Submitted by: ODNI FLPO HLT Group
Presenter: Elaine Marsh
Topic: Return on Investment for Government Human Language Technology Systems

Over the years, the government has translated reams of material, transcribed decades of audio, and processed years of text. Where is that material now? How valuable would it be to have that material available to push research and applications and to support foreign language training? Over 20 years ago, DARPA funded the Linguistic Data Consortium (LDC) at the University of Pennsylvania to collect, catalog, store and provide access to language resources. Since that time, the LDC has collected thousands of corpora in many different genres and languages. Although the government has access to the full range of LDC data through a community license, until recently corpora specific to government needs were usually deleted soon after they were created. In order to address the need for a government-only catalog and repository, the Government Catalog of Language Resources was funded through the ODNI, and an initial prototype has been built. The GCLR will be transferred to a government executive agent who will be responsible for making improvements, adding corpora, and maintaining and sustaining the effort.

The purpose of this talk is to present the model behind GCLR, to demonstrate its purpose, and to invite attendees to contribute and use contents. Background leading up to the current version will be presented. Use cases of parallel corpora in teaching, technology development and language maintenance will also be covered. Learning from the LDC on how corpora are used, and linking with the LDC will be part of future directions to enable government applications to utilize these resources.