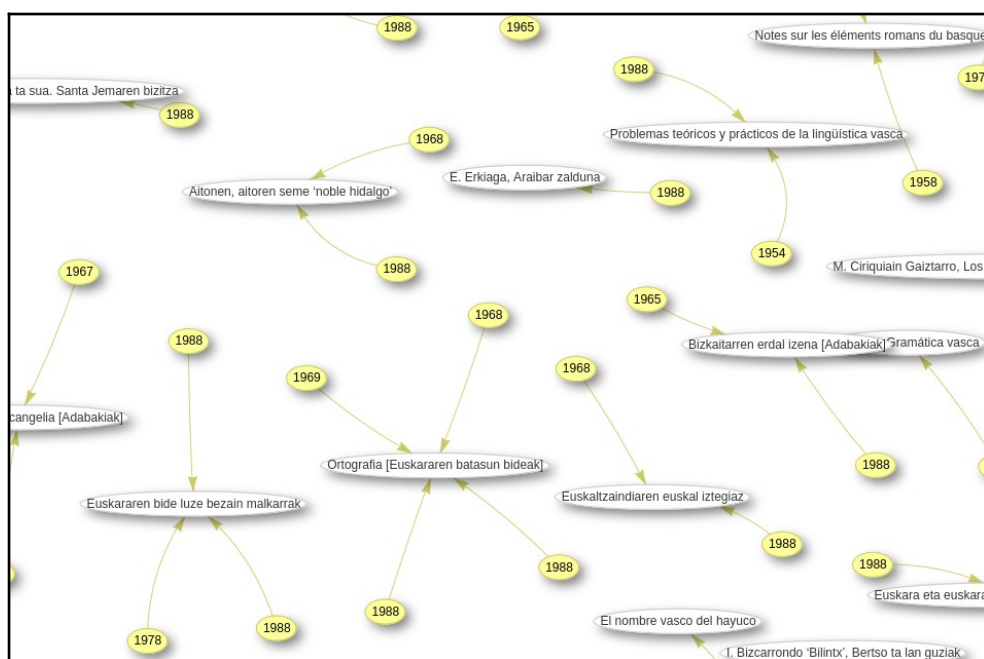
20 Ikus <https://ojs.ehu.eus/index.php/ASJU/index>.

21 Ikus <https://aldizkariak.ueu.eus/index.php/uztarok>.

22 Ikus https://wikibase.inguma.eus/w/index.php?title=Ziterauzi_KM#Zitazio-erlazioak.

23 Iturria: https://wikibase.inguma.eus/wiki/Ziterauzi_KM#OCC-ko_argitalpenen_aurrekoak.

lehenagokoek nahiz *OC*ko argitalpenek kasu gehien-gehienetan ez dute makinak interpreta lezakeen identifikatzailerik (DOI, *handle* edo antzekorik), aipatzen duen artikuluen bibliografia-atalean agertu litekeena. Horrek aipamenen identifikazio eta kudeaketa automatizatua zaildu egiten du. Horri gehitu behar zaio aipamenen erredakzioan formatuen arau eta estilo aniztasuna, makinarentzako ulergaitza dena. Erregistroen identifikazioak eskuzko lan handia eskatzen du, ondorioz.



1. irudia

*OC*ko bildumako testuen aurreko argitalpenak

Hala ere, identifikatze-lan horretan laguntzeko metodo semi-automatiko bat frogatu dugu, GROBID tresnak²⁴ eskaintzen duena. Tresna hori PDF argitalpena egituratzen saiatzen da, lanaren atalak isolatuz (izenburua, egilea eta afiliazioa, laburpena, testu-gorputza eta bibliografia), eta bibliografia-atala sarreretan zatituz. Sarrera bakoitzaren barne-egitura ere zehazten saiatzen da, hau da, egileen izenak, argitalpenaren izenburua, euskarria, orrialdeak, argitalpenaren urtea eta abarrekoak etiketatzen. DOI identifikatzailea ageri bada, tresnak ez du arazorik: DOI hori aurkituko du, eta Crossref²⁵ web-zerbitzuaren bitartez ebatziko du aipatutako argitalpena zein den, eta haren metatadua formatu egituratuan datu-base hartatik lortzeko moduan egongo da.²⁶ Kasu horretan, PDFtik irakurritako informazioa baztertuko du,

24 Ikus <https://grobid.readthedocs.io/en/latest/>.

25 Ikus <https://www.crossref.org/>.

26 Ikus <https://api.crossref.org/works/10.1075/cilt.131.05mic> helbidean CrossRef datu-baseak Mitxelenaren liburu-atal baten inguruan dituen datuak.

eta Crossref datu-baseko informazioa erabili. Halako identifikatzailerik ematen ez denean, eta aurrean dugun datu-multzoan (*ASJU* eta *Uztaro* aldizkarietako artikuluetan Mitxelenaren lanei egiten zaizkien aipamenak) hori horrela da kasu gehien-gehienetan, GROBIDek proposatzen duen sarreraren zatiketak zuzena izan behar du, erabiltzaileak hortik abiatuta dagokion argitalpenaren identifikatzailea eskuz zehaztu ahal izateko (Google Scholar,²⁷ Dialnet,²⁸ Inguma,²⁹ OpenAlex³⁰ eta abarreko zerbitzuen bitartez).

GROBID tresnarekin froga batzuk egin ondoren, Mitxelenaren aipamenak aurkitzeko metodoa baztertu egin dugu, tresnak aipamen asko aurkitzen ez zituela ikusita. Hau da, aipamenak dituen testua eskuz aztertzea beharrezkoa izango litzateke edozein kasutan. Bizkorragoa eta eraginkorragoa izan da, beraz, “Mitxelena” eta “Michelena” hitzak bi aldizkarietako artikuluen guztietan automatikoki bilatzea, eta ondoren hitz horien agerpenak bisitatzea, Mitxelenaren lanetako baten aipamena ote duten zehazteko. Aipamena aurkituz gero, aipatutako lanaren identifikatzailea bilatu dugu Wikibaseko erabiltzaile-interfazea erabiliz. Wikibaseko interfazeak 2. irudian bezala erakusten du bilaketaren emaitza; aipatutako idazlanaren izenburua bilaketa-eremuan sartu eta agertzen diren iradokizunen artean egokia hautatuz ezartzen da erlazioa.



2. irudia

Aipamen-erlazioa eskuz definitzen, Wikibase interfazean

ASJU eta *Uztaro* aldizkarietako artikuluen nahiz Mitxelenaren argitalpenen metadatuak Wikibase berean ditugunez, bi aldizkari horietan Koldo Mitxelenak dituen aipamen guztiak Wikibaseko identifikatzaileak erabiliz errepresentatu ditugu. Horrela, aipamen-erlazio bakoitza, teknikoki, bi identifikatzaileen arteko erlazioa da, Wikibase beraren baitan, 1. taulan adibide batez irudikatzen dugun bezala.

27 Ikus <https://scholar.google.com/>.

28 Ikus <https://dialnet.unirioja.es/>.

29 Ikus <https://www.inguma.eus/>.

30 Ikus <https://openalex.org/>.

Aipamen-erlazio bat hirukote semantiko gisa

Aipamen-erlazioak aipatutako argitalpenaren argitaraldi zehatzarekin lotu ditugu. Bestalde, argitalpen bakoitza *OC*ko argitalpenei lotu diegunez, Mitxelenaren testu bakoitzeko argitaraldiek dituzten aipu-erlazioak batzeko moduan gaude, *OC* argitalpena *pibot*-elementutzat hartuz, 3. irudiak³¹ erakusten duen bezala.



Aipamen-grafoaren bisualisazioa, aipuak OC-itemei baturik

9

Horrela sortutako datu-multzoan zeharkako galdeketak egin daitezke; adibidez, Mitxelenaren lanen ibilbideak haien artean konparatzeko galdeketak, lan horiek aipatzen dituzten ikertzaileak eta aipatutakoaren nahiz aipatzen duen lanaren argitalpen datak aintzat hartuz.

Azkenik, Koldo Mitxelenaren lanen aipamenak dituzten artikuluen metadatuak eta OCko lanen erreferentzien metadatuak Wikidatan gehitu ditugu, bi multzo hauen arteko aipamen-erlazioa barne (hau da, aipu guztiak OCko argitaraldietara “birbidalita”).³² *Scholia*,³³ datu bibliografikoen kudeaketa bisualerako tresna, Wikidatako datuez elikatzen da eta bibliometriako oinarritzko azterketak eskaintzen ditu. Horrela, esperimentuaren emaitza orokor batzuk erakuts ditzakegu.³⁴

5. Mitxelenak aipatzen dituen pertsonak: OC izendegiaren berregituratzea

Mitxelenaren lana atzigarriago eta argitalpenen barruko elementu batzuk aurkigarriago bihurtzeko urratsak egin ditugu: Mitxelenak bere artikuluko ugarietan aipatzen dituen pertsonak dira gaia. Idazlan batean, testuan zehar edota bibliografian, ageri diren pertsona-izenak aurkitu eta erauzteko teknika automatikoak badaude; horri dagokion ataza *entitate izendunen erauzketa* (*named entity recognition, NER*) izenpean da ezaguna. Hala ere, nekeza izaten da entitateak izendatzen dituzten hitzak testu batetik erauztea eta lortutako emaitzak zuzentzea, tresna automatikoaren emaitzak ez baitira erabat egokiak izango. Hau da, aipatzen diren entitate batzuk ez dira lortutako zerrendan egongo, eta, bestalde, zerrendako batzuk ez dira entitate-izenak izango. Lortutako zerrendak eskuz zuzentzea da aukera bat.³⁵

32 Hemen [idatzitako](https://wikibase-inguma.eus/wiki/KM_Wikidata_export#Wikidatan_idatzitako_aipu-erlazioak) [erlazioen](https://wikibase-inguma.eus/wiki/KM_Wikidata_export#Wikidatan_idatzitako_aipu-erlazioak) zerrenda: https://wikibase-inguma.eus/wiki/KM_Wikidata_export#Wikidatan_idatzitako_aipu-erlazioak.

33 Ikus <https://scholia.toolforge.org/>.

34 Hemen eskuragarria: <https://scholia.toolforge.org/author/Q2457804#citation-statistics>.

35 Euskarazko Hizkuntza Eredu Handia (Pérez *et al.* 2025) horrelako esperimentuak egiteko erabilgarria denean, eta hori aurtengo urterako espero da, harekin esperimentuak egitea interesgarria izango da.

<i>GH</i> X, 443 n. 7; XIII, 64, 181, 272, 330	Goitia, J. XIII, 454
Giacomelli, G. I, 139 n. 67	<i>Goiz-argi</i> XIII, 399, 415
Gide, A. XIII, 370	<i>Goizian goizik</i> (canción) XIII, 18 n. 32, 108 n. 7
Giedion-Welcker, C. XIV, 322	Goizueta, J. M. ^a XIII, 60, 175
Giese, W. III, 39; VI, 350, 367; VIII, 87 n. 8, 287-288, 629, 659; IX, 103, 129, 161, 172; X, 444	Gómez, G. XII, 98
Gifford, D. J. VI, 27; VII, 650 n. 2; VIII, 6; IX, 337; X, 380; XI, 1, 1 n. 1, 7, 11, 336 n. 13, 345, 506 n. 2, 609; XII, 58, 202 n. 1; XIII, 17 n. 27, 106 n. 2	Gómez, P. XIV, 162
Gil, J. V, 284 n. 12	Gómez Moreno, M. I, 363; III, 7, 18-19, 48, 51, 74, 80, 85 n. 1, 90, 96-98, 99 n., 103-104, 106, 108, 109 n. 14, 110, 112, 118 n. 27, 123, 130, 138, 155, 158, 161, 163-164, 166, 166 n. 13, 167, 167 n. 16, 169, 169 n. 19, 170, 170 n. 20, 171, 174, 174 n. 4, 175 n., 176, 177 n. 7, 178, 178 n. 11, 196, 214; IV, 6, 31, 75, 95, 112, 114, 161, 164; V, 71, 97, 122, 168, 168 n. 9, 169 n. 11, 172, 192, 218 n. 19, 226
Gil, L. II, 49 n. 4; XI, 585	
Gil Reicher, G. XI, 90, 98; XIII, 11, 21, 98, 112	
Gili Gaya, S. VII, 394	
Gilliéron, J. I, 161 n. 4, 207; VIII, 392, 415 n. 1; X, 192	

4. irudia

Obras Completas, “Índice de nombres” (lagina)

Kasu honetan, zorionez, ez dugu bide luze horretatik jo beharrik izan, 2011ko Mitxelenaren *Obras Completas* bildumak izendegi-eranskina baitakar (XV. liburukian datorren “Índice de nombres”). Hura eskuz ekoitzi egin zuten argitalpenaren arduradunek, antza denez, *Fonética Histórica Vasca* laneko izendegi eta indizea *OC* guztira hedatuz. Paperezko argitalpenean tokia aurrezteko beharra dela medio, izenak laburturik agertzen dira, baita agertokien zerrenda ere, izen bakoitzarekin batera ematen dena. 4 irudian ikusten den moduan, pertsona gehienak abizen bakarraz eta lehenengo hizkietara mugatutako ponte-izenez deskribatzen dira, eta agertokiak bolumenaren erromatar zenbakiaz eta orrialdearen arabierazko zenbakiaz. Horrek erabilgarritasun mugatua dakarkio erabiltzaileari: Laburtutako izenaz jakin behar du pertsonak identifikatzen, eta zein testutan eta zer testuingurutan pertsona hori aipatzen ote den jakiteko, adierazitako bolumenaren orrialdera jo behar du, aurretik jakin gabe zer idazlan izango den orrialde horietan aurkituko duena. Bigarren arazoa konpontzea erraza da, lehenengoa, aldiz, ez hainbeste.

Erabiltzaileak izenak desanbiguatzeko izan ditzakeen zalantzei konponbidea emateko, entitate izendun horiei izen hutsaz gainerako metadatuak gehitzeko, eta, aldi berean, aipamenen testuinguruetara bide zuzenago bat eskaintzeko asmoz, izendegia paperezko eredutik atera eta eraldaketari ekin diogu, Inguma Wikibase datu-basean integratzeko. Laneko urratsak honako hauek izan dira:

- (1) *Obras Completas* bildumako artikuluei dagozkien PDF fitxategiak TXT bihurtu eta orrialdeka zatitzea, orrialde bakoitza bere orrialde-zenbakiz aurkigarri geratzeko.
- (2) Izendegia duen PDF argitalpena TXT formatura bihurtzea.

- (3) Izendegia egituratzea, erregela bitartez, TXT dokumentuaren lerro bakoitzean izen bakarra eta bere agertokien kodeak izan arte. 2 861 izenetako zerrenda da emaitza. Gehien-gehienak pertsona-izenak dira, eta gutxi batzuk testu historiko (anonimoen) izenak (*Cantar de la quema de Mondragón* edo *Refranes y Sentencias* gisakoak), edo aldizkari-izenak (*FLV*, *RIEV*).
- (4) TXT fitxategiko lerroak banatzea, bi zutabetako taula ekoizteko, horrela lehen zutabearen entitatearen izena, eta bigarren zutabearen agertokien kodeak izateko. Orain, datu-multzoa makina bitartez interpretatzeko moduan gaude, makinari adierazita zutabe bakoitzean zer aurkituko duen. Horregatik, Berners-Lee-ren eskemaren arabera, hiru izardun datu-multzo bihurtu dugu *OC*ko izendegia: datuek orain egitura dute, datu-kategoria biak (izenak eta agertokiak) banandurik agertzen direlako, bakoitza bere zutabearen.
- (5) Inguma Wikibase datu-basean, izen bakoitzerako entitate bat sortzea, eta izendegiko datuak hari eranstea; honako datu hauek, zehazki:
- Agertokiari dagokion idazlana (*OC*ko atala). Inguma Wikibasean *OC* bildumaren atal guztien metadatuak bagenituenez (ikus goiko 3. atala), erregela batek agertokiaren kodeak (bolumena, orrialdea) dagokion artikuluekin erlazionatu eta orrialde-zenbakiak dagokien artikuluen azpian jaso, 5. irudiak erakusten duen moduan.
 - Artikulu horri dagozkion agertokien zerrenda (kodeak). Denera, *OC*ko atalek 13 218 aipu-erlazio dituzte, testu berean entitate berak dituen aipamenak aipamen bakartzat hartuta. Mitxelenaren lan bereko entitate bereko aipamen guztiak zenbaketan onartuz gero, ordea, 21 633 dira. 5. irudiko adibidean, lan berean 4 aldiz aipatzen du Mitxelenak Gómez Moreno.

El vasco y las lenguas antiguas de España y Francia	
orrialdea(k)	V: 168, V: 168 n. 9, V: 169 n. 11, V: 172
testuingurua	[V: 168] Este endeble castillo de naipes se hundió, más que por el ataque de orden teórico —bien fundado, por otra parte— procedente de Zyhlarz, por el descubrimiento debido a don Manuel Gómez-Moreno de un nuevo sistema de lectura conforme al cual los signos ibéricos sólo en parte son alfabéticos —es decir, representación de vocales o consonantes aisladas—, lo que deja un residuo de 15 signos que corresponden a grupos de consonante más vocal.

5. irudia

Gómez Moreno izenaren aipamen bat testuinguruan (Wikibase)

- Mitxelenaren idazlan horretan, adierazitako orrialdean, izenak duen testuingurua. Dagokion izena duen esaldi osoa erauzten saiatzen da erregela bat, adierazitako orrialde zehatzetik (ikus *Gómez-Moreno* izenaren³⁶ aipamen bat Mitxelenaren “El vasco y las lenguas antiguas...” lanean, Wikibase interfazean erakusten den bezala, 5. irudian). Lehenengo saiakeran, 21 633 aipu-erlazioetatik 19 407 testuinguruz hornitu ahal izan dira (% 89,7). Saiakera horretan, entitate gutxi batzuei baino ez zaie grafia alternatiboa duen bigarren izenik gehitu (adib., *Leizarraga*-ri *Leiçarraga* gehitu diogu). Bigarren saiakera egiteko, testuingurua aurkitu gabe geratu diren entitateei sistematikoki gehitu diegu *OC*ko izendegian duten parentesi arteko ezizena (*Txillardegí*, *Orixe*, etab.), eta beste askori abizena grafia alternatibo batean gehitu diegu eskuz (*Álvarez*-i *Alvarez*, *Aquesolo*-ri *Akesolo*, eta *Artetxe*-ri *Arteche*),³⁷ bilaketa gehitutako izenekin errepikatzeko. Oraingo honetan, aipamenen % 91,8 testuinguruarekin deskribatuta dago datu-basean; 131 izen (% 4,5) geratzen dira testuinguru bat ere aurkitu gabe, tartean pertsona-izenak ez diren asko.
- (6) Izena Wikidatan edo euskarazko Wikipedian bilatzea, adierazitako pertsona deskribatzen duen Wikidata entitatea zehazteko. *Open Refine* aplikazioa³⁸ erabili dugu, izen literalak Wikidatako entitateekin lotzeko eskaintzen duen metodo semi-automatikoa eskaintzen duena, baita Wikidatako eta euskarazko Wikipediaren online interfazeak ere. Open Refine kasu honetan mugaturik geratu da oso, ponte-izenik gabe anbigutasun gehiegi baitago hainbat eta hainbat abizenetan (ikus 6. irudian maiztasun handiko *Iturriaga* abizenarentzako lotura-iradokizuak). Hala ere, maiztasun gutxiko abizenei lotzeko hautagaiak Open Refine bitartez bizkor aurkitzea posible izan da (adib. *Haritschelhar*, *J.*).³⁹

36 Gómez Moreno deskribatzen duen entitateak honako helbidea du: <https://wikibase.inguma.eus/entity/Q46308>.

37 Ikus https://wikibase.inguma.eus/wiki/KM_OOCC_izendegi_eta_hiztegiak#OOCC-ko_izendegiko_entitateak_izen_alternatiboekin.

38 Ikus <https://openrefine.org>.

39 Jean Haritschelhar edo Haritxelhar: <http://www.wikidata.org/entity/Q3068794>.



6 .irudia

Open Refine aplikazioan, *Iturriaga* abizenarentzako iradokizunak

Euskarazko Wikipedian izenak bilatzea eraginkorra izan da maiztasun handiko abizenen kasuetan (*Barbero, Wagner, Iturriaga*), abizen bereko hautagai gutxiago ematen duelako Wikidatak baino, eta euskalaritzarekin lotura duen hautesleak euskarazko Wikipedian bere izeneko orrialdea izango duelako. Wikipedian deskribatzen den entitate orok bere identifikatzailea du Wikidatan, baina Wikidatan deskribatzen den entitate guztiek, ordea, ez dute Wikipedian artikulua. Bilaketa bizkortzeko, egokiena izan da pertsonaren ponte-izenak (eta bigarren abizena) *OC*ko agertokietan aurkitzea; kasu gehienetan testuingururen batek izen osoa ematen du, entitatea desanbiguatzen duena. Horretarako, Wikibasen jasotako testuinguruak taula batean jaso ditugu, izen osoa edo desanbiguate-atazan lagungarria izan daitekeen bestelako informazioa bertan bilatu ahal izateko; Wikidatako identifikatzaile egokia lortutakoan, taulan bertan anotatu dugu, gero Wikibasean jasotzeko. Oro har, izen osoa lortuta, pertsona deskribatzen duen Wikidata-entitatea aurkitu egiten da; ikus 7. irudian Wikidatako interfazeko bilaketa-emaitzetan emaitza bat hautatzeko unea.

Manuel Gómez-Moreno
Manuel Gómez-Moreno Martínez Spanish archaeologist and historian (1870–1970)
Manuel Gómez-Moreno González Spanish painter (1834-1918)
Manuel Gómez Moreno botanist

7. irudia

Wikidataren erabiltzaile-interfazean bilaketa-emaizak, entitate egokia eskuz hautatzeko

(7) Emaizak ikusarazteko datu-base galdeketa definitzea. SPARQL lengoaia erabiliz, Wikibaseko edukiak lor daitezke. Galdeketa batek baldintzak eta aldagaiak ditu; baldintzak betetzen dituzten hirukote semantikoek osatuko dute emaitza. Gainera, emaitza hori hainbat bisualizio-eratan erakuts daiteke, taulak, mota anitzeko grafikoak eta mapak tartean. Oraingo honetan, ondorengoak definitu ditugu, eta ikerketa-galderaren arabera, berriak definitu ditzake edonork. (3) adibidekoak, adibidez, Mitxelenak gehien aipatzen dituen pertsonak erakusten ditu: aipatuena, Azkue, eta haren ondoren, Larramendi eta Schuchardt.

- (1) [OCko izendegiko entitateen agerpenak testuinguruan 1](#) (testuak baturik)
- (2) [OCko izendegiko entitateen agerpenak testuinguruan 2](#) (erlazioak banan-banan)
- (3) [OCko izendegiko entitateen agerpenak \(barra-grafikoa\)](#)
- (4) [OCko izendegiko entitateen agerpenak \(grafo bisualizazioa\)](#)
- (5) [OCko lanetan zenbat izendun entitate aipatzen diren](#)

6. Pertsonak Wikidatan

Eskuartean dugun gaiari dagokionez, hau da, Koldo Mitxelenaren lana ezagutzeko bideak zabaltzeari begira, pertsonak deskribatzen dituzten Wikidatako entitateek balio gehigarria dakarte, OC bezalako paperezko argitalpenaren mugak gainditzen dituenak. Jarraian aurkeztuko dugunak txosten hau idazteko unean ditugun datuak ditu oinarri; oraindik, OCko izendegiko pertsonetako batzuk baino ez daude Wikidatarekin lotuta. Zehazki, izendegiko 2 855 entitateetatik, 454 baino ezin izan ditugu Wikidatarekin lerrotatu, hau da, % 16. Mitxelenaren lanean 7 aipamenetik gora duten guztiak multzo horretan daude. % 16 ratio hori handitzeko asmoa dugu ahal bezain laster, aukera izanez gero, % 100era eramateko. Eskuzko lana eskatuko du asmo horrek, pertsona-izenak, goian azaldu bezala, anbiguoak izaten direlako, ponte-izen osoak ez baditugu batez ere. Hala ere, uste dugu 454 horien gainean jada emaitza interesgarri batzuk ditugula, lanean jarraitzea justifikatzeko lain. Hasteko, datu

esanguratsu bat: 454 horiek Wikidatarekin lotzeko, bederatzi entitate baino ez dugu Wikidatan sortu behar izan (% 2), beste guztiak (% 98) aurretik sortuak baitziren.⁴⁰

Wikidatako pertsona-entitate bakoitzak, besteak beste, pertsona hori deskribatzen dituen Wikipedia artikuluetara du lotura. Artikulu horien eduki entziklopedikoa ez dago hirukote semantikoetan egituraturik, eta horregatik ezin da datu-base galdeketen bitartez aztertu. Baina artikulua existitzen diren edo ez bada datu interesgarria, hizkuntza ezberdinetako Wikipedietan hutsuneak zenbatzeko, adibidez, eta horren arabera artikulua sortzeko lehentasunak definitzeko. 2. taulak lau hizkuntzako Wikipedietan ageri diren lehen hamar hutsuneak zerrendatzen ditu, Mitxelenaren lanean duten aipamen-kopuruaren arabera ordenaturik. Euskal Filologiak Wikipedietan duen isla hobetzeko asmoz, horietatik has liteke artikulua sortzeko lana. Pertsonak Mitxelenaren lanean duten aipamen-kopurua Euskal Filologian duten garrantziaren adierazletzat onartzen badugu, ikusten da euskarazko Wikipedian entitate esanguratsuenek artikulua dutela, beste hiruretan aipatuenetariko batzuk ere deskribatu gabe geratzen diren bitartean.

2. taula

Wikipediatako hutsuneak (2025ko urtarrilean): pertsonak (Wikidatan duten izen lehenetsiaz) eta aipamen-kopuruak Mitxelenaren lanetan

<i>Euskarazko Wikipedia</i>	<i>Gaztelaniazko Wikipedia</i>	<i>Ingeleseko Wikipedia</i>	<i>Frantsesezko Wikipedia</i>
Jean-Michel Iribarren (36)	René Lafon (118)	Manuel Larramendi (144)	Nicolaus Landuchius (74)
José de Jesús María Araquistain (35)	Silvain Pouvreau (71)	René Lafon (118)	Esteban de Garibay y Zamalloa (72)
Johannes Hubschmid (30)	Karl Bouda (62)	Luis Villasante (81)	Karl Bouda (62)
Fidel Fita y Colomé (28)	Ángel Irigaray (60)	Nicolaus Landuchius (74)	Ángel Irigaray (60)
Ulrich Schmoll (27)	Julien Vinson (54)	Silvain Pouvreau (71)	Juan Antonio Moguel (56)
Alfred Ernout (25)	Juan Gorostiaga (44)	Karl Bouda (62)	Xabier de Lizardi (54)
Hans Vogt (25)	Gerhard Bähr (43)	Ángel Irigaray (60)	José María Lacarra (50)
Joseph Greenberg (22)	Seber Altube (42)	Joanes Etxeberri (60)	Rafael de Micoleta (49)
Michel Lejeune (21)	Christianus Cornelius Uhlenbeck (42)	Juan Ignacio de Iztueta (56)	Justo Gárate (47)
Luciano Serrano (21)	José de Jesús María Araquistain (35)	José María Lacarra (50)	Pedro Añibarro (42)

40 Hauexek dira sortu behar izan ditudanak: G. N'Diaye, A. Legarda, M. Palomar Lapesa, J. J. Beloqui, T. H. Wilbur, J. J. M. Araquistain, I. Zumalde, F. Echebarria Ibarra, J. Arambillaga. Baliteke pertsona horiek jada Wikidatan egotea, nahiz eta guk ez aurkitu. Horrela balitz, sortu berri dugun entitatea aurretik existitzen zenarekin batzea (bietatik bat egitea, eta baten URIa betiko bestera birbidaltzea) litzateke ohiko jokabidea.

Bestetik, Mitxelenak aipatzen dituen pertsonen datu biografikoak ikus daitezke, tartean bakoitzaren jaiotze-data, jarduera eta abarrekoak. Adibide gisa, aipatutako metadatu horiek biltzeko galdeketa ekartzen dugu hona, 8. irudiko taula ekoizten duena,⁴¹ baita 9. irudia sortzen duena ere, aipaturiko pertsonen jaiolekuak erakusten dituen mapa.⁴² Esan bezala, ikerketa-galderaren arabera, bestelako datu-base galdeketak defini daitezke.

Q45316	Araquistain J. J. M. ^a	 wd:Q132366803	José de Jesús María Araquistain		1700ko urtarrilaren 1a	Itziar	apaiz katoliko; lexikografo
Q45317	Araquistain J. V.	 wd:Q11686614	Juan Benanzio Arakistain	euskal idazlea	1828ko urtarrilaren 1a	Deba	idazle; jabetza-erregistratzaile
Q45315	Araquistain L.	 wd:Q2074856	Luis Araquistáin		1886ko ekainaren 18a	Bárcena de Pie de Concha	idazle; politikari; diplomazialari; kazetari; eleberrigile

8. irudia

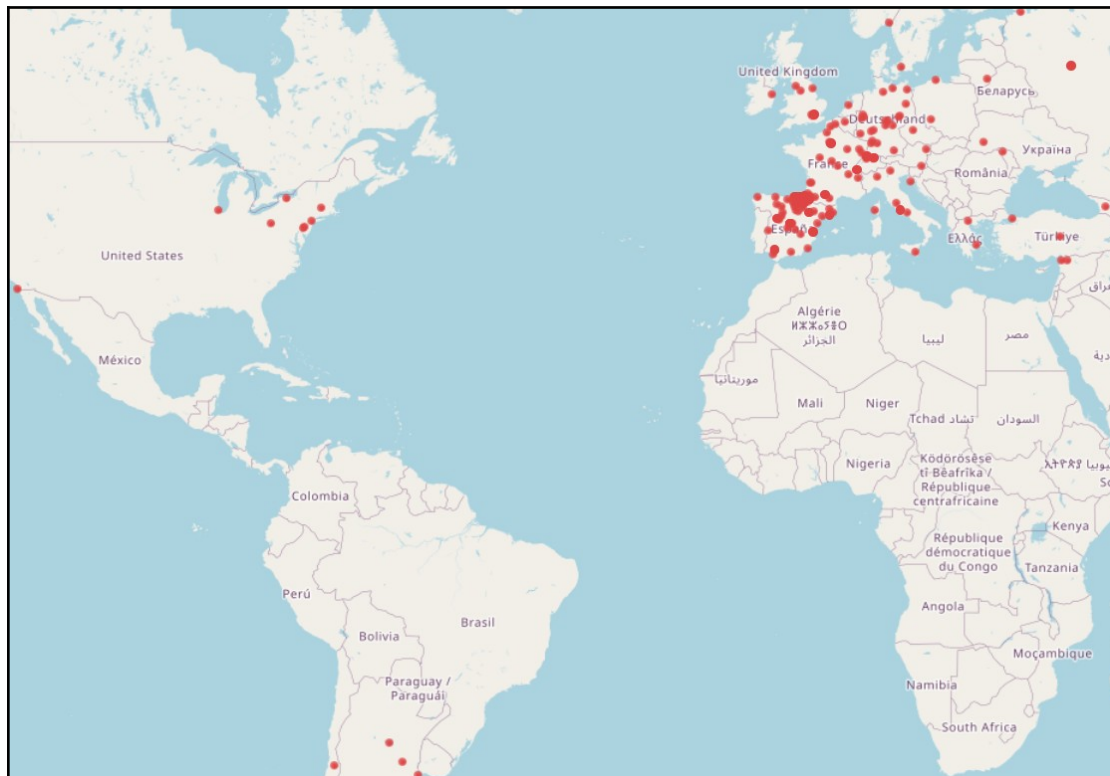
Entitateen metadatu batzuk, Wikidatatik ekarriak

7. Ariketaren emaitza: Mitxelenaren lanetako aurkibideak

Software librea erabiliz, Mitxelenaren lana ezagutu eta aztertzeko bide batzuk zabaltzea izan da aurkeztutako esperimenduaren helburua, baita hura aberasteko ere: Mitxelenaren lanak dituen lotura batzuk esplizitu bihurtu ditugu makinak ulertzen duen eta era sistematikoan datu-base galdeketa bidez esplota daitekeen forma batean. Zehazki, aipamen-erlazioak prozesatu ditugu, bi norabidetan: bi aldizkaritako lanetatik Mitxelenarengana, eta Mitxelerarengandik beste pertsonengana, Euskal Filologiako lehen edo bigarren mailako iturrien egileengana. Ez dugu jakintza berririk sortu, baina bai egin dugu dagoen jakintza aurkigarriagoa, elkarreraginkorragoa, berrerabilgarriagoa.

41 Ikus emaitza guztiak hemen: https://wikibase.inguma.eus/wiki/KM_OOCC_izendegi_eta_hiztegiak#OOCC-ko_izendegiko_wikidatarekin_lerrokatutako_pertsonen_buruzko_datu_batzuk.

42 Ikusi hemen: https://wikibase.inguma.eus/wiki/KM_OOCC_izendegi_eta_hiztegiak#OOCC-ko_izendegiko_wikidatarekin_lerrokatutako_pertsonen_jaiolekuak.



9. irudia

Pertsonen jaiolekuak, Wikidatatik ekarriak

Aipamen-erlazioak landu ditugu hemen, eta esperimentu bat baino ez da izan, hau da, ez ditugu datu-multzo handiak landu; inondik ere ez ditugu Mitxelena aipatzen duten idazlan guztiak batu, bi aldizkaritara mugatu baitugu esperimentera. *OC* izendegia ere ez dugu osorik Wikidatarekin lerrokatu oraindik, baina zazpi aipamenetik gora dituztenak badiugu, oinarri gisa.

8. Ondorioak

Aurrera begira, antzeko metodologiaz egin litezkeen “FAIR-ifikazio” ariketa batzuk aipatzen saiatuko gara, Mitxelenaren idazlanen edukiak aztergai izanik. Baina urrunagora jo baino lehen, aipatu behar dugu *OC*k, izendegiaren antzera, hiztegi-eranskina ere badakarrela, Mitxelenak aztertzen dituen euskal hitzak zerrendatzen dituen, agertokiekin batera. Oso interesgarria izango litzateke hiztegi haren sarrerak agertokiko testuinguruarekin batera identifikatzea eta aurkigarri bihurtzea. Kasu honetan, tamalez, *OC*ren PDF argitalpenak ez dakar erregela tribialen bitartez egituratu litekeen formaturik, sarrerak eta azpisarrerak baititu (eta azpisarrerak interpretatzeko, sarrera nagusiarekin duen erlazioa esplizitu bihurtu beharko litzateke). Edozein kasutan, hau da, indizea eraldatzeak eskuzko lana eskatzen badu ere, norabide horretan pentsatzen jarraitzeak merezi duela uste dugu.

Hemen aurkezten ditugun emaitzak beste euskalari askoren idazlanak indexatzeko berrerabil daitezkeela esan beharrik ez dago, idazlan horietan aipatuko dituzten entitate asko jada jasota ditugu eta. Entitateen identifikatzaileak berrerabiliko genituzke, noski. Horrela, zer entitate (zer pertsona ez ezik, zein lan, zein toki, etab. ere) zein euskalariren lanean eta non aipatzen den neurtu ahalko genuke.

Antzeko metodologiaz lantzeko beste arlo interesgarri bat Mitxelenak eta, printzipioz, euskalaritzan ari den edonork erabiltzen duen Hizkuntzalaritza eta Filologia arloetako terminologia jasotzea litzateke. Izendegiaren kasuan bezala, bi aukera dugu, eta biak konbina daitezke: (a) testuetan terminoen bila abiatzea, terminoen erauzketa automatikoaren bitartez, edo (b) aurrez definituriko termino-bilduma baten sarrerak testuetan bilatzea. Datu-basean, testuak deskribatzen dituzten entitateei ondorengo metadatuak gehituko genizkieke: zer termino agertzen den testuan eta zenbat instantzian, bakoitza bere agertokiari lotuta. Terminok (unitate lexikoak) hizkuntzaz hizkuntza (gaztelaniaz eta euskaraz) lotuko genituzke izendutako *kontzeptuak* deskribatzen dituzten entitate-multzo bat *pibot* gisa erabiliz. Bestetik, terminoek izendatzen dituzten kontzeptuen artean erlazioak ezarri ahalko lirateke, erlazio kontzeptualak, alegia (adibidez: *ergatibo* eta *kasu gramatikal* terminoez izendatutako kontzeptuek duten erlazioa, edo *aitzin-euskara* eta *berreraiketa*). Horrelakorik egiteko lan-fluxuak inplementatu izan ditugu beste nonbait.⁴³ Euskalaritzako argitalpenak helburu horiekin prozesatzea etorkizun hurbileko proposamena izan daiteke. Hainbat ikerketa-galdera erantzuten lagunduko lukeen baliabidea izango litzateke emaitza; honako galderak, esaterako: noiztik erabiltzen den termino jakin bat, zeintzuk diren termino horren itzulpen-ordainak, edo nork erabiltzen duen, eta beste zein terminorekin batera.

Aipatu bezala, PDF formatuko iturriak erabili ditugu aurkeztutako esperimentuetan: *ASJU* eta *Uztaro* aldizkarien testu osoak, alde batetik, eta *OC* PDF bilduma, bestetik. PDF euskarriak mugak ditu: aldizkarian idazten duenak bazekien zehazki Mitxelenaren zer lan aipatu duen, eta *OC*ko indizea egin duenak bazekien zehazki zer entitate aipatzen duen Mitxelenak zer testutan. Informazio hori, ordea, anbiguo geratu da PDF euskarrian. Horrelako galerak ekiditen saiatzea zuzena litzateke, identifikatzaile unibokoak erabiliz. Mitxelenaren lanak aldizkari-artikulu batean aipatzean horrela egiteko, Inguma Wikibaseko identifikatzaileak erabiltzea aukera bat da, edota Wikidatakoa, aipatzen dena *OC*koa bada.

43 Azkenik, ENEOLI proiektuan (*European Network on Lexical Innovation*). ENEOLI Wikibase, proiektuan erabiltzen duguna, honako helbide honetan dago: <https://eneoli.wikibase.cloud>. Aurretik, *Elexis* Horizon proiektuan, ikus <https://lexbib.elex.is>.

Gaur egun OC bezalako bilduma argitaratzean, atal mailako DOI identifikatzaileak hasiera batetik esleitzea bide egokia litzateke, horiek aipamena egitean erabili ahal izateko.

Edizioak prestatzeari begira, horrelako ondorioa plazaratzen dugu: egile-hiztegia, pertsonen indizea, edo beste edozein motatako indizea paperezko argitalpena eskatzen duen formatuan mugatu beharrean datu-base batean jasoko bagenu, erabilgarritasun handia lortuko genuke hasiera batetik, entitateen anbiguotasunik gabe, testuingurua eskuz bilatzeko beharrik gabe, datu-base galdeketa aurreratuak ahalbidetzeko eraldaketa-lanik burutu behar izan gabe. Datu Lotuen paradigmak, gainera, elkarreraginkortasuna bermatzen du, lan honetan erakusten saiatu garen moduan.

9. Erreferentzia bibliografikoak

- Astigarraga, Aitzol, David Lindemann & Marije Bidaguren. 2025. Ziterauzi: euskarazko artikulua akademikoetatik zitazioak erauzteko tresna-katea. In Begoña Altuna, Jon Alkorta, Xabier Arregi, Jose Mari Arriola, Ainara Estarrona, Aritz Farwell, Joseba Fernandez de Landa, Xabier Goenaga & Mikel Iruskieta (arg.), *CLARIAH-EUS. Zientzia Sozialak eta Humanitate Digitalak gaur egun*, 87-99. [Leioa]: EHU. <https://doi.org/10.5281/zenodo.15052532>.
- Bilbao, Gidor, Ricardo Gómez, Joseba A. Lakarra, Julen Manterola, Céline Monoule & Blanca Urgell. 2011. Lazarraga eskuizkribuaren edizioa eta azterketa. *Lazarraga eskuizkribuaren edizioa eta azterketa*, v.1.2. Gasteiz: EHU. <https://www.ehu.eus/monumenta/lazarraga/>.
- Estarrona, Ainara, Izaskun Etxeberria, Ander Soraluze, Ricardo Etxepare & Manuel Padilla-Moyano. 2022. The first annotated corpus of historical Basque. *Digital Scholarship in the Humanities* 37:2. 391-404. <https://doi.org/10.1093/llc/fqab066>.
- Heibi, Ivan, Arianna Moretti, Silvio Peroni & Marta Soricetti. 2024. The OpenCitations Index: Description of a Database Providing Open Citation Data. *Scientometrics* 129 (12): 7923–42. <https://doi.org/10.1007/s11192-024-05160-7>.
- Lindemann, David & Mikel Alonso. 2025. Testu historikoak wiki-plataformetan, Datu Lotu gisa. In Begoña Altuna, Jon Alkorta, Xabier Arregi, Jose Mari Arriola, Ainara Estarrona, Aritz Farwell, Joseba Fernandez de Landa, Xabier Goenaga & Mikel Iruskieta (arg.), *CLARIAH-EUS. Zientzia Sozialak eta Humanitate Digitalak gaur egun*, 101-115. [Leioa]: EHU. <https://doi.org/10.5281/zenodo.15052564>.
- Lindemann, David, Aitzol Astigarraga, Marije Bidaguren, Emilio Delgado, Galder Gonzalez & Kepa Sarasola. 2023. Inguma eta Wikidata uztartuz, euskarazko zientziaren ezagutza-graforantz. In David Lindemann (arg.), *Miren Azkarateri esker onez*, 193-210. [Leioa]: EHU. <https://doi.org/10.5281/zenodo.7902516>.

- Lindemann, David, Marije Bidaguren & Aitzol Astigarraga. 2024. Koldo Mitxelenaren Zitazio-Graforantz. CLARIAH-EUS III. workshopen aurkeztutako posterra. Bilbo, 2024-11/21. <https://doi.org/10.5281/zenodo.14221905>.
- Martin, Sébastien, Muriel Foulonneau & Slim Turki. 2013. 1-5 Stars: Metadata on the openness level of open data sets in Europe. In Emmanouel Garoufallou & Jane Greenberg (arg.), *Metadata and semantics research*, 234-245. Berlin: Springer. https://doi.org/10.1007/978-3-319-03437-9_24.
- Mitxelena, Koldo. 2011. *Obras completas (ASJUren Gehigarriak 54-68)*, 15 lib. Donostia & Gasteiz: “Julio Urkixo” Euskal Filologi Mintegia, Gipuzkoako Foru Aldundia & EHU (Joseba A. Lakarra & Iñigo Ruiz Arzalluzen argit.).
- Passarotti, Marco Carlo & Francesco Mambrini. 2021. Linking Latin: Interoperable lexical resources in the LiLa Project. In Erica Biagetti, Chiara Zanchi & Silvia Luraghi (arg.), *Building new resources for historical linguistics*, 103-124. Pavia: Pavia University Press. <https://zenodo.org/records/5994271>.
- Pérez, Naiara, Julen Etxaniz, Oscar Sainz, *et al.* 2025. Latxa Euskarazko Hizkuntza-Eredua. *Ekaia* ale berezia 2025. 13-27. <https://doi.org/10.1387/ekaia.26338>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3. 160018. <https://doi.org/10.1038/sdata.2016.18>.