Perspectives on the Internet Version of the Language and Culture Atlas of Ashkenazic Jewry:
Prepared for the Inauguration of Internet Access

Judith L. Klavans, Ph. D.
University of Maryland, formerly of Columbia University in the City of New York

A Decade of Consequences: Although the Archives of the Yiddish language have been collected for a period of nearly sixty years, my experience and exposure have only existed for just over a decade. However, I believe that I was fortunate to have come into contact at a critical juncture in the history of this valuable and important archival collection. I open my talk with a recollection of this moment that helped contribute to the acceptance by Columbia University of the archives for the subsequent preservation that was essential to its survival. The ultimate access, of course, is fully electronic,which I address in my final points. But first, I will set the context for this fortunate moment, and then present the consequences.

The Center for Research on Information Access (CRIA) at the Columbia University Libraries: In the fall of 1994, the central administration of Columbia University funded an innovative research center, to be housed in the distinguished Libaries of Columbia University. The goal of this new Center was to build interdisciplinary links between the humanities and digital technologies, such as digital collections and computer based access techniques. This then nascent field, which came at the time that the Internet reached the wider public, was called °Digital Libraries° and encompassed many kinds of interpretations. At the time, I was a researcher in computational linguistics in the Department of Computer Science at Columbia… for those of us in computer science, digital libraries meant databases, but I soon found out that for those in the libraries °Digital Libraries° meant digital collections. Given that I had an interdisciplinary background myself, the Provost°s office asked if I would lead this new center in October of 1994. With my Ph.D. in theoretical linguistics, followed by a postdoctoral fellowship in psycholinguistics and code.switching, and then by nearly a decade at the IBM TJ Watson Research center in computational natural language processing, the opportunity to take on the challenge of building the Center for Research on Information Access at Columbia University was an honor. And so I accepted the offer.

The offices of the new Center were located next to Ms. Carol Mandel, a distinguished and knowledgeable librarian with a background in technical services, who was the Deputy University Librarian to Dr. Elaine Sloan, the University Librarian. As a linguist, and as a computational linguist, I had never worked inside the administrative offices of a library. I was a researcher and a user of libraries, but I was ignorant of the full workings of how librarians perform the valuable services that make them the keys to storing, restoring, and providing access to information.

On Selecting Contributions to the Libraries: Within the first few weeks of starting my new position, I learned that one of the major challenges of the University Libraries was making careful selections about which archival material to accept into the collections. With such a distinguished and extensive faculty, these decisions are ongoing and require that the libarians exercise careful intellectual discrimination. For each page, or box, or file that is accepted, the libraries staff must decide the cost of preserving, cataloging, and storing. Thus, not all donations can be accepted. Even if accepted, decisions on providing the level of descriptive access (e.g. by the box, by the page) must be made judiciously and selectively. These decisions require the input of many librarians and scholars, lest the librarian judge the value of a particular collection accidentally incorrectly.

The LCAAJ Collection at the Columbia University Libraries: In November, 1994, Ms. Mandel requested that I join her for a meeting with Mr. Robert Neumann and Dr. Mikhl Herzog, who had come to her to discuss the archives. After the dissolution of the Department of Linguistics in 1989, it was unclear what the fate of the archives would be. They were widely recognized as an essential and unique resource in the community of Yiddish language and culture scholars, but the size and technical challenges of this archive were formidable. This meeting turned out to be critical to the archives. Ms. Mandel had consulted me as a linguist, and I was honored to be able to serve. Since my initial studies in linguistics had been in comparative philology at Harvard University, I was deeply familiar with the ground-breaking research of Professor Uriel Weinreich on languages in contact. Furthermore, my interest in the Yiddish language, quite outside any of my education, meant that I was fully aware of the creativity involved in the LCAAJ. The profound realization that at this moment, on November 24, 1994, a decision was made for the Libraries at the Columbia University in the City of New York to accept this collection continues to bring deep intellectual and personal satisfaction. However, satisfaction aside, the true hard work had just begun.

A Vision of Internet Access: As a computational linguist, accustomed to working with multilingual digital information, the natural first step was to consider application to the National Science Foundation for a project in the digital libraries program to provide electronic access to the data. However, the condition of the material required extensive funding for preservation. This funding was applied for by Dr. Janet Gertz, the director of the preservation division of the Libraries. Her description of this ten-year effort is given at this meeting, and her contribution as an expert in preservation (as well as her own background in linguistics) cannot be underestimated. Frequent visits from Dr. Robert Neuman and Dr. Ulrike Kiefer provided ongoing guidance and inspiration. Dr. Mikhl Herzog, the hub of this valuable project, continued to generate intellectual guidance (as well as Yiddish humor) to move us towards our goal. With the very hard work of many, we are now realizing the achievement of Internet access. This is the vision which many of us had years ago, although my joining in this vision came later than many of my colleagues.

Broad Impact Across Fields and Across Languages: Before I close, I want to address the issue of why Internet access is essential to the wide use of this unique collection. We are all cognizant of how the Internet has changed the way research is performed. However, the building of large text corpora, and particularly corpora with associated speech, is a relatively new aspect of internet data. Archives of spoken language data, such as news broadcasts, abound. However, there are relatively few carefully constructed spoken language archives available. There are even fewer which have been so carefully transcribed and annotated with explicit linkages between speech and written text. And there are far fewer, if any, culturally indexed material complete with associated linguistic data. At the Center for Advanced Study of Language at the University of Maryland, where I have just moved to build a technology use program, we will be undertaking the collection of dialects across several languages of research interest, such as Chinese. The principles of the LCAAJ will be invaluable to drive this research. The LCAAJ infrastructure is universal, and promises to create a new standard for the collection of linguistic data from other languages and cultures. The uses of the archives by anthropologists, linguists, politicians, psychologists, and analysts from many fields will impact the building of additional databases of spoken languages. Such an effort has international implications at many levels, some of which have yet to be envisioned. The existence of this data on the internet contributes to the possibility that these visions will be realized.