

AQUAINT: Data and Scenarios

6-Month Workshop, Monterey, CA
13 June 2002





Objectives

- Acquire/develop datasets, auxiliary resources and scenarios that can be used to extend the range of Q/A problems addressed by AQUAINT
- Dimensions of the extension
 - Multiple genre
 - Multiple media
 - Fixed-domains
 - Knowledge bases
 - Intel problems and/or analogs



Data Resources

- LDC (deal signed)
 - New 3 Gbyte collection of English newswire from Jun'98 through Sep'00, taken from NYT, AP, etc.
- CNS (in final negotiation)
 - Coherent body of text and structured data, together with auxiliary resources, in the Nonproliferation domain
- EELD (in discussion)
 - Text and structured incident data for Nuclear Smuggling and Contract Killing events
- MiTAP (under consideration)
 - Multi-lingual, multi-media content related to Infectious Diseases



Data Resources (cont.)

- Federal Govt. data
 - PTO
 - SEC
 - Medline

- Digital libraries

- Microsoft Tech Support data

- Google API



Auxiliary Resources

- Gazetteers
 - Sundheim/SPAWAR
- Thesauri, KBs, ontologies
- Question/Answer sets



Scenarios and Case Studies

- WMD scenarios
 - Will be working with CNS to develop scenarios and/or leverage existing case studies
- CT scenarios (options)
 - JMIC classroom exercises
 - EELD al Qaeda incidents
- Other scenarios (options)
 - NIMD



Next Steps

- Continue search for interesting data & resources
 - Augment CNS WMD data
 - Subsets of the TREC data
 - FBIS, Lexis/Nexis
 - WMD KBs
 - SEC filings, legal documents
 - Automatically generated “transaction” data
- Scenario development
 - Focus on fixed-domains
 - Tasks that require structured responses
 - Problems of interest to IC users
- Transition into MITRE testbed