해부 · 생물인류학 제 39 권 제 1 호
Anat Biol Anthropol Vol. 39, No. 1 (2026) pp. 15~29
https://doi.org/10.11637/aba.2026.39.1.15

Review Article

생물인류학 연구에서 eHRAF 활용에 대한 방법론적 검토: eHRAF World Cultures와 eHRAF Archaeology 텍스트 코퍼스 기반 비교 연구 소개

권지헌[1,2], 박한선[1,2]

[1]서울대학교 사회과학대학 인류학과, [2]서울대학교 사회과학대학 인류학과 진화인류학 교실

Methodological Review of eHRAF Utilization in Biological Anthropology Research: An Introduction to Text Corpus-Based Comparative Studies Using eHRAF World Cultures and eHRAF Archaeology

Jihun Kwon[1,2], Hanson Park[1,2]

[1]*Department of Anthropology, Seoul National University College of Social Science*
[2]*Laboratory of Evolutionary Anthropology, Seoul National University College of Social Science*

Abstract : Biological anthropology has experienced rapid methodological advances in ancient DNA, stable isotope analysis, medical imaging, and osteological reconstruction, enabling high-resolution inferences about past human population structure, mobility, diet, disease burden, and life history variation. Yet higher-resolution data do not automatically yield higher-resolution explanations, because similar biological signals may reflect distinct selection pressures and exposure environments under different social institutions, subsistence strategies, warfare patterns, mortuary practices, medical behaviors, and age and gender norms. Much of the contextual information needed to constrain such interpretations exists in textual sources such as ethnographies and excavation reports, but it is difficult to retrieve systematically by concept and transform into comparable analytical units that can be linked to quantitative bioanthropological data. This review examines methodological strategies for using eHRAF World Cultures and eHRAF Archaeology, the online databases of Human Relations Area Files (HRAF), not merely as background references but as reproducible text corpora for bioanthropological research. We describe the logic of paragraph-level subject indexing and concept-based retrieval enabled by the Outline of Cultural Materials (OCM), and propose decision rules for selecting and integrating eHRAF World Cultures and eHRAF Archaeology using a triangulation

이 연구는 서울대학교 신임교수 연구정착금으로 지원되는 연구비에 의하여 수행되었음.
저자(들)는 '의학논문 출판윤리 가이드라인'을 준수합니다.
저자(들)는 이 연구와 관련하여 이해관계가 없음을 밝힙니다.
Received: February 23, 2026; **Revised:** March 23, 2026;
Accepted: March 25, 2026
Correspondence to: 박한선 (서울대학교 사회과학대학 인류학과)
E-mail: hansonpark@snu.ac.kr

ISSN 2671-566X (Online)

approach to control for ecological fallacies and time-averaging, given their distinct units of analysis (cultures versus archaeological traditions). We further synthesize four analytical approaches, context retrieval, variable construction, middle-range theory building, and design scaffolding, and provide a stepwise workflow covering query logging, extraction unit definition, codebook development, intercoder reliability assessment, aggregation rules, unit alignment across cultural, temporal, and spatial scales, and strategies to address non-independence (Galton's problem). We conclude by discussing key limitations, including representativeness, recording bias, temporal mismatch, and licensing constraints, and outline future directions such as leveraging OCM as labels for weakly supervised natural language processing (NLP) and artificial intelligence (AI) to enhance scalable and transparent biocultural inference.

Keywords : electronic Human Relations Area Files (eHRAF), Cross-cultural research, Outline of cultural materials, Biocultural synthesis

서 론

생물인류학은 분자유전학과 생화학, 영상의학, 통계적 인과추론 도구의 발전에 기반해 과거 인류 집단의 유전적 구성, 이동, 식생활, 질병 부담, 생애사(life history) 변이를 매우 세밀하게 추정할 수 있게 되었다(본고에서는 전통적 체질인류학과 분자 단위의 현대 생물인류학을 포괄하는 넓은 의미로 '생물인류학'을 사용한다). 그러나 데이터의 정교화가 설명의 정확도를 보장하지는 않는다. 동일한 유전적 신호와 골격 지표가 서로 다른 사회 제도, 생계 전략, 전쟁 양상, 매장 방식, 의료 행위, 성별 및 연령 규범 아래에서 전혀 다른 선택압과 노출 환경을 반영할 수 있기 때문이다. 이러한 문화적 맥락(예: 생계경제 구조, 혼인 체계, 생태적 기근 수준) 정보의 결핍은 명확한 통계적 신호에도 불구하고 해석을 빈약하게 만든다.

그런데 문제는 해당 맥락 정보가 대개 텍스트 형태로 존재한다는 점이다. 민족지 단행본, 발굴 보고서, 지역사 자료, 학위논문, 1차 자료 번역본 등 텍스트는 이미 풍부하지만, 특히 의학 기반의 생물인류학 연구자는 특정 개념을 기준으로 체계적으로 검색하고, 비교 가능한 단위로 변환하여 다른 정량 자료와 결합하는 맥락 기반 접근에서 한계를 경험하는 일이 많다. 이 간극을 메우는 대표적 인프라로서 체계적 비교문화 데이터베이스의 역할이 주목받아 왔다.

전자인간관계지역파일(electronic Human Relations Area Files, eHRAF)은 이러한 역할을 수행해 생물인류학 연구의 맥락적 한계를 보완할 수 있는 대표적인 비교문화 데이터베이스 시스템이다[1-4]. 인간관계지역파일(Human Relations Area Files, HRAF)의 온라인 버전인 eHRAF는 선별된 1차 문헌의 본문 텍스트를 제공하는 동시에, 각 문단(paragraph)을 문화자료개요(Outline of Cultural Materials, OCM) 등의 코드로 색인하여 키워드가 아닌 개념 단위의 검색을 지원한다. 특히 OCM 주제와 불리언(Boolean) 연산, 키워드를 결합한 매우 정교한 개념 기반 검색이 가능하다[1-4].

본 논문에서는 해부생물인류학 저널 독자를 대상으로 eHRAF를 '자료'로 활용하는 방법을 논의하고자 하며, 크게 두 가지 주제에 천착할 것이다. 첫째, eHRAF는 단순 사례 참고문헌이 아니라, OCM 색인을 통해 재현 가능한 방식으로 텍스트를 추출 및 코딩하여 통계 모형에 직접 투입할 수 있는 구조화된 '텍스트 코퍼스(text corpus)'이다. 둘째, eHRAF를 구성하는 eHRAF World Cultures와 eHRAF Archaeology는 서로 다른 분석 단위를 갖지만 연구 질문을 기준으로 통합적으로 활용할 때 생물문화적 통합(biocultural synthesis)의 기반을 제공할 수 있다[1-4].

본고에서는 다음의 세 가지 과제를 순차적으로 다룬다. 첫째, eHRAF를 처음 접하는 독자를 위해 HRAF와 eHRAF의 역사, 데이터 구조, 검색 논리를 소개한다. 둘째, 생물인류학 연구에서 eHRAF 활용이 가능한 이유를 데이터 구조와 분석 논리 차원에서 제시한다. 셋째, eHRAF 기반 연구를 수행하는 실질적 방법과 요령을 단계별 워크플로 형태로 정리한다.

본 론

1. HRAF와 eHRAF의 구축 배경 및 국제 비교문화 데이터베이스 현황

생물인류학 연구에서 민족지 연구 데이터는 골격과 유전자 자료 이면의 문화적 맥락을 복원하는 데 필수적인 역할을 한다. 이러한 맥락 정보들을 보존하고 연구자가 체계적으로 활용할 수 있도록 돕는 학술적 데이터베이스의 필요성은 오래전부터 인류학계에서 제기되어 왔다. HRAF는 1937년

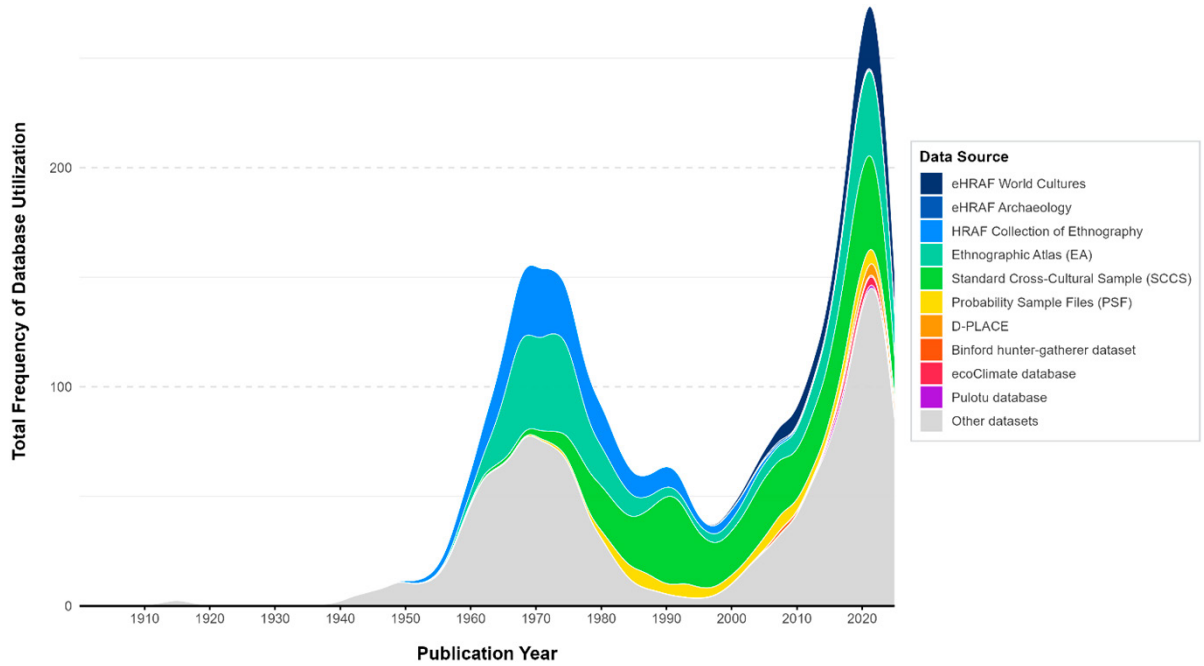

Fig. 1. Temporal trends in the utilization frequency of HRAF-related samples and cross-cultural databases, 1900~2025. The first peak (1960s~1970s) reflects the utilization of the paper-based HRAF Collection of Ethnography and the Ethnographic Atlas (EA). The second peak (2010s~present) is driven by the online eHRAF World Cultures and the Standard Cross-Cultural Sample (SCCS), highlighting the transition to digital archives. Data source: Explaining Human Culture (HRAF, https://hraf.yale.edu/ehc).

예일대학교의 조지 피터 머독(George Peter Murdock)과 동료들이 시작한 비교문화 조사(Cross-Cultural Survey) 프로젝트에서 출발하여 1949년 공식 설립된 비영리 컨소시엄이자 데이터베이스 명칭이다. 1950년대에는 OCM 개정판과 세계문화개요(Outline of World Cultures, OWC)가 정비되었으며, 종이와 마이크로피시(microfiche)의 형태로 사용되던 기존 자료는 1990년대 이후로 CD를 거쳐 점차 전자화되기 시작했다. 그 결과, 1997년에는 온라인 서비스인 eHRAF World Cultures가 시작되었고, 이후 고고학자 피터 페레그린(Peter Peregrine) 등에 의해 전 세계 고고학 및 민족고고학 기록을 체계화한 eHRAF Archaeology가 구축되면서 현재의 형태에 이르렀다[1-9].

최근에는 D-PLACE (Database of Places, Language, Culture and Environment) 등 다양한 국제적 비교문화 데이터베이스가 널리 활용되고 있다[10-24]. 이들 데이터베이스는 대부분 연구자들에 의해 사전 코딩된 변수(coded variables)를 제공하여 신속한 대규모 정량 분석에 특화되어 있다는 장점이 있다. 그럼에도 eHRAF는 이들과 구별되는 고유한 강점을 지닌다. 문화권 간 비교를 수행할 수 있도록 검색과 표본 선택을 위한 동의어 사전(thesaurus) 및 메타데이터를 함께 제공해 연구자가 직접 원문 텍스트의 맥락을 확인하고, 개별 연구 질문에 맞추어 유연하게 변수를 조작적으로 정의할 수 있는 체계적인 '원문 텍스트 코퍼스'를 제공하기 때문이다[1-4]. Table 1은 생물인류학 및 비교문화 연구에서 eHRAF와 상호보완적으로 연계할 수 있는 주요 국제 데이터베이스의 현황을 사회·문화, 언어, 유전, 통합의 네 가지 차원으로 분류하여 보여준다.

예를 들어, 언어 및 유전 데이터베이스는 공간적 비독립성 문제를 통제하고 유전자와 문화 사이의 공진화를 추적하기 위한 계통적 기준선을 제공하며, 통합형 데이터베이스는 해당 다층적 데이터를 유기적으로 엮어내는 네트워크 허브의 역할을 수행한다.

Fig. 1은 HRAF 관련 표본 및 비교문화 데이터베이스의 연도별 활용 빈도를 나타낸 것이다. 1960~70년대에는 종이 기반 HRAF Collection of Ethnography와 Ethnographic Atlas (EA) [7,25,26]가 비교문화 연구의 주된 자료원이었으며, 해당 시기가 비교문화 연구의 첫 번째 확장기에 해당한다. 1980~90년대에는 활용 빈도가 일시적으로 감소하는데, 이는 비교문화 방법론에 대한 이론적 비판, 특히 갈톤의 문제(Galton's problem)와 민족지 단위의 자의성에 대한 논쟁이 집중되었던 시기와 겹친다. 그러나 2000년대 이후 온라인 eHRAF 서비스가 본격화되면서 다시 활용 빈도가 급증하였고, 그와 동시에 D-PLACE 등 코드형 데이터베이스[12-14,23,24]의 사용도 크게 늘어났다. 이러한 추세는 디지털 인프라의 확충이

Table 1. Major external databases complementing eHRAF for biocultural and cross-cultural comparative research

Category	Database	Primary reference	Coverage	Core data features
Socio-cultural	eHRAF World Cultures[†]	HRAF (https://ehrafworldcultures.yale.edu/)	360+ cultures worldwide	Full-text ethnographies indexed by OCM subject codes at the paragraph level
	eHRAF Archaeology[†]	HRAF (https://ehrafarchaeology.yale.edu/)	100+ archaeological traditions	Full-text archaeological reports indexed by OCM subject codes at the paragraph level
	Hunter-Gatherer Dataset[a]	Binford (2001)	339 hunter-gatherer societies	Coded variables on subsistence dependence, mobility, group size, and environmental parameters
	Seshat: Global History Databank	Turchin et al. (2015) (https://seshatdatabank.info/)	400+ polities	Coded variables on social complexity, warfare, ritual, and institutional hierarchy
	Pulotu: Database of Austronesian Religions	Watts et al. (2015) (https://pulotu.com/)	137 Austronesian cultures	Coded variables on supernatural beliefs, ritual practices, and social structure
	Database of Religious History (DRH)	Slingerland & Sullivan (2017) (https://religiondatabase.org/)	800+ religious traditions worldwide	Crowd-sourced coded variables on religious beliefs, practices, and institutions
Linguistic & expressive culture	Glottobank	Glottobank consortium (Max Planck Institute) (https://glottobank.org/)	2,400+ languages	Lexical cognate sets, morphosyntactic features, and language phylogenies
	World Atlas of Language Structures (WALS)	Dryer & Haspelmath (2013) (https://wals.info/)	2,600+ languages	Structural features (phonological, grammatical, typological)
	Expanded Natural History of Song (NHS)	Bertolo et al. (2025) (https://zenodo.org/records/15725182)	1000+ audio recordings	Cross-cultural music corpus with audio recordings and behavioral metadata
	Global Jukebox	Wood et al. (2022) (https://theglobaljukebox.org/)	5,700+ song recordings worldwide	Coded variables on song style, performance structure, and dance
Genetic	1000 Genomes Project (IGSR)	1000 Genomes Consortium (2015) (https://www.internationalgenome.org/)	2,504 individuals from 26 populations	Whole-genome variation data (SNPs, indels, structural variants)
	Allele Frequency Net Database (AFND)	Gonzalez-Galarza et al. (2020) (http://www.allelefrequencies.net/)	1,600+ populations (>10M individuals)	Immune gene allele frequencies (HLA, KIR, MHC) at allele, haplotype, and genotype levels
	Simons Genome Diversity Project (SGDP)	Mallick et al. (2016) (https://www.simonsfoundation.org/simons-genome-diversity-project/)	300 genomes from 142 populations	High-quality WGS (43× mean depth) targeting underrepresented populations
Integrated	Database of Places, Language, Culture and Environment (D-PLACE)	Kirby et al. (2016) (https://d-place.org/)	1,900+ societies	Linked cultural traits, environmental variables, and language phylogenies
	Genes and Languages Together (GeLaTo)	Barbieri et al. (2022) (https://gelato.clld.org/)	390+ populations	Linked genetic diversity metrics and linguistic classification for population-pair comparison

Categorized into socio-cultural, linguistic, genetic, and integrated domains, these repositories provide the essential quantitative metrics, linguistic phylogenies, and genetic baselines needed to contextualize eHRAF's qualitative corpus. Integrating these external datasets enables researchers to rigorously control for spatial non-independence (Galton's problem) and robustly test hypotheses regarding gene-culture coevolution (GCC).

[†]Subscription required. All other databases are open access.

[a]R package available: binford [11].

비교문화 연구의 양적 성장을 견인했음을 보여주는 동시에, eHRAF가 단순한 데이터베이스가 아니라 현재의 과학적 연구에서 활발히 사용되는 도구임을 확인시켜 준다.

2. eHRAF 데이터의 분석 단위와 표본 추출(Sampling) 전략에 따른 분류

eHRAF는, 전 세계 360개 이상의 문화 컬렉션을 포함하는 eHRAF World Cultures와 100개 이상의 고고학적 전통(tradition)을 포함하는 eHRAF Archaeology를 대표 데이터베이스로 제공하고 있다. 전자는 민족지 문헌을 '문화(culture)'의 단위로 조직하며[1,3], 후자는 고고학 문헌을 고고학적 '전통'의 단위로 조직한다. 후자에서 '문화' 대신 '전통'이 분석 단위가 되는 이유는, 고고학 자료가 언어, 친족, 이데올로기 등 사회문화적 범주로 직접 정의되기 어렵고, 시간적 지속성과 공간적 연속성, 물질문화와 생계 기반을 중심으로 단위를 구성하기 때문이다. 전통은 넓은 지리적 범위와 비교적 긴 시간에 걸쳐 지속되는 물질문화 및 생활양식의 연속성을 가리키는 분석 단위이며, 개별 유적이나 개체 수준 자료를 집단 수준의 경향성으로 해석할 때 유용하다[1,27].

이러한 분석 단위의 차이와 더불어, 비교 연구에서 필수적으로 고려해야 하는 것은 관찰값의 대표성 문제와 지리적, 계통적 인접성에 의한 공간적 비독립성의 문제, 즉 갈톤의 문제(Galton's problem)이다[28-30]. eHRAF World Cultures는 이를 표본 추출 단계에서 완화하기 위해 다양한 층화 및 확률 표본을 제공한다. 대표적 표본으로 주로 사용되는 표준비교문화표본(Standard Cross-Cultural Sample, SCCS)은 Murdock과 White가 EA에 수록된 약 1,260개 집단을 약 200개의 표집 지역(sampling province)으로 구분한 뒤, 각 지역에서 민족지 자료가 가장 풍부한 사회를 대표로 선정하여 구성한 186개 집단의 표본이다[31]. 이러한 설계는 공간적 비독립성 문제를 표본 추출 단계에서 완화하려는 목적을 갖는다. 그 외에도 활발히 사용되는 확률표본파일(Probability Sample Files, PSF)은 Naroll 등이 데이터의 품질과 지리적 분포를 기준으로 층화 무작위 추출을 수행하여 구축한 확률 표본이며[32], 최근에는 단순무작위표집(simple random sampling) 전략을 기반으로 eHRAF World Cultures 수록 집단에서 28개를 무작위 추출한 단순무작위표본(Simple Random Sample, SRS)도 제공되고 있다.

그리고 eHRAF Archaeology 역시 고고학전통개요(Outline of Archaeological Traditions, OAT) [2,27]에 기반해 무작위로 46개의 전통을 추출한 SRS를 제공하고 있다. 아직은 eHRAF World Cultures에 비해 통용되는 표본의 수가 많지 않다. 하지만 eHRAF World Cultures와 유사하게 OAT를 기반으로 유연하게 표본과 데이터 추출이 가능해, 생물인류학을 비롯한 고고학 비교 연구에서의 폭넓은 활용 가능성이 존재한다.

연구자는 연구 질문의 성격과 요구되는 표본 크기에 따라 이들 표본 중 적합한 것을 선택하거나, 전체 eHRAF 컬렉션에서 자체 기준에 따라 세부 표본을 구성할 수 있다.

3. eHRAF 비교문화 텍스트 코퍼스의 개념 검색 표준 색인: OCM, OWC, OAT

표본 추출과 함께, '문단 단위 주제 색인'은 eHRAF를 분석 가능한 텍스트 코퍼스로 만들어주는 핵심적 기능을 수행한다. 일반적인 데이터베이스 검색이 서지 정보 혹은 전체 텍스트의 단순 단어 일치에 의존하는 반면, eHRAF는 전문 코더가 문맥을 읽고 특정 주제를 부여한 단락 단위로 결과를 반환한다. 이는 특정 자료 전체를 읽어야만 파악되는 맥락 내용을 비교 가능한 단위로 쪼개어 찾아주는 질적 자료의 '위치 지시자(pointer)' 역할을 한다. 따라서 연구자는 전체 텍스트를 처음부터 끝까지 읽지 않고도 관련된 텍스트 단락을 집중적으로 검토하여 동일한 주제를 여러 문화나 전통에서 병렬적으로 비교하는 맥락 회수가 가능해진다[1-4].

eHRAF는 이를 위해 두 가지 핵심 분류 체계를 사용한다. 첫째는 연구 질문을 개념화하는 색인 체계인 OCM이다. OCM은 1930년대부터 Murdock과 그의 동료들이 개발한 이래 700개 이상의 범주를 제공해 온 방대한 통제 어휘(controlled vocabulary)의 집합이다[33]. eHRAF는 문헌의 단락마다 하나 이상의 OCM 코드를 부여하여 검색의 기본 단위로 삼는다. OCM은 자료 편찬자에게는 문헌을 체계적으로 분류하는 공통 언어이며, 연구자에게는 서로 다른 집단에서 동일한 주제를 찾을 수 있게 해주는 검색 언어이다. 또한, 상위 범주는 넓은 주제를, 하위 범주는 구체적인 주제를 포괄하는 계층적(hierarchical) 구조를 갖는다(예: 매장과 장례는 OCM 764로 색인). 연구자는 특정 연구 질문을 OCM 주제로 번역하여 텍스트를 추출하며, 이때 OCM은 개념의 범위를 통제하고 일반 검색어(keyword)는 세부 맥락을 정밀화하는 보조 역할을 수행한다[33]. 둘째는 각각 표본을 식별하고 구성하는 OWC와 OAT이다[2,27,34]. OWC는 eHRAF World Cultures의 수많은 문화 집단들을 표준화하여 분류하고 식별하는 색인이며[34], OAT는 고고학 영역의 전통들을 분류·식별하는 색인으로 eHRAF Archaeology의 명시적인 표본틀(sampling frame)로서 기능한다[2,27].

결론적으로 생물인류학 연구자는 OWC와 OAT를 통해 분석할 '집단(표본)'을 결정하고, OCM을 통해 텍스트에서 회수할 '개념(변수)'을 설계하게 된다. 이러한 대응 관계의 구체

Table 2. Selected Outline of Cultural Materials (OCM) codes bridging textual cultural data with bioanthropological indicators

Research theme	OCM code	Official OCM descriptor	Biological anthropology use case (examples)
Subsistence & diet	146	Nutrition	Diet composition, food distribution and sharing, taboos, famine responses, contextual variables for stable isotope interpretation, dental caries, growth and stress markers
	223	Fowling	Animal-resource use, hunting technique variation, labor intensity and risk exposure, contextual covariates for diet reconstruction and trauma/activity interpretations
	224	Hunting and trapping	Hunting strategy, mobility, energy budgets, weapon use and injury risk, linking trauma patterns to subsistence and division of labor
	225	Marine hunting	Marine dependence, seasonality, long-range mobility and risk exposure, contextualizing marine isotope signals and stress indicators
	226	Fishing	Aquatic subsistence, settlement near water and health risks, labor intensity, exposure contexts, supporting dietary and life-history inference
	227	Fishing gear	Technological level, labor division, differential risk exposure, supporting interpretation of specific trauma patterns and musculoskeletal stress markers
	260	Food consumption	Consumption norms, preparation practices, sharing rules and taboos, constructing behavioral "intake-context" variables for nutritional stress and health inequality
	262	Diet	Main food resources, seasonality, subsistence diversification, famine buffering, core contextual variables for isotope-based diet inference and skeletal stress interpretations
Physical activity & trauma	304	Mutilation	Intentional body modification (cranial deformation, dental modification, amputation), differentiating cultural modification from pathology, coding prevalence and norms
	310	Activities	Labor and activity types, workload intensity, sex and age division of labor, contextual covariates for musculoskeletal stress markers and activity reconstruction
	411	Weapons	Technical and social conditions of violence, weapon types and use contexts, distinguishing warfare/interpersonal violence/accidents in trauma pattern interpretation
Population structure & gene flow	516	Freedom of movement	Residence and mobility constraints by sex/age, variables relevant to gene flow, mobility proxies, and integration with genetic/isotopic mobility inference
	580	Marriage	Endogamy/exogamy and marriage norms, cultural constraints on gene flow and population structure, contextualizing ROH/inbreeding and diversity metrics

Table 2. Continued

Research theme	OCM code	Official OCM descriptor	Biological anthropology use case (examples)
Population structure & gene flow	582	Regulation of marriage	Consanguinity rules (e.g., cousin marriage, prohibitions), operational variables for mating structure and population homozygosity/ kinship inference
	613	Lineage	Matrilineal/patrilineal descent rules, linking social organization to uniparental markers and kin-structured interpretations
Health & mortuary context	750~757	Sickness; Preventive medicine; Bodily injuries; Theory of disease; Medical therapy (and related codes)	Skeletal pathology context, trauma treatment and healing practices, surgery marks, medicinal practices, ethnomedical classification as covariates for paleopathological inference
	760~769	Death; Dying; Burial practices and funerals; Special burial practices; Mortuary specialists; Cult of the dead (and related codes)	Mortuary treatment, secondary burial, special/elite vs deviant burials, taphonomy and preservation contexts, aDNA recovery conditions, mediating assumptions in kinship and inequality inference

These codes exemplify how qualitative ethnographic and archaeological texts can be systematically operationalized into behavioral and contextual variables (e.g., subsistence patterns, mating structures, violence, and mortuary treatments) to interpret skeletal, isotopic, and genetic data. In actual studies, researchers should consult the detailed scope notes in the eHRAF OCM Reference and explicitly document their code-combination rules, coding units, and aggregation strategies.

적 사례를 Table 2에 제시하였다(전체 OCM의 세부 내용은 Table 1에 제시된 각 eHRAF 링크의 'Browse Subjects'에서 확인할 수 있다).

4. eHRAF 텍스트 코퍼스 데이터화를 위한 네 가지 방식

eHRAF의 텍스트 코퍼스를 분석 가능한 데이터로 변환하는 방식은 크게 네 가지로 구분된다.

첫째, 맥락 회수이다. 골격이나 유전 자료가 제시하는 생물학적 패턴을 해석할 때, 그 이면에 존재하는 문화적 배경 정보(예: 매장 방식, 집단 간 폭력의 형태, 질병 치료 관행, 혼인과 이동 규범 등)를 텍스트에서 회수하여 인과적 추론의 타당성을 점검하는 기능이다. 골격 지표만으로 무리한 인과적 서사를 구성하는 오류를 방지하는 핵심 역할을 한다. 예를 들어, Ensor 등은 사후 매장 위치(postmortem location)의 문화 간 변이를 확인하기 위해 OCM 590~610 범위의 친족 관련 주제와 'cemeter*' 키워드를 결합하여 검색을 수행했다. 이를 통해 68개 문화권의 115개 문헌에서 161개 문단을 확보하고, 그중 28개 문화권의 실제 매장 관행 자료를 추출함으로써 골격 지표 해석의 전제를 엄밀하게 교차 검증한 바 있다[35].

둘째, 변수화(Variable Construction)이다. 텍스트에 기술된 특정 관행의 유무, 빈도, 규범성, 예외 조건 등을 연구자의 조작적 정의에 따라 코드화하여 비교 가능한 척도로 변환하는 작업이다. OCM 코드가 관련 내용의 위치를 지시하더라도, 이를 실제 변수로 추출하기 위해서는 명시적인 코드북 설계와 코더 간의 신뢰도(Inter-coder reliability) 평가가 필수적으로 수반되어야 한다. 그 활용 사례로, Hewlett과 Winn은 모성 외 수유(allomaternal nursing)의 문화 간 분포를 추정하기 위해 현지 관찰 자료와 함께 eHRAF에서 '영유아 돌봄(Infant Care, OCM 862)' 주제를 중심으로 문헌 서베이를 결합하여 행동 변수를 구축하였다[36]. 또한 Hrnčíř는 수렵채집 사회의 무기 사용에 대한 정량 비교 연구에서 SCCS 표본을 대상으로 '무기(Weapons, OCM 411)' 주제와 구체적인 무기 명칭(예: club, throwing stick 등) 키워드를 결합한 다층 검색을 통해, 단순 키워드 검색의 한계를 극복하고 명시적 코딩 규칙 기반의 변수화를 수행했다[37].

셋째, 중범위 이론(Middle-Range Theory)의 구축이다. 생물인류학과 고고학은 발굴된 물질적인 지표로부터 과거의 사회적 제도나 행위를 역추적해야 하므로 중범위 이론에 대한 의존도가 높다[38,39]. eHRAF World Cultures가 제공하는 광범위한 민족지 비교 자료는 특정 매장 패턴이나 골격의 외상(trauma) 분포가 실제로 어떤 형태의 친족 구조, 폭력 조직 방식, 혹은 돌봄 체계와 연관되어 발현되는지에 대한 경험적인 범위 근거를 제공하여 물질과 행위를 연결하는 이론적

틀의 구축을 지원한다.

넷째, 연구 설계 보조(Design Scaffolding)이다. 본격적인 대규모 데이터 코딩이나 통계 분석에 앞서, 연구자가 설정한 핵심 개념의 생태적 타당성을 점검하는 도구로 활용된다. 예컨대 '영양'이라는 변수가 단순한 칼로리 섭취량인지, 분배와 기근 대응(OCM 146, 260)을 포함하는지 혹은 '질병 치료'라는 변수가 실제로 어떤 하위 관행들로 구성되는지 등을 사전에 파악할 수 있다. 이를 통해 골격 및 유전 자료에서 측정할 생물학적 지표와 텍스트에서 코딩할 문화적 지표 간의 대응 관계를 한층 정교하게 설계할 수 있다.

5. eHRAF 기반 연구의 단계별 접근 절차와 전략

eHRAF 기반 연구의 재현성을 확보하기 위한 단계별 접근 절차와 전략을 아래와 같이 제시한다. 이들 단계별 접근의 핵심 원칙은 다음과 같다. 첫째, eHRAF는 읽기 자료가 아니라 코딩 가능한 텍스트 코퍼스이며, 검색식과 코딩 규칙을 논문에 보고하여 분석의 재현성을 반드시 확보해야 한다. 둘째, eHRAF World Cultures와 eHRAF Archaeology는 경쟁 관계가 아니라 상호 보완 관계이며, 연구 질문이 요구하는 분석 단위에 따라 전략적으로 선택하거나 통합하여 사용해야 한다. 셋째, eHRAF는 자동으로 정량 변수를 제공하지 않으므로, 연구팀이 코드북을 설계하고 코더 간 신뢰도를 함께 보고해야 한다[9,40].

6. 연구 질문과 분석 단위의 설정

앞서 논의한 분석 단위의 차이를 간과할 경우, 자료 연결 단계에서 시공간적 불일치, 생태적 오류(ecological fallacy), 표본 중복 등의 치명적인 문제가 발생할 수 있다. 이를 방지하고 단계별 데이터 수집의 과정을 보조하는 전체 3단계의 의사결정 워크플로를 아래 Fig. 2에 제시하고자 한다.

연구자가 다루려는 질문이 현대 또는 근현대 사회의 생태, 건강, 생계, 사회 조직(예: 돌봄, 혼인 규범, 질병 관념과 치료, 종교적 믿음과 금기 등)과 직접 관련될수록 eHRAF World Cultures가 적합하며, 특정 고고학 전통 또는 유적군의 정착, 경제, 사회 계층, 매장 및 의례, 물질문화와 관련될수록 eHRAF Archaeology가 더 적합하다[1,4]. 그러나 생물인류학 영역에서는 골격 자료의 해석을 민족지 비교로 보정하거나 장기적 시계열을 복원하려는 경우, 동일 연구 안에서 두 데이터베이스가 동시에 필요해지는 경우가 많다. 그렇기에 생물인류학 연구에서 특정 행동(예: 매장)의 변이를 해석할 때, 민족지 기반 분산을 eHRAF World Cultures에서 추정하고, 고고학적 전통에서 관측되는 실제 양상을 eHRAF Archaeology에서 추출하는 삼각측량 기반의 통합적 설계 방식으로 증거 다층화 작업을 수행할 수 있다(Fig. 2).

이를 위해, 연구자는 우선 '무엇을 관측치로 볼 것인가'를 결정해야 한다. 다시 말해, 가장 먼저 연구 질문을 보다 구체적인 구성 요소들로 분해하는 작업(Stage 1)부터 시작해야

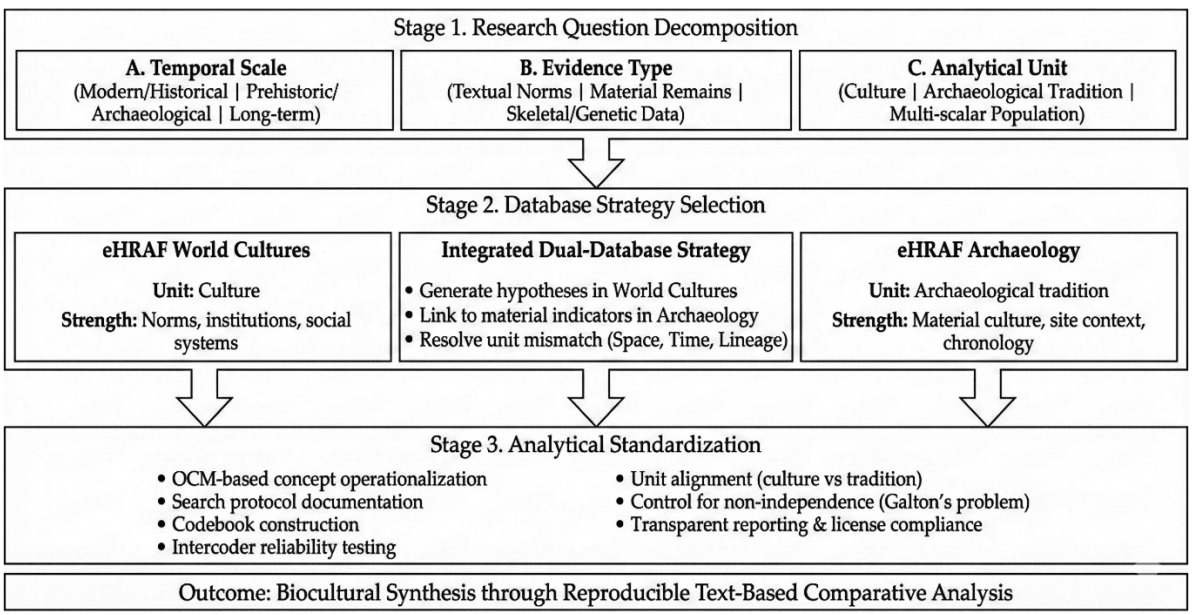

Fig. 2. A three-stage analytical workflow for integrating eHRAF databases in bioanthropological research. This framework outlines the systematic protocol to translate comparative texts into reproducible biocultural data. Stage 1 deconstructs the research question to determine the appropriate temporal scale, evidence type, and analytical unit. Stage 2 guides database selection, including an integrated dual-database triangulation strategy. Stage 3 standardizes variable operationalization, inter-coder reliability, and non-independence control.

한다. 예를 들어, '외상 빈도가 높은 집단에서 폭력의 조직 양상이 어떠했는가'라는 질문은 외상(신체 손상), 폭력/전쟁, 무기, 사회조직, 매장 맥락 등의 개별 요소로 분해할 수 있다. 이후 각 요소를 OCM 코드 후보로 매핑하고, 상위-하위 OCM 코드 계층 구조를 활용해 검색 범위를 조정해 나간다. 이때, OCM은 '관련 단락이 있을 법한 위치'를 좁혀 주지만, 최종적으로 사용될 변수는 연구자가 선택하고 정의해야만 한다[9,33,40].

이렇게 질문의 분해와 데이터베이스 선택(Stage 2)이 끝난 후, 추출된 텍스트 코퍼스를 실제 생물학적 지표와 연결하는 단계(Stage 3)에서는 '단위 정렬(Unit Alignment)'의 문제가 핵심적인 과제로 대두된다. eHRAF World Cultures의 단위는 과거 특정 시점의 민족지 기록 속 집단에 가까운 반면[41], eHRAF Archaeology의 단위는 시간적, 공간적으로 넓은 분산을 가진다. 따라서 특정 유적의 골격 표본 데이터를 별도의 처리 없이 고고학 전통 단위의 텍스트, 혹은 인류학 문화 단위의 텍스트와 대응시킬 경우 일종의 시간 평균화(time-averaging) 오류가 발생할 수 있다. 이를 통제하기 위해 시기 범위를 좁힐 수 있는 하위 문헌을 정밀하게 선별하고, 유적의 교류권과 전통 내 위치를 명시하며, 동일 유적에 대한 복수의 보고서가 있을 경우 명시된 체계적 기준을 바탕으로 중복을 제거하는 단위 정렬 과정이 선행되어야 한다.

단위 정렬이 끝난 후에는 추출된 단위와 증거의 '중복(Duplication)' 문제를 통제해야 한다. eHRAF의 검색 결과는 기본적으로 단락 단위로 반환되는데, 이 단락을 그대로 1차 데이터 포인트로 사용할 경우 표본의 크기는 커지지만 동일 문서나 저자에서 기인한 증거의 과대 대표(overrepresentation) 문제가 필연적으로 발생한다. 따라서 연구의 분석 단위를 문단, 문헌, 혹은 문화(전통) 중 무엇으로 설정할지에 대한 명시적인 기준이 필요하며, 연구자들은 이 문제를 제어하기 위해 크게 두 가지 전략을 활용한다.

첫째, 단락을 1차 관측치로 간주하는 '상향식 집계' 방식이다. 개별 단락을 분절된 '증거 조각'으로 보아 단락 수준에서 세밀하게 코딩을 수행하거나, 연속된 단락들을 하나로 묶어 문서 단위의 텍스트 기록으로 요약한 뒤 이를 다시 문화와 전통 수준으로 집계한다. 이 방식은 민족지의 자체 맥락과 예외 상황을 정교하게 다룰 수 있다는 장점이 있으나, 자료가 풍부한 문화의 단락(관측치)이 과도하게 늘어나 표본이 부풀려지는 문제가 있다.

둘째, 문화 또는 전통 자체를 1차 관측치로 간주하는 '직접 축약' 방식이다. 수집된 텍스트들을 해당 단위에 대해 완전히 집계(aggregation)하여 명시적 기준 하에 특정 관행의 '존재 여부', '우세한 규범', '허용 조건'과 같은 변수로 축약해버리는 전략이다. 이 방식은 다문화 간 비교 통계 분석에 매우 유리하여 현재 많은 연구에서 지속적으로 채택하고 있으나, 문헌 내부에 존재하는 서술의 이질성과 미세한 시간적 변화 양상이 소실된다는 뚜렷한 단점이 있다.

연구 목적에 따라 어떤 전략을 취하든 핵심은 통제 규칙의 투명한 공개이다[40]. 대표적인 예시로, Lightner 등은 질적인 비교문화 텍스트를 코딩 가능한 단위로 전환하는 과정에서 단락의 중복을 어떻게 통제했는지 표본 추출, 코딩, 집계의 세부 규칙을 투명하고 명확하게 보고하여 분석의 타당성을 입증한 바 있다[42].

7. OCM과 키워드 결합 기반 다층 개념 검색 전략

이와 같은 OCM 중심 검색은 개념 기반 맥락 회수를 가능하게 하지만, 세부 행위의 미세한 차이를 포착하기 위해 키워드와의 조합이 필요하다. 그러나 단순 키워드 중심으로만 접근하면 효율이 급격히 떨어진다. 예를 들어 'burial'만으로 검색하면 매장과 무관한 맥락(예: 매립, 은폐 등의 비유적 용법)이 대량으로 혼입되는데, 이는 eHRAF를 처음 사용하는 연구자가 가장 자주 겪는 시행착오 중 하나이다. eHRAF는 문단 단위의 검색 환경이므로, 긴 문맥에서만 등장하는 키워드가 누락될 수 있으며, 다의어 때문에 잡음이 급증할 수 있다. 그렇기에 OCM 주제와 불리언 연산, 키워드를 결합한 개념 기반 검색이 수행되어야 한다. 예를 들어, '모유수유'는 breastfeeding, nursing, colostrum, wet nurse 등 여러 표현이 가능하며, '매장'은 burial, cremation, exposure, ossuary 등 다양한 행위를 포괄한다. 따라서 1) 연구 질문을 가장 가까운 OCM 주제로 번역하여 텍스트를 넓게 회수하고, 2) 자료가 과다할 경우 구체적 키워드로 좁히며, 3) 와일드카드 키워드(*)로 누락을 보완하는 '3단계 다층 검색 전략'이 유용하다. 또한, 검색 결과를 재현이 가능하도록 만들기 위해서는 사용한 데이터베이스, 표본 범위(SCCS, EA, PSF 등), 검색식, 필터, 중복 제거 규칙 등을 상세한 로그로 기록하고 보고해야 한다[2-4,27,33,40,42].

8. 코드북 구축과 코더 간 신뢰도의 확보

확보된 텍스트 코퍼스를 분석 목적에 맞는 명목형 또는 서열형 변수로 정량화하는 작업이 이어진다[9,40]. 이때 발생할 수 있는 코딩의 임의성 문제는 연구에서 가장 취약해지기 쉬운 지점으로, 해당 과정에서 코딩의 임의성을 통제하지 못하면 결과의 신뢰도는 심각하게 훼손된다. 동일 문단을 두 연구자가 읽었을 때 같은 코드가 부여되는지, 코드의 정의가 일관적인지, 결측과 모호한 사례를 어떻게 처리했는지 등이 불명확하면 결과의 재현이 불가능해지기 때문이다. 그렇기에, eHRAF 기반 연구에서는 조작적 정의에 기반한 개념적 등가

성(conceptual equivalence)의 확보와 함께 아래와 같은 코딩 프로토콜을 준수하는 것이 권장된다[40].

첫째, 각 변수의 정의, 포함 기준, 제외 기준, 경계 사례 처리 기준, 예시 문단, 결측 처리 규칙에 대한 내용을 모두 명시적으로 포함하는 코드북을 작성한다. 둘째, 확보된 표본 중 임의로 선정된 문화 또는 전통 20~50개 문단에 대한 파일럿 코딩을 진행하면서 코드북을 수정한다. 파일럿 코딩 없이 본 코딩에 진입하면 중반에 변수 정의를 재조정해야 하는 경우가 흔하므로, 이 단계를 생략하지 않는 것이 권장된다. 셋째, 2인 이상의 코더가 전체의 일정 비율(예: 20%)을 독립적으로 코딩한다[40]. 코딩 결과의 객관성을 확보하기 위해 코더 간 일치도를 산출하여 함께 보고해야 한다. 명목 변수에는 Cohen's kappa, 서열 또는 혼합 수준 변수에는 Krippendorff's alpha가 널리 사용되며, 범주 분포가 극단적으로 편향된 경우에는 Gwet's AC1도 고려할 수 있다[9,40]. 신뢰도는 완벽할 필요가 없지만, 낮은 신뢰도는 변수 정의가 불안정하다는 신호이므로 코드북을 수정하고 재코딩하는 과정이 필요하다. 넷째, 코더 간 불일치 문단을 어떻게 논의하고 최종값을 확정했는지, 합의 후 코드북을 어떻게 업데이트했는지와 같은 합의 절차의 기록을 남긴다. 투명성과 재현성을 담보하기 위해 코더의 수와 학문적 배경, 코딩 단위, 훈련 절차, 불일치 해결 절차, 신뢰도 지표, 코드북 수정 이력 등을 보관 및 보고하여야 하며, 최종 코딩 결과와 합의 절차는 반드시 논문에 명시되어야 한다[9,40,42].

9. 공간적 자기상관성의 통제와 비교

eHRAF 자료에 기반한 세계 여러 집단 간 비교문화 연구는 공간적 자기상관성에 따른 비독립성 문제로 인한 해석상의 어려움을 동반한다. 개별 관찰값의 비독립적 유사성은 공유된 생태적 조건만이 아니라, 공간적으로 인접한 집단 간의 확산(diffusion), 계통적으로 공유되는 조상 집단(ancestor group)의 존재 가능성 등의 이유로 인해서도 나타날 수 있다. 인류학자 에드워드 타일러(Edward Burnett Tylor)의 선구적 비교문화 연구에 대한 갈톤의 비판 이후, 인류학자들은 이 문제를 해결하기 위해 다양한 방법을 개발했다[9,28-30,32].

현재 고려 가능한 통제 방법으로, eHRAF World Cultures에서는 사전 표본 추출 과정에서 공간적 독립성을 일정 수준 확보한 기존의 SCCS, PSF, 혹은 SRS 등을 표본틀로 사용하는 것이 있다. eHRAF Archaeology에서는 마찬가지로 기존 SRS를 표본틀로 사용하거나, 전통의 시간적 중첩이 만들어내는 직렬 자기상관(serial autocorrelation)을 고려해 시계열 단위 또는 전통의 중간 시점만을 대표값으로 사용하는 방식 등을 활용할 수 있다. 그러나 해당 표본틀의 사용만으로는 비독립성 문제를 완벽히 통제할 수 없기에, 현재의 연구에서는 추가로 지역 더미 혹은 지리적 거리 기반 가중치 행렬을 공변량으로 포함하거나, 지역 또는 언어 계통을 무작위 효과(random effect)로 처리하는 혼합 모형(mixed effect model)이 계통 비교 방법(phylogenetic comparative methods)과 함께 사용된다[43-45].

진화생물학에서 주로 사용되는 계통 비교 방법이 생물인류학 연구에 도입된 것은, Cavalli-Sforza 등의 선구적인 유전-언어 상관 연구와 Mace, Pagel 등의 문화 계통 연구에 기반한다[46-50]. 해당 접근법은 언어 계통수(linguistic tree)를 대리 지표(proxy)로 삼아, 전 세계 집단의 문화적, 고고학적 유사성이 우연인지, 공통 조상으로부터의 계승인지, 혹은 전파에 의한 것인지 그 시나리오를 베이지안 통계에 기반해 추론하는 강력한 도구이다[46-50]. 이러한 계통 비교는 eHRAF의 텍스트에서 추출 및 코딩된 변수를 외부 데이터베이스(예: 언어 식별자, 생태적 변수들을 제공하는 D-PLACE, 혹은 각 언어 식별자 집단의 유전적 거리에 대한 정보를 제공하는 GeLaTo)와 결합하는 방식으로 구현이 가능하다. D-PLACE와 GeLaTo 같은 통합형 데이터베이스는 이들 영역을 가로지르는 연결 인프라로 기능한다(Table 1).

만약 표본이 특정 대륙 또는 문화(혹은 전통) 유형에 편중된다면, 결과는 보편적 일반화가 아니라 해당 범주의 조건부 일반화로 보고되어야 한다. 따라서 표본 선택의 근거, 포함 및 제외 기준, 표본의 분포를 논문에 기술하는 것이 필요하다[40].

10. 재현성을 위한 분석과 보고의 투명성 요건

eHRAF 기반 비교문화 연구의 재현성을 온전히 확보하려면, 연구에 사용된 텍스트 자체뿐만 아니라 데이터가 추출되고 가공된 '알고리즘적 절차' 전체를 투명하게 공유해야 한다. 연구자는 논문 작성 시 최소한 다음의 정보들을 명시적으로 보고해야 한다[40]. 여기에는 연구 질문과 분석 단위, 사용한 eHRAF 데이터베이스 및 접속 일자, 표본 선택 방식(SCCS, PSF 등)과 포함 기준, 상세한 다층 검색식(OCM 및 키워드 조합), 최종 확보된 문헌 및 문단 수, 코드북 요약 및 신뢰도 지표, 집계 규칙과 비독립성 통제 모형 등이 포함된다. 또한 eHRAF 문헌을 인용할 때는 원출판 정보와 eHRAF 시스템 내 접근 정보를 병기하는 것이 원칙이다. 이러한 최소 보고 요건들을 체계적으로 점검하고 실제 데이터 추출에 바로 활용할 수 있도록, 체크리스트와 통합형 코드북 템플릿을 Table 3에 제시하였다.

11. eHRAF 기반 연구의 한계와 전망

eHRAF 텍스트 코퍼스를 활용한 비교 연구에는 몇 가지 방법론적 한계가 수반된다. 첫째, 민족지 텍스트 내 '관찰의 부

Table 3. Standardized reporting matrix and ethnographic metadata extraction template for eHRAF-based cross-cultural research

Research stage	Methodological reporting criteria (Reproducibility checklist)	Source metadata & ethnographic extraction fields (Codebook elements)
Stage 1: Search & sampling protocol	[ ] Database & Sampling Frame: Explicit selection of eHRAF World Cultures/Archaeology and sampling frame (e.g., SCCS).	• Society Identifiers: OWC culture code, exact society name, and geographic region.
	[ ] Multilayered Search Logic: Documentation of applied OCM descriptors (e.g., 146 Nutrition), keyword combinations, wildcards (*), and Boolean operators.	• Document Filtering: Strict alignment of selected ethnographies with the designated time- and place-foci.
Stage 2: Source evaluation & contextualization	[ ] Temporal Alignment: Procedures for aligning the focal year of the ethnography with external data to prevent time-averaging.	• Bibliographic Metadata: Full document citation, author credentials, and specific publication year.
	[ ] Source Quality Control: Protocols for assessing the reliability of the ethnographic record and author bias.	• Temporal Context: Exact focal year of observation matching the comparative sample criteria.
		• Search Log: Date of extraction and coder identification.
Stage 3: Variable operationalization (Core codebook)	[ ] Operational Definitions: Clear definitions of the target variable (e.g., distinguishing natural vs. supernatural attributions).	• Raw Ethnographic Extract: Exact direct quotation of the retrieved paragraph serving as evidence, including the specific page number.
	[ ] Coding Scales & Boundary Rules: Explicit integer code assignments (e.g., 1 = Rare, 2 = Common, 99 = Not enough information).	• Assigned Integer Code: The specific numerical value assigned to the extract based on the codebook.
	[ ] Confidence Rating System: Implementation of a formalized scale to quantify coder certainty.	• Confidence Score: A rating from 1 (very unconfident) to 4 (very confident).
		• Summary Statement: A brief qualitative memo summarizing the rationale in bold.
Stage 4: Reliability & aggregation control	[ ] Inter-coder Reliability (ICR): Number of independent coders, training procedures, and specific ICR metrics (e.g., Cohen's Kappa).	• Coder Identification: IDs of primary and secondary coders.
	[ ] Discrepancy Resolution: Standardized protocol for resolving coding disagreements, including the reporting of preferred intermediate codes prior to consensus.	• Intermediate & Consensus Code: Documentation of initial preferred codes and the final agreed-upon variable value.
	[ ] Aggregation & Spatial Control: Rules for aggregating paragraph-level data into a single culture-level point, and statistical models applied to control for Galton's problem.	• Aggregated Culture-Level Score: The finalized data point representing the entire society.
		• Non-independence Proxy: External identifiers used for spatial control (e.g., glottocode).

Serving a dual purpose as a reproducibility checklist and a codebook template, this matrix systematizes the minimum reporting requirements for eHRAF-based cross-cultural research across four stages: search protocol, source evaluation, variable operationalization, and reliability control.

재'를 '행동의 부재'로 오인하는 오류를 경계해야 한다. 기록자의 관심사, 체류 기간, 정보제공자의 접근성에 따라 기록의 공백은 필연적으로 발생하므로, 단순한 기술의 누락은 '결측'으로, 명시적 부재 보고만 '없음'으로 코딩하는 보수적 원칙이 요구된다[51]. 둘째, 동일 문헌 내 반복 서술에 의한 표본 과대 대표와 문단 단위 결과의 문화 내 비독립성은 사전에 설정된 집계 규칙으로 통제되어야 한다. 셋째, eHRAF World Cultures의 민족지 기록 시점과 eHRAF Archaeology 표본의 시점 간 불일치가 해석을 왜곡할 수 있으므로, 시간 범위를 명시하고 민감도 분석을 병행해야 한다[41]. 넷째, eHRAF 원문 텍스트는 라이선스상 대량 배포가 제한되므로, 원문 인용은 필요 최소한으로 제한하되 코드북, 검색 로그, 집계 규칙, 분석 코드 등 절차 자료를 적극 공개하여 재현성을 확보해야 한다[40].

시스템 차원에서는, 개별 문헌 및 단락 단위의 DOI 연계 강화와 개방형 API를 통한 프로그래밍 기반 검색·추출 환경의 구축이 기대된다. 나아가 가장 주목할 과제는 대규모 언어 모델(LLM) 및 자연어 처리(NLP) 기술과의 결합이다. 단어 임베딩을 활용한 민족지 텍스트의 의미 구조 분석이나, LLM 기반의 구조화된 추출을 통해 수동 코딩의 병목을 완화하는 시도가 이미 진행되고 있으며[52-55], OCM 코드는 이러한 자동화 과정에서 약한 지도학습의 라벨로 기능할 수 있다. 라이선스 범위 내에서 이를 검증하는 연구가 필요하다.

결 론

eHRAF 기반 비교 연구의 핵심 장점은 OCM 색인을 통한 개념 기반 검색으로 체계적 맥락 회수가 가능하다는 점이다. 여기에 SCCS, PSF 등 표본 정보와 계통 비교법을 결합하면 공간적 비독립성을 통제하는 엄밀한 연구 설계가 가능하며, 문단 단위 텍스트 코퍼스는 코드형 데이터셋이 놓치기 쉬운 예외와 조건까지 변수 설계에 반영할 수 있게 해준다. 특히 eHRAF World Cultures와 eHRAF Archaeology를 통합하는 삼각측량 설계는, 분석 단위의 차이에서 비롯되는 시간 평균화와 생태적 오류를 완화하면서 생물문화적 통합을 이끌어내는 유력한 전략이 될 수 있다. 본고에서 제시한 보고 매트릭스 겸 코드북 템플릿(Table 3)은 이러한 절차의 재현성을 실무적으로 뒷받침하기 위한 것이다.

다만 민족지 텍스트의 기록 편향, 라이선스 제약, 시간 불일치 등의 한계는 여전하며, 연구자는 코드북, 검색 로그, 집계 규칙 등 절차 자료의 투명한 공개를 통해 저작권을 존중하면서도 재현성을 확보해야 한다. 향후에는 OCM 코드를 약한 지도학습의 라벨로 활용하는 자연어 처리 및 대규모 언어 모델 기반의 자동 코딩이 본격화될 것으로 예상되며, 라이선스 범위 내에서 이를 검증하는 연구가 필요하다.

결론적으로, eHRAF는 생물인류학에서 문화적 설명 변수를 구축하는 유용한 텍스트 코퍼스이며, 그 활용의 성패는 데이터베이스 자체에 대한 이해보다 연구 설계의 엄격함, 특히 단위 정합성과 재현성 보고에 달려 있다.

REFERENCES

1. Roe SK. A brief history of an ethnographic database: The HRAF collection of ethnography. Behav Soc Sci Libr. 2007; 25:47-77. https://doi.org/10.1300/J103v25n02_03
2. Peregrine PN. Cross-cultural comparative approaches in archaeology. Annu Rev Anthropol. 2001;30:1-18. https://doi.org/10.1146/annurev.anthro.30.1.1
3. Ember M, Ember CR. Testing adaptational explanations of culture: The utility of eHRAF World Cultures and eHRAF Archaeology. General Anthropology Bulletin [Internet]. 2009; 16 [cited 2026 Mar 22]. Available from: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=15371727&asa=N&AN=64301253&h=KT7pQ1aA69PgXY6pt5NmZdKjUkLnyh4Zay9QQ5ZG0s13zM3DcdDZ%2F637k9658B8%2FG2UMRRKU0idWvAeaBDoHAg%3D%3D&crl=c.
4. Fischer MD, Ember CR. Big data and research opportunities using HRAF databases. In: Chen SH, editor. Big Data in Computational Social Science and Humanities [Internet]. Cham: Springer International Publishing; 2018 [cited 2026 Feb 2]. p. 323-36 (Computational Social Sciences). Available from: http://link.springer.com/10.1007/978-3-319-95465-3_17.
5. Murdock GP. Feasibility and implementation of comparative community research: With special reference to the human relations area files. Am Sociol Rev. 1950;15:713. https://doi.org/10.2307/2086603
6. Ford CS. Human relations area files: 1949-1969 a twenty-year report. Behav Scie Notes. 1970;5:1-61. https://doi.org/10.1177/106939717000500101
7. White DR, Brudner-White LA. The Murdock legacy: The ethnographic atlas and the search for a method. Behav Sci Res. 1988;22:59-81. https://doi.org/10.1177/106939718802200107
8. Ember M. Evolution of the human relations area files. Cross-Cult Res. 1997;31:3-15. https://doi.org/10.1177/106939719703100101
9. Ember CR, Ember M. Cross-cultural research methods. 2nd ed. Rowman Altamira; 2009.
10. Binford LR. Constructing frames of reference: an analytical

method for archaeological theory building using ethnographic and environmental data sets [Internet]. University of California Press; 2001 [cited 2026 Feb 10]. Available from: https://books.google.com/books?hl=ko&lr=&id=I_kkDQAAQBAJ&oi=fnd&pg=PR11&dq=constructing+frames+of+reference&ots=Jcx-hk-glv&sig=DrM2sWpeM_jlWultB4iBUZ8dd08.
11. Marwick B, Johnson A, White D, Eff EA. binford: Binford's Hunter-Gatherer Data [Internet]. 2016 [cited 2026 Feb 10]. p. 0.1.0. Available from: https://CRAN.R-project.org/package=binford.
12. Turchin P, Brennan R, Currie T, Feeney K, Francois P, Hoyer D, et al. Seshat: The global history databank. Cliodynamics [Internet]. 2015;6 [cited 2026 Feb 23]. Available from: https://escholarship.org/uc/item/9qx38718.
13. Watts J, Sheehan O, Greenhill SJ, Gomes-Ng S, Atkinson QD, Bulbulia J, et al. Pulotu: Database of Austronesian supernatural beliefs and practices. PLoS One. 2015;10:e0136783.
14. Slingerland E, Sullivan B. Durkheim with data: The database of religious history. J Am Acad Relig. 2017;85:312-47.
15. Diessel H, Dryer MS, Haspelmath M. The world atlas of language structures online. 2013.
16. Bertolo M, Snarskis M, Kyritsis T, Yurdum L, Bainbridge CM, Atwood S, et al. The expanded natural history of song discography, a global corpus of vocal music. Open Mind. 2025;9:844-63.
17. Wood AL, Kirby KR, Ember CR, Silbert S, Passmore S, Daikoku H, et al. The global jukebox: A public database of performing arts and culture. PLoS One. 2022;17:e0275469.
18. Consortium 1000 Genomes Project. A global reference for human genetic variation. Nature [Internet]. 2015 [cited 2026 Feb 23]. p. 68. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4750478/.
19. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Res. 2010;39(suppl_1):D913-9.
20. González-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MHT, Silva ALS da, et al. Allele frequency net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res. 2015;43:D784-8.
21. Gonzalez-Galarza FF, McCabe A, Santos EJM dos, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. 2020;48:D783-8.
22. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538:201-6.
23. Kirby KR, Gray RD, Greenhill SJ, Jordan FM, Gomes-Ng S, Bibiko HJ, et al. D-PLACE: A global database of cultural, linguistic and environmental diversity. PLoS One. 2016;11:e0158391. https://doi.org/10.1371/journal.pone.0158391
24. Barbieri C, Blasi DE, Arango-Isaza E, Sotiropoulos AG, Hammarström H, Wichmann S, et al. A global analysis of matches and mismatches between human genetic and linguistic histories. Proc Natl Acad Sci USA. 2022;119:e2122084119. https://doi.org/10.1073/pnas.2122084119
25. Murdock GP. Ethnographic Atlas, Installments I-XXVII. Ethnology. 1962 1971;1-10.
26. Gray JP. A corrected ethnographic atlas. World Cultures. 1999;10:24-85.
27. Peregrine PN. Outline of archaeological traditions [Internet]. HRAF; 2001 [cited 2026 Feb 23]. Available from: https://hraf.yale.edu/wp-content/uploads/2020/12/Outline-of-Archaeological-Traditions-Intro.pdf.
28. Naroll R. Two solutions to Galton's problem. Philos Sci. 1961;28:15-39. https://doi.org/10.1086/287778
29. Naroll R. Galton's problem: The logic of cross-cultural analysis. Soc Res. 1965;428-51.
30. Ember M. An empirical test of Galton's problem. Ethnology. 1971;10:98-106.
31. Murdock GP, White DR. Standard cross-cultural sample. Ethnology. 1969;8:329. https://doi.org/10.2307/3772907
32. Naroll R. The proposed HRAF probability sample. Behav Sci Notes. 1967;2:70-80. https://doi.org/10.1177/106939716700200202
33. Murdock GP. Outline of cultural materials [Internet]. 1961 [cited 2026 Feb 10]. Available from: https://eric.ed.gov/?id=ED044998.
34. Murdock GP. Outline of world cultures [Internet]. 1969 [cited 2026 Feb 10]. Available from: https://eric.ed.gov/?id=ED044997
35. Ensor BE, Irish JD, Keegan WF. The bioarchaeology of kinship: Proposed revisions to assumptions guiding interpretation. Curr Anthropol. 2017;58:739-61. https://doi.org/10.1086/694584
36. Hewlett BS, Winn S. Allomaternal nursing in humans. Curr Anthropol. 2014;55:200-29. https://doi.org/10.1086/675657
37. Hrnčíř V. The use of wooden clubs and throwing sticks among recent foragers: Cross-cultural survey and implications for research on prehistoric weaponry. Hum Nat. 2023;34:122-52. https://doi.org/10.1007/s12110-023-09445-3
38. Raab LM, Goodyear AC. Middle-range theory in archaeology: A critical review of origins and applications. Am Antiq. 1984;49:255-68.
39. Binford LR. Archaeology as anthropology. Am Antiq. 1962;28:217-25.
40. Slingerland E, Atkinson QD, Ember CR, Sheehan O, Muthu-

krishna M, Bulbulia J, et al. Coding culture: Challenges and recommendations for comparative cultural databases. Evolut Hum Sci. 2020;2:e29. https://doi.org/10.1017/ehs.2020.30
41. Divale WT. Temporal focus and random error in cross-cultural hypothesis tests. Behav Sci Res. 1975;10:19-36. https://doi.org/10.1177/106939717501000102
42. Lightner AD, Heckelsmiller C, Hagen EH. Ethnoscientific expertise and knowledge specialisation in 55 traditional cultures. Evol Hum Sci. 2021;3:e37.
43. Syme KL, Garfield ZH, Hagen EH. Testing the bargaining vs. inclusive fitness models of suicidal behavior against the ethnographic record. Evol Hum Behav. 2016;37:179-92.
44. Agey E, Morris A, Chandy M, Gaulin SJC. Arranged marriage often subverts offspring mate choice: An HRAF-based study. Am Anthropol. 2021;12:861-78. https://doi.org/10.1111/aman.13656
45. Mori S. Estimating polygyny rates among hunter-gatherers: A statistical model for historical source criticism with a Yamana case study. Lett Evol Behav Sci. 2024;15. https://doi.org/10.5178/lebs.2024.116
46. Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes [Internet]. Princeton university press; 1994 [cited 2026 Feb 23]. Available from: https://books.google.com/books?hl=ko&lr=&id=FrwNcwKaUKoC&oi=fnd&pg=PA270&dq=The+history+and+geography+of+human+genes&ots=HpaTWeEFdc&sig=uWVasaprB57Pn5WujkPL0Z4mjQk.
47. Mace R, Pagel M, Bowen JR, Otterbein KF, Ridley M, Schweizer T, et al. The comparative method in anthropology [and comments and reply]. Curr Anthropol. 1994;35:549-64. https://doi.org/10.1086/204317
48. Mace R, Holden CJ. A phylogenetic approach to cultural evolution. Trends Ecol Evol. 2005;20:116-21.
49. Mace R, May Zhang H. Cross-cultural comparative methods for testing evolutionary hypotheses [Internet]. 2023 [cited 2025 Dec 31]. Available from: https://academic.oup.com/edited-volume/45648/chapter/406029982.
50. Fortunato L, Holden C, Mace R. From bridewealth to dowry?: A bayesian estimation of ancestral states of marriage transfers in Indo-European groups. Hum Nat. 2006;17:355-76. https://doi.org/10.1007/s12110-006-1000-4
51. Bliege Bird R, Codding BF. Promise and peril of ecological and evolutionary modelling using cross-cultural datasets. Nat Ecol Evol. 2021;6:6-8. https://doi.org/10.1038/s41559-021-01579-w
52. Fischer MD. HDNS-I: Infrastructure for knowledge linkages from ethnography of world societies. NSF Award. 2020;20:24286.
53. Alfano M, Cheong M, Curry OS. Moral universals: A machine-reading analysis of 256 societies. Heliyon [Internet]. 2024;10 [cited 2026 Mar 22]. Available from: https://www.cell.com/heliyon/fulltext/S2405-8440(24)01971-6?uuid=uuid%3A605c3966-75a1-4c1e-8d2f-e8ef9b0599c6.
54. Dubourg E, Thouzeau V, Baumard N. A step-by-step method for cultural annotation by LLMs. Front Artif Intell. 2024;7:1365508.
55. Syme KL, Motos N, Placek CD. Generating units of cultural analysis with large language models: methods and validation for scalable cross-cultural research. R Soc Open Sci [Internet]. 2026;13 [cited 2026 Mar 22]. Available from: https://royalsocietypublishing.org/rsos/article/13/2/251766/480262.

간추림 : 생물인류학은 고대 DNA, 안정동위원소, 영상의학, 골격형태 분석의 고도화로 과거 인류 집단의 이동, 식생활, 질병, 생활사 변이를 정밀하게 추정할 수 있게 되었으나, 동일한 생물학적 신호가 서로 다른 사회 제도와 생계 전략, 폭력 양상, 매장 관행, 의료 행위 아래에서 상이한 선택압과 노출 환경을 반영할 수 있다는 점에서 해석의 맥락 결핍 문제가 지속된다. 이러한 맥락 정보는 주로 민족지와 발굴 보고서 등 텍스트로 존재하지만, 개념 기준의 체계적 회수와 비교 가능한 단위로의 변환이 어렵다. 본 종설은 인간관계지역파일(Human Relations Area Files, HRAF)의 온라인 데이터베이스인 eHRAF World Cultures와 eHRAF Archaeology를 생물인류학 연구에서 배경 참고문헌이 아니라 재현 가능한 텍스트 코퍼스(text corpus)로 활용하는 방법론을 검토한다. eHRAF의 문화자료개요(Outline of Cultural Materials, OCM) 색인을 활용한 개념 기반 검색의 논리를 서술하고, 서로 다른 분석 단위(문화 단위 대 고고학적 전통 단위)를 고려하여 시간 평균화 및 생태적 오류를 통제하는 '삼각측량(triangulation)' 기반의 데이터베이스 통합 원칙을 제시한다. 또한 맥락 회수, 변수화, 중범위 이론(middle-range theory) 구축, 연구 설계 보조라는 네 가지 활용 방식을 정리하고, 검색식 기록, 추출 단위 설정, 코드북 작성, 코더 간 신뢰도 평가, 집계 규칙, 단위 정렬(unit alignment), 갈톤의 문제(Galton's problem)와 같은 비독립성 통제, 보고 최소요건을 포함한 단계별 워크플로를 제안한다. 마지막으로 표본 대표성, 기록 편향, 시간 불일치, 라이선스의 제약 등을 한계로 논의하고, OCM을 약한 지도학습(weak supervision)의 라벨로 활용하는 등 인공지능과 자연어처리 기반의 확장 가능성을 제시한다.

찾아보기 낱말 : 전자인간관계지역파일, 교차문화 연구, 문화자료개요, 생물문화적 통합