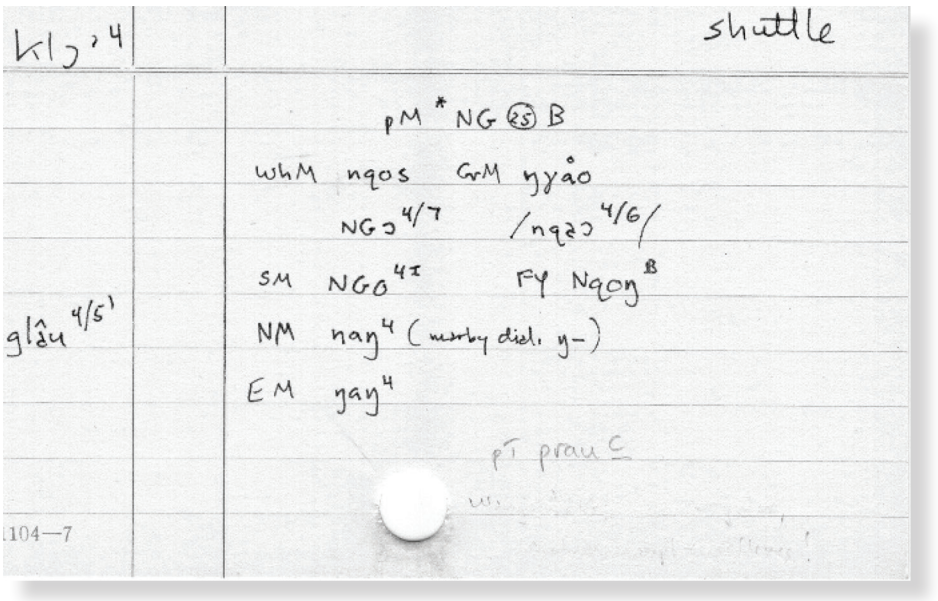
Archiving descriptive language data



JUDITH KAPLAN EXPLORES THE POSSIBILITY OF A NEW GOLD STANDARD FOR ARCHIVING THE WORLD'S ENDANGERED LANGUAGE DATA.

TWO HANDSOME MAHOGANY BOXES, LONG AND NAR- row, sit on a high shelf in a home basement that is prone to flooding in the American Midwest. Release one of their tiny silver latches, and inside you will find a collection of vocabulary cards annotated in an impeccable hand. These cards constitute the sum total of known field data on Biao Min, a language of the Hmong-Mien family spoken by some 21,000 people in southern China. Compiled during a four-month research trip to Guangxi Province in 1982, they were filed away in a closet and forgotten for more than 10 years. Only in 2001 did these notecards come to the attention of the broader research community. Since that time, linguists and data curators have used them alongside other similarly vulnerable materials in a demonstration project called E-MELD that is designed to create both a repository and an infrastructure for the management of past, present, and future language data.

As this contingent and layered history suggests, problems of data management in linguistics are not new. They extend back at least as far as the early twentieth century, when a well-known fieldwork imperative took hold in American anthropology. Pushing beyond what had until then been a narrow focus on Indo-European language and culture, Franz Boas, Edward Sapir, and their students set out into the field—first with phonographs, then with tape recorders—to capture, transcribe, analyze, and ideally revitalize a host of Indigenous American languages (Darnell 2001). This work was dedicated to future generations of researchers and speakers alike. The resulting collections were preserved in text and audio formats by institutions like the American Philosophical Society and Indiana University's Archive of the Languages of the World. It was a race to fix the characteristics of thousands of languages before they changed beyond recognition or disappeared entirely (often at the hands of government acculturation programs). Speed and efficiency in the field were prioritized over any kind of long-term or systematic archival strategy (Swadesh 1954). Such work became a cornerstone of graduate education, a focal point of government programming, and a rallying cause for some speech communities. It gave rise to numerous rich, though unruly, collections of language data. Many of these, the Boas Collection not least of all, are undergoing rapid digitization today.

World War II put the brakes on this early boom in descriptive linguistics: prominent researchers left fieldwork in progress to join the war effort. At the 1944 meeting of the Linguistic Society of America, 80 out of a total 96 members in attendance reported that they were involved in "military crucial work" (Martin-Nielsen 2010). Defense funding flowed into the discipline, which rapidly gained institutional prominence and moved away from its roots in anthropology. This coincided with a shift in theoretical emphasis, from historical particularities to linguistic universals—a move that continued through the postwar period (Harris 1993). But concern with the loss of linguistic diversity—not unlike contemporaneous trends in the biological realm—by no means disappeared. By the 1990s, linguists were visibly in the field again, raising the profile of documentary and data-driven research within the discipline as a whole.

Marking this development, in 1992 the Linguistic Society of America (LSA) formed a Committee on Endangered Languages and their Preservation, which issued its policy statement on "The Need for the Documentation of Linguistic Diversity" shortly thereafter. This statement reflected the spirit of the day, justifying its recommendations via the benefits inductively to be won for "the study of universal grammar and linguistic typology." Expressing a level of disciplinary self-confidence that would have been unthinkable in Boas' day, the Committee intervened "for the sake of the future of linguistics, with the intent of enriching and preserving" the field. Specifically, it called upon academic departments to "support the documentation and analysis of the full diversity of the languages which survive in the world today," giving highest priority to those facing extinction and/or featuring highly divergent characteristics. Significantly, Committee members further recommended that data be "systematically preserved in a network of repositories which also regulate the availability of this documentation."[1] Such work was incentivized through the conferral of graduate degrees, hiring, promotion, and tenure priorities. It was reinforced over time by a number of organizations including the Endangered Language Fund, UNESCO, the Foundation for Endangered Languages, the Indigenous Language Institute, Terralingua, the Resource Network for Linguistic Diversity, the DOBES Archive, the Rosetta Project, and the Hans Rausing Endangered Languages Project.

Almost a quarter-century after the LSA first

1 http://www.linguisticsociety.org/sites/default/files/lsa-stmt-documentation-linguistic-diversity.pdf

published their recommendations, linguists and data curators are trying to wrangle the collections born of the last 100 years of "salvage" linguistics into some kind of order. The goal of projects like E-MELD and the Open Language Archives Community is just that: "to aid in the development of infrastructure for linguistic archives" (E-MELD 2000). For the architects of E-MELD, the mission is to address two serious problems facing documentary linguistics today: the rapid loss of linguistic diversity (current opinion estimates that roughly half of the languages spoken in the world today will disappear by the end of the century) and the rapid proliferation of independent digitization initiatives.[2] Governing here boils down to the cultivation of "best practices" for intermedial translation, and the development of metadata linking heterogeneous resources and concepts to one another. Furthermore, reinforcing a logic of collective and distributed effort in the digital preservation of language data, it extends to the relations among researchers who are expected to share resources and custodial responsibility. Such interoperability ideally holds out the promise for direct communication—across individual languages, technological platforms, and research traditions—without leveling linguistic diversity.

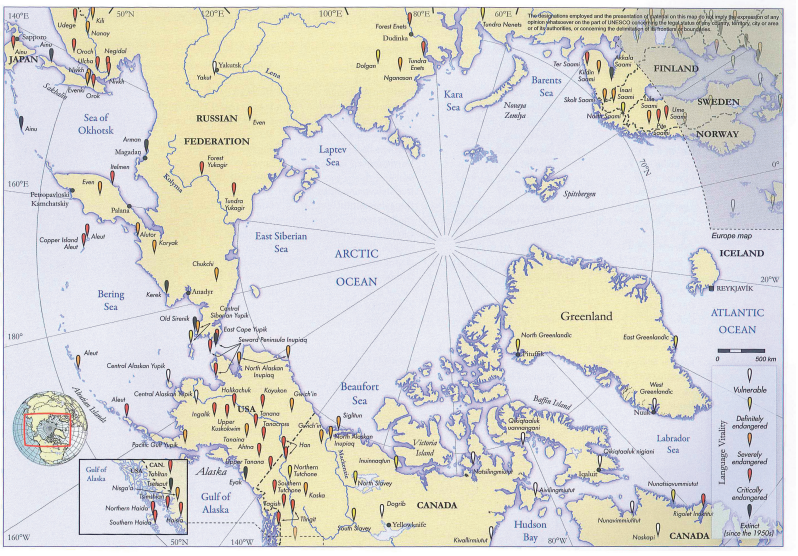
While E-MELD primarily addresses the needs of stakeholders in endangered languages research, in practice the project is as much about the protection of endangered archival materials: those two mahogany boxes. There is the sense that these can be revitalized through digitization. Ten case studies are featured in the project's "school of best practices," which is available to researchers around the world through LINGUIST List, the discipline's central online forum. Here, the project explores the nuts and bolts of moving between various media and a universally accessible web archive; the challenge being to move literally "From Notecards to the Web," "Shoebox to the Web," "Filemaker Data to the Web," prior integrative efforts like "TASX to the Web," and audio recordings on "Cassette to the Web." In the case of Biao Min, the task was to standardize digitization of the notecards—which maintain a window onto the comparative history of the Hmong-Mien family, an impression of the cultural life of Biao Min speakers, and vital characteristics of the language itself—in such a way that the resulting images would hold up for long-term preservation. Project members made choices about archival image format, user interface, data entry, and storage on the basis of this model collection that were meant to be generalizable. Moreover, they also applied themselves to the creation of resource metadata that would make the language intelligible within the framework of a hoped-for total linguistic archive.

The problems E-MELD faces derive from the logic of distributed responsibility—the need to host online archives, which can be quite sizable, at various sites—and the notion that data only have value when they can be found. What common infrastructure will allow linguists to identify relevant resources for a given study across archives? For example, what if a language of interest goes by different classifications or names in different collections (e.g., Lappish vs. Sami)? What if different structural tags are used by linguists in different traditions (e.g., possessive vs. genitive)? What if different systems of presentation are used (e.g., chronological vs. frequency-based vs. alphabetical)? And what if the resources themselves are submitted in formats that are wholly incommensurable (e.g., incompatible software tools; textual vs. recorded vs. video samples)? These are the kinds of questions motivating the search for a new total governing infrastructure.

Metadata, in this case, can be of two types: those that pertain to language *resources*, or to the languages *themselves*. The latter category is of rich conceptual interest because it blends top-down (theoretical) and bottom-up (descriptive) commitments about the characteristics of natural human language. Rather than hierarchically imposing a

2 http://www.unesco.org/new/en/culture/themes/endangered-languages/

BELOW: The UNESCO Atlas of Endangered Languages — Arctic Circumpolar.



Atlas of the World's Languages in Danger

Ahtna (USA)
Ainu (3) (JPN; RUS)
Aivilingmiutut (CAN)
Akkala Saami (RUS)
Aleut (3) (RUS; USA)
Alutor (RUS)
Arman (RUS)
Baraba Tatar (RUS)
Carrier (CAN)
Central Alaskan Yupik (2) (USA)
Central Siberian Yupik (2) (RUS; USA)
Chukchi (RUS)
Chulym Turk (RUS)
Copper Island Aleut (RUS)
Dogrib (CAN)
Dolgan (RUS)
East Cape Yupik (RUS)
East Greenlandic (GRL)
Even (2) (RUS)
Evenki (RUS)
Eyak (USA)
Forest Enets (RUS)
Forest Yukagir (RUS)
Gitksan (CAN)
Gwich'in (2) (CAN; USA)
Haisla (CAN)
Han (2) (CAN; USA)
Holikachuk (USA)
Inari Saami (FIN)
Ingalik (USA)
Inuinnaqtun (CAN)
Itelmen (RUS)
Kaska (CAN)
Kerek (RUS)
Kildin Saami (RUS)
Kili (RUS)
Kivallirmiutut (CAN)
Koryak (RUS)
Koyukon (USA)
Lule Saami (NOR; SWE)
Michif (CAN)
Nanay (CHN; RUS)
Naskapi (CAN)
Natsilingmiutut (CAN)
Negidal (RUS)
Nganasan (RUS)
Nisga'a (CAN)
Nivkh (2) (RUS)
North Alaskan Inupiaq (3) (CAN; USA)
North Greenlandic (GRL)
North Saami (FIN; NOR; RUS; SWE)
North Slavey (CAN)
Northern Haida (CAN; USA)
Northern Selkup (RUS)
Northern Tutchone (CAN)
Nunatsiavummiutut (CAN)
Nunavimmiutitut (CAN)
Old Sirenik (RUS)
Oroch (RUS)
Orok (RUS)
Pacific Gulf Yupik (USA)
Pite Saami (NOR; SWE)
Qikiqtaaluk nigiani (CAN)
Qikiqtaaluk uannangani (CAN)
Rigolet Inuktitut (CAN)
Seward Peninsula Inupiaq (4) (RUS; USA)
Siglitun (CAN)
Skolt Saami (FIN; NOR; RUS)
South Slavey (CAN)
Southern Haida (CAN)
Southern Tutchone (CAN)
Tagish (CAN)
Tahltan (CAN)
Tanacross (USA)
Tanaina (USA)
Tanana (USA)
Ter Saami (RUS)
Tlingit (2) (CAN; USA)
Tsetsaut (CAN)
Tsimshian (CAN)
Tundra Enets (RUS)
Tundra Nenets (RUS)
Tundra Yukagir (RUS)
Udege (RUS)
Ulcha (RUS)
Ume Saami (SWE)
Upper Kuskokwim (USA)
Upper Tanana (CAN; USA)
West Greenlandic (GRL)
Yakut (RUS)

new annotation system on top of the many already in existence—these being more or less suited to the particularities of the individual languages described—linguists and data curators have developed an "ontology" capable of linking extant, and possibly future, strategies for language analysis. This flexible and decentralized governing strategy has facilitated the recognition of new collective kinds: new groups of languages can now be compared, and therefore defined; new communities of researchers can interact and share their data; and new assemblages of archival objects can be brought together under the big tent of information.

Central to all this is the 2010 General Ontology of Linguistic Description (GOLD), which provides a formalized account of the most basic categories and relations used in linguistic description (GOLD 2010; Farrar and Langendoen 2003). With its roots in Scott Farrar's 2003 doctoral dissertation, GOLD allows linguists to search and compare *within* relevant resources (once these have been identified) using a standardized search vocabulary. To take an easy example, if a linguist wanted to look comprehensively within a corpus of glossed texts for examples of past-tense morphemes, he or she could invoke the GOLD term *PastTense* in a query, taming a babel of alternatives used in other linguistic markup schemes (e.g., *Past, PST, RemotePast, HodiernalPast*, and so on). A reduced ontology—one with ultimate compatibility with the Semantic Web—thus enables more languages, resources, and linguists to come together in a streamlined comparative framework.

The 100-year history I have just flown over reveals the emergence of a new *disciplinary* collective, one that is being defined—beyond Indo-European studies, Americanist anthropology, or the endangered languages community—by the web-based archiving of language data. Linguistics has been characterized as a field that depends on "second sourcing" its data: borrowing is widely accepted, as linguists understand that language learning and fieldwork are too labor intensive to be replicated continually from scratch (Lewis, Farrar, and Langendoen 2006). Thus, linguistics is cumulative, cooperative, and conservative with respect to data. E-MELD further illustrates how the desire to digitize and make web archives openly available has occasioned new methods of governance, ranging over increasingly general linguistic populations. But governing in this case has more to do with flexibility than control: taxonomy has been rejected in favor of ontology. With a radically simplified conceptual structure that articulates what are thought to be universal features of human language, recent adventures in linguistic data curation attempt to figure a new species-level population from the ground up. Whether or not these efforts will deliver a new gold standard remains to be seen. Historians and science studies scholars can ask in the meanwhile, what systems of value underpin contemporary efforts to archive endangered language data, and for whom do they apply?

JUDITH KAPLAN *is currently pursuing her interests in the history of the human and historical sciences (linguistics in particular) as a postdoctoral fellow at the Max Planck Institute for the History of Science, Berlin.*

BIBLIOGRAPHY

Darnell, Regna. 2001. *Invisible Genealogies: A History of Americanist Anthropology*. Lincoln, NE, and London: University of Nebraska Press.

E-MELD. 2000. Electronic Metastructure for Endangered Languages Data Grant Proposal. http://emeld.org/documents/E-MELD.html (accessed 3/10/15).

Farrar, Scott, and D. Terence Langendoen. 2003. "A Linguistic Ontology for the Semantic Web." *GLOT International* 7(3):97–100.

GOLD. 2010. *General Ontology for Linguistic Description (GOLD)*. Bloomington, IN: Department of Linguistics (The LINGUIST List), Indiana University. http://linguistics-ontology.org/

Harris, Randy Allen. 1993. *The Linguistics Wars*. New York: Oxford University Press.

Lewis, W., S. Farrar, and T. Langendoen. 2006. "Linguistics in the Internet Age: Tools and Fair Use." In *Proceedings of the EMELD '06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. http://emeld.org/workshop/2006/papers/lewis.pdf (accessed 4/12/15).

Martin-Nielsen, Janet. "'This War for Men's Minds': The Birth of a Human Science in Cold War America." *History of the Human Sciences* 25(5):131–155.

Swadesh, Morris. 1954. "Linguistic Time Depths of Prehistoric America: Penutian." Grant Report. Philadelphia, PA: American Philosophical Society.