# Why should you care about UIMA?

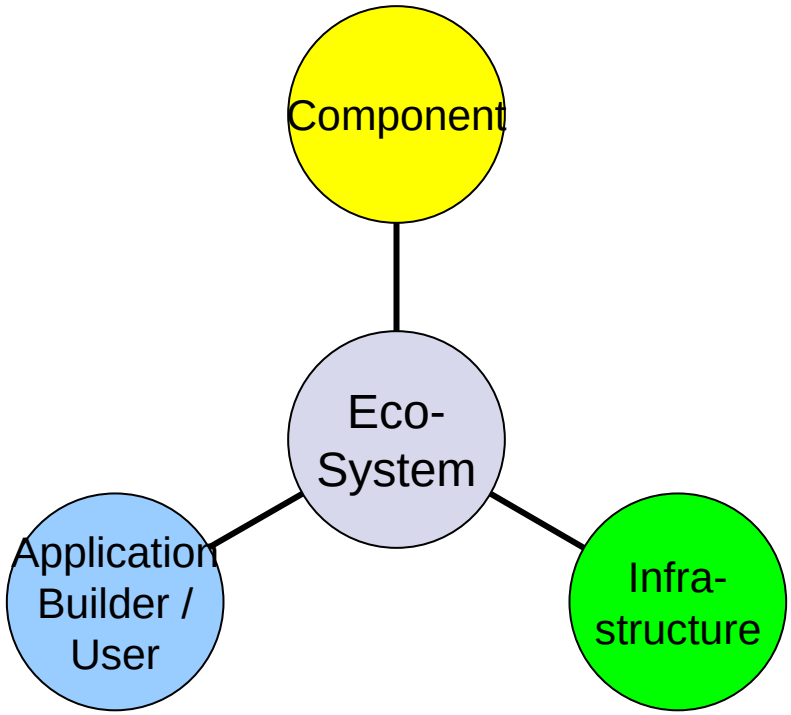
Marshall Schor
schor@apache.org

Apache UIMA Committer

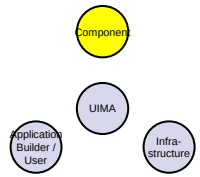
ApacheCon US 2007

Apache UIMA is an Apache Incubator Project

# What color is your hat?

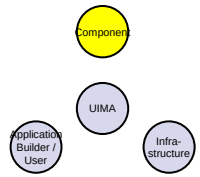


# Components



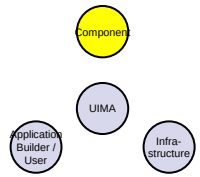

- Components often do complex analytics
  - Language ID
  - Tokenizers, Parsers
  - Named Entity Recognizers (Persons, Places, Countries, Companies, Time/Date, Phone numbers etc.)
  - Relation detectors (owns, occurs-during, located-in, is-CEO-of, etc.)
  - Sentiments
  - Not necessarily Text (audio / video)

# Components



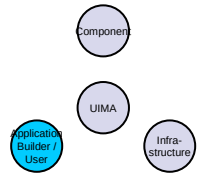

- UIMA Sandbox – something “out of the box”
  - Whitespace Tokenizer Annotator
  - Snowball Annotator
  - Regular Expression Annotator
  - Dictionary Annotator
  - Tagger Annotator (hidden Markov Model)
  - PEAR Packaging ANT Task
  - CAS Editor (tooling for manually annotating corpora)
  - PEAR Packaging Maven Plugin
  - Feature Structure Variables (experimental extension to framework)

# Component EcoSystem



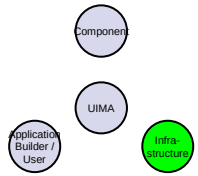

- Repositories
  - Carnegie Mellon University hosts a repository of UIMA components: http://uima.lti.cs.cmu.edu
  - Jena University also has started a collection of UIMA components: http://www.julielab.de/content/view/122/179/
- Sampling of Commercial Companies
  - Nstein
  - Temis
  - InfoExtract
  - RASP4UIMA
- Research Consortiums
  - Sapir (EU consortium) Search in Audio Visual Content using Peer-to-peer IR
  - TAO (EU consortium) UIMA enablement by Mondeca Open source infrastructure for Transitioning Applications to Ontologies
  - National Centre for Text Mining — NaCTeM
  - Mayo Clinic
  - Software Environment for the Advancement of Scholarly Research
  - University Workshop: GLDV CONFERENCE 2007 Society for Computational Linguistics and Language Technology - Universität Tübingen http://incubator.apache.org/uima/gldv07.html
- Standardization – http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=uima

# Solution Building



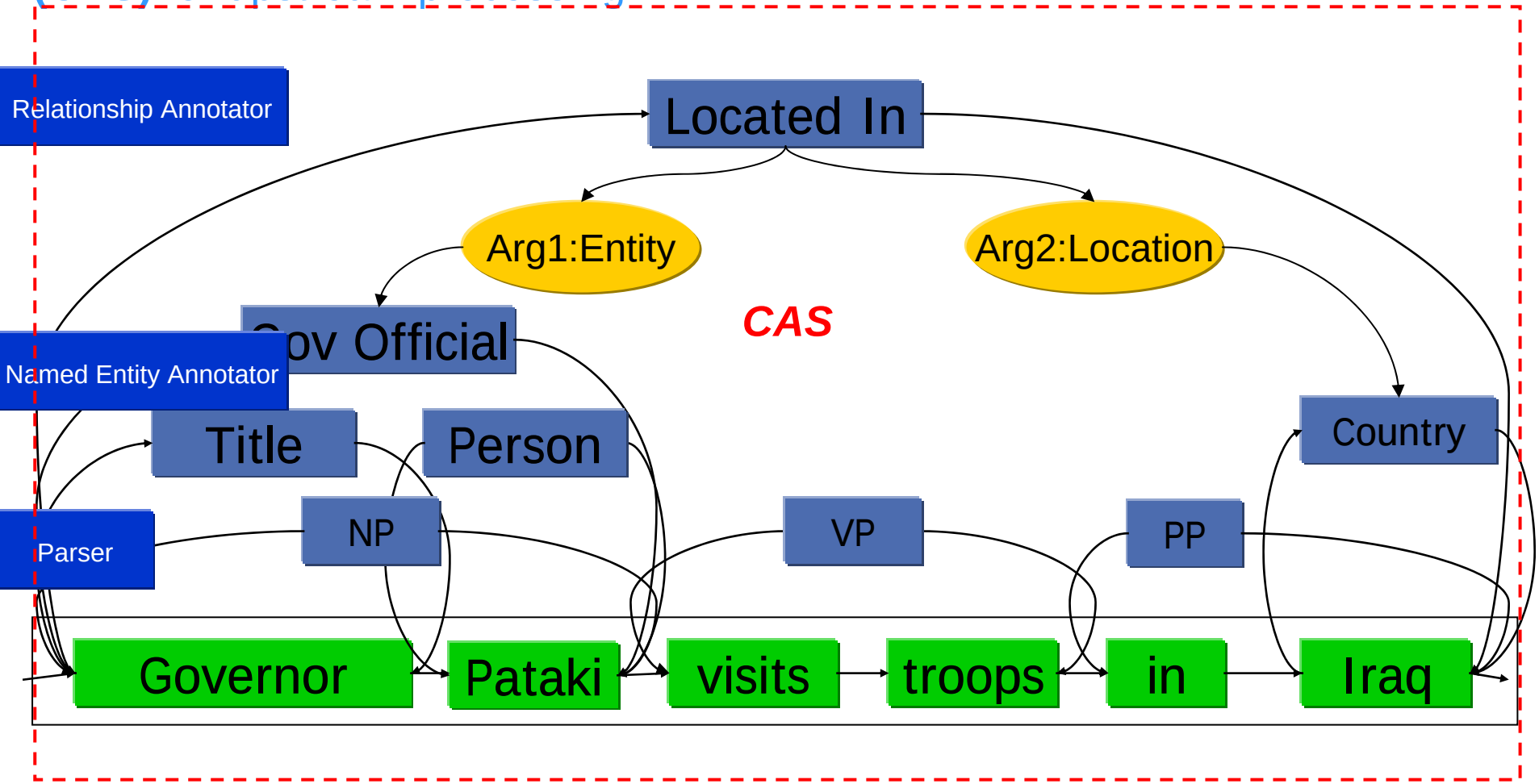

- TALES
  - From Video Feeds in foreign languages to Semantic Index of news stories
  - Speech to text, translation, named entity recognition, semantic indexing, with video retrieval
  - TALES video
  - TALES demo

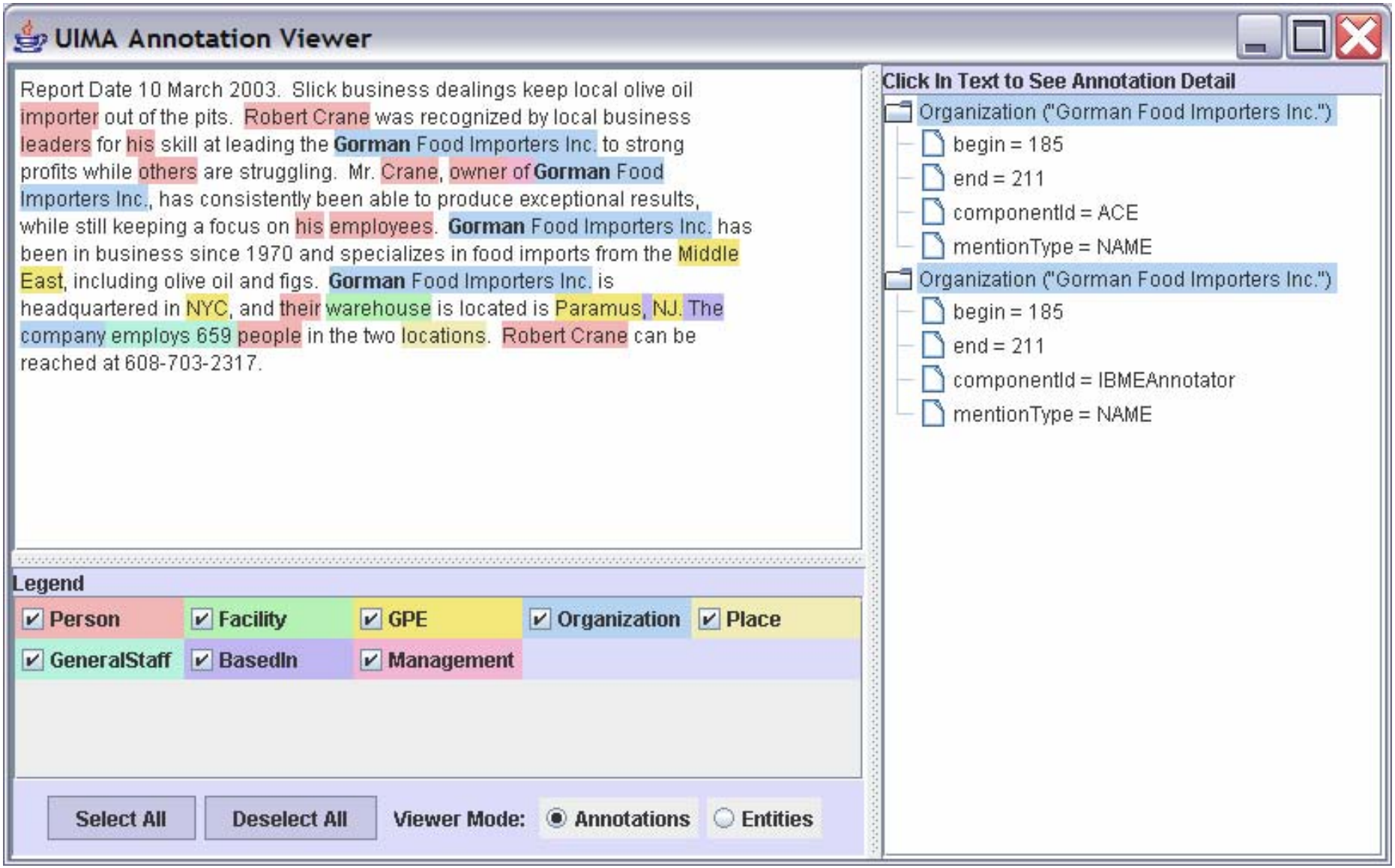
# Infrastructure




- UIMA inside IBM's:
  - Data Warehouse Edition - http://www-306.ibm.com/software/data/db2/dwe/unstructured-data-analysis.html
  - OmniFind Analytics Edition - ftp://ftp.software.ibm.com/software/data/ECM/WP/IBM_ECM_Omnifind_AE_WP.pdf
  - UIMA-Stage in DataStage (proof-of-concept)
- UIMA inside Temis's infrastructure - http://www.temis.com/
- UIMA inside InfoExtract - http://infoextract.com/id69.html
- New Scale-out prototype using Apache ActiveMQ - http://cwiki.apache.org/UIMA/uimaasdoc.html

UIMA's Basic Building Blocks are **Annotators – t**hey iterate over an artifact to discover new information and update the **Common Analysis Structure (CAS)** for upstream processing




UIMA Annotation Viewer
Report Date 10 March 2003. Slick business dealings keep local olive oil importer out of the pits. Robert Crane was recognized by local business leaders for his skill at leading the Gorman Food Importers Inc. to strong profits while others are struggling. Mr. Crane, owner of Gorman Food Importers Inc., has consistently been able to produce exceptional results, while still keeping a focus on his employees. Gorman Food Importers Inc. has been in business since 1970 and specializes in food imports from the Middle East, including olive oil and figs. Gorman Food Importers Inc. is headquartered in NYC, and their warehouse is located is Paramus, NJ. The company employs 659 people in the two locations. Robert Crane can be reached at 608-703-2317.
Click In Text to See Annotation Detail
Organization ("Gorman Food Importers Inc.")
begin = 185
end = 211
componentId = ACE
mentionType = NAME
Organization ("Gorman Food Importers Inc.")
begin = 185
end = 211
componentId = IBMEAnnotator
mentionType = NAME
Legend
Person
Facility
GPE
Organization
Place
GeneralStaff
BasedIn
Management
Select All
Deselect All
Viewer Mode:
Annotations
Entities

# Text Analysis in DB2 Warehouse

## Design Time

Configure annotators
(controlled Vocabulary, regular expressions taxonomies)

Define Dataflow
Create metadata

## Runtime

Transformation engine

Warehouse application

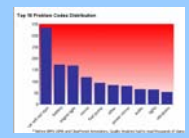
UIMA

Configured annotators

## Analysis / Reporting

Reporting OLAP

Data Mining



Intelligent Miner

COGNOS

SAS

BUSINESS OBJECTS

MicroStrategy

Cube views

## Deployment preparation

UIMA Tools

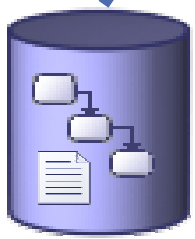
Package Analysis eng.

Design studio

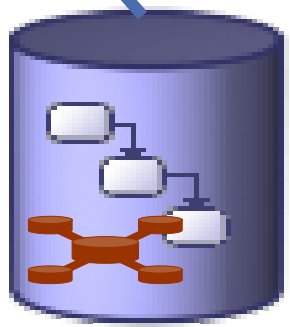
Generate Warehouse application

UIMA annotator, dictionaries

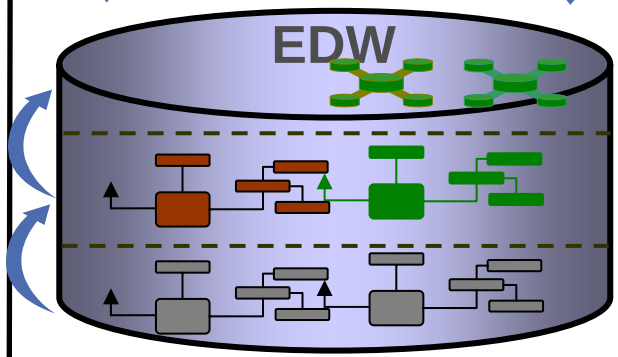
Warehouse application



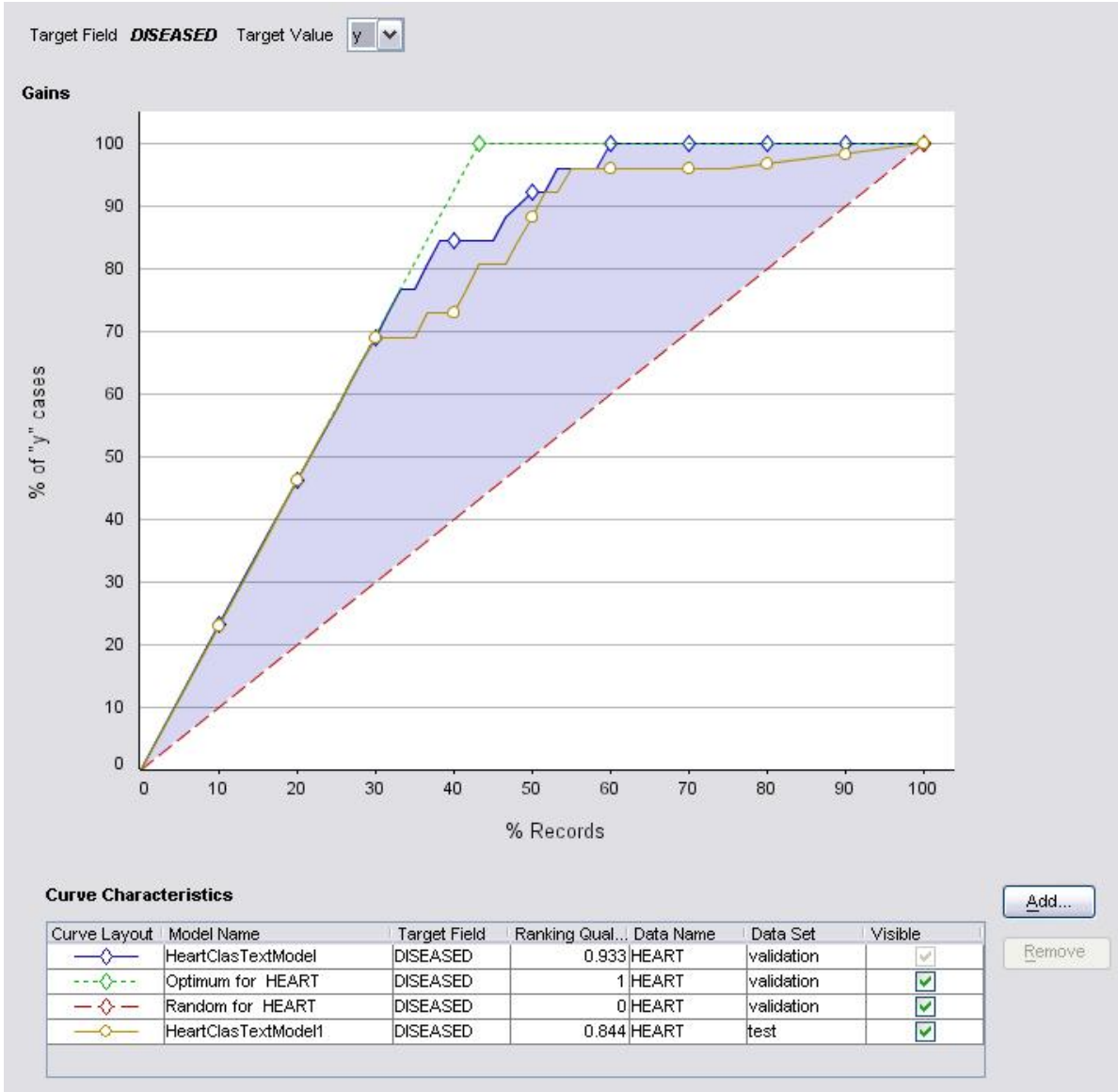
**Source DB**
**(with text fields)**



**EDW**



# The Classification model – Gains Chart




**Curve Characteristics**

| Curve Layout | Model Name | Target Field | Ranking Qual... | Data Name | Data Set | Visible |
|---|---|---|---|---|---|---|
| | HeartClasTextModel | DISEASED | 0.933 | HEART | validation | ☑ |
| | Optimum for HEART | DISEASED | 1 | HEART | validation | ☑ |
| | Random for HEART | DISEASED | 0 | HEART | validation | ☑ |
| | HeartClasTextModel1 | DISEASED | 0.844 | HEART | test | ☑ |

Add...

Remove

IBM Information On Demand 2007

IBM

# Extracting Knowledge from Unstructured Information using DB2 Warehouse 9.5

*Peter Bendel, IBM, peter_bendel@de.ibm.com*

*Session 1191A Dynamic Warehousing*

Act.Right.Now.

IBM INFORMATION ON DEMAND 2007
October 14 – 19, 2007
Mandalay Bay
Las Vegas, Nevada

# TEMIS and UIMA: Initial decision

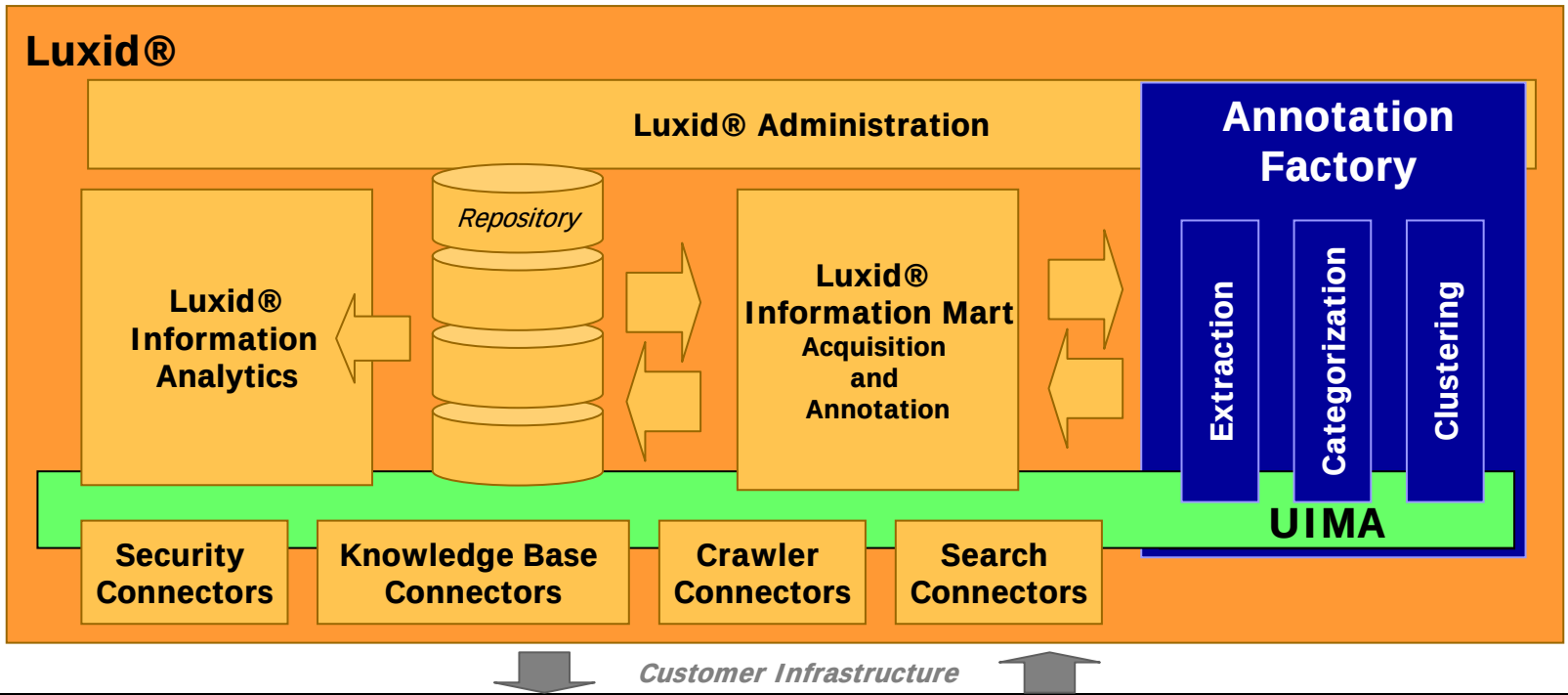
- Temis decided to embrace UIMA early 2006
  - As a standard to follow and support
  - And to use the IBM Java UIMA framework
  - Middleware of TEMIS Luxid servers
  - Started with IBM version 1.3, Switching now toward Apache version in next major release
- A thoughtful decision
  - Deep internal testing for the framework
  - Excellent documentation
  - IBM commitment
  - Customer requests for an open architecture
- Mastering and extending
  - Use of standard API
  - Extend when needed
    - Administration
    - Scalability

# TEMIS and UIMA: current involvements

- Integrated as the backbone of LUXID
  - Since mid 2006
  - Long term commitment
- Founder member of the OASIS UIMA specification committee
- Involvements in the Apache UIMA dev
  - Just started
  - Will contribute different modules in the next months (starting with generic scripting annotator)
  - Committed to participate to the integration of transport layer alternative

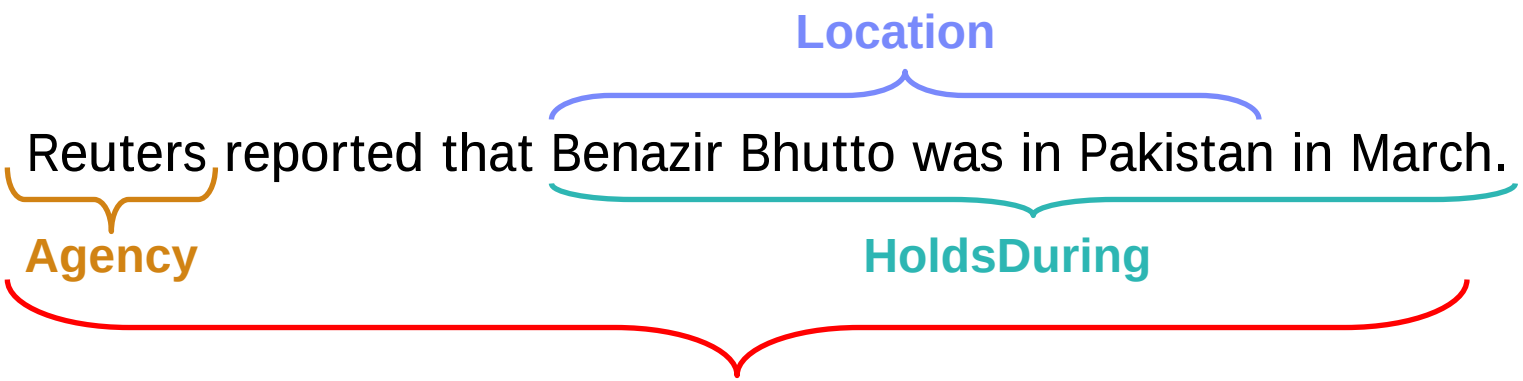
# Luxid® is built on top of UIMA

- Sources
  - Sources are UIMA sources
- Annotators
  - Annotators are UIMA annotators
- Workflow
  - Workflows are UIMA workflows
- Content Storage
  - Consumers are UIMA consumers
- Knowledge representation
  - A unique generic TEMIS UIMA TypeSystem represents metadata and annotations across Luxid



# XMLFragments – the overlapping issue

- Annotators may cause overlapping




```
<Reports>
      <Agency>Reuters<Agency>
      reported that
      <Location><HoldsDuring>Benazir Bhutto was in
      Pakistan</Location></Reports>
      in March
 </HoldsDuring>
```

Overlap

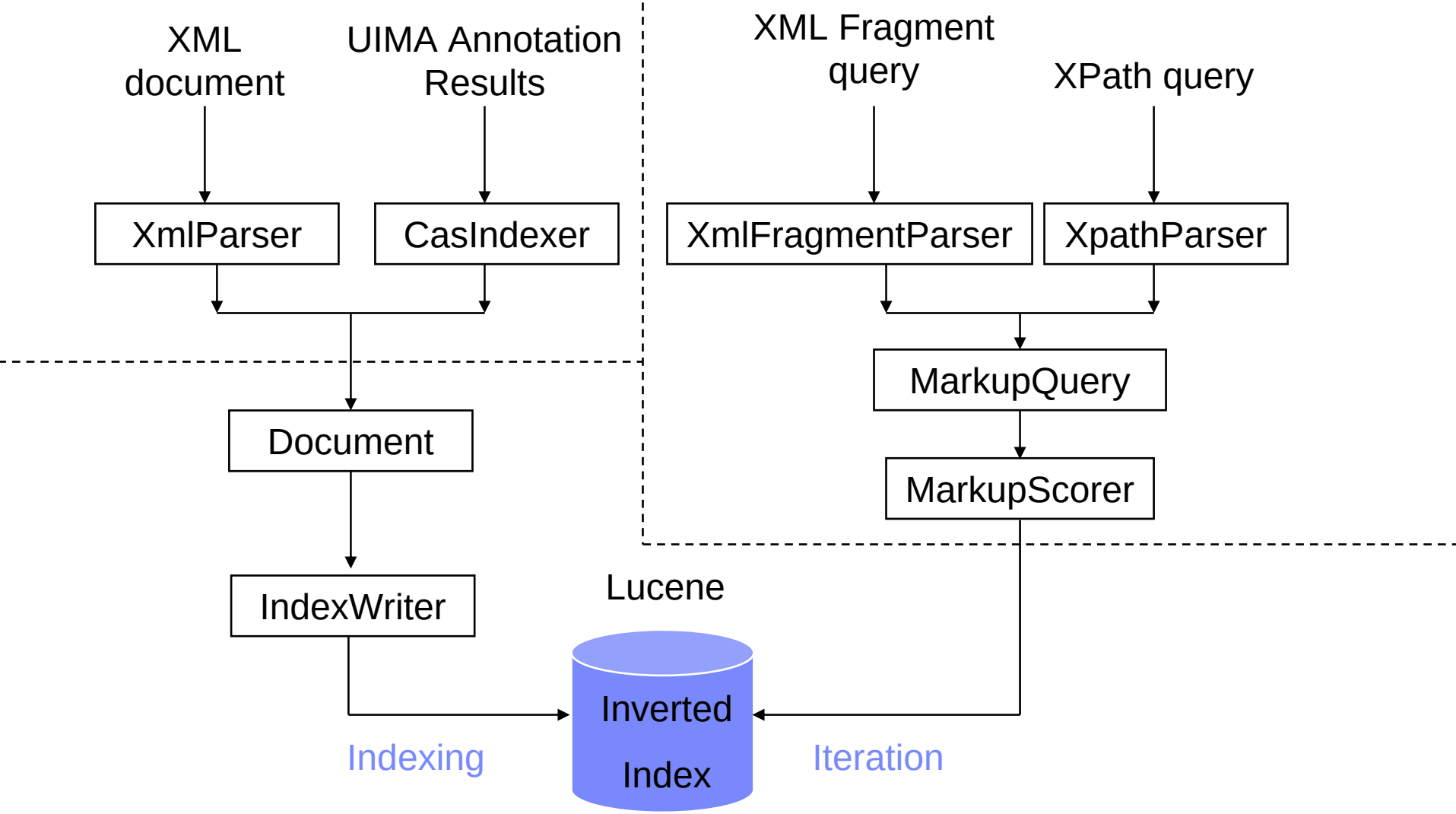
# XMLFragments – the overlapping issue - solution

- XMLFragment was extended by two operators
- "*Extending the XML Fragment Model to Support Querying over Annotated Text*", Mass et al., *SIGIR* workshop, 2004

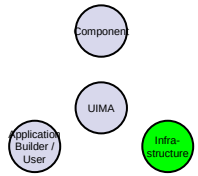
- Intersection Operator <T1 * T2> </T1 * T2>
- Concatenation Operator <T1 + T2> </T1 + T2>

- Query : '**<Report+HoldsDuring>Pakistan March Reuters</Report+HoldsDuring>**'

# Semantic Search Architecture



# Infrastructure




- Opportunity: Integrate *Apache Tika*
  - As configurable UIMA annotator or other?
- Opportunity: Consider incorporating UIMA-based analytics into your infrastructure
  - Allow your users access to the growing set of UIMA components
  - Apache UIMA Core team will help
- Opportunity: Ruby/Rails, Groovy/Grails integration ?
  - making it very easy to incorporate UIMA analytics into web applications

# IBM Funds UIMA Innovation Awards


- Encouraging components and community building
- ~20 awards in 2006 and 2007
- Activities around:
  - Research
  - Creating / evaluating annotators
  - Corpora
  - Tooling
  - Teaching
  - Infrastructure (for example, Grid based scaleout)
  - Semantic Web

# Apache UIMA symbiotics

- Apache Community possible involvement
  - Integrate UIMA where appropriate
  - Help us extend / align with Apache projects
- UIMA world-wide adoption (both commercially and academically) increases leverage for component work

# Help us grow!

**Thank You!**