Intelligent Content Engineering.

Klaus Netter (DNC GmbH)
Hannes Meyer (RC AG)

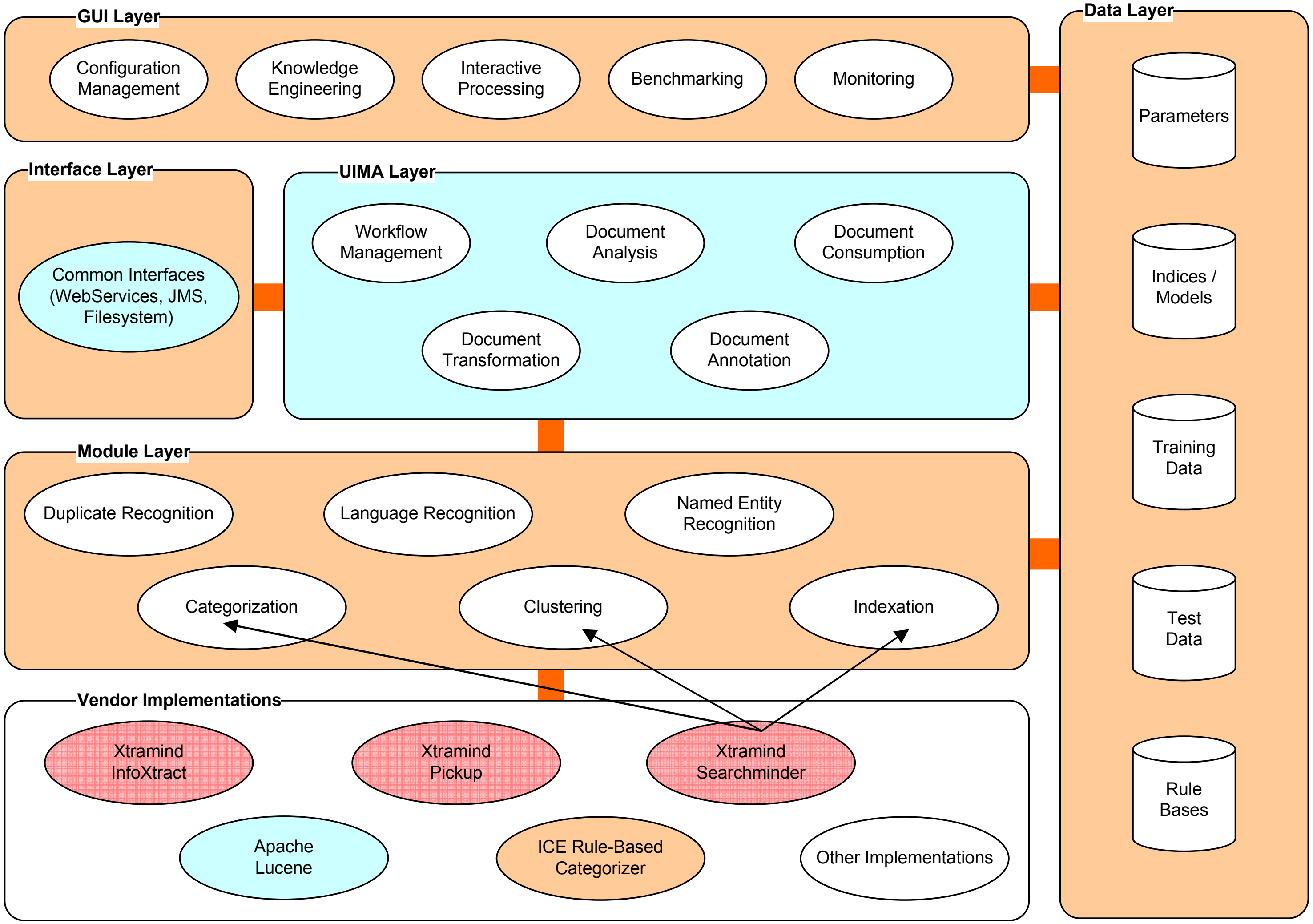
UIMA Workshop
GLDV Frühjahrstagung
Tübingen 2007

What is ICE?

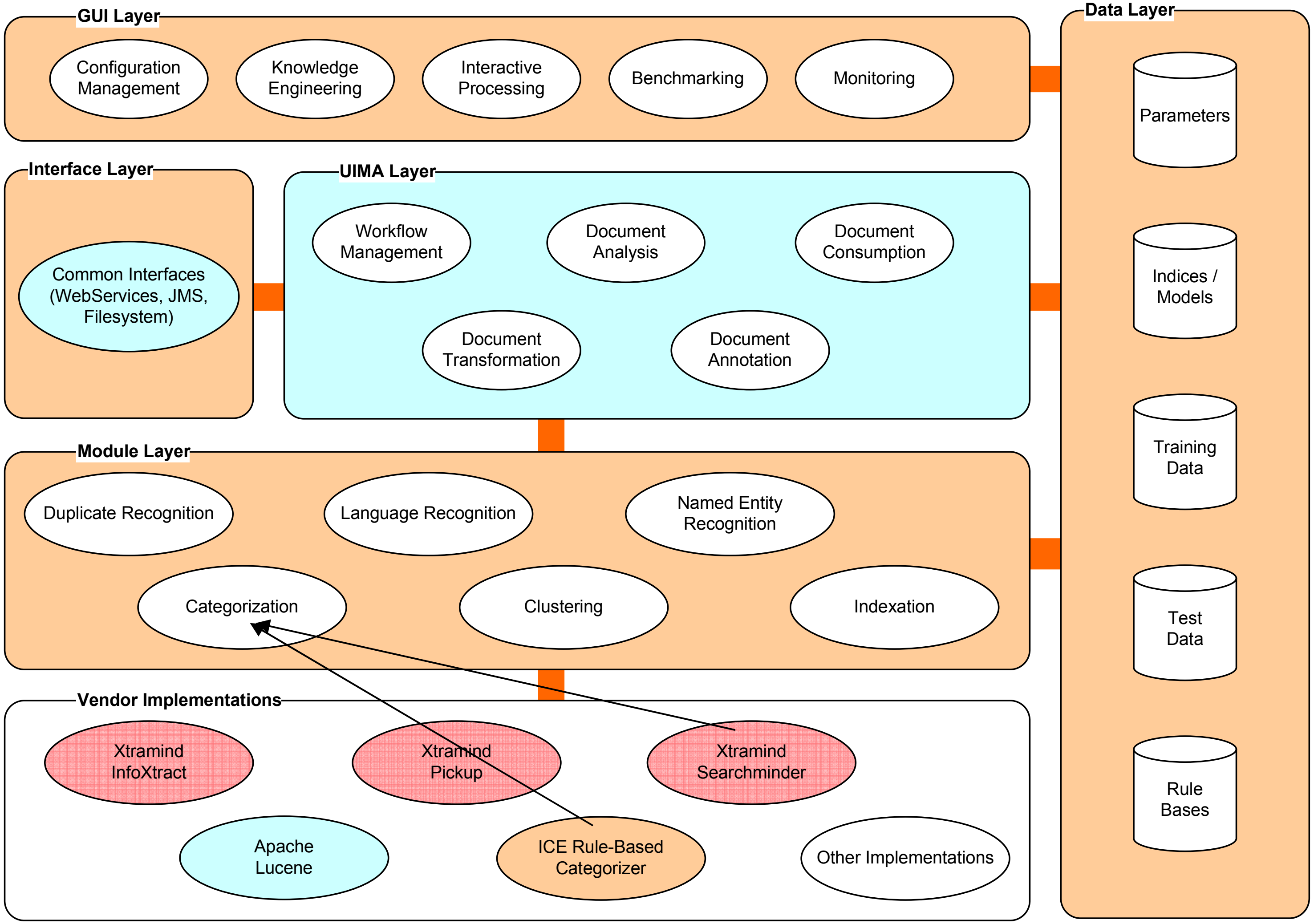
- ICE (Intelligent Content Engineering) is an industrial strength platform for the automatic disclosure and analysis of textual content with methods from Language Technology and Artificial Intelligence.
- ICE enables clients to model the knowledge of their particular domains without specific expertise in content analysis and text mining.

- ICE comprises different functional areas for
  - the development and maintenance of knowledge models
  - the testing and benchmarking of analysis functions
  - the processing of textual content in different workflows

Architecture



Architecture

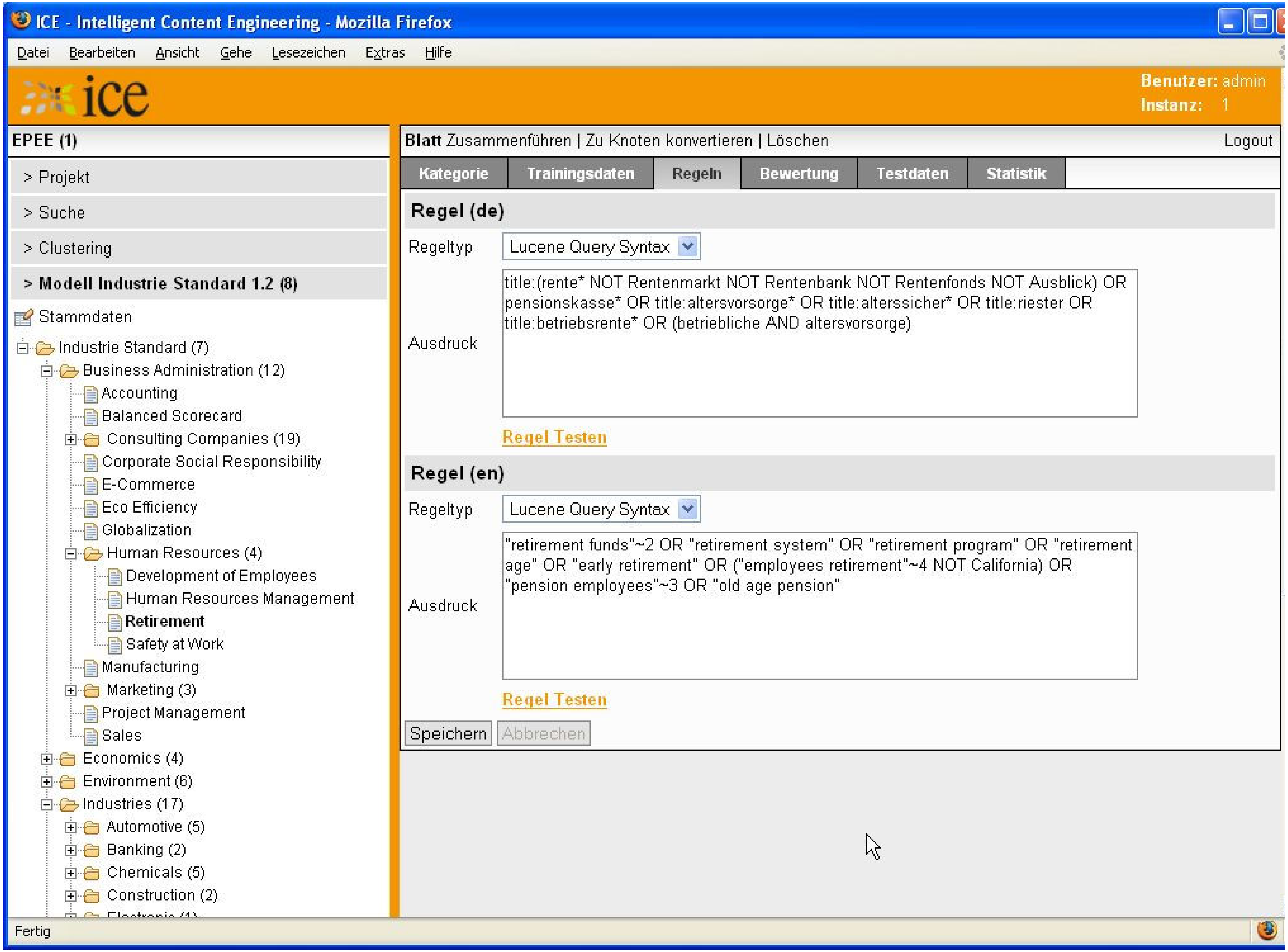


Development of Knowledge Models

- Models contain the knowledge necessary for processing components (e.g. category models, language models, filter rules, etc.)
- Models are developed largely independent of specific components and their implementation
- Models are based on abstract hierarchical tree structures
- Trees can be manipulated with all possible kinds of transformations (renaming, insertion, deletion, deactivation, shifting, merging, conversion) at the level of nodes or leafs
- Categories, rules, etc. are defined at the level of leafs
- Models are developed on the basis of versions
- Changes to models are registered and logged

Example: Category Models

- Multilingual models for parallel development in different languages
- Combination of rule-based and example-based category definitions
- Rule-based definitions with different variants
  - Lucene-based definitions with ranking
  - SQL-based with binary category assignment
- Example-based definitions for self-learning categorizers
  - Collection and maintenance of training material
  - Definition of test material
- Combination and individual weighting of category definitions
- Statistics for categorization results



Regel (de)

Regeltyp: Lucene Query Syntax

Ausdruck:

```
title:(rente* NOT Rentenmarkt NOT Rentenbank NOT Rentenfonds NOT Ausblick) OR
pensionskasse* OR title:altersvorsorge* OR title:alterssicher* OR title:riester OR
title:betriebsrente* OR (betriebliche AND altersvorsorge)
```

Regel (en)

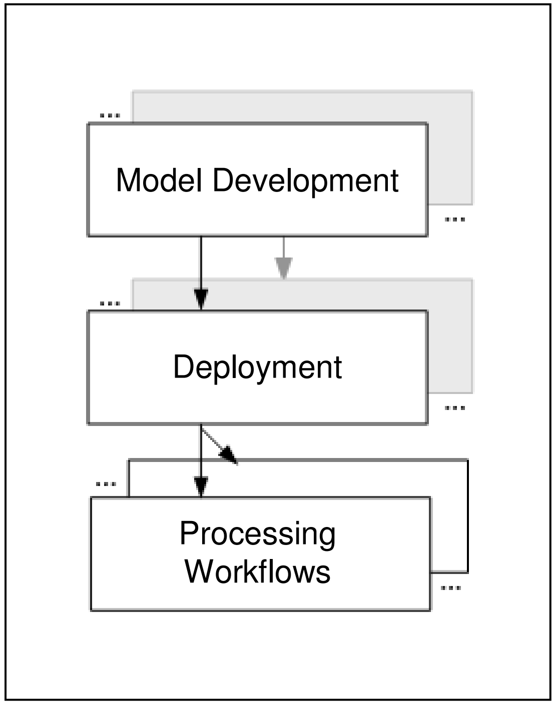
Regeltyp: Lucene Query Syntax

Ausdruck:

```
"retirement funds"~2 OR "retirement system" OR "retirement program" OR "retirement
age" OR "early retirement" OR ("employees retirement"~4 NOT California) OR
"pension employees"~3 OR "old age pension"
```

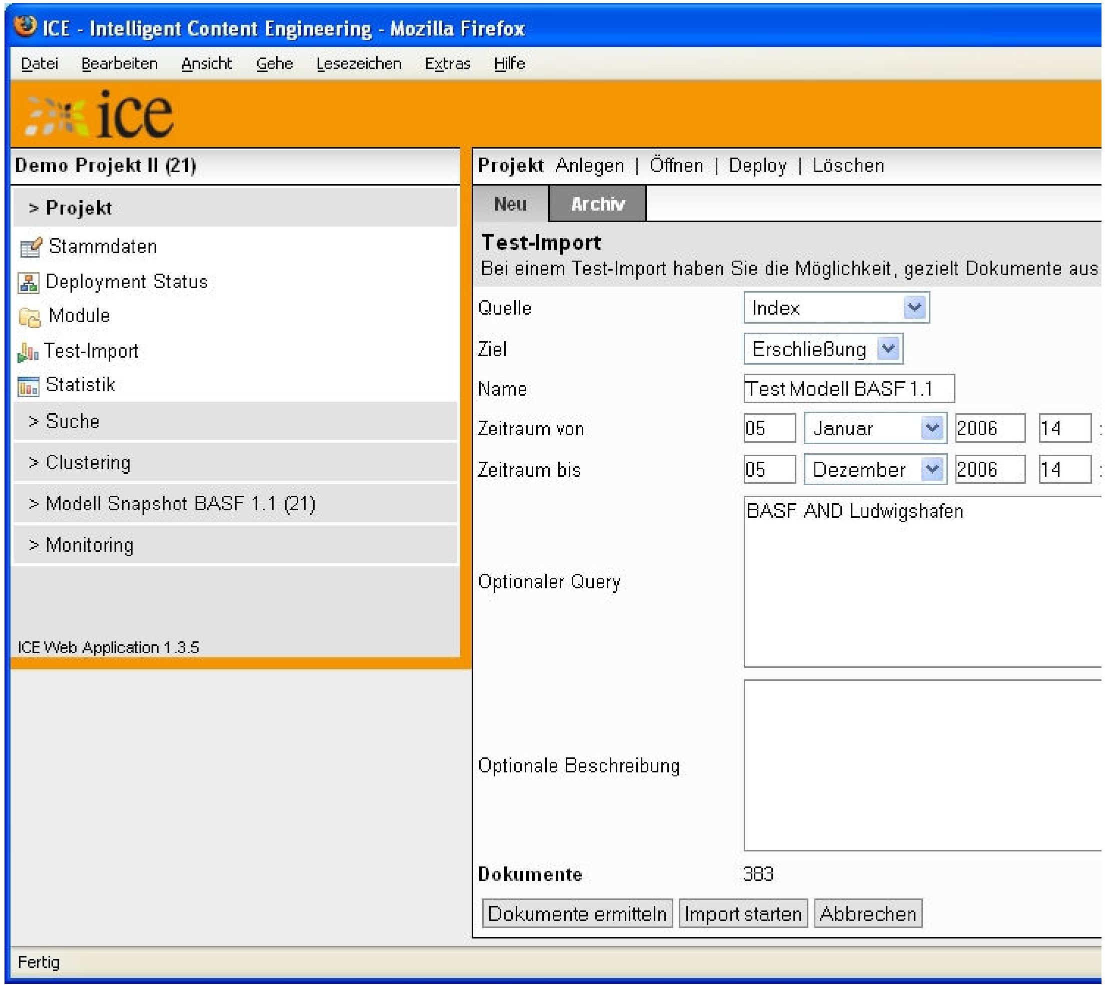
Processing Workflows

- Processing is organized into individual projects in an ICE instance
- Each project can have a different workflow or sequence of analysis steps
- Projects can be run in a productive or test mode
- Models or versions of models can be deployed in different projects
- Models can be shared across distributed instances of ICE



Testing and Statistics

- ICE provides a test environment for the evaluation of processing results
- Test-Imports can be selected from external or internal data sources
- Result data and statistics are collected at the level of
  - complete corpora or test imports
  - processing modes or components
  - individual rules, categories, etc.



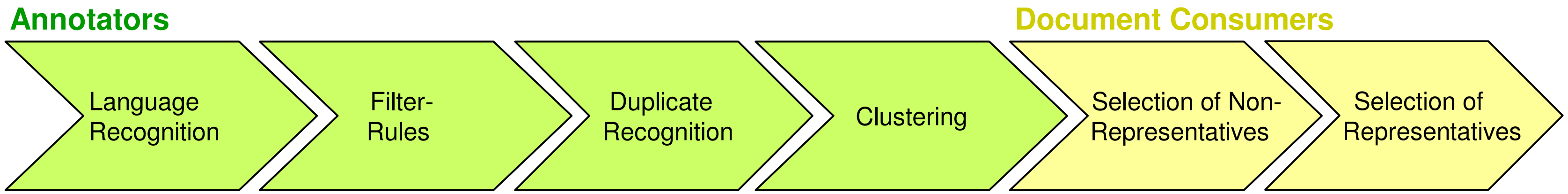
Test-Import

Bei einem Test-Import haben Sie die Möglichkeit, gezielt Dokumente aus

Quelle	Index
Ziel	Erschließung
Name	Test Modell BASF 1.1
Zeitraum von	05 Januar 2006 14 :
Zeitraum bis	05 Dezember 2006 14 :
Optionaler Query	BASF AND Ludwigshafen
Optionale Beschreibung	
Dokumente	383

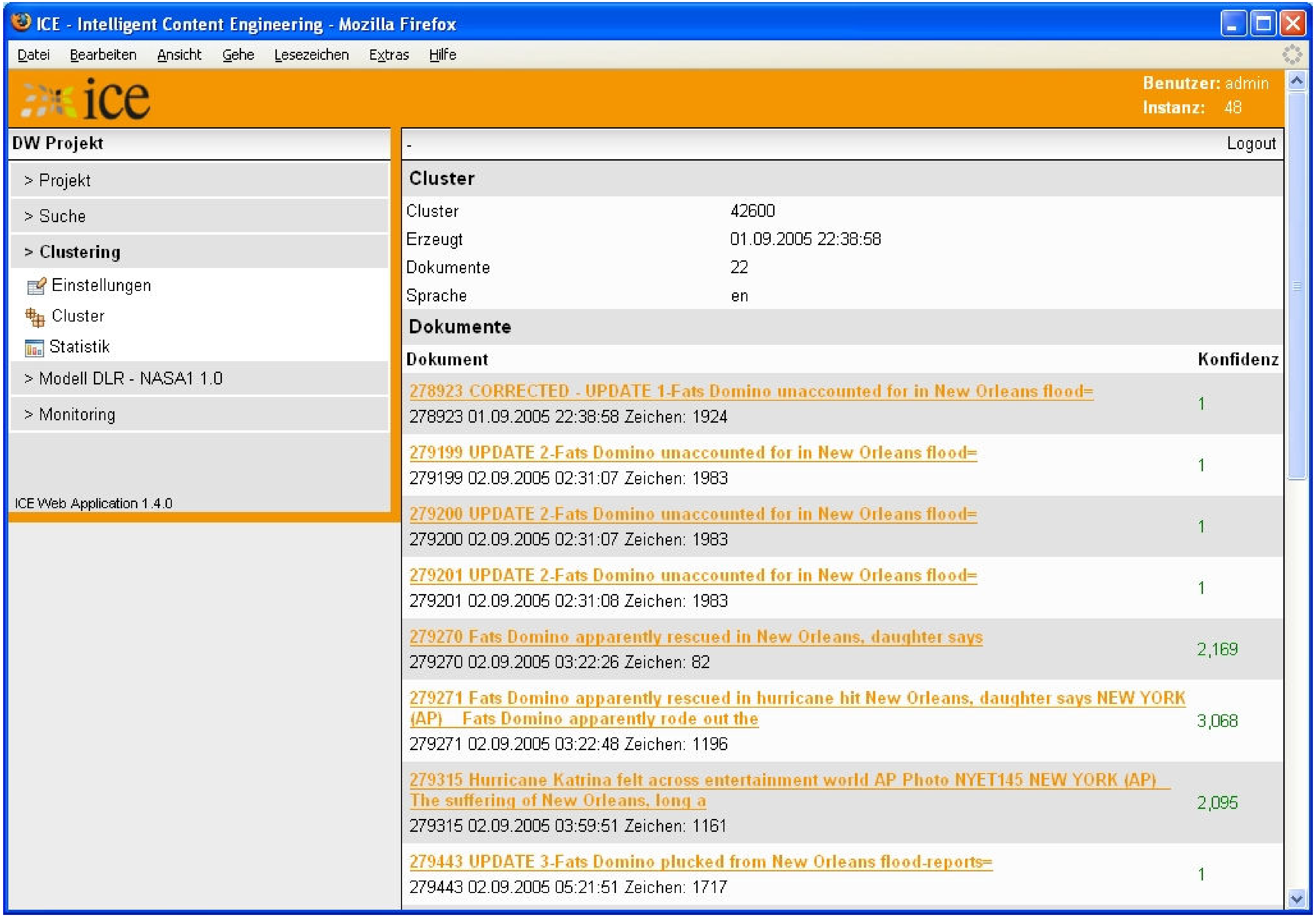
Intelligent Content Engineering.

Application Example: Clustering of newswire reports



Objective:

- Analysis of continuous flow of newswire reports and detection of topic threads across different agencies
- Condensation of document volume through selection of cluster representatives
- Parallel processing for different languages
- Long term archiving of substantially reduced original information



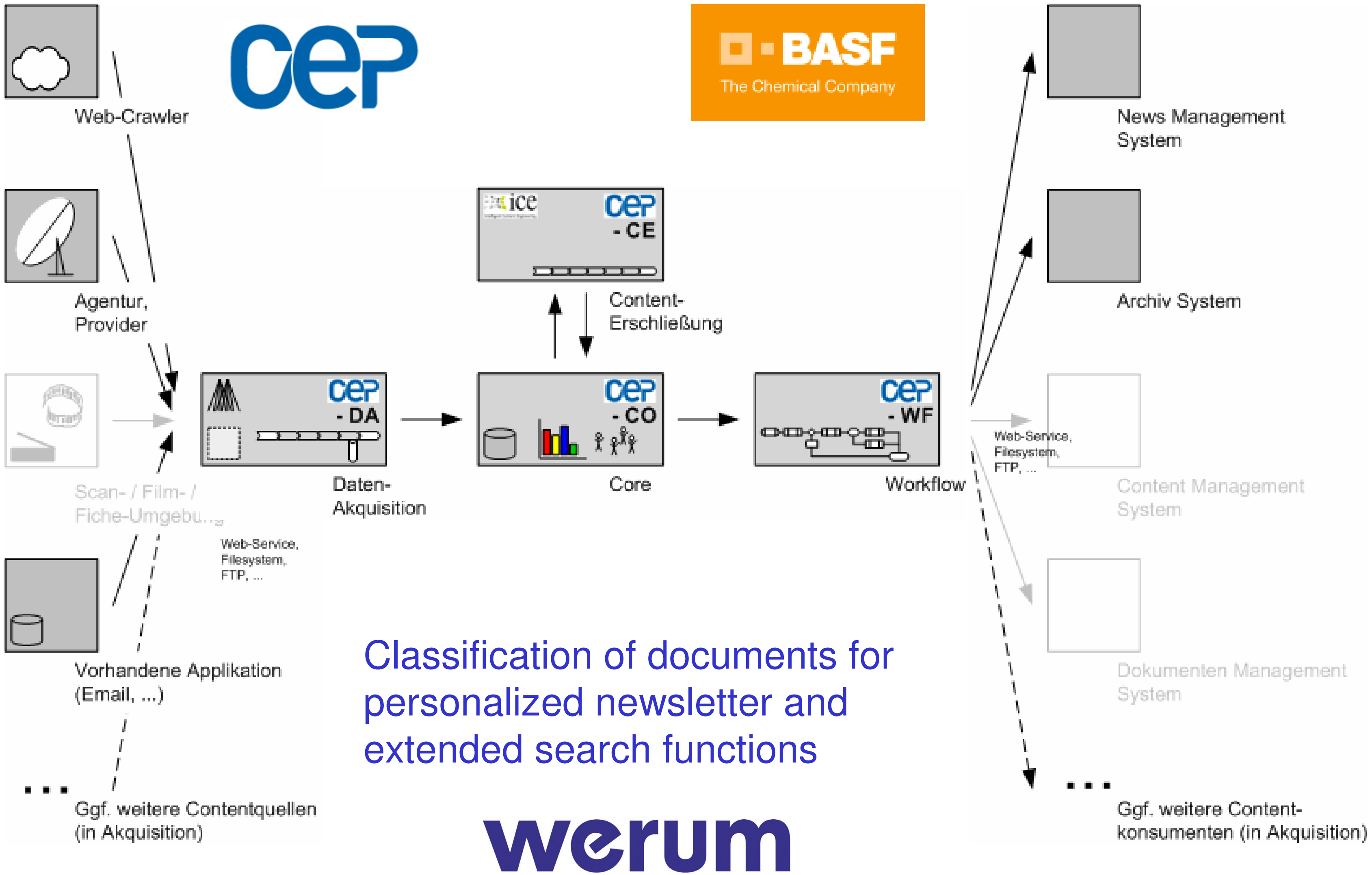
Cluster

Cluster	42600
Erzeugt	01.09.2005 22:38:58
Dokumente	22
Sprache	en

Dokumente

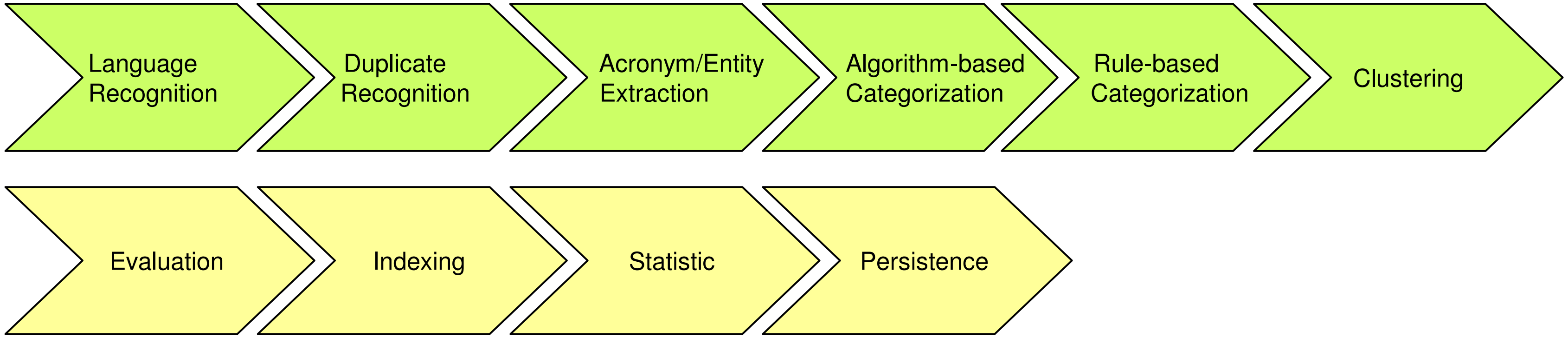
Dokument	Konfidenz
278923 CORRECTED - UPDATE 1-Fats Domino unaccounted for in New Orleans flood= 278923 01.09.2005 22:38:58 Zeichen: 1924	1
279199 UPDATE 2-Fats Domino unaccounted for in New Orleans flood= 279199 02.09.2005 02:31:07 Zeichen: 1983	1
279200 UPDATE 2-Fats Domino unaccounted for in New Orleans flood= 279200 02.09.2005 02:31:07 Zeichen: 1983	1
279201 UPDATE 2-Fats Domino unaccounted for in New Orleans flood= 279201 02.09.2005 02:31:08 Zeichen: 1983	1
279270 Fats Domino apparently rescued in New Orleans, daughter says 279270 02.09.2005 03:22:26 Zeichen: 82	2,169
279271 Fats Domino apparently rescued in hurricane hit New Orleans, daughter says NEW YORK (AP) Fats Domino apparently rode out the 279271 02.09.2005 03:22:48 Zeichen: 1196	3,068
279315 Hurricane Katrina felt across entertainment world AP Photo NYET145 NEW YORK (AP) The suffering of New Orleans, long a 279315 02.09.2005 03:59:51 Zeichen: 1161	2,095
279443 UPDATE 3-Fats Domino plucked from New Orleans flood-reports= 279443 02.09.2005 05:21:51 Zeichen: 1717	1

Application Example: News Management


Web-Crawler
Agentur, Provider
Scan- / Film- / Fiche-Umgebung
Vorhandene Applikation (Email, ...)
...
Ggf. weitere Contentquellen (in Akquisition)
Web-Service, Filesystem, FTP, ...
cep - DA
Daten-Akquisition
cep - CE
Content-Erschließung
cep - CO
Core
cep - WF
Workflow
Web-Service, Filesystem, FTP, ...
News Management System
Archiv System
Content Management System
Dokumenten Management System
...
Ggf. weitere Content-konsumenten (in Akquisition)

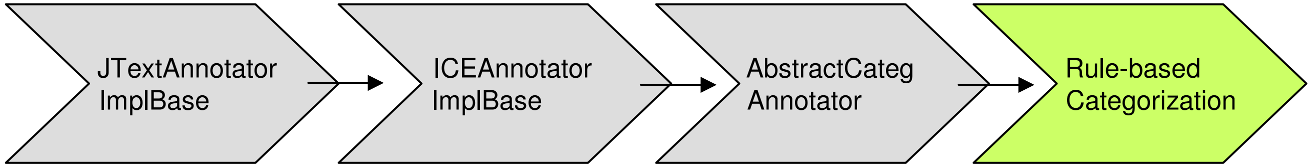
Classification of documents for personalized newsletter and extended search functions

Annotators/Consumers



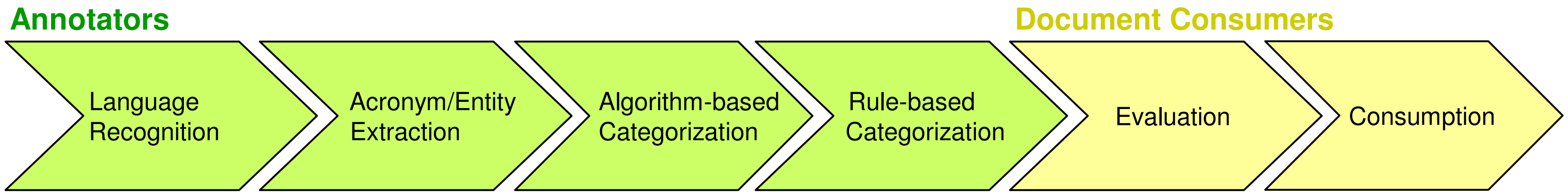
- Annotators are adding metadata to documents, e.g.
  - Language
  - Categorization information
  - Extracted information
- Consumers are reading document and previously added metadata, e.g.
  - Store document and metadata to a database
  - Index document

Annotator Example



- Extends ICE Annotator
- References a resource XML with rule definitions
- Definitions in different formats
  - Lucene Query Syntax
  - SQL Syntax
  - Regular Expressions
- Annotates document with category annotation
  - Category ID
  - Confidence
  - Categorization Type

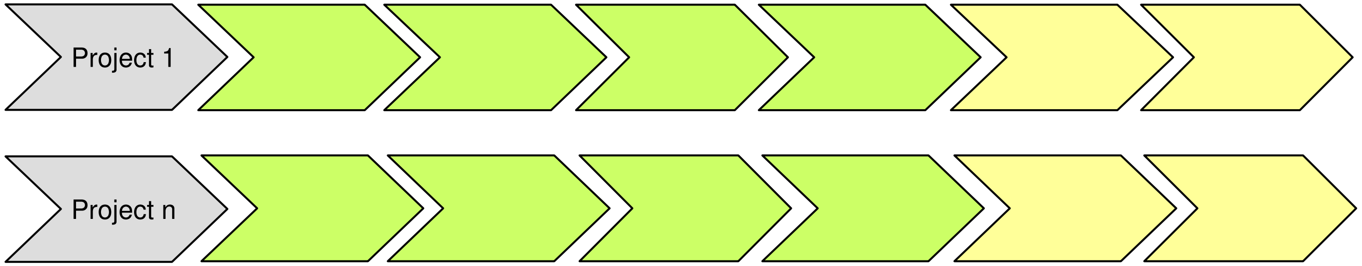
Aggregate Analysis Engine Example


Annotators
Document Consumers
Language Recognition
Acronym/Entity Extraction
Algorithm-based Categorization
Rule-based Categorization
Evaluation
Consumption

- Language Recognition on the basis of document text
- Extraction and Categorization is performed language dependent:
  - Training documents for example-based categorization
  - Specific rule expressions
- The evaluation document consumer decides on the basis of categorization confidence whether metadata about categorization is kept or discarded
- Consumption takes care about:
  - Indexing
  - Persistence
  - Statistics

The Project Approach

Processing Workflow



- Simultaneous processing of documents
- Different configurations per project
  - Combination of Annotators/Consumers
  - Rule sets
  - Training documents

Core Technologies

- Java 6
- RedHat JBoss Application Server
  - Serving interfaces (JMS, Webservices, RMI)
  - Serving ICE Administration GUI (Web Application)
- IBM UIMA
  - Apache UIMA currently in test phase
- Apache Lucene
  - Indexing/Search/Categorization
- Hibernate
- Xtramind Mindset
  - Categorization
  - Information Extraction

Benefits using UIMA

- The separation of code and configuration makes it easy to extend ICE without creating a new release
  - Change resources in UIMA descriptors
  - Change class declaration to use a specific Annotator/Consumer on the classpath
- Ready blueprints for tasks to standardize development
  - Annotator
  - Document Consumer
  - Collection Processing Engine
- Tooling support with the Eclipse IDE plug-in
- Testing made easy because no special container is needed

Authors

Klaus Netter

DNC Dr. Netter Consulting GmbH

info@dn-c.de

www.dn-c.de

Hannes Meyer

Rebel Creations AG

hme@rc.ag

www.rc.ag