

、国政事情にかかわる文書の研究、翻訳、転写、報告、分析に非常に重  
はたします。文中での微妙な単語選択、使用により、全体的な意味合いが  
化するのを言語分析で扱うことができます。本イベントでは、諜報分析者、防  
言語学者、システムアーキテクト、または言語テクノロジーの専門家の方々  
多言語ソフトウェアの最新技術情報を紹介します。訳者及び分析者の方々に  
ode、多言語のシステム設計、デジタル科学捜査のチュートリアル、デスクトップ  
などの提供いたします。アラビア固有名称一致、中国語形態素解析、及び  
語オンラインチャットなど多様な言語分析の最新トピックをお届けします。情報  
文書認識、メディア検索、デジタル科学捜査の現場からの言語分析テクノロジー  
況をぜひご自分の目でお確かめください。

إطلاع شامل في تكنولوجيا الاستخبار والتناظرات الرقمية - إطلاع شامل في تكنولوجيا  
ماتة وتطبيقه في مجال تحقيقات البيانات الرقمية - إتقان في علم اللغات والعلوم المعلوماتية وتطبيقه في مجال تحقيقات البيانات  
اطلاع شامل في تكنولوجيا الاستخبار والتناظرات الرقمية - إتقان في علم اللغات  
ماتة وتطبيقه في مجال تحقيقات البيانات الرقمية - إطلاع شامل في تكنولوجيا  
اطلاع شامل في علم اللغات والعلوم المعلوماتية وتطبيقه في مجال تحقيقات  
اطلاع شامل في تكنولوجيا الاستخبار والتناظرات الرقمية - إتقان في علم  
ماتة وتطبيقه في مجال تحقيقات البيانات الرقمية - إطلاع شامل في تكنولوجيا

[www.basistech.com/conference](http://www.basistech.com/conference)



**BASIS**  
TECHNOLOGY

# GOVERNMENT USERS CONFERENCE

June 14, 2006 →

The George Washington University,  
Marvin Center, Washington, D.C.



150 CambridgePark Drive  
Cambridge, MA 02140

T 617.386.2000

F 617.386.2020

Dear Colleague,

You are invited to join your fellow language and technology professionals at Basis Technology's Government Users Conference and Workshop. We have assembled a full day of talks, demonstrations, and tutorials showcasing the state-of-the-art in Human Language Technology (HLT). This is a unique opportunity to meet the experts, network with your peers, acquire new skills, and plug-in to the fast-growing Basis Technology user community.

## CONFERENCE HIGHLIGHTS

**Keynote Address** — Learn from Basis Technology's CTO how linguistic enhancements to National Harmony provide a new framework for large-scale multilingual DOCEX, and how these innovations can serve as a model for future systems throughout the USG.

**Specialized Tracks** — Conference content has been organized into two tracks, one with emphasis on language and the other with emphasis on technology. Take advantage of tutorials on Unicode, multilingual system design, digital forensics, and desktop tools for translators and analysts. Explore complex linguistic topics in depth, including Arabic name-matching, Chinese segmentation, and Persian online chat. Get a first-hand look at results from real-world, deployed HLT systems for information retrieval, document triage, media exploitation, and digital forensics.

**Networking Opportunities** — Meet fellow system architects, software engineers, computational linguists, and program managers who are building the next generation of HLT systems. Learn what approaches are working, and what's not.

**Training Sessions** — Gain "hands on" experience with Basis Technology products, and bring home tools that will make a difference in your job, right now.

**Feedback Sessions** — Your opportunity to "meet the developers", get a sneak peek at what's in store for later this year, and give your input on future product directions.

Whether you are an intelligence analyst, a military linguist, a system architect, or an HLT practitioner, this event will provide you with the information you need to understand and take advantage of the latest developments in multilingual software technology.

## REGISTRATION

For more information or to register, please visit [www.basistech.com/conference](http://www.basistech.com/conference) or send e-mail to [conference@basistech.com](mailto:conference@basistech.com) or call 617-386-2050.

This conference is organized for the benefit of and provided at no charge to employees of the U.S. Government. Individuals outside the U.S. Government may also attend, however a contractor participation fee of \$495 per person will apply. Contractor registrations received before May 20 will receive a \$100 discount.

**Spaces are limited. Register today!**

## CONFERENCE AGENDA

7:30

*Registration & Breakfast*

8:30

Welcome & Keynote Address  
TBA

9:00

Linguistic Enhancements to National Harmony  
Benson Margulies

9:45

*Break*

### 🕒 TECHNOLOGY TRACK

### 🕒 LANGUAGE TRACK

10:00

Tutorial — Understanding Unicode 5.0  
Ed Schwalenberg

Tutorial — Introduction to the Basis  
Transliteration Assistant  
Melissa Lucius & Bushra Zawaydeh

11:00

Tutorial — Designing Large-Scale  
Multilingual Systems  
Benson Margulies

Tutorial — Introduction to the Basis  
Arabic Editor and Workbench  
Mary Galvin & Bernard Greenberg

12:00

*Lunch Break*

12:45

What Language is That?  
Using the Rosette Language Identifier  
Nobuo Otsuka

Looking Behind the Name:  
Etymology of Arabic Names  
Zina Saadi

1:30

Putting the Rosette Global Name  
Matcher to Work  
David Murgatroyd & Benson Margulies

Romanized Persian in Online  
Communications  
Sabiha Imran

2:15

*Break*

2:30

Building Applications with the  
Rosette Linguistics Platform  
Steve Cohen & Ken Glidden

Reverse Transliteration of Arabic:  
Recovering the Original Text  
Eva Sanchez & Arnie Chien

3:15

A Crash Course in Digital Forensics  
Brian Carrier

Orthographic Variations in Arabic Corpora  
Bushra Zawaydeh & Zina Saadi

4:00

*Break*

4:15

What Every Software Developer  
Needs to Know About Arabic  
Bernard Greenberg

Chinese Language Analysis:  
Solving the Chinese Puzzle  
Joe Ho

5:00

*Closing Discussion & Feedback*

## PLENARY TALK

---

### Linguistic Enhancements to National Harmony

---

**Benson Margulies, CTO, Basis Technology**

The National Harmony database is a major government repository of intelligence documents from a wide variety of sources. Historically, National Harmony contained primarily scanned images, primarily in English. Recently, enhancements have been made to store electronic text in Unicode, and National Harmony has begun to accumulate text in many foreign languages, including Arabic, Chinese, Korean, and Persian.

To facilitate search and retrieval of this data, Basis Technology has integrated key software components into National Harmony, including multilingual full-text search, cross-language keyword search, and word-by-word lexicographic analysis. This talk will describe these components and their integration into the National Harmony architecture, with specific emphasis on examples and test cases in Arabic.

## TECHNOLOGY TRACK

---

### Tutorial — Understanding Unicode 5.0

---

**Ed Schwalenberg, Director of Engineering Services, Basis Technology**

Unicode makes it easy for computer systems to support virtually all the world's written languages. This tutorial will provide a gentle introduction to the basic concepts of the Unicode 5.0 standard, including characters, encodings, transcoding, byte ordering, and the common UTF-8 and UTF-16 transformation formats. Also covered will be practical information about support for Unicode in popular operating systems, computer languages, and protocols. Finally, we will review some of

the problems that Unicode doesn't solve, such as fonts, translation, and the rendering of complex scripts such as Arabic, Persian, and Thai.

This talk is recommended for software or language professionals at all levels of experience.

### Tutorial — Designing Large-Scale Multilingual Systems

---

**Benson Margulies, CTO, Basis Technology**

Applications built for national defense and intelligence missions face an ever-increasing volume of data in foreign language. Traditional approaches, including discarding non-Latin text or preserving it only as bitmaps, fail to produce actionable intelligence. These applications must gain the sophistication needed to accurately analyze and triage multilingual information.

Foreign language documents pose challenges for the entire document-management pipeline: identifying the format, extracting text, indexing, search, retrieval, and display. While commonly-used technologies work much better than they did a few years ago, there are still many ways to build systems that fail to handle foreign text. This presentation will provide an overview of the problem and point up some of the more important issues and traps.

This “firehose style” tutorial will draw upon over ten years of Basis Technology experience building large-scale multilingual systems for commercial and government customers. Bring your toughest problems, a thick notebook, and a passion for writing software which can competently handle foreign language.

This talk is recommended for software developers, software architects, and technical program managers.

---

## What Language is *That*? Using the Rosette Language Identifier

---

**Nobuo Otsuka, Software Engineer,  
Basis Technology**

Differentiating languages which share the same writing system is a difficult task, both for humans and for machines. For example, each of the distinct languages Arabic, Kurdish, Pashto, Persian, and Urdu share a common script. And to the untrained eye, it may be impossible to determine which of these languages appear in any given document. This problem also exists for other language families, including:

- Chinese, Japanese, and Korean
- Malay and Indonesian
- Danish, Swedish, Norwegian, and Icelandic
- Czech, Croatian, Slovak, and Slovenian

This talk will present an overview of the Rosette Language Identifier (RLI) and discuss the techniques RLI uses to automatically identify the language and encoding of a block of text. It will also explain how language and encoding identification is an essential stage in the process of working with unstructured multilingual text.

This talk is recommended for experienced software developers, software architects, and technical program managers.

---

## Putting the Rosette Global Name Matcher to Work

---

**David Murgatroyd, Software Engineer,  
Basis Technology**

Do your users encounter unfamiliar foreign names with widely varied spellings? Is searching, matching, and sorting foreign names a constant headache? If so, the Rosette Global Name Matcher (GNM) SDK is for you. This software toolkit enables any application to perform cross-script searching of a set of names, which may appear in a variety of written forms. They

may be in a foreign writing system, such as Arabic or Korean, or they may appear in transliterated form using the Latin alphabet.

GNM allows, for example, a database of Arabic names to be searched either in Latin script or in Arabic script, or a database of Korean names to be searched either in Latin script or in Hangul. The search result is a quantitatively ranked list of hits, giving its users the flexibility to handle ambiguous or inexact matches. This presentation will examine the capabilities of the GNM SDK and explore possible client architectures for applying it, such as relational databases, web services, or standalone applications written in C++ or Java.

This talk is recommended for experienced software developers, software architects, and technical program managers.

---

## Building Applications with the Rosette Linguistics Platform

---

**Steve Cohen, VP Products, Basis Technology  
& Ken Glidden, Director of Product  
Engineering, Basis Technology**

The Rosette Linguistics Platform (RLP) is a robust, high performance software library that adds linguistic capabilities to existing applications for text search, mining and analysis. It includes tools for low-level analysis—such as segmentation and morphology—as well as sophisticated entity extraction to find key information such as names, places, and dates. The first part of this talk will review the capabilities of RLP, the applications for which it can be used, and the techniques it employs. The second part will focus on how RLP can be integrated and used in existing systems, and how it can be tuned for each system's requirements. Topics include:

- The merits of statistical versus lexical approaches in named entity extraction
- How RLP combines different techniques for analysis

- How users can enhance precision and recall with RLP
- The RLP SDKs for C++ and Java
- Planned improvements for the upcoming RLP 5.5 release

This talk is recommended for experienced software developers, software architects, and technical program managers.

---

## A Crash Course in Digital Forensics

---

**Brian Carrier, Ph.D., Director of Digital Forensics, Basis Technology**

Digital Forensics is the science of analyzing computers and other digital devices to gain information regarding past events. Forensic analysis of digital media has been practiced by law enforcement for many years and has recently become common in corporate environments, intelligence gathering, and civil law suits. This talk will provide an overview of key topics in digital forensics, including the investigation process; analysis techniques and tools; and some examples. It will also provide information on new forensics products being developed at Basis Technology and how linguistic analysis techniques will be incorporated into these products.

This talk is recommended for software or language professionals at all levels of experience.

---

## Large-Scale Digital Forensics

---

**Brian Carrier, Ph.D., Director of Digital Forensics, Basis Technology**

Digital Forensics is the science of analyzing computers and other digital devices to gain information regarding past events. While the quality and number of analysis tools have increased over the years, so has the amount of data that must be analyzed. This presentation will address the large-scale data issues associated with digital forensics from the perspectives of

increasing individual disk sizes and increasing numbers of interrelated disks in multiple languages.

The presentation will outline the current and future digital forensics solutions that Basis Technology provides as well as how Basis Technology's linguistic analysis tools will be used for these digital forensic products. The Advanced File Format (AFF) is an open format for storing digital evidence so that it can be easily analyzed by multiple tools. The Media Exploitation Kit (MEK) is an automated acquisition program that can copy the contents of a hard disk with no user interaction. We will also describe our automated data reduction system that requires no user interaction and analyzes disk images to identify unknown and high priority files. This system will also be able to correlate data and artifacts between multiple computers.

This talk is recommended for software or language professionals with some prior forensic analysis experience.

---

## What Every Software Developer Needs to Know About Arabic

---

**Bernard Greenberg, Chief Technology Innovator, Basis Technology**

Software developers are increasingly faced with the challenge of adapting information systems to handle data written in or derived from Arabic. This talk will provide a concise introduction to the Arabic language and writing system for those with no experience in the subject matter at all. Basic issues of the script (the writing system); Modern Standard Arabic (MSA) orthography (how letters are used together); morphology (how words are composed); and grammar will be presented, as well as geographical and historical context (specifically, relation to other languages). The basics of the representation of Arabic in computers, as well as bidirectional text processing, will be outlined. The key problems of dealing with Arabic text in desktop and network

environments will be enumerated and discussed. Some of it will assume some basic experience with the representation of data in computers.

This talk is recommended for software or language professionals at all levels of experience.

---

## Multilingual Geographic Names: Mapping the World in Unicode

---

**Carl Hoffman, CEO, Basis Technology**

For hundreds of years, cartographers have struggled with the problem of rendering foreign place names into familiar writing systems, often with awkward or confusing results. Do we write Tokyo, Toukyou, or Tokio? Peking or Beijing? Mosul or Mawsil? Today, a new generation of Geographic Information Systems (GIS) is emerging, capable of representing place names in Unicode and easing the task of managing databases of authoritative Latin spellings. This presentation will review some of the computational linguistic issues associated with multilingual place names, and demonstrate tools recently developed by Basis Technology to support multilingual GIS, including phonetic search, cross-script name matching, and automatic transliteration.

## LANGUAGE TRACK

---

### Tutorial — Introduction to the Basis Transliteration Assistant

---

**Melissa Lucius, Program Manager, Basis Technology & Bushra Zawaydeh, Ph.D., Computational Linguist, Basis Technology**

Professional translators wage a daily struggle to render Arabic names into Romanized forms, because these names have so many possible (and conflicting) transliterations. Translators employed by the U.S. government are commonly instructed to transliterate names according to one or more formal standards, including the U.S.

Board on Geographic Names (BGN), the Standard Arabic Technical Transliteration System (SATIS), and the Intelligence Community (IC) standard. To simplify the task of working with each of these standards, Basis Technology has developed the Transliteration Assistant (XA), a Microsoft Office plug-in which enables translators to quickly and consistently produce accurate transliterations of Arabic names.

XA takes input which may either be Arabic text or arbitrary Latin text. Each name is carefully analyzed, one word at a time. Computational linguistic techniques and a large cross-script database of Arabic names are employed to produce a list of candidates which may correspond to each input word. After the translator then selects the correct choice, the complete name both in Arabic script and in formal transliteration is inserted into the document. XA also provides a powerful batch transliteration feature which can be used for automatically translating large lists of names (i.e. “phone books”).

This talk is recommended for software or language professionals at all levels of experience.

---

### Tutorial — Introduction to the Basis Arabic Editor and Workbench

---

**Mary Galvin, Software Analyst, Basis Technology & Bernard Greenberg, Chief Technology Innovator, Basis Technology**

This tutorial will provide a broad overview of Basis Technology’s Arabic Editor and Linguist’s Workbench, a powerful and flexible text editing and analysis system. Arabic Editor is best known for providing a simple method for entering and editing Arabic text using a standard “QWERTY” keyboard. However, its plug-in architecture allows it to incorporate many other utilities which make it a powerful tool for professional linguists, including phonetic (or “fuzzy”) search of Arabic text; integrated Arabic-to-English dictionaries; word analysis; automatic vocalization, and automatic transliteration.

This talk is recommended for software or language professionals at all levels of experience.

---

## Romanized Persian in On-line Communication

---

**Sabiha Imran, Computational Linguist, Basis Technology**

The advent of the World Wide Web and the rapid growth of computer mediated communication (CMC) through diverse on-line venues—such as chatrooms, blogs, emails, instant messaging systems, and bulletin boards—has led to an increase in the volume of online text in foreign languages and scripts. Based on research recently conducted by Basis Technology on Persian CMC, this presentation will give an overview of Romanized Persian in two types of “informal” CMC: bulletin boards and instant messaging. Focusing on language choice and language representation in on-line settings, we will present a linguistic analysis of online communication in Romanized Persian by the native speakers of the language. We will also discuss common patterns of linguistic behavior, such as cross-lingual code-switching; the use of English and Romanized Persian terms, acronyms, and phrases interchangeably. We will review some of the computational linguistic issues associated with Persian CMC, such as those caused by the colloquial Persian variants based on the speaker’s cultural background and educational level.

This talk is recommended for language professionals with some experience in online text analysis.

---

## Chinese Language Analysis: Solving the Chinese Puzzle

---

**Joe Ho, Principal Software Engineer, Basis Technology**

The Chinese language has been evolving for thousands of years. Following the establishment

of the People’s Republic of China, the mainland Chinese government simplified the writing system from the traditional style used in Taiwan and Hong Kong. Pan-Chinese trade, cultural exchanges, and the Internet have added further variation to an already complex language and orthography. Hence, analyzing modern Chinese text is a challenging task.

The presentation will survey the problems associated with automatic processing of Chinese. We will review the various Chinese character sets and encoding systems; input methods and transliteration; and the solutions offered by Basis Technology’s Chinese Language Analyzer and Named Entity Extractor.

This talk is recommended for software or language professionals at all levels of experience.

---

## Orthographic Variations in Arabic Corpora

---

**Bushra Zawaydeh, Ph.D., Computational Linguist, Basis Technology & Zina Saadi, Computational Linguist, Basis Technology**

This presentation will discuss the different kinds of Arabic orthographic issues that Basis Technology’s Arabic linguists have encountered and handled while building various software solutions for Arabic text analysis. Examples of variants that will be reviewed are typographic, morphological, phonological, cross-lingual, and spelling errors. The kinds of variations that are found in a corpus depend on the type of corpus. The kinds of corpora may be: (1) Modern Standard Arabic (MSA) corpora, such as a corpus of news articles, or a list of names written in Arabic script; (2) Colloquial Arabic written in Arabic script, such as a corpus of Arabic telephone speech, or a corpus developed for Arabic speech recognition, or a corpus in Arabic forums or chat rooms on the Internet; or (3) Colloquial Arabic written in Roman letters, such as a corpus of Arabic chat room speech written in Roman letters or Arabic. In this talk, we will present orthographic variations (that could be

viewed as challenges) which we encountered through our work on these three different kinds of corpora.

This talk is recommended for language professionals with some experience in Arabic information processing.

---

## Looking Behind the Name: Etymology of Arabic Names

---

**Zina Saadi, Computational Linguist,  
Basis Technology**

This presentation will give some samples of various linguistic rules that contributed to the evolution of certain famous Arabic names. It will sample different types of names as well as the influence of various foreign languages; regional and social impacts; and language evolution. We will focus mainly on the surface structure of full names and the deep linguistic structure behind each element that composes the full name. Examples will be chosen from various affiliations including ethnic (i.e. Kurdish, Persian, Berber); geopolitical (from both the Arab and Islamic worlds); and dialectal (i.e. Levantine, Moroccan, Egyptian, and Gulf).

This talk is recommended for language professionals with some experience in name analysis.

---

## Reverse Transliteration of Arabic: Recovering the Original Text

---

**Eva Sanchez, Director of Project  
Management, Basis Technology &  
Arnie Chien, Senior Software Engineer,  
Basis Technology**

Basis Technology's Arabic Reverse Transliterator converts Romanized Arabic text back into its original script. When Arabic speakers communicate using computers they often lack the tools or the training to type Arabic letters and therefore use the English alphabet to write phonetically. This form of communication

is especially popular in message boards and chat systems. This software module developed by Basis Technology converts Romanized Arabic text into Modern Standard Arabic (MSA) and may be integrated into machine translation systems to build larger analysis systems. This presentation will showcase the functionality of the Arabic Reverse Transliterator and its ability to handle the complexities of analyzing phonetic script.

This talk is recommended for software or language professionals with some experience in Arabic text analysis.



150 CambridgePark Drive  
Cambridge, MA 02140

T 617.386.2000

F 617.386.2020

## FAX / MAIL REGISTRATION FORM

### Basis Technology Government Users Conference and Workshop

The George Washington University, Marvin Center, Washington, D.C.

June 14, 2006 — 8:30 am to 5:30 pm

Name \_\_\_\_\_

Title \_\_\_\_\_

Organization \_\_\_\_\_

Mailing Address \_\_\_\_\_

City, State, Zip \_\_\_\_\_

Phone \_\_\_\_\_

Fax \_\_\_\_\_

E-mail \_\_\_\_\_

### Registration Rates

- N/C — U.S. Government Employee
- \$395.00 — Contractor (payment received on or before May 20, 2006)
- \$495.00 — Contractor (payment received after May 20, 2006)
- \$295.00 — Academic (current faculty or student ID required)

Please check the talks that you are most interested in attending:

- Tutorial — Understanding Unicode 5.0
- Tutorial — Designing Large-Scale Multilingual Systems
- What Language is *That*? Using the Rosette Language Identifier
- Putting the Rosette Global Name Matcher to Work
- Building Applications with the Rosette Linguistics Platform
- A Crash Course in Digital Forensics
- Large-Scale Digital Forensics
- What Every Software Developer Needs to Know About Arabic
- Multilingual Geographic Names: Mapping the World in Unicode
- Tutorial — Introduction to the Basis Transliteration Assistant
- Tutorial — Introduction to the Basis Arabic Editor and Linguist's Workbench
- Looking Behind the Name: Etymology of Arabic Names
- Romanized Persian in Online Communications
- Reverse Transliteration of Arabic: Recovering the Original Text
- Orthographic Variations in Arabic Corpora
- Chinese Language Analysis: Solving the Chinese Puzzle

Fax completed form to 603-658-4542 or register online at [www.basistech.com/conference](http://www.basistech.com/conference).  
For more information, please contact Stacy Pantazopoulos at 617-386-2050.