REVISTA DE LLENGUA I DRET
JOURNAL OF LANGUAGE AND LAW
#85
www.rld.cat

# INTEROPERABILITY WITHOUT ERASURE: LAW, LANGUAGE, AND THE FUTURE OF THE EUROPEAN DATA SPACE

Joaquín David Rodríguez Álvarez*

## Abstract

This article interrogates the legal and linguistic foundations of interoperability within the European data space, contending that current semantic frameworks risk undermining the European Union's commitment to pluralism. While interoperability is often framed as a technical objective, the paper reveals its deeper implications for legal meaning-making and epistemic justice. Standards like the Data Catalogue Application Profile (DCAT-AP) and the Semantic Interoperability Community (SEMIC), though designed to facilitate cross-border data exchange, embed dominant legal categories that marginalise alternative worldviews, particularly those of minority and indigenous communities. Drawing on legal hermeneutics, semiotics, and decolonial theory, the article proposes a model of federated interoperability grounded in multilingual ontologies, dynamic consent protocols, and relational data governance. It argues that true legal interoperability must move beyond uniformity towards a *jurisgenerative* architecture, one that enables semantic alignment without erasure. In doing so, the European data space can evolve into a digital common that honours the linguistic and legal diversity it purports to serve.

Keywords: multilingual systems; artificial intelligence; semantic protocols; EU law

## *INTEROPERABILITAT SENSE ESBORRAMENT: DRET, LLENGUA I EL FUTUR DE L'ESPAI EUROPEU DE DADES*

### *Resum*

*Aquest article qüestiona els fonaments legals i lingüístics de la interoperabilitat dins l'Espai europeu de dades, i sosté que els marcs semàntics actuals corren el risc de soscavar el compromís de la Unió Europea amb el pluralisme. Si bé la interoperabilitat sovint es planteja com un objectiu tècnic, l'article revela unes implicacions més profundes en la creació de conceptes legals i en la justícia epistèmica. Estàndards com el Perfil d'aplicació del catàleg de dades (DCAT-AP) i la comunitat d'interoperabilitat semàntica (SEMIC), tot i que estan dissenyats per facilitar l'intercanvi transfronterer de dades, incorporen categories legals dominants que marginen visions del món alternatives, particularment les de les comunitats minoritàries i indígenes. Sobre la base de l'hermenèutica legal, la semiòtica i la teoria descolonial, l'article proposa un model d'interoperabilitat federal fonamentat en ontologies multilingües, protocols de consentiment dinàmic i una governança de dades relacional. Argumenta que la veritable interoperabilitat legal ha d'anar més enllà de la uniformitat cap a una arquitectura jurisgenerativa que permeti l'alineació semàntica sense esborrament. D'aquesta manera, l'espai de dades europeu pot evolucionar cap a un bé digital comú que respecti la diversitat lingüística i legal a la qual pretén servir.*

*Paraules clau: sistemes multilingües; intel·ligència artificial; protocols semàntics; dret de la Unió Europea*

* Joaquín David Rodríguez Álvarez, Professor at the Department of Administrative Law, Universitat Autònoma de Barcelona. joaquin.rodriguez@uab.cat 0002-0524-4501

Article received: 23.07.2025. Reviews: 24.11.2025 and 17.12.2025. Final version accepted: 23.01.2026.

**Recommended citation**: Rodríguez Álvarez, Joaquín David. (2026). Interoperability without erasure: Law, language, and the future of the European data space. *Revista de Llengua i Dret, Journal of Language and Law*, *85*, 168–187. https://doi.org/10.58992/rld.i85.2026.4509



CC BY NC ND

## Contents

1 Introduction

2 Theoretical framework: pluralism, semiotics, and relational infrastructure in the European data space

2.1 The semiotics of standardisation: unmasking epistemic violence

2.1.1 Coloniality of knowledge: the epistemic violence of interoperability

2.2 Towards federated pluralism: a synthesised approach

3 Treaties, charters, and regulations, the legal backbone of EU multilingualism

3.1 Institutional and regulatory responses: policy experiments and doctrinal implications

4 The foundational role of linguistics in AI and data systems

5 Conventional management of multilingual systems: architectures and limitations

6 Conclusion: towards a jurisgenerative future for multilingual AI in Europe

6.1 Recommendations

6.2 Limitations and future research

7 References

## 1 Introduction

As the European Union (EU) advances its vision of a unified European data space, the promise of interoperability – seamless data exchange across borders, sectors, and systems – has become a central legal and technological goal.[1] Yet, this pursuit of integration risks reproducing a deeper epistemic and linguistic homogenisation. While data infrastructures are designed to optimise efficiency and standardisation, they often do so at the expense of linguistic diversity, cultural nuance, and minority knowledge systems.

This article argues that the challenge of data interoperability is not merely technical or legal, but fundamentally linguistic. Semantic standards such as DCAT-AP, Core Vocabularies, and the EU's Semantic Interoperability Community (SEMIC)[2] initiative aim to harmonise metadata and ontologies across jurisdictions. However, these frameworks often encode dominant epistemologies and legal categories such as property, consent, or identity, that do not translate easily into minority,[3] indigenous,[4] or non-Western languages and worldviews (Escobar, 2015; 2020). As a result, the very infrastructures meant to enable inclusion may inadvertently erase the linguistic and legal pluralism they claim to support.

This tension brings a particular urgency to the context of language endangerment. Of the world's approximately 7 000 languages, nearly half are at risk of extinction within this century (UNESCO, 2024). Many of these languages are not only vehicles of cultural identity, but also repositories of ecological knowledge, legal traditions, and relational ontologies. Their disappearance is not merely a loss of words, but rather a loss of worlds. As planetary-scale computational systems increasingly mediate global governance (Bratton, 2016), the risk is that endangered languages will be excluded from the data architectures that shape global decision-making (Chuang, 2018; Jutel, 2022; Mejías & Couldry, 2019; Thatcher et al., 2016).

The European data space must therefore evolve beyond a model of semantic standardisation towards one of legal and linguistic co-design. This means embedding multilingual[5] ontologies, dynamic consent protocols, and Indigenous data governance principles into the very fabric of its infrastructure (Poblet et al., 2019). Interoperability must not come at the cost of epistemic flattening. Instead, it should be reimagined as a pluralistic architecture that enables data to flow across systems while respecting the linguistic and legal specificity of the communities it represents.

1 See the ISA² - Interoperability solutions for public administrations, businesses and citizens website.

2 DCAT-AP, or the Data Catalogue Application Profile, is a metadata standard developed by the European Commission to facilitate the sharing and discovery of public sector data across Europe. It is based on the W3C's DCAT vocabulary and is tailored to the needs of European data portals. Its main function is to ensure that datasets published by different countries or institutions can be described in a consistent way, making them easier to find, access, and reuse across borders and sectors. SEMIC, short for Semantic Interoperability Community, is an initiative also led by the European Commission. It focuses on promoting semantic interoperability among EU Member States by developing and maintaining common data models, vocabularies, and guidelines. SEMIC supports the creation of digital public services that can work together seamlessly, even when they are built by different countries or organisations. It provides the conceptual and technical foundations for initiatives like DCAT-AP, ensuring that data exchanged across systems retains its meaning and legal relevance. In essence, DCAT-AP is a practical tool for describing datasets, while SEMIC is the broader community and framework that supports the development of such tools to enable meaningful data exchange across Europe.

3 *Minority language* is used to refer to a language traditionally used within a state by a group of its nationals that is numerically smaller than the rest of the population and different from the state's official language(s); this usage excludes dialects of the official language(s) and the languages of recent migrants. This follows the legal policy sense set out in the *European Charter for Regional or Minority Languages* (Art. 1) and the treatment in the scholarly literature.

4 *Indigenous* is used in line with the United Nations practice: there is no single universal legal definition, but widely accepted criteria emphasise selfidentification; historical continuity with precolonial societies; a distinct cultural, linguistic and institutional identity; and a special attachment to lands and territories.

5 *Multilingual* in the present article refers to the ability to use, support, or provide content in more than one language.

## 2 Theoretical framework: pluralism, semiotics, and relational infrastructure in the European data space

The present theoretical framework aims to synthesise semiotic theory, legal hermeneutics, decolonial epistemology, and Indigenous data sovereignty to conceptualise interoperability without erasure. It posits that data federation must evolve beyond technical standardisation into a legally pluralistic and linguistically just architecture, resisting epistemic homogenisation while enabling cross-system exchange.

### 2.1 The semiotics of standardisation: unmasking epistemic violence

Barthes's analysis in *Mythologies* (2009) reveals that technical standards function as linguistic hegemonies by transforming culturally specific concepts into seemingly universal truths. In doing so, Barthes distinguishes between two layers of meaning:

- denotation: the literal, surface-level definition (e.g., "consent" as agreement to data processing); and
- connotation: the cultural, historical, and political associations embedded within a term (e.g., "consent" implying *individual autonomy*, a cornerstone of Western liberal jurisprudence).

Technical standards such as DCAT-AP and SEMIC operate by elevating mainstream legal-epistemic categories, such as "property", "consent", or "identity", to denotative universality. For instance, when SEMIC defines "property" as an exclusive, transferable right, it presents this Roman-law conception as a neutral global standard, and in so doing erases the connotations of relational stewardship found in Sámi tradition. Similarly, DCAT-AP's "consent" field assumes a transactional agreement between an individual and an institution, rendering invisible the collective decision-making of Indigenous communities, e.g., *siida* governance in Sámi societies.

This process of semiotic flattening, stripping terms of their cultural connotations while naturalising mainstream frameworks, is what Barthes calls "myth". Myth does not lie; it *purifies* historically contingent concepts into timeless facts, making them appear self-evident.

#### *2.1.1 Coloniality of knowledge: the epistemic violence of interoperability*

Walter Mignolo's concept of the coloniality of knowledge (Mignolo, 2007a; Mignolo, 2007b) elucidates the way in which this semiotic flattening perpetuates epistemic hegemony, allowing us to understand how interoperability infrastructures like SEMIC or DCAT-AP:

- universalise mainstream categories, framing common European juridical concepts as globally applicable;
- peripheralise alternative epistemes, forcing non-mainstream worldviews into ill-fitting schemas;
- erase connoted meanings, reducing complex, context-dependent ideas to technical abstractions.

For example, the Sámi term *Eallinbiras*, denoting a holistic stewardship of the living environment encompassing ecological, spiritual, and legal responsibilities, becomes "environment" in a SEMIC ontology. Thus, the jurisprudential depth and relational essence of the term are erased.

This is not mere mistranslation; it is epistemicide (De Sousa Santos & Meneses, 2014), the systematic destruction of alternative knowledge systems through technological standardisation. The erasure of connoted meanings has tangible repercussions such as, for example, the invalidation of Indigenous land claims by geospatial databases that recognise only individual "property" rights, not relational custodianship; or the fact that low-resource languages (e.g., Breton, Sardinian, or Inuktitut) lack the computational tools to express their legal concepts, which accelerates their digital obsolescence.

Liu and Yang's work also critiques the *cultural* dimensions of this dynamic. By highlighting the way in which anthropological field recordings of ethnic minorities were appropriated for sci-fi soundtracks (e.g., *Close Encounters of the Third Kind*), the artists reveal how marginalised communities are doubly commodified: their labour powers energy grids, while their cultural heritage is mined for Western creative capital (Chuang, 2018).

To counter this, data ontologies must be deconstructed and reimagined:

- replacing rigid fields with negotiable boundary objects (Star & Griesemer, 1989): semantic containers that allow divergent concepts (e.g., "property" and *kaitiakitanga*) to coexist without reduction;
- implementing multilingual semantic lifting: using RDF triples or knowledge graphs to map relationships between terms across ontologies while preserving cultural connotations; and
- embedding dynamic hermeneutics: context-aware interpretation protocols where AI systems recognise that "consent" in a Sámi context requires negotiating communal norms.

Accordingly, both Barthes's and Mignolo's theories expose interoperability standards as connotative battlegrounds rather than neutral tools. When the European data space enforces SEMIC or DCAT-AP without critical scrutiny, it risks becoming an engine of epistemic homogenisation. Deconstructing these standards is an ethical imperative: only by centring linguistic and legal pluralism can interoperability avoid becoming another frontier of coloniality: *"Myth deprives the object of which it speaks of all history. […] It organises a world without contradictions because it is without depth" (Barthes, 2009).* The "depth" erased is the connoted sign: the living pulse of cultural meaning.

Antoni Milian i Massana's work on multilingual EU law (Milian i Massana, 2010; Milian-Massana, 2004, 2008) demonstrates the importance of legislating to protect non-official languages. A multilingual reality capable of constructing legal equivalences requires dynamic hermeneutics, not literal translation (Ioriatti, 2023). When data schemas reduce context-dependent legal concepts (e.g., "trust" or "stewardship") to static technical terms, they sever law from its linguistic and cultural lifeworld. Federated data governance must therefore embrace jurisdictional pluralism, wherein legal semantics are negotiated through dialectical engagement with community-specific ontologies.

Building upon Indigenous critiques of data colonialism (Tuck & Yang, 2014; Kukutai & Taylor, 2016), this framework centres on relational accountability as the cornerstone of interoperability. Indigenous data governance models (e.g., the CARE Principles,[6] or the Te Mana Raraunga Charter)[7] reject extractive data

6 The CARE Principles for Indigenous Data Governance are a set of guidelines developed to ensure that data involving Indigenous peoples is handled in a way that respects their rights and interests. CARE stands for:

- Collective benefit: data should be used in ways that support the well-being of Indigenous communities. This includes promoting inclusive development and innovation, and ensuring that data use results in tangible benefits for Indigenous peoples.
- Authority to control: Indigenous peoples must have the right to control the collection, access, and use of data relating to them. This principle supports Indigenous sovereignty and self-determination.
- Responsibility: those handling Indigenous data must act responsibly, ensuring that data is used ethically and with respect for Indigenous values, cultures, and worldviews.
- Ethics: data governance must prioritise the rights and dignity of Indigenous peoples, ensuring that data practices are grounded in ethical frameworks that reflect Indigenous perspectives.

These principles complement the FAIR (Findable, Accessible, Interoperable, Reusable) principles by emphasising people and purpose over mere data management. They were developed by the Global Indigenous Data Alliance (GIDA) to promote Indigenous data sovereignty.

7 The Te Mana Raraunga Charter is a foundational document created by the Māori Data Sovereignty Network in *Aotearoa*/New Zealand. It outlines the principles and values that guide the governance and use of Māori data. The Charter emphasises that data is a living *tāonga* (treasure) and holds strategic value for Māori communities.

There follows a brief summary of its key principles:

practices by embedding protocols for dynamic consent, collective benefit, and intergenerational ethics. Integrating these principles into the European data space necessitates infrastructural support for multilingual ontologies, i.e., semantic networks that preserve epistemic nuance across languages (Fernández-Llamazares et al., 2024). Such an approach aligns with Escobar's (2020) call for pluriversal[8] design, where technology sustains, rather than erases, diverse legal-linguistic worlds.

### 2.2 Towards federated pluralism: a synthesised approach

The synthesis of these strands yields a vision of interoperability as relational translation, rejecting the false binary between fragmentation and homogenisation by positioning data federation as a continuous hermeneutic practice. Legal categories become negotiable boundary objects (Star & Griesemer, 1989) translated contextually across systems while retaining community-specific meanings. Infrastructure itself transforms into a federated legal-linguistic ecosystem, embedding protocols for:

- multilingual semantic lifting: mapping concepts across ontologies without reduction;
- dynamic consent gateways: enabling granular data reuse permissions tied to linguistic context; and
- relational stewardship: distributing governance across epistemic communities.

Therefore, law should present itself as interoperability infrastructure that reimagines interoperability as technical alignment rather than jurisgenerative[9] practice (Cover, 1983), whereby data infrastructures become sites for sustaining legal-linguistic plurality. By centring on these approaches, the European data space can evolve into a pluriversal federation, wherein data flows without erasure, and difference is the condition of connection. Before proceeding, however, it is necessary to outline the legal architecture that frames multilingualism as a fundamental European value.

## 3 Treaties, charters, and regulations, the legal backbone of EU multilingualism

The EU's commitment to linguistic diversity forms a constitutive element of its legal identity, articulated through foundational treaties, binding charters, and specialised regulations that collectively establish a robust, though imperfect, framework for preserving multilingualism. This architecture has evolved to address

- *Rangatiratanga* (authority): Māori have the right to control the collection, access, and use of their data. This reflects self-determination and jurisdiction over their own information.
- *Whakapapa* (relationships): data must be understood in context, respecting its origins and connections. This includes disaggregating data to reflect diverse *iwi* and *hapū* identities and ensuring its future relevance.
- *Whanaungatanga* (obligations): There is a responsibility to balance rights and ensure accountability in data use, especially when it involves Māori communities.
- *Kotahitanga* (collective benefit): data should be used to benefit Māori collectives, build capacity, and strengthen connections within and across communities.
- *Manaakitanga* (reciprocity): data practices must be respectful, requiring informed consent and ethical engagement.
- *Kaitiakitanga* (guardianship): Māori are the guardians of their data, and its use must align with their ethical standards and cultural values.

The Charter aligns with international Indigenous data sovereignty movements and complements frameworks such as the CARE Principles.

8 *Pluriversal* refers to the idea that multiple worlds, ontologies, and ways of knowing can coexist without being reduced to a single universal framework. It challenges the modern Western notion of a singular, dominant, "universal" worldview, and instead affirms the legitimacy of diverse epistemologies, especially those of Indigenous, non-Western, and marginalised communities.

9 The term *jurisgenerative* comes from legal scholar Robert Cover, who introduced it in his 1983 essay *Nomos and Narrative*. In essence, jurisgenerative refers to the capacity of communities to create, interpret, and sustain their own legal meanings and normative worlds, outside or alongside formal state law.

emergent challenges in the digital and artificial intelligence landscape, creating intersecting obligations for institutions, developers, and Member States.

The Treaty on European Union (TEU) enshrines respect for "rich cultural and linguistic diversity" as a constitutional obligation under Article 3(3), framing it as intrinsic to the Union's identity (European Union, 2007). This principle is operationalised through Article 22 of the Charter of Fundamental Rights, which explicitly mandates respect for linguistic diversity alongside cultural and religious pluralism (European Union, 2012). Crucially, Article 21 of this Charter prohibits discrimination based on language, establishing negative and positive obligations: Member States must refrain from linguistically discriminatory acts while actively ensuring equitable treatment. Article 165(2) of the consolidated version of the Treaty on the Functioning of the European Union (TFEU) emphasises that "Union action shall be aimed at developing the European dimension in education, particularly through the teaching and dissemination of the languages of the Member States", with Article 165(1) stating this by done "while fully respecting […] cultural and linguistic diversity" (2016).

Regulation No 1/1958 (as amended) codifies the practical manifestation of these principles by designating 24 official languages and guaranteeing citizens the right to communicate with EU institutions in any official language, with replies provided in the same language. However, this protection remains circumscribed, as non-official languages, such as Catalan, Basque, Galician, Sámi, and Sardinian, receive markedly weaker safeguards under EU law. Their protection derives primarily from national legislation or voluntary initiatives like the European Language Equality Network (ELEN), rather than supranational mandates (Commissioner for Human Rights, 2019).[10] The Council of Europe's European Charter for Regional or Minority Languages (ECRML) (Council of Europe, 1998) and Framework Convention for National Minorities (FCNM) (Council of Europe, 1992) provide additional instruments, but their enforcement mechanisms remain intergovernmental rather than integrated into EU competence.

When it comes to the digital landscape of the Union, especially in relation to artificial intelligence (AI), the AI Act become the referral text (Shaping Europe's digital future, 2025), adopted on 21 May 2024 (Regulation (EU) 2024/1689), establishing what is often described as the first comprehensive regulatory framework for artificial intelligence, implementing a risk-stratified model with direct implications for linguistic diversity, as its provisions intersect with language rights across three dimensions:

- Transparency obligations (Art. 52). Generative AI systems, including multilingual large language models (LLMs) such as ChatGPT must disclose training data sources and label AI-generated content. This enables users to identify machine-generated translations or summaries, preserving accountability in multilingual outputs. A trend that is materialised through initiatives such as the European Digital Infrastructure Consortium for the Alliance for Language Technologies (ALT-EDIC) (Commission Implementing Decision (EU) 2024/458)[11] and the Language Data Space (European Language Data Space, n.d.-b), for low-resource languages where training data scarcity exacerbates error risks, this transparency facilitates error detection and correction (Shaping Europe's digital future, 2025).
- High-risk systems (Annex III). Artificial intelligence deployed in contexts where linguistic errors could violate fundamental rights – such as legal interpretation, migration control, public services, or employment screening – faces stringent requirements. Providers must ensure high-quality datasets, rigorous bias mitigation protocols, and human oversight mechanisms. These obligations

10 See the European Language Equality Network (ELEN) website.

11 See the Alliance for Language Technologies EDIC (ALT-EDIC) website.

are particularly critical for languages like Irish or Maltese, where limited digital corpora increase algorithmic distortion risks.

- General-purpose AI (Article 53). Providers of foundational models must document training data and comply with copyright law, with enhanced evaluations for models deemed systemically risky. The systemic risk designation, determined by the AI Office based on capability benchmarks, implicitly prioritises multilingual models whose outputs could disproportionately affect minority language communities.

Complementing the AI Act, the EU has developed specialised tools to operationalise linguistic diversity in digital infrastructures:

- eTranslation. A secure hybrid AI system funded under the Digital Europe Programme (Regulation (EU) 2021/694) that processes texts between all 24 official languages plus Ukrainian, Icelandic, Norwegian, with Moldovan and Serbian to be added in the near future. Unlike commercial tools, it prioritises low-resource languages (e.g., Irish or Maltese) by allocating computational resources to minimise accuracy gaps. Data processing occurs within EU data centres, with inputs deleted within 24 hours to align with Article 8 of the Charter (data protection) and the General Data Protection Regulation (GDPR).
- Language Data Space (LDS). Launched through ALT-EDIC in 2024, this initiative addresses the acute shortage of European language data for training LLMs. By creating a centralised marketplace for multilingual data, it aims to counter the dominance of English-centric corpora and prevent digital minoritisation. Seventeen Member States participate directly, with nine others as observers, reflecting a pan-European commitment to linguistic sovereignty (European Language Data Space, n.d.-b).
- Interoperability standards: The SEMIC Core Vocabularies and DCAT-AP metadata frameworks facilitate cross-border data exchange. Future implementing acts or guidance could require dynamic ontology mapping for highrisk systems, ensuring that semantic standards can accommodate culturally specific concepts such as *Eallinbiras* (holistic stewardship of the living environment) without epistemic erasure during translation.

Furthermore, the European Parliament consistently amplifies linguistic protections through resolutions and programmatic interventions. For example, the European Parliament Resolution of 11 September 2018 on language equality in the digital age demanded a coordinated strategy for multilingual digital markets, urging investment in language technologies for marginalised languages. The Creative Europe Programme (Regulation (EU) 2021/818) explicitly funds translations of literary works under Article 3, linking linguistic diversity to cultural preservation.

Despite this architecture, structural challenges persist:

- The non-official language paradox: while instruments like the Minority SafePack Initiative[12] receive parliamentary support, EU competence remains constrained (Milian-Massana, 2004; 2008). Protections for languages such as Catalan or Sámi depend on Member State discretion, creating a fragmented landscape (European Citizens' Initiative, n.d.).
- The AI literacy deficit: Article 4 of the AI Act mandates literacy programmes to ensure informed human oversight, but resources for minority-language communities remain underdeveloped.

12 See the Minority SafePack Initiative website.

- The Court of Justice of the EU (CJEU) increasingly references the Charter in linguistic discrimination cases, but its jurisdiction remains bounded by subsidiarity, it cannot adjudicate purely domestic language policies.

Therefore, we can understand that the EU's legal architecture embeds linguistic diversity as a non-negotiable value across legislative, executive, and judicial domains. By integrating explicit safeguards into the AI Act and supporting data infrastructures like the LDS, it acknowledges that algorithmic systems are not linguistically neutral but rather shape epistemic hierarchies. However, the subordination of non-official languages to national prerogatives reveals a constitutive tension (Milian-Massana, 2008): the EU protects *state* languages vigorously while leaving *stateless* languages[13] vulnerable. Future efficacy hinges on strengthening cross-border enforcement for minority languages and allocating Digital Europe funding to accelerate LLM training for marginalised linguistic communities. Only through such measures can "unity in diversity" transcend motto to materialise as digital reality.

### 3.1 Institutional and regulatory responses: policy experiments and doctrinal implications

Across Europe, linguistic justice in AI is no longer merely a normative claim but an emerging infrastructure and governance agenda. Two vectors are especially salient: (i) public AI infrastructures built to secure multilingual parity, and (ii) languagestatus decisions within the EU that condition data pipelines, tooling, and institutional uptake. Read through established legaltech doctrine, these choices determine whether "interoperability" supports pluralism, or enacts erasure (Lessig, 2003).

With respect to the first vector – public multilingual AI infrastructures – it is worth underscoring Spain's ALIA initiative,[14] coordinated by the Barcelona Supercomputing Center (BSC) and fully publicly funded. The initiative aims to deliver open, transparent language models for Spanish and the coofficial languages (Catalan, Basque, and Galician), with verification by the Spanish Agency for the Supervision of Artificial Intelligence (AESIA) and alignment with the EU AI Act's transparency ethos. In 2025, the BSC announced ALIA40B, a foundation model trained on MareNostrum 5 across 35 European languages, and released models and training assets under permissive licences, enabling scrutiny and downstream finetuning by public bodies and SMEs (BSC, 2025). These features embed linguistic equity by design, rather than treating it as a byproduct of scale. ALIA's public portal and communications emphasise universal access for administrations, universities, and companies; meanwhile, the Interoperable Europe portal documents model availability and early pilots, reinforcing the project's aim to democratise multilingual AI resources (Vilas Gómez, 2025).

However, that promise must be paired with realism about current limitations. The ALIA40B model card explicitly cautions that the base model has not undergone instruction tuning and may produce biased or unsafe alignment outputs; early thirdparty testing notes uneven guardrails, shorter context windows, and variable performance in complex tasks, precisely the kind of "known unknowns" that open releases help the community to diagnose (BSC, 2025). Beyond national projects, ALTEDIC offers a panEuropean scaffold that will federate language and multimodal data across Member States, develop and evaluate multilingual models, and prioritise languages with fewer than 10 million speakers, in tandem with the LDS.[15] As such, ALTEDIC complements ALIA's national focus with crossborder datasets, governance, and certification.

13 *Stateless language* is used to refer descriptively to a language without a sovereign nationstate in which it is the statewide official language, even if it enjoys coofficial status in substate jurisdictions (e.g., within an autonomous community). In Europe, many such languages fall under the European Charter for Regional or Minority Languages, which protects languages "traditionally used" by nationals and distinct from the state's official language(s); the Charter makes clear that languages may be official regionally without being official at the state level.

14 See the ALIA: The public AI infrastructure in Spanish and co-official languages website.

15 See the European Language Data Space (LDS) website.

Secondly, with respect to the other vector – official status and the EU language regime – it is essential to recognise that decisions on legal status shape incentives in ways that no purely technical solution can. An amendment to Regulation No 1/1958 to include, for example, Catalan, Basque and Galician, would require unanimity in Council. Since Spain tabled the proposal in 2023, repeated debates have ended without consensus due to legal and budgetary reservations, despite Madrid's offer to fund costs. Council documentation explains the unanimity requirement and current administrative arrangements; press reporting through 2025 records ongoing postponements and a split among capitals (Hernández-Morales, 2025). The practical consequence is immediate: status outcomes influence translation or interpretation commitments, terminological databases, procurement defaults and evaluation focus, thereby modulating what data are curated and which models are tuned for institutional use.

These institutional moves sit squarely within classic and contemporary techlaw doctrine. As Lawrence Lessig argues, digital architecture regulates: "code is law", and values can be embedded or displaced through design. In the AI context, the relevant architectures are models and datasets, hence the necessity of law regulating code regulating law (i.e., legal obligations that shape technical systems which, in turn, mediate legal meaning). In parallel, Floridi (2026) demonstrates that AI constitutes a new form of agency that must be ethically governed, via both binding rules and softethics instruments that translate principles into design choices; in multilingual settings, this requires attention to dataset composition, representation harms, and power asymmetries rather than assuming neutrality. Similarly, Veale and Zuiderveen Borgesius (2021) show that the AI Act's riskbased regime fails to adress enforcement and harmonisation challenges; for language rights, this implies that linguistic equity must be operationalised (e.g., perlanguage data quality, bias reporting, and evaluation) rather than presumed.

Finally, two influential studies underline why these guardrails matter. Bender et al. (2021) caution that everlarger language models amplify environmental, financial and representational harms to absent robust curation and documentation; open, auditable model releases are one way to mitigate risks. While Joshi et al. (2021) document the systematic underrepresentation of most of the world's estimated 7 000 languages across datasets, models and conference outputs, challenging "languageagnostic" claims and urging prioritised support for lowresource languages.

Therefore, the main implication for this article of these policy experiments and doctrinal implications is that, if interoperability is to avoid erasure, the policy experiments and doctrinal insights described above must be made operational in the European data space. This would entail (a) the tying of access to EU infrastructures (e.g., ALTEDIC/LDS) to languageequity deliverables and published perlanguage benchmarks; (b) the incorporation of linguistic equity audits in highrisk deployments under the AI Act; and (c) recognition that status law (e.g., Regulation No 1/1958) and technology law (e.g., AI Act, data spaces) converge at the point where code regulates legal meaning. Without these guardrails, the "semantic lifting" celebrated by data engineers will continue to elevate some languages while dropping the rest. But why is it so important to discuss multilingualism in the age of AI?

## 4 The foundational role of linguistics in AI and data systems

Linguistics is indispensable to artificial intelligence and data management because language constitutes the primary medium through which AI interprets reality and data acquires meaning, especially when it comes to semantic layered technologies. AI systems, particularly LLMs and knowledge graphs, depend fundamentally on linguistic structures to parse, classify, and generate information (Kumar, 2024). When these systems process legal documents, environmental records, or cultural archives, they do not engage with raw "data", but with linguistically mediated representations of human knowledge. Consequently, the semantic frameworks embedded in AI – such as ontologies, taxonomies, and entity relationships – are

inherently shaped by the grammatical logic, lexical hierarchies, and epistemic assumptions of the languages in which they are trained (Pragya, 2025). Barthes's (2009) semiotic insights reveal that these frameworks function as connotative systems, silently privileging certain worldviews (e.g., Western juridical concepts of "ownership") while marginalising others (e.g., Indigenous notions of "custodianship"). Without rigorous linguistic scrutiny, as outlined above, AI perpetuates what De Sousa Santos and Meneses term *epistemicide*: the systematic erasure of alternative knowledge systems through technological standardisation (2014).

Therefore, maintaining multilingual infrastructures is not merely a technical obstacle but a profound sociotechnical dilemma rooted in the tension between interoperability and epistemic diversity. First, semantic untranslatability (Darawsheh et al., 2025) poses a core challenge: many legal and cultural concepts are linguistically embedded and resist direct mapping across ontologies. For instance, the Catalan civil law term *pàtria potestas* (parental authority) or the Sámi concept of *Eallinbiras* (holistic stewardship of the living environment) cannot be reduced to English-centric equivalents without significant hermeneutic loss (Huang & Wang, 2024).

Second, asymmetries in computational resources exacerbate linguistic injustice. While dominant languages (e.g., English and French) benefit from vast training datasets and sophisticated natural language processing (NLP) tools, endangered and minority languages lack the corpora necessary for robust AI modelling, creating a feedback loop of digital obsolescence. Moreover, UNESCO (2024) notes that 50 % of global languages face extinction, a crisis that will likely be accelerated by AI's linguistic biases. Third, infrastructural path dependence locks in colonial hierarchies. Existing data standards (e.g., Schema.org, DCAT-AP) embed monolingual assumptions, forcing multilingual systems into cumbersome workarounds such as "tagging" minority languages as metadata rather than treating them as primary semantic frameworks. This reduces linguistic diversity to a technical afterthought rather than a design imperative.

Thus, AI exacerbates these challenges by operationalising linguistic bias at scale. Natural language processing tools trained on hegemonic language corpora inevitably reproduce syntactic and semantic norms that exclude dialectal variations, oral traditions, and context-dependent meanings. For example, in the context of New Zealand, where a great effort is under way on the part of the Māori community to preserve their cultural heritage in the digital age, studies have found that translation systems often flatten juridical nuance, converting the relational Māori term *kaitiakitanga* (guardianship entailing spiritual reciprocity) into the transactional English "guardianship", thus distorting Indigenous legal philosophies (Ruckstuhl, 2023).

Accordingly, we can assume that automated semantic alignment in knowledge graphs prioritises computational efficiency over cultural fidelity, coercing diverse epistemes into reductive Western categories (e.g., classifying the Basque *auzolan* or communal labour as "volunteer work"). Generative AI hallucinates or misrepresents low-resource languages due to training data scarcity, further alienating linguistic communities from data sovereignty.

These issues expose a central paradox: the very systems designed to "unify" data across languages often function as engines of epistemic homogenisation. As Mignolo (2007a) argues, such technologies reinforce a "monolingual paradigm of modernity", wherein interoperability becomes synonymous with cultural assimilation.

However, addressing these challenges will require a reconceptualisation of data systems as multilingual legal-linguistic ecosystems that includes:

- the embedding of dynamic hermeneutics by adopting a model of context-aware legal interpretation, where AI ontologies support fluid concept mapping across languages without reductionism;

- relational data stewardship, whereby Indigenous principles like the CARE guidelines are integrated to ensure data governance respects linguistic context, collective consent, and intergenerational ethics; and
- pluriversal semantic architectures, developed through interoperable infrastructures that treat each language as a distinct epistemic jurisdiction, with "boundary objects" enabling negotiation across ontological differences.

In essence, linguistic diversity in AI is not so much a localisation problem as it is a question of epistemic democracy. Only by centring linguistics as core infrastructure, not an add-on, can data systems enable interoperability without erasure.

## 5 Conventional management of multilingual systems: architectures and limitations

Multilingual systems, such as those deployed by supranational entities like the European Union or global technology platforms, typically operate through hierarchical governance models that prioritise operational efficiency over linguistic equity (Laaksonen, 2022). These systems generally adhere to three interconnected management paradigms, each of which entail significant epistemic and practical constraints (Tan et al., 2020).

The most prevalent approach is centralised standardisation with peripheral translation. This model designates a dominant language, typically English, as the pivotal schema for core ontologies, data structures, and legal taxonomies. Minority languages are then accommodated through post-hoc translation layers. For instance, the EU's Interactive Terminology for Europe (IATE) database[16] – the official terminology repository for EU institutions since 2004 – relies primarily on English and French as source languages for structuring legal concepts, with other official EU languages treated as derivative targets. This is evident in IATE's architecture, where EU-specific terms (e.g., *acquis communautaire*) are first defined in English/French before multilingual dissemination, embedding Anglo-Francophone juridical assumptions into all translations. Similarly, machine translation systems like the EU's eTranslation service exacerbate this hierarchy through linguistic cascading (e.g., Finnish→English→Greek), a process that leverages neural models trained on institutional translation memories but inherently amplifies semantic drift.[17] As confirmed by technical documentation, eTranslation processes over 200 million pages annually using this pivot-language approach, prioritising computational efficiency despite known accuracy trade-offs for non-pivot languages.[18] This architecture reinforces linguistic hegemony by masking cultural assumptions embedded in the pivotal language, a tension that is codified in EU law. While Regulation No 1/1958 mandates equality among 24 official languages, internal workflows permit drafting in "procedural languages" (chiefly English/French), creating a *de facto* linguistic stratification that conflates operational pragmatism with epistemic neutrality. As Barthes's (2009) semiotic analysis reveals, the apparent neutrality of such technical standards functions as a modern myth, a dynamic particularly visible in metadata ecosystems like Schema.org,[19] where English-centric ontologies force non-English content into predefined semantic hierarchies that obscure culturally distinct knowledge structures.

A second common strategy relies on parallel corpora and static equivalence, institutionalised through juridical-linguistic governance frameworks. Institutions like the United Nations and European Union maintain legally binding multilingual texts under principles such as Regulation No 1/1958 in the case of the EU, which

16 See the IATE homepage for the standard view of the search engine.

17 See the Digital Europe - AI-based multilingual services website.

18 Read more about the eTranslation project in the LuxProvide website.

19 See the Schema.org website.

mandates equal authenticity across all official languages. This operationalises what Curran (2019) identifies as the "decoding of legal presences" in transnational systems, treating linguistic diversity as a procedural challenge rather than a substantive value.

Governance mechanisms include mandatory juridical-linguistic revision, exemplified by the United Nations (UN) sentence-aligned parallel corpus (1990–2014), containing over 20 million sentence pairs across six languages, all manually translated to ensure denotative parity.[20] This process employs controlled vocabularies (e.g., IATE terminology database) wherein terms such as "consumer" become *consument* in Dutch or *καταναλωτής* in Greek, enforcing lexical conformity while reducing linguistic variation to metadata tags (e.g., "pt-PT" for European Portuguese) rather than integral semantic infrastructure.

The critical limitation, as Sacco's (1991) legal formants theory reveals, is the systematic privileging of denotative equivalence over connotative meaning, erasing culturally embedded epistemologies that resist lexical standardisation. When the Sámi concept *Eallinbiras* (holistic stewardship of the living environment) is reduced to "environment" in EU environmental directives, the jurisprudential depth of the concept, encompassing human-non-human relational ethics, is erased – as observed in the LingLaw Project's analysis of the Services Directive's transposition (CORDIS, 2024).

This exemplifies De Sousa Santos and Meneses's (2014) epistemicide, which the EU inadvertently perpetuates despite Article 3 of the TEU's "unity in diversity" principle. As the Semantic Spaces project (CORDIS, 2024) notes, such reductionism arises because EU legal concepts, detached from national cultural substrates, function as "hybrid" constructs that prioritise harmonisation over ontological richness.

Technological platforms frequently deploy a third paradigm: modular localisation, which decouples linguistic layers from operational infrastructure through abstraction frameworks like .PO (portable object) files in software development. This architecture enables third-party translation but relies on asymmetric validation ecosystems where crowdsourcing platforms (e.g., Wikipedia, Mozilla) exhibit pronounced resource disparities: Breton localisers must navigate three to four moderation tiers, whereas English commands dedicated NLP-assisted quality control agents.

Algorithmic disambiguation compounds the inequity: multilingual transformers like Helium 1 resolve polysemy through statistical prevalence in training corpora, where Sardinian constitutes <0.1 % of the 2TB EU dataset, despite equal official status. This operationalises digital minoritisation by treating languages as cosmetic overlays rather than integral semantic systems. Consequently, endangered languages face computational resource deserts: Breton lacks morphological analysers for verb conjugation in .PO file contexts; Sardinian's absence from lexical databases such as IATE forces ad hoc neologisms for technical terms.

The EU's Digital Language Diversity Project identifies this as a *vicious cycle* where inadequate tools depress digital usage, which in turn starves NLP training corpora, accelerating obsolescence. Survey data reveals that 74 % of Sardinian speakers avoid written digital communication due to unstandardised orthography and a lack of spell-checkers (Russo & Soria, 2017), while Karelian users report psychological barriers from stigmatised digital interactions (Salonen, 2017). This epistemic exclusion mirrors colonial linguistic hierarchies, as platforms prioritise "localisability efficiency" over pluriversal knowledge systems.

These conventional models share fundamental flaws that undermine linguistic justice. Epistemic reductionism compresses nuanced legal and cultural concepts into the granularity of dominant ontologies. Hermeneutic rigidity prioritises lexical alignment over contextual meaning. The Catalan civil law term *coses comunes* (denoting indivisible common elements in co-owned property) maps mechanically to "common

20 See the United Nations Parallel Corpus website.

areas" in anglophone systems, obscuring its distinct legal history. And finally, resource asymmetry creates self-perpetuating inequities: UNESCO documents that over 50 % of online content exists in just 20 languages, while AI training data for Basque or Irish Gaelic remains critically scarce. This scarcity entrenches a vicious cycle where under-resourced languages become computationally invisible.

Notwithstanding, progressive frameworks are challenging these paradigms by centring linguistic justice as infrastructural imperative. Dynamic ontology mapping, exemplified by the Open Language Archives Community,[21] enables cross-lingual concept alignment without pivotal languages, using web technologies such as semantic or RDF triples to preserve context. Community-driven stewardship models, such as those advanced by the Māori Data Sovereignty Network,[22] assert the autonomy of Indigenous languages: data in *te reo Māori* – the Māori language – operates under its own semantic rules. Pluriversal AI initiatives such as Masakhane reject pivot-language dependencies, developing neural machine translation for African languages through participatory research, defining themselves as "a grassroots NLP community for Africa, by Africans".[23] The CARE Principles for Indigenous Data Governance further provide ethical scaffolding, emphasising collective benefit and relational accountability.

But the route is still long, and conventional multilingual management remains constrained by a monolingual logic of control. Transcending this requires the reimagining of systems as federated ecosystems of meaning, where interoperability emerges from continuous hermeneutic negotiation rather than static equivalence. As legal scholar Robert Cover (1983) theorises, law flourishes in "jurisgenerative" spaces where communities co-create normativity. Data infrastructures must evolve into such spaces, enabling worlds to converse without surrender, and languages to flow without erasure. This entails the embedding of relational protocols – multilingual semantic lifting, dynamic consent gateways, and distributed epistemic stewardship – into the architectural core of systems like the European data space. Only through such structural pluralism can interoperability uphold its promise: connection without conquest, and unity without uniformity.

## 6 Conclusion: towards a jurisgenerative future for multilingual AI in Europe

This article has demonstrated that what travels under the banner of "interoperability" in the European data space is neither neutral nor merely technical. It is a juridicallinguistic practice whose prevailing implementation tends to privilege denotative alignment over connotative depth, producing semiotic flattening of culturally embedded legal concepts such as *property*, *consent*, *stewardship,* or *identity*. Read through legal hermeneutics and decolonial theory, this flattening operates as a vector of epistemic homogenisation, up to and including epistemicide, whenever semantic standards, knowledge graphs and AI systems normalise one worldview while relegating others to metadata or "edge cases". In this way, multilingual ontologies are not auxiliary tools, but rather sites where legal meaning is made, with direct consequences for access to rights, accountability and remedy.

At the same time, the Union's constitutional and regulatory architecture already encodes pluralism: the Treaties, the Charter, Regulation No 1/1958 and the AI Act collectively articulate linguistic diversity as a constitutive value, even if protection remains uneven for nonstateofficial and stateless languages. Against this backdrop, the article offers a jurisgenerative reading of interoperability: rather than enforcing uniformity, infrastructures can be designed to sustain negotiated copresence, through multilingual semantic lifting, dynamic hermeneutics and relational data stewardship, so that data flows without erasure and legal categories travel across communities without being reduced to their dominant equivalents.

21 See the Open Language Archives Community website.

22 See the Te Mana Raraunga website.

23 See the Masakhane website.

Moreover, the EU's commitment to "unity in diversity" must transcend rhetorical invocation and materialise as digital multilingualism by design. This requires structural investment in low-resource language technologies, legal recognition of stateless languages within AI governance frameworks, and the institutionalisation of community-led data protocols. Initiatives like the LDS and ALT-EDIC mark important steps, but their efficacy hinges on whether they treat linguistic diversity as a constitutive value or a peripheral concern.

Ultimately, the future of European AI must be pluriversal: a federated ecosystem where data flows without erasure, and where semantic interoperability is achieved not through reduction, but through negotiated co-presence. In such a future, languages are not filtered through dominant ontologies, but rather recognised as epistemic jurisdictions in their own right. As Lessig (2003) famously argues, "code is law" and, in the age of AI, the architectures of interoperability are not neutral tools but normative frameworks that shape whose knowledge counts. Only by embedding legal and linguistic pluralism into the code itself can the European project fulfil its promise and become a multilingual polity where difference is the condition of connection, as opposed to a monolingual technocracy.

## 6.1 Recommendations

To reconfigure the European data space as a pluriversal federation that resists epistemic erasure, substantive structural reforms are imperative. First, the AI Act requires targeted amendments to enforce epistemic pluralism. Article 53 should mandate dynamic ontology mapping for high-risk AI systems, compelling foundational models to preserve culturally embedded concepts, such as the Sámi *Eallinbiras,* as distinct semantic classes within knowledge graphs, rather than subsuming them under reductionist Western categories. This necessitates the implementation of negotiable boundary objects (Star & Griesemer, 1989) to enable cross-lingual alignment; for instance, linking "property" to *kaitiakitanga* via contextual RDF predicates such as *owl:equivalentProperty* with provenance annotations. Concurrently, Article 52 must expand to enforce linguistic equity audits, requiring transparency reports that quantify algorithmic performance gaps for low-resource languages like Irish or Maltese, with levied fines directly funding corrective R&D through the Digital Europe Programme.

Second, the Language Data Space must evolve into a federated ecosystem centred on relational governance. Devolving authority to epistemic communities, such as the Sámi Parliament or the Institute of Catalan Studies (IEC), would empower them to curate language-specific ontologies using RDF/SKOS[24] frameworks that preserve connotative meanings. For example, encoding *Eallinbiras* as an intersectional class combining *dct:environment* and custom *sami:relational_stewardship* properties. This architecture must integrate dynamic consent gateways: smart contracts that enforce communal decision-making protocols, such as triggering *siida*-based approvals for Sámi ecological data reuse. SEMIC Core Vocabularies should be retooled to support multilingual semantic lifting, allowing DCAT-AP fields to resolve to community-defined concepts through *skos:closeMatch* relationships, thus avoiding connotative flattening.

Third, epistemic reparations require funding and compliance mechanisms to be institutionalised. A greater percentage of Digital Europe Programme budgets (Regulation (EU) 2021/694) should be allocated to developing NLP tools for critically endangered languages – such as Sardinian morphological parsers or Basque named-entity recognition models – and supporting participatory AI initiatives modelled on Masakhane or Te Mana Raraunga. SEMIC and DCAT-AP compliance must be contingent on adherence to

24 RDF (Resource Description Framework) is a standard model for representing data on the web using triples – each consisting of a subject, predicate, and object – to describe relationships between resources in a structured, machine-readable way. SKOS (Simple Knowledge Organisation System) builds on RDF to provide a framework for representing controlled vocabularies such as thesauri, taxonomies, and classification schemes. It allows concepts to be defined with preferred and alternative labels, and to be linked through hierarchical (broader/narrower) or associative (related) relationships. Together, RDF and SKOS enable the semantic organisation and sharing of knowledge across systems in a consistent and interoperable manner.

relational stewardship principles, requiring proof of implementation for frameworks like the CARE Principles when handling minority community data.

Finally, structural safeguards demand a reform of language status and oversight. Regulation No 1/1958 must extend "official language" recognition to territorially rooted languages such as Catalan and Sámi, mandating their integration into eTranslation pipelines, IATE terminological databases, and SEMIC's high-risk ontologies under the AI Act. A European agency for linguistic justice should be established to enforce violations of Article 21 of the Charter of Fundamental Rights of the European Union in AI systems and maintain public ledgers documenting epistemic distortions in legal databases.

In embracing a jurisgenerative future, the European Union must recognise that multilingualism is a democratic imperative as well as a technical challenge. The infrastructures of AI and data governance must evolve from instruments of standardisation into vessels of cultural co-presence, where languages, laws, and lifeworlds converge without collapse. This transformation demands more than policy tweaks; it calls for a paradigmatic shift in how Europe conceives of knowledge, authority, and connection. By embedding pluralism into the very syntax of its digital architectures, the EU can chart a path towards a federated epistemic common future where diversity is not tolerated but cultivated, and where every language, no matter how small, is a sovereign contributor to the shared future of European intelligence.

## 6.2 Limitations and future research

This study is EUcentred and focuses on selected standards and infrastructures; it does not present modelbymodel quantitative evaluations across the full spectrum of minoritised languages, nor does it quantify computational or administrative tradeoffs for the proposed architectural changes. The analysis is primarily documentary and theoretical rather than ethnographic or experimental. Future work should conduct comparative, perlanguage audits of foundation models in legally salient tasks; prototype and evaluate participatory ontology design with affected communities; longitudinally track whether emerging initiatives deliver measurable parity for lowresource languages; and examine how potential reforms to the EU language regime interact with technical governance in data spaces and the AI Act – testing, in practice, whether jurisgenerative interoperability can be sustained at scale.

**Statement on the use of artificial intelligence**

The author declares that artificial intelligence tools have been used in carrying out this research and in the process of writing the article. During the research phase, Microsoft Copilot was used for translation and English improvement tasks. All results generated using this tool have been verified and validated by the author, who take full responsibility for them.

The author confirms that all scientific content, interpretations, and conclusions are the result of his intellectual contribution, and that AI has been used in accordance with current ethical and editorial policies.

## 7 References

Barthes, Roland. (2009). *Mythologies*. Vintage Classics.

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, & Shmitchell, Shmargaret. (2021). On the dangers of stochastic parrots: Can language models be too big?. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bratton, Benjamin H. (2016). *The stack: On software sovereignty*. The MIT Press. https://doi.org/10.7551/mitpress/9780262029575.001.0001

Barcelona Supercomputing Center (BSC). (2025, January 21). *ALIA, Europe's first public, open and multilingual AI infrastructure*. Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS).

Chuang, Liu. (2018). *Bitcoin mining and field recordings of ethnic minorities* [3-channel video installation]. Tate, London, United Kingdom.

Commission Implementing Decision (EU) 2024/458 of 1 February 2024 on setting up the European Digital Infrastructure Consortium for the Alliance for Language Technologies (ALT-EDIC). (OJ L, 2024/458, 7.2.2024)

Commissioner for Human Rights. (2019, October 29). *The Commissioner's Human Rights Comments: Language policies should accommodate diversity, protect minority rights and defuse tensions*. Council of Europe.

Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union. (OJ C 202, 7.6.2016, pp. 1–388)

CORDIS. (2024, June 18). *Semantic spaces, multilingualism and diversity in the law: The challenge of legal knowledge management within the European Union (LingLaw)*. European Union.

Council of Europe. (1998). *Framework Convention for the Protection of National Minorities (ETS No. 157)*.

Council of Europe. (1992). *The European Charter for Regional or Minority Languages*.

Cover, Robert. (1983). *The Supreme Court, 1982 Term -- Foreword: Nomos and Narrative*. Yale Law School.

Curran, Vivian Grosswald. (1998). Cultural immersion, difference and categories in U.S. comparative law. *The American Journal of Comparative Law*, *46*(1), 43–92. https://doi.org/10.2307/841078

Darawsheh, Kareem, Hamamra, Bilal, & Abuarrah, Sufyan. (2025). Addressing untranslatability: ChatGPT's compensatory strategies for translating verbified proper nouns in English–Arabic contexts. *Traduction et Langues*, *24*(1), 298–326. https://doi.org/10.52919/translang.v24i01.1037

De Sousa Santos, Boaventura, & Meneses, María Paula. (2014). *Epistemologías del Sur*. Akal.

Escobar, Arturo. (2015). *Transiciones*: A space for research and design for transitions to the pluriverse. *Design Philosophy Papers*, *13*(1), 13–23. https://doi.org/10.1080/14487136.2015.1085690

Escobar, Arturo. (2020). *Pluriversal politics: The real and the possible*. Duke University Press.

European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028(INI)). (OJ C 433, 23.12.2019, pp. 42–49)

European Union. (2012). Charter of Fundamental Rights of the European Union.

European Union. (2007). Treaty on European Union.

Fernández-Llamazares, Álvaro, Fa, Julia E., Brockington, Dan, Brondízio, Eduardo S., Cariño, Joji, Corbera, Esteve, Farhan Ferrari, Maurizio, Kobei, Daniel, Malmer, Pernilla, Márquez, Guadalupe Yesenia H., Molnár, Zsolt, Tugendhat, Helen, & Garnett, Stephen T. (2024). No basis for claim that 80% of biodiversity is found in Indigenous territories. *Nature*, *633*(8028), 32–35. https://doi.org/10.1038/d41586-024-02811-w

Floridi, Luciano. (2026). Artificial intelligence as a new form of agency. In Sven Nyholm, Atoosa Kasirzadeh & John Zerilli (Eds.), *Contemporary Debates in the Ethics of Artificial Intelligence* (pp. 17–33). Wiley. https://doi.org/10.1002/9781394258840.CH2

Hernández-Morales, Aitor. (2025, May 27). EU countries pass on Spanish bid to make Catalan an official language in Brussels. *Politico*.

Huang, Jiaqi, & Wang, Guifang. (2024). A study on translation strategies of legal English texts from the perspective of functional equivalence theory. *Frontiers in Humanities and Social Sciences*, *4*(7), 37–43. https://doi.org/10.54691/PYCVVC29

Ioriatti, Elena. (2023). Common contexts of meaning in the European legal setting: Opening Pandora's box?. *International Journal for the Semiotics of Law*, *36*(1), 275–291. https://doi.org/10.1007/s11196-022-09919-0

Joshi, Pratik, Santy, Sebastin, Budhiraja, Amar, Bali, Kalika, & Choudhury, Monojit. (2021). *The state and fate of linguistic diversity and inclusion in the NLP world*. arXivLabs, Cornell University. https://doi.org/10.48550/arXiv.2004.09095

Jutel, Olivier. (2022). Blockchain humanitarianism and crypto-colonialism. *Patterns*, *3*(1), 100422. https://doi.org/10.1016/j.patter.2021.100422

Kukutai, Tahu, & Taylor, John. (2016). Indigenous data sovereignty: Toward an agenda. Australian National University Press. http://doi.org/10.22459/CAEPR38.11.2016

Kumar, Pranjal. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*(10). https://doi.org/10.1007/S10462-024-10888-Y

Laaksonen, Jenni. (2022). Translation, hegemony and accounting: A critical research framework with an illustration from the IFRS context. *Critical Perspectives on Accounting*, *89*, 102352. https://doi.org/10.1016/J.CPA.2021.102352

Lessig, Lawrence. (2003). Law regulating code regulating law. *Loyola University Chicago Law Journal*, *35*(1).

Mejías, Ulises, & Couldry, Nick. (2019). Colonialismo de datos: repensando la relación de los datos masivos con el sujeto contemporáneo. *Virtualis*, *10*(18), 78–97.

Mignolo, Walter D. (2007a). Delinking: The rhetoric of modernity, the logic of coloniality and the grammar of de-coloniality. *Cultural Studies*, *21*(2–3), 449–514. https://doi.org/10.1080/09502380601162647

Mignolo, Walter D. (2007b). Introduction: Coloniality of power and de-colonial thinking. *Cultural Studies*, *21*(2–3), 155–167. https://doi.org/10.1080/09502380601162498

Milian-Massana, Antoni. (2004). Le régime juridique du multilinguisme dans l'Union européenne. Le mythe ou la réalité du principe d'égalité des langues. *Revue juridique Themis*, *38*, 211–260.

Milian-Massana, Antoni. (2008). Languages that are official in part of the territory of the Member States: Second-class languages or institutional recognition in EU law?. In Xabier Arzoz (Ed.),

*Respecting Linguistic Diversity in the European Union* (pp. 191–229). John Benjamins Publishing Company. https://doi.org/10.1075/wlp.2.13mil

Milian i Massana, Antoni. (2010). *Drets lingüístics per a tothom. Estudis de dret lingüístic*. Lleonard Muntaner.

Poblet, Marta, Casanovas, Pompeu, & Rodríguez-Doncel, Víctor. (2019). Legal linked data ecosystems and the rule of law. *Linked Democracy*. *SpringerBriefs in Law* (pp. 87–126). Springer, Cham. https://doi.org/10.1007/978-3-030-13363-4_5

Pragya, Apoorv. (2025). Generative AI and epistemic diversity of its inputs and outputs: Call for further scrutiny. *AI and Society*, *40*(4), 2341–2342. https://doi.org/10.1007/s00146-024-02097-6

Publications Office of the European Union. (n.d.). *DCAT Application profile for data portals in Europe (DCAT-AP)*. European Union.

Regulation No 1 determining the languages to be used by the European Economic Community. (OJ P 017, 6.10.1958, p. 385)

Regulation (EU) 2021/694 of the European Parliament and of the Council of 29 April 2021 establishing the Digital Europe Programme and repealing Decision (EU) 2015/2240. (OJ L 166, 11.5.2021, pp. 1–34)

Regulation (EU) 2021/818 of the European Parliament and of the Council of 20 May 2021 establishing the Creative Europe Programme (2021 to 2027) and repealing Regulation (EU) No 1295/2013. (OJ L 189, 28.5.2021, pp. 34–60)

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). (OJ L, 2024/1689, 12.7.2024)

Ruckstuhl, Katharina. (2023). Data as Taonga: Aotearoa New Zealand, Māori data sovereignty and implications for protection of treasures. *Journal of Intellectual Property & Entertainment Law* (Symposium issue), *12*(3).

Russo, Irene, & Soria, Claudia. (2017). Sardinian – a digital language?. In Claudia Soria, Irene Russo & Valeria Quochi (Eds.), *Reports on Digital Language Diversity in Europe*. Erasmus+ National Agency.

Sacco, Rodolfo. (1991). Legal formants: A dynamic approach to comparative law. *The American Journal of Comparative Law*, *39*(1), 1–34. https://doi.org/10.2307/840669

Salonen, Tuomo. (2017). Karelian – a digital language?. In Claudia Soria, Irene Russo & Valeria Quochi (Eds.), *Reports on Digital Language Diversity in Europe*. Erasmus+ National Agency.

Shaping Europe's digital future. (2025, March 20). *Commission welcomes new initiative to support European cultural and linguistic diversity in artificial intelligence*. European Commission.

Star, Susan Leigh, & Griesemer, James. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Tan, Zhixing, Wang, Shuo, Yang, Zonghan, Chen, Gang, Huang, Xuancheng, Sun, Maosong, & Liu, Yang. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5–21. https://doi.org/10.1016/j.aiopen.2020.11.001

Thatcher, Jim, O'Sullivan, David, & Mahmoudi, Dillon. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space*, *34*(6), 990–1006. https://doi.org/10.1177/0263775816633195

Tuck, Eve, & Yang, K. Wayne. (2014). Unbecoming claims: Pedagogies of refusal in qualitative research. *Qualitative Inquiry*, *20*(6), 811–818. https://doi.org/10.1177/1077800414530265

UNESCO. (2024, July 25). *A digital future for indigenous languages: Insights from the Partnerships Forum*. UNESCO.

Veale, Michael, & Zuiderveen Borgesius, Frederik. (2021). Demystifying the draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/CRI-2021-220402

Vilas Gómez, Álvaro. (2025, April 30). *Spanish authorities release ALIA AI models: Spanish Government promotes open access to its ALIA AI models*. Interoperable Europe; European Commission.