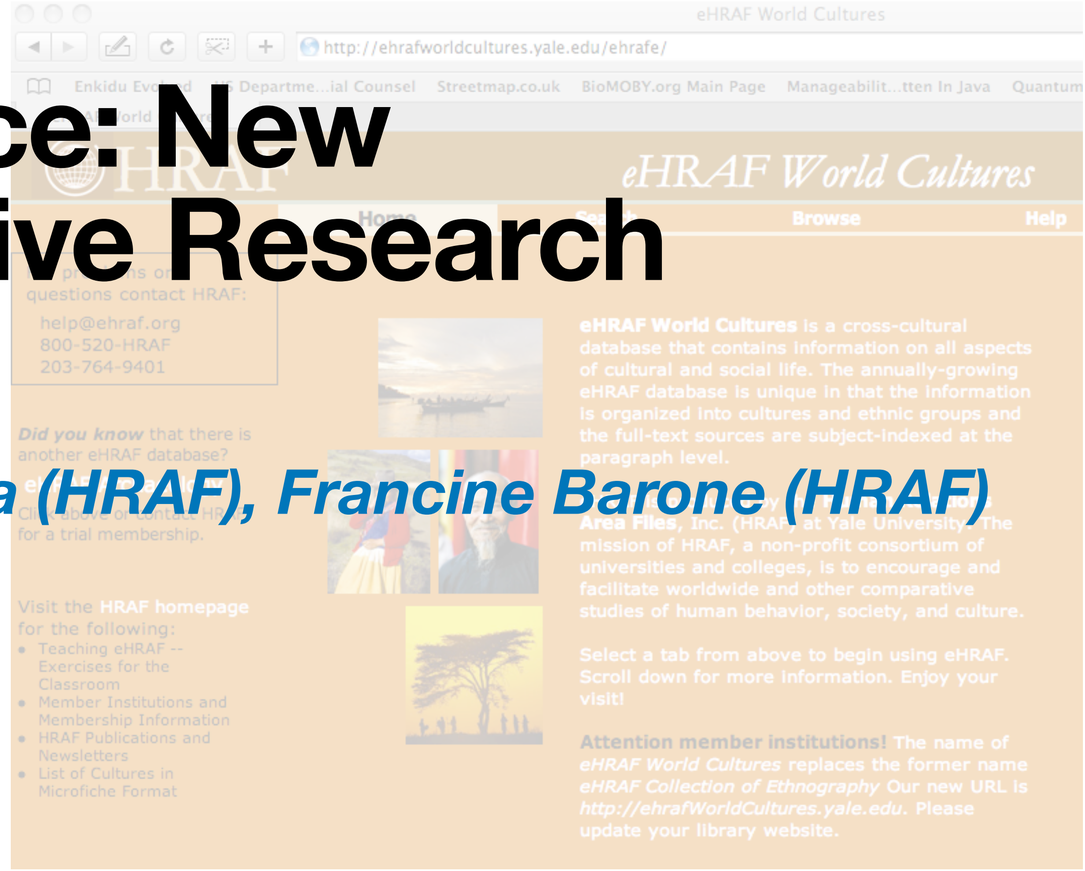
Ethnographic Data Science: New Approaches to Comparative Research



Michael Fischer (Kent/HRAF), Sridhar Ravula (HRAF), Francine Barone (HRAF)
Human Relations Area Files, Yale University

June 10th 2022 - RAI: Anthropology, AI and the Future of Human Society

Michael Fischer: mike@hrafarc.org
Sridhar Ravula: sridhar.ravula@yale.edu
Francine Barone: francine.barone@yale.edu

HRAF: https://hraf.yale.edu

HRAF

iKLEWS

(Infrastructure for Knowledge Linkages from Ethnography of World Societies)

- iKLEWS is a Human Relations Area Files (HRAF) project underwritten by the National Science Foundation *Human Networks and Data Science Infrastructure* programme.
- iKLEWS is using data science to create semantic infrastructure and ethnographic research services for a growing ethnographic database (eHRAF World Cultures),
  - roughly 800,000 pages from
  - 7,000 ethnographic documents covering
  - 361 world societies, each at several time points in the ethnographic present.

HRAF

iKLEWS

(Infrastructure for Knowledge Linkages from Ethnography of World Societies)

- Improve interoperability with external databases
- Develop tools to work with our databases for researchers casual to expert:
  - tools to broaden and narrow search with greater insight into meaning
  - tools to summarise, visualise and navigate the contents of large search results amounting to hundreds or thousands of pages.
  - tools to extract structured data from ethnographic text

HRAF

iKLEWS

(Infrastructure for Knowledge Linkages from Ethnography of World Societies)

- We aim to support researchers who seek to understand the range of possibilities for human understanding, knowledge, belief and behaviour:
  - to address work in anthropological theory,
  - to explore the relationship between human evolution and human behaviour,
  - to inform real-world problems we face today, such as: climate change; violence; disasters; epidemics; hunger; and war.

HRAF

Human Relations Area Files

- Founded 1949
- Mission: to encourage and facilitate the cross-cultural study of human culture, society, and behavior in the past and present.
- Curates knowledge of day to day life of peoples of different cultures recorded in ethnographic writing.
- Initially using paper – now digital.
- Key metadata -
- Ethnonyms – Outline of World Cultures - OWC
- Descriptors – Outline of Cultural Materials - OCM

Human Relations Area Files

- Since inception the HRAF collection of ethnography has included topical metadata for each entry in each document.
- These entries roughly correspond to paragraphs, but may include images, figures, lists, tables, etc.
- We refer to entries as **S**earch and **R**etrieval **E**lements, or **SREs**.
- Each SRE in each ethnographic work is assigned classificatory terms by an anthropologist, one or more of 790 drawn from **O**utline of **C**ultural **M**aterials (**OCM**)
- OCMs are organised as major and minor topics.

Human Relations Area Files: Outline of World Cultures

OWC	EHRAF WORLD CULTURES NAME	REGION	SUBREGION	SUBSISTENCE TYPE	PSF	SRS	SCCS
SI04	Abipón	South America	Southern South America	hunter-gatherers			Yes
RI03	Abkhazians	Asia	Caucasus	pastoralists			Yes
NK04	African Americans	North America	Regional and Ethnic Cultures	commercial economy			
AB06	Ainu	Asia	East Asia	hunter-gatherers			Yes

Human Relations Area Files: Outline of Cultural Materials

150 BEHAVIOR PROCESSES AND PERSONALITY

151 SENSATION AND PERCEPTION

152 DRIVES AND EMOTIONS

153 MODIFICATION OF BEHAVIOR

154 ADJUSTMENT PROCESSES

155 PERSONALITY DEVELOPMENT

156 SOCIAL PERSONALITY

157 PERSONALITY TRAITS

158 PERSONALITY DISORDERS

159 LIFE HISTORY MATERIALS

430 EXCHANGE AND TRANSFERS

431 GIFT GIVING

432 BUYING AND SELLING

433 PRODUCTION AND SUPPLY

434 INCOME AND DEMAND

435 PRICE AND VALUE

436 MEDIUM OF EXCHANGE

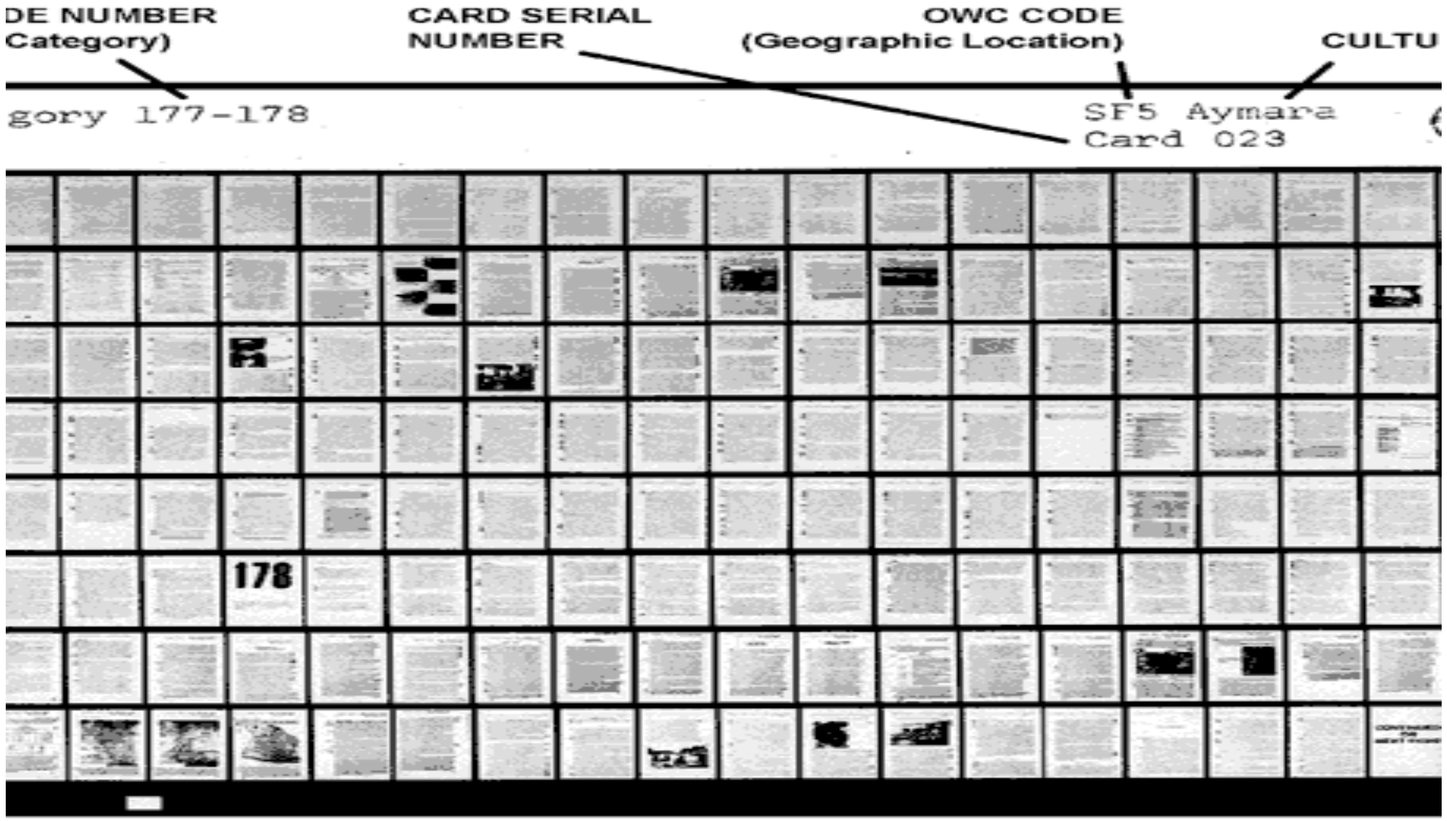
437 EXCHANGE TRANSACTIONS

438 INTERNAL TRADE

439 EXTERNAL TRADE

Technological Steps for HRAF

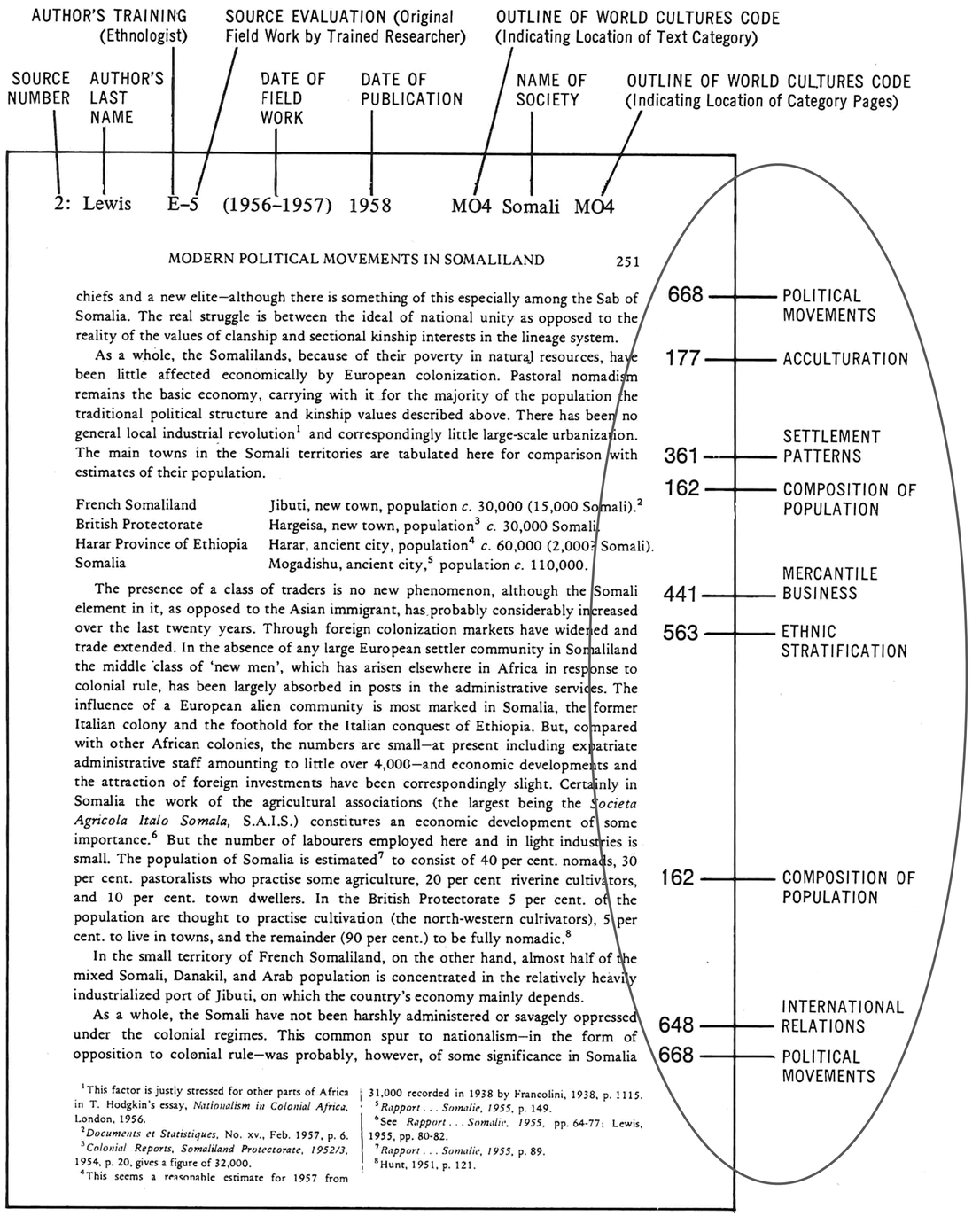
1949-1959. The file drawers organized by OWC contained full-texts and every subject category included all relevant pages from all documents



Microfiche cards-1959-1993

HRAF

Sample File Page, Annotated



Anthropologically-trained analysts subject-index to the paragraph-level with 3-4 digit numbers serving as a short-hand

Sample of HRAF Text - Expert Judgements

```
</p>
<p pageEid="or19-025-00714" xml:id="or19-025-00724" ocms="423
613" dispocms="423 613">
    <p.ocm>
        423 613
    </p.ocm>
    These gifts, in addition to maintaining a balance between
    population and resources, enhance the potential of separate
    groups of children to split apart and form separate
    lineages. Lineages such as these may continue to reside in
    the same district or village and to maintain friendly and
    cooperative relations with each other. Goodenough (1950)
    refers to such a collection of lineages as a ramage. When
    Trukese females migrate to other villages or islands they
    may found separate but related lineages. Members of lineages
    so related may have the option of membership in either
    lineage. In such cases Goodenough (1950) refers to the
    collective entity as a sub-sib.
</p>
<p pageEid="or19-025-00714" xml:id="or19-025-00725" ocms="613
614 192" dispocms="613 614 192">
    <p.ocm>
        613 614 192
    </p.ocm>
    Finally, all the lineages on different islands which bear
    the same name consider themselves to be somehow related,
    though completely unable to trace the alleged relationship.
    Generally speaking, lineage members so related have tended
    to avoid marriage with each other, but to extend a degree of
    hospitality when visiting one another's home islands.
    Goodenough (1950) refers to each of these large groups of
    lineages as sibs. Traditionally, however, Trukese have not
    distinguished these several levels of lineage organization
    by the use of distinct lables. Although they have borrowed
    the term family (
    <highlight xml:id="or19-025-00726" rend="underline">
        faamenii
    </highlight>
```

Initial metadata for iKLEWS

ocms:: #304 #567; type::p; pageEid::fa08-002-010227; prevPage::fa08-002-010167; nextPage::fa08-002-010278; sreid::fa08-002-010241; sreprev::fa08-002-010219-0; srenext::fa08-002-010250; parent::fa08-002-009963; section::fa08-002-009963; sectpar::fa08-002-009156; sectgpar::fa08-002-008623; division::fa08-002-000203; culture::Bambara; coverage::1500-1923; place::Mali; page::209; roll::; hdoc::fa08-002; title::The Bambara of Ségou and Kaarta: an historical, ethnographical and literary study of a people of the French Sudan; byline::Charles Monteil; pub.date::1924; pub.lang::English translation from French; field.date::1902-1923; pub.type::Monograph; owcs::fa08; mainowc::fa08; samples::SCCS;

{{304}} Among the Kouloubali the __wolo-so__ have a characteristic tattoo formed by three broad incisions from temple to lower jaw on both sides of the face. Far from feeling humiliated by this mark, every __wolo-so__ is as proud of it as though it were a sign of noble origin. Many free men have adopted this tattoo in order to pose as members of the great Bambara family.

Human Relations Area Files: iKLEWS

- eHRAF is very fast at retrieving relevant ethnography, but fundamentally uses same methods as HRAF's paper files in 1949.
- There are no aids to analysing the material once found; researchers read the results of search and apply own methods.
- iKLEWS introduces advanced methods of working with text through a framework to deploy analytic tools and improve search.
- Tools will support researchers from beginner to advanced, through web apps or Jupyter notebooks, supplied by HRAF or constructed by the researcher.

HRAF

Human Relations Area Files: iKLEWS

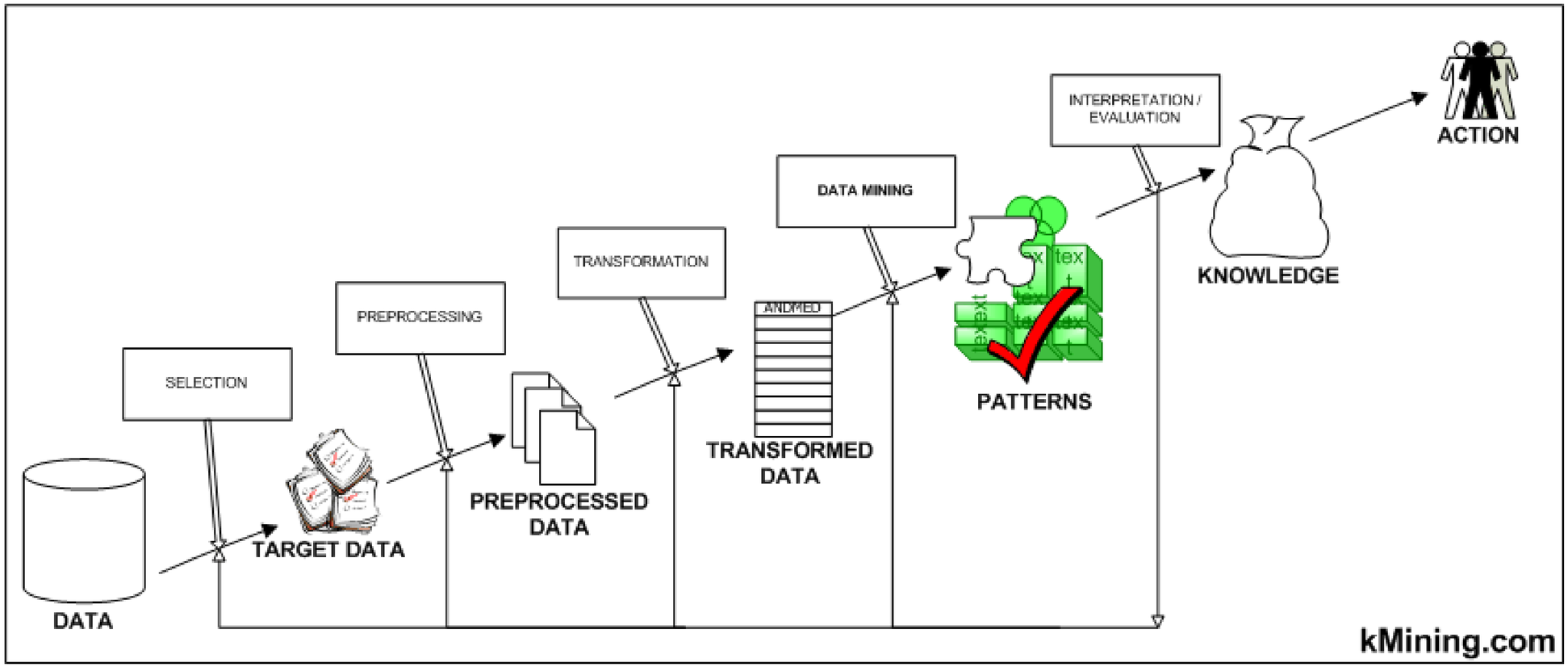
- Our infrastructure supports investigating a wide range of topics: social emotion & empathy, economics, politics, use of space & time, morality, or music & songs, examples using prototypic tools preceding this project.
- We are applying pattern extraction and linguistic analysis through deep learning, NLP and other ML tools supporting a logical framework.
- Some of the methods used can be applied to build a bridge between rather opaque (or ‘dark’) deep learning outcomes and more transparent logic driven narratives, and thus easier to generalise results.
- We will apply these results as new metadata and infrastructure so that researchers can operate in real time and we can scale up using less processor intensive algorithms than most ML and NLP methods require.

Human Relations Area Files: iKLEWS

- These outcomes will result in improved relevance of search results though identifying new and finer grained topics in each SRE in addition to those associated with the OCM;
  - establishing semantic representations of SREs in the texts with links between SREs so that researchers can follow topic trails effectively;
  - and provide tools for management, analysis, visualisation, and summarisation of results, researcher-initiated data mining and pattern identification, based largely on precomputed data.
- These will assist researchers to identify and test hypotheses about the societies they investigate.
- Researchers can access data and analytic capabilities directly through a Jupyter notebook run on the researcher's computer, or using a web application such as Kaggle or Google's Collaboratory.

DETECTING ORDER

- Data mining - knowledge from information
  - Collecting Data
  - Transforming Data
  - Lots of working approaches for identifying patterns in data



Identifying significance

- ML/NLP/NN evaluate the importance of a word is to a document in a collection or corpus

- Importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus

Goldilocks words - not too common or too rare

Word2Vec: method based on neural networks

Culture

Word	Similarity
cultural_tradition	0.86
cultural_pattern	0.84
cultural	0.82
traditional_culture	0.82
other_culture	0.82
native_culture	0.81
world_view	0.81
subculture	0.80
cultural_system	0.79
western_culture	0.78

Intelligence

Word	Similarity
intellect	0.77
capability	0.75
shrewdness	0.74
aptitude	0.72
talent	0.72
intelligent	0.72
cleverness	0.72
superior_intelligence	0.71
intellectual_ability	0.71
wisdom	0.70

Myth

Word	Similarity
mythology	0.89
creation_myth	0.85
origin_myth	0.84
legend	0.84
tale	0.83
mythological	0.82
origin_myths	0.80
mythical	0.79
mythic	0.77
these_storie	0.77

Example: Word2Vec

Deference

Word	Similarity
respect	0.86
respect_due	0.76
politeness	0.76
social_superior	0.76
deference_toward	0.74
obedience	0.73
deferential_behavior	0.73
respectful	0.73
extreme_deference	0.71
submissiveness	0.71

Compute

Word	Similarity
calculate	0.87
computation	0.78
estimate	0.77
calculation	0.75
tabulate	0.74
an_approximate	0.72
rough_estimate	0.71
by_multiplye	0.68
approximate	0.68
an_estimate	0.67

Computer

Word	Similarity
electronic	0.67
data_processe	0.65
computerized	0.65
typewriter	0.64
programming	0.63
software	0.63
programmer	0.62
dbase	0.61
xerox	0.61
data_processing	0.61

Word2Vec: Magic and Power: Good vs Bad

Word	Similarity
magical	0.72
charm	0.69
magical_power	0.69
object	0.65
magic_power	0.64
magical_charm	0.64
supernatural_power	0.64
medicine	0.64
purpose	0.64
requisite	0.63

Word	Similarity
black_magic	0.80
sorcery	0.75
evil	0.74
witchcraft	0.72
sorcerer	0.70
witch	0.70
evil_spirit	0.70
spell	0.70
malignant	0.69
magical	0.69

Word	Similarity
ability	0.75
prestige	0.74
wealth	0.72
capacity	0.72
political_power	0.71
control_over	0.70
not_only	0.69
spiritual_power	0.69
benefit	0.69
advantage	0.69

Word	Similarity
powerless	0.71
evil	0.69
fear	0.68
danger	0.64
evil_power	0.64
cause	0.64
threat	0.62
machination	0.62
malevolence	0.62
supernatural_power	0.61

Example: Word2Vec - Love

Love

Word	Similarity
affection	0.82
he_love	0.78
she_love	0.77
compassion	0.75
pity	0.75
hate	0.72
lover	0.71
passion	0.71
pleasure	0.70
happy	0.69

Positive love

Word	Similarity
affection	0.75
pleasure	0.72
devotion	0.69
generosity	0.67
desire	0.67
enjoyment	0.66
companionship	0.66
satisfaction	0.65
faithfulness	0.64
tender_affection	0.64

Negative love

Word	Similarity
hate	0.78
pity	0.70
unhappy	0.69
anger	0.67
grieve	0.66
jealous	0.66
shame	0.66
angry	0.65
wicked	0.65
torment	0.65

shy love

Word	Similarity
affection	0.74
like	0.74
admire	0.72
pleasure	0.71
affectionate	0.71
happy	0.69
he_love	0.68
fond	0.68
polite	0.68
good_natured	0.67

Topics- Geography (OCM 130)

using LDA (latent Dirichlet allocation)
and NER (Named Entity Recognition)

Weather variables captured in topics

SRE	Text	Topic_ID	TopicWeight	TopicWords
rv03-001-001138	february – 31.0 – 12.9 – 46.7 –38.2 – 24.7 – 51.4 – 40.2 –12.7 – 53.5 – 40.7 – 23.7 – 51.8 – 38.9 – 27.1 – 49.5	2	0.590772	may rain water season dry rainfall_cardinal quantity loc date
rv03-001-001138	february – 31.0 – 12.9 – 46.7 –38.2 – 24.7 – 51.4 – 40.2 –12.7 – 53.5 – 40.7 – 23.7 – 51.8 – 38.9 – 27.1 – 49.5	7	0.024604	climate average mean temperature rainfall_quantity person date cardinal org

Composite text broken into multiple topics

Text	Topic_ID	TopicWeight	TopicWords
situated in the middle of the eurasian continent the republic of kazakstan is the ninth largest country in the world. geographically kazakstan is distinguished by vast steppe areas some deserts and vast mountain ranges in the south and southeast. it has a continental climate the precipitation is low and strong winds are characteristic for most parts of the republic a factor which provides the impetus for attempts to develop wind power on a large scale setting in the open landscape. northern kazakstan with its black soils has a relatively good supply of water with large lakes and river systems. the southern part of the country on the contrary has a shortage of water. rivers such as syr darya ural chu irtysh sarysu and ili provide water for the valleys. kazakstan borders on the northern and northeastern shores of the caspian sea as well as the northern aral sea. other well-known lakes include zaysan and balkhash. the latter and the aral sea today present major environmental problems and could be said to amount to ecological catastrophes.	2	0.369462	may rain water season dry rainfall_cardinal quantity loc date
	9	0.284411	side though ice large drift water net place_cardinal time
	19	0.333716	areas rivers mountains climate_quantity person date loc cardinal org

Topics: Marriage- Nuptials (OCM 585)

Text	Topic_ID	TopicWeight	TopicWords
in the evening gavara men place the idol of gairamma into the palanquin which has been gaily decorated with coloured paper. the gairamma pot with its live rice plants is carried on a seating board pīta on the head of a gavara woman who is desirous of having children. the procession begins led by stick and kōya dancers with the hired band then the drummers—jangam village barbers the mala village servant and village madigas—and finally the gairamma palanquin carried by four washermen. 2. note the parallel with gavara weddings in which there are palanquins also carried by four washermen. alongside the palanquin goes the gairamma pot. a pressure lamp is carried by a barber and a fire torch by a washerman. it is not a required feature of the procession but often jangams ring a bell and blow on a conch shell acts associated with worship of gairamma's husband siva. 3. it will be recalled that a bell and a conch are also used by jangams auspiciously to dispatch a deceased person to siva's heavenly abode after he or she has been cremated.	10	0.378753	groom bridegroom party women guests wedding bride_time date cardinal
	40	0.608825	take food wedding bride see face may also_ordinal cardinal

Marriage- Termination (OCM 586)

Text	Topic_ID	TopicWeight	TopicWords
such cases were known in shin bagh. there is no importance given to the fact that the girl might dislike her husband. her main fear is that he might not like her and shame her and her lineage by marrying again.	15	0.151113	girl married wife said husband father would_date cardinal person
	25	0.095708	wife divorce marriage widow husband pay woman bride_ordinal cardinal
	42	0.705381	rate family wife cases divorce husband adultery_law percent cardinal

Human Relations Area Files: iKLEWS

- We will expand capacity to advance secondary comparative, cross-cultural, and other ethnographic research and extend this capacity to a much wider constituency of researchers by exposing
  - metadata,
  - computer assisted text analysis methods and
  - data management tools,
  - with guided means to leverage these through interactive web applications and JupyterLab templates together with interactive exemplars and training materials
- These services will make practical the inclusion of ethnographic and cross-cultural analysis in other kinds of research, scientific or applied.

- In the context of this session, *Programming anthropology: coding and culture in the age of AI*,
  - while we have not attempted an ethnography of the methods we have used, or the processes of their application,
  - we have been very mindful of how our choice of methods might influence or bias future research carried out on the infrastructure we are developing.
- the impact of these massive additions to eHRAF is significant, particularly the impact on researchers, how and what they can research, and bias we could be introducing to research outcomes.

Human Relations Area Files: iKLEWS

- One obvious problem with data science and NLP methods is that these are:
- either based directly on statistical methods, and each thus is associated with a probability of error, often 10% or more,
- or are based on non-deterministic ‘deep learning’ through adaptations of ML, including neural networks, which are more or less opaque with respect to their internal operation. Only the outcomes are explicable, and these imbue substantial margins of error in application, 10-20%, which is regarded as ‘good-enough’ by most of the ML community.

HRAF

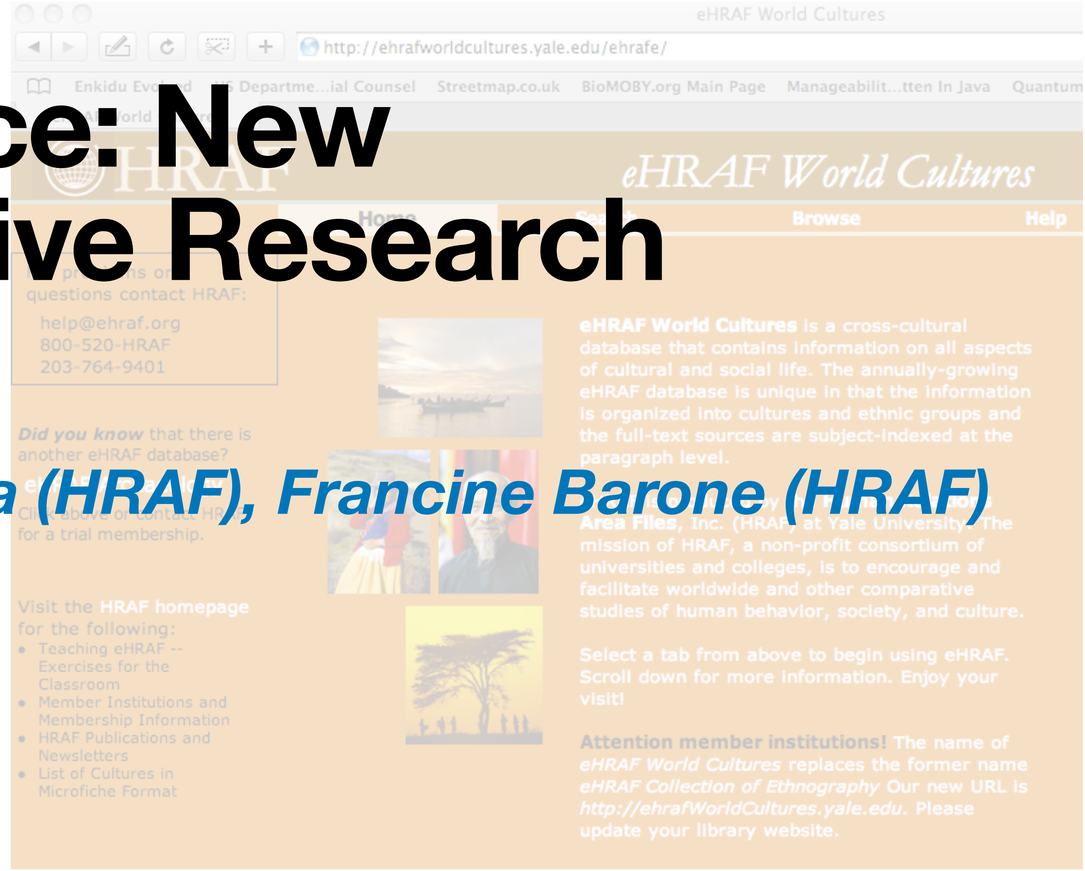
Human Relations Area Files: iKLEWS

- Our ignorance is, in principle, specific to demonstrating the detail of evolving relationships
- Thus the inability to fully understand how a history of interactions leads from the original context to the outcomes of algorithmic application.
- But the algorithm is available, the constituents known, and knowledge of the underlying logic and specified constituent properties is available.
- Armed with the latter elements, one gains an understanding of the process, if not an absolute understanding.
- Less than ideal, but not dissimilar to the limits of ethnological precision of an expert ethnographer. The limits of analytic processes, and the error rates associated, can be documented and applied to use.

Human Relations Area Files: iKLEWS

- From a developer's, and an anthropologist's, point of reference, it is important that we provide constant reference to these limitations, given the propensity of researchers to accept the outcomes of such procedures as definitive, or are over-suspicious.
- But just as imperative is the need to provide similar guidance for the content of the ethnographies themselves.
- Over time ethnological conceptions and perceptions have changed, been added or eliminated. Focal interests and themes change. Objectives change. Biases change.
- Early sources were often produced by missionaries, administrators, or tourists. Even then, eHRAF's coverage of a society might range from the 16th century to the 21st century, covering a range of different perspectives from many different roles.
- This guidance does not take the form of specific guidance on specific documents, but rather of tools and procedures to assist in evaluating content against other ethnographic sources, and identifying critical assumptions in the text.

Ethnographic Data Science: New Approaches to Comparative Research



Michael Fischer (Kent/HRAF), Sridhar Ravula (HRAF), Francine Barone (HRAF)
Human Relations Area Files, Yale University

June 10th 2022 - RAI: Anthropology, AI and the Future of Human Society

Michael Fischer: mike@hrafarc.org
Sridhar Ravula: sridhar.ravula@yale.edu
Francine Barone: francine.barone@yale.edu

HRAF: https://hraf.yale.edu

HRAF