**OpenKE - Technology**

Open Source Knowledge Enrichment Database  
Laboratory for Analytic Sciences  
las-openke@googlegroups.com

---

**Focused Web Crawling**

- Tunable and Configurable Web Crawler
- Page Data Model
  - Provenance Capture
  - Metadata Capture
  - Policy Support Data Header
- Multiple Source Types (Web, Forums, Search APIs)
- Web Crawling Configuration (Depth, Breadth, Relevancy, Site)
- Leveraging Structured Data Within Pages
- Policy (robots.txt, data header)
- Javascript and HTML challenges
- Access and Auditing Concepts for Policy

**OpenKE Web Crawling Job Config Example**

---

**Domain Learning and Discovery**

- Dictionary and PESTLE Annotations
- Regular Expressions / Relevancy Tuning
- Domain Discovery Capability
  - Search APIs and “Session” Result Comparison
  - Indexing Session Corpus via Text Rank
  - Annotation Analysis
    - Topic Modeling (LDA)
    - Data Source Learning
      - Page Crawl Progression (Page History)
      - Dynamic Content Challenge
      - Source Data Freshness

**OpenKE Domain Discovery Index View**

---

**Unstructured Text, Analytics, and Summarization**

- Document Handling Text Extraction (Apache Tika / POI)
- Diffbot Extraction
- Analytics
  - Text Summarization
  - Topic Modelling
  - WordCloud
  - Publish Date
  - Voyant Experiment
  - RASOR/Olympics Indications and Warnings Demo
  - ElasticSearch and Kibana Analysis
  - Annotated Field Search
  - Visualizations (Tuning/Optimize Crawl, OpenKE Usage)
- Future Analytic Framework

**OpenKE Web Crawling Rio Olympic I&W Tuning**

---

**Interfaces with External Tools**

- IBM Watson Content Analytics
- LAS Instrumentation
- PNNL Knowledge Graph
  - https://github.com/streaming-graphs/NOUS
- SAS
- Voyant

---

**Knowledge Graphs**

- “Holistic”
  - Facts
  - Events
  - Causal
  - Connections
  - Meta-data
- Intelligence / Analytic Tasks
  - Discovery
  - Behavioral Modeling
  - Network Discovery
- Analytic Support
  - Data / Evidence Gathering
  - Prediction
  - Model Generation: AI Planning
  - Ontology Development
  - Knowledge Base

---

**OpenKE Technical Framework**

- **Platform:**
  - Java
  - Hortonworks Data Platform
- **Storage:**
  - Accumulo
  - ElasticSearch
  - HDFS
  - OrientDB
  - PostgreSQL
- **Open Source Libraries:**
  - Crawler4J
  - jsoup
  - Apache Tika
  - Apache POI
  - Tabula
  - Stanford CoreNLP
  - Python NLTK
  - Python Gensim
  - University of Washington: OpenIE
  - d3.js
- **Other:**
  - Docker
  - Kibana
  - ApacheSpark
  - Apache Zeppelin
  - Tor2Web

---

**OpenKE Web Crawling Rio Olympic I&W Tuning**

---

**OpenKE - Technology**

Open Source Knowledge Enrichment Database  
Laboratory for Analytic Sciences  
las-openke@googlegroups.com