



cybertap

The Use of
Search Engines
for Massively Scalable
Forensic Repositories

www.cybertapllc.com/

www.cybertapllc.com

transforming data into knowledge

John H. Ricketson
jricketson@cybertapllc.com
jricketson@dejavutechnologies.com
+1-978-692-7229



cybertap

Who Is cybertap?

- We provide a **forensic platform for cyber investigations based on search engine technology**
 - External Threats & Viruses
 - Hacking
 - Financial Fraud
 - Electronic Warfare
 - Security Event Analysis
- **Markets**
 - Law Enforcement
 - Cyber Security, for both Government and Commercial Enterprises
 - Banking / Trading
 - Electronic Commerce
 - Call Centers



cybertap

Forensic Evidence

- **Documents – Computer Forensics / eDiscovery / “data-at-rest”**
 - Disk Scrubbing, Dead Boxes, etc
 - Shared Repositories like Dropbox or Sharepoint
- **Archives**
 - Email
 - Instant Messaging
- **Web**
 - Downloaded HTML pages
- **Financial Information**
 - Invoices, Credit cards, Private Information
 - Electronic Trades
- **Log Files from Network Devices**
- **Cell Phone Call Records**

- **Real-Time Network Transactions – “data-in-motion”**
 - Packet Captures
 - Network Streams



cybertap

Forensic Data

- Archival in Nature
- Non-Transactional
- Contains Meta-data & Content & Extracted Intelligence
 - Meta-Data
 - Document Attributes
 - Author, Dates, Printers, Macros, Document Edits, File reference
 - Network Attributes
 - Addressing Endpoints, ID's, Domains, Protocol Headers
 - Content
 - Message Content
 - Body Content
 - Media Streams
 - Extracted Intelligence
 - Electronic Persona (epersona)
 - Geo-location
 - Correlations and links among heterogeneous data



cybertap

Search Engines Provide

- Non-Transactional repository of archival data
- Meta-data descriptors for network and document attributes
 - Delineating meta-data from content in search queries is crucial
 - Show me all email documents from this.IP address to that.IP
 - Show me all IM messages from this.ID to that.ID containing “nitrate”
- The ability to search in free form any meta-data or content item
- Massively scalable forensic repositories
 - 10+ Billion documents representing Terabytes of data
- Sub-second search times
- Cross reference and correlation of all data with a single search
- Document parsing
- Leverage these **FREE**, rich-functionality, forensic repositories
 - <http://lucene.apache.org/solr/>
 - <http://tika.apache.org/>



cybertap

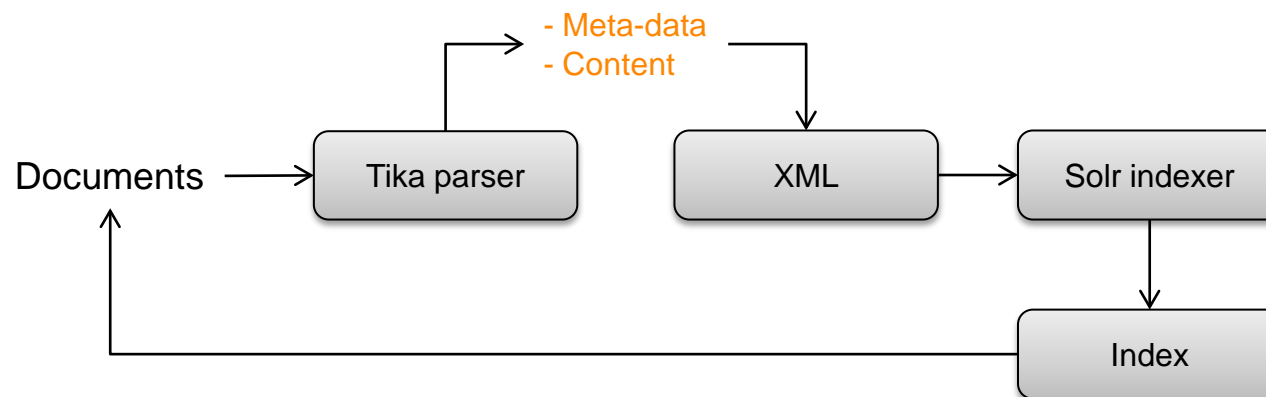
Tika Provides

- **Meta-data, MIME, Language & Content**
 - HyperText Markup Language (html)
 - XML and derived formats
 - Microsoft Office document formats
 - OpenDocument Format
 - Portable Document Format (pdf)
 - Electronic Publication Format
 - Rich Text Format (rtf)
 - Compression and packaging formats (zip, tar, etc.)
 - Text formats
 - Audio formats
 - Image formats
 - Video formats
 - Java class files and archives
 - The mbox format
- **Identifies documents by content ONLY**
 - File extension \neq Content



cybertap

Apache Solr & Tika Open Source Projects





cybertap

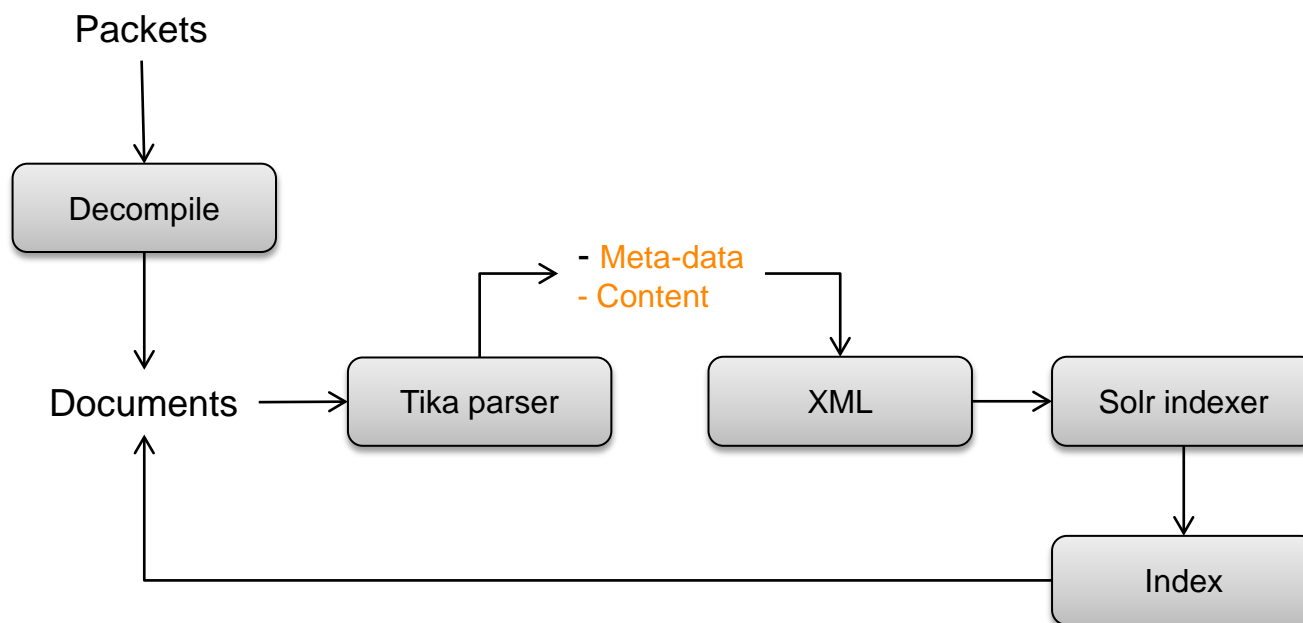
Demonstration of Document Searches

- Importing Documents
- Importing E-mail
- Searching both Meta-data & Content



cybertap

Real-Time Network Packet Ingestion





cybertap

Open Document Extraction & Enrichment

- **VoIP telephone call processing – wav**
 - Including Speech-to-Text, Voice Identification, Voice Recognition
- **Video processing – video, flash**
 - Including OCR, Speech-to-Text, closed captioning, and multi-frame analysis.
- **Image processing – jpeg, pdf, gif, etc**
 - Including OCR, facial recognition, flesh tone detection, etc.
- **Natural Language Processing - text**
 - Including language translation, text entity extraction, proper name disambiguation, summarize conversations or identify people by writing style.
- **Automated Zero-Day Malware/Virus Detection**
- **Steganographic detection**
- **Decryption**
- **ePersona Background Checks**



cybertap

Demonstration of Packet Searches

- Importing Packets
- Searching both Meta-data & Content
- IMAP - HTTP - VoIP - Facebook
- Reconstruction
- Content Extraction
- Electronic Persona Identification (epersona)

