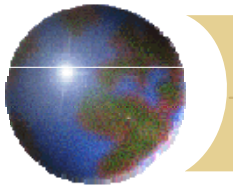




# *International Panel*

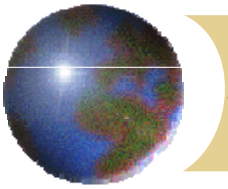
*IPRO Innovations 2009*

*May 14, 2009      Phoenix, AZ*



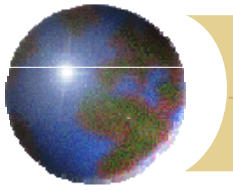
# *Agenda*

- Legal and Practical Perspectives on EU Data Protection and Privacy
- Cultural Considerations when Conducting E-Discovery Abroad
- Tips and Tricks on Reviewing the Multilingual Document Set
- Unicode and other Language Related Challenges



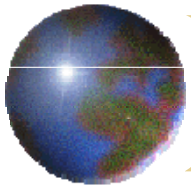
# *Legal and Practical Perspectives*

- Defining privacy and data protection
- “Safe Harbor” certification—requirements and limitations
- Restrictions on transfer of personal data
- Business Practices vs. Data privacy
- Issues regarding global movement and consolidation of data centers as it impacts the processing and transfer of personal data
- Issues regarding fact and work product cross-jurisdictional access to personal data via the internet



# *Agenda*

- Cultural Considerations when Conducting E-Discovery Abroad
- Tips and Tricks on Reviewing the Multilingual Document Set
- Unicode and other Language Related Challenges



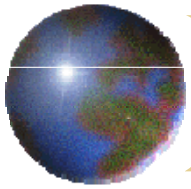
# *Non English Language Documents*

## ASCII vs. Unicode

- Computers only understand numbers—0's and 1's..
- ASCII designed to allow humans to communicate with computers.
- Invented for teletypes
- Original ASCII character set limited to 127 characters.

A -> 0100 0001



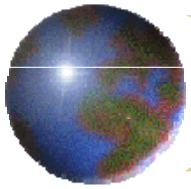


# *Non English Language Documents*

## Printable ASCII Characters

**0 1 2 3 4 5 6 7 8 9 a b c d e f g  
h I j k l m n o p q r s t u v w  
x y z  
A B C D E F G H I J K L M  
N O P Q R S T U V W X Y Z  
~ ! @ # \$ % ^ & \* ( ) \_ + ` -=  
[ ] \ { } | ; ' : " , . / < > ?**





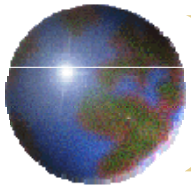
# *Non English Language Documents*

The bottom line...

- Chinese language has 65,000+ symbols
- Unicode assigns numbers to every possible character set.
- UTF-8 has become defacto Unicode standard to represent multi-byte languages.



**E-Discovery processing and review software must support Unicode!**

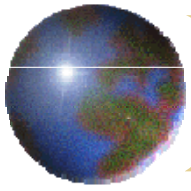


# *Non English Language Documents*

## Non English Language Tokenisation

- Western search based on spaces and punctuation.

アーティストコープに、ようこそ！  
オウンサウンドアーティストコープは、世界中の訪問者をいつも暖かく歓迎致します。この店では素晴らしい真のカナダ人のアートやクラフト製品が見られます。アーティストコープは、1994年、オウンサウンドのアーティスト達によって、この地域で開店致しました。この店の作品は皆、この地域で造られたもので、私達の活動は、この活気に満ちたアート社会での見本ともされています。アーティストコープは、また、アート、クラフトの町増進の為、案内や宣伝、紹介や教育などにも、携わっております。

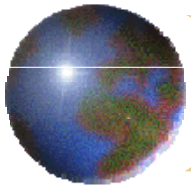


# *Non English Language Documents*

## Non English Language Tokenisation

- Some languages often don't use spaces or punctuation.

アーティストコープに、ようこそ！  
オウンサウンドアーティストコープは、世界中の訪問者をいつも暖かく歓迎致します。この店では素晴らしい真のカナダ人のアートやクラフト製品が見られます。アーティストコープは、1994年、オウンサウンドのアーティスト達によって、この地域で開店致しました。この店の作品は皆、この地域で造られたもので、私達の活動は、この活気に満ちたアート社会での見本ともされています。アーティストコープは、また、アート、クラフトの町増進の為、案内や宣伝、紹介や教育などにも、携わっております。



# *Non English Language Documents*

Non English Language Tokenisation

**The dog ate my dinner before I could stop him next time I will  
put him out before I eat**

The dog ate my dinner before I could stop him.  
Next time I will put him out before I eat.

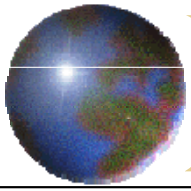
**裁判所はどこにありますか？**

Where is the courthouse?

小草的成长，  
离不开您的呵护

The grass's up-growing  
can't leave your blessing  
and good care.





# *Non English Language Documents*

Chinese

中國人

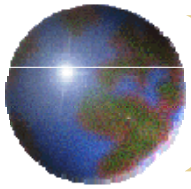
Middle country person

China

中國

Middle country

アーティストコープに、ようこそ！  
オウンサウンドアーティストコープは、世界中の訪問者をいつも暖かく歓迎致します。この店では素晴らしい真のカナダ人のアートやクラフト製品が見られます。アーティストコープは、1994年、オウンサウンドのアーティスト達によって、この地域で開店致しました。この店の作品は皆、この地域で造られたもので、私達の活動は、この活気に満ちたアート社会での見本ともされています。アーティストコープは、また、アート、クラフトの町増進の為、案内や宣伝、紹介や教育などにも、携わっております。



# *Non English Language Documents Docs*

The bottom line...

- Western search based on spaces and punctuation.
- CJK languages based on symbols rather than characters.
- CJK languages often don't use spaces or punctuation.
- Words may consist of one or multiple symbols

アーティストコープに、ようこそ！  
オウンサウンドアーティストコープは、世界中の訪問者をいつも暖かく歓迎致します。この店では素晴らしい真のカナダ人のアートやクラフト製品が見られます。アーティストコープは、1994年、オウンサウンドのアーティスト達によって、この地域で開店致しました。この店の作品は皆、この地域で造られたもので、私達の活動は、この活気に満ちたアート社会での見本ともされています。アーティストコープは、また、アート、クラフトの町増進の為、案内や宣伝、紹介や教育などにも、携わっております。

**E-Discovery processing and review software should tokenize non English languages!**