The Future Of OSINT

Bridging the OSINT Capability Gap Through Collaboration

Andy Lasko

October 12th 2011

This briefing is classified

UNCLASSIFIED
Who am I?

• Andy Lasko
• Consulted on **dozens** of the IC’s Largest OSINT Programs
• 100’s of Private Sector OSINT programs
• Technical Alliance Manager, Kapow Software
  – Premier OSINT Collection Platform since 1998
  – Booth 205
What is OSINT?

Finding, selecting, and acquiring information from **publicly available sources** and analyzing it to produce **actionable intelligence**.

- **Media**: newspapers, magazines, radio, television etc.
- **Web-based communities and user generated content**: social-networking sites, video sharing sites, wikis, blogs etc.
- **Public Data**: government reports, budgets, demographics, hearings, legislative debates, press conferences, speeches, marine and aeronautical safety warnings, environmental impact statements and contract awards.
- **Professional and Academic**: conferences, professional associations, academic papers, and subject matter experts.
- **Geospatial Open Source**: maps, atlases, gazetteers, port plans, navigation data, human terrain data, environmental data, commercial imagery etc.
Why Is **OSINT** The Internet Important?

The growth of social media, social networking sites, media sharing sites, and their ease of access through various devices.

– Whether its riots in Egypt, political protest in Iran or terror group recruitment, OSINT provides a relatively cheap and immediate form of intelligence for the community.

  • Al Jazeera reporter Dan Nolan tweeted during Egyptian clashes on 2 February: "Soldiers left 4 tanks outside museum. Now anti gov. protesters sitting on top. Main battle about 100m further toward gala st."

We must collect now!
How Good is Our OSINT Capability?

• Lack Defined Processes
  – Unreliable Data, Sub-Par Processes
• Lack of Automation
  – Wasted Time, No Re-Use
• Overwhelmed by Unstructured Content
  – Over focus on Machine Learning and AI
  – Neglecting Structure in Unstructured Enrichment
  – Ignoring Structure to Influence the Enrichment Pipeline
• Improper Priorities
  – OSINT is a low priority compared to other INTs.
  – Programs invest too heavily on manual efforts
  – Programs focus on making sense of messy collected data
What Do We Need to Do?

- Automate the collection process
- Get more structure into your pipeline
- Remove noise from the data
- Improve accuracy of the data pipeline
- Leverage multiple ontologies
- Seamlessly discover information across structured and unstructured data
- Crowdsourceto improve enrichment
- Push OSINT services to the people
Automate the Collection Processes

• Deploy On-Line, On-Demand OSINT Services
  – Rapid Service Creation
    • Data is changing, too many sources, changing environment
  – On-Line
    • Leverage these services across the enterprise
  – On-Demand
    • Initiate new data collections
    • Query Enriched Content

• Evaluate and Refine Processes

• Invent New Processes
Demonstration
Finding Structure In the Unstructured

• Broad Crawls
  – Use common data
    • H1, H2, Metadata tags – title, keywords

• Targeted URL Crawls
  – Use the HTML tags to find structure on targeted crawls
    • Relationships, many to ones, dozens of data points
  – Requires an Extraction Browser

• Always keep raw data
Remove Noise From The Data

- Remove advertising through pattern matching

- Don’t load Noise

- Crowdsourcing, feedback loops, systems that learn based on user behavior
Improve Accuracy of the Data Pipeline

- Use the Structured Data Points to help the Pipeline’s Accuracy
- Allow the Pipeline to make recursive calls
  - Re-collect or collect new content and call other portions of the pipeline as your workflow see’s fit.
- Trust, trustworthy data, leverage less trustworthy data
  - An OSINT phone number lead to the death of Abu Musab al-Zarqawi, former al Qaeda in Iraq leader
  - A Google search on an IP address of interest returned a link to GhostNet’s central management console.
- Teach Your Pipeline Applications
  - NLP technologies have used data collected to learn
Leverage Multiple Ontologies

• Use Ontologies to Influence the Pipeline
  – Human Terrain Mapping Example of a news story
• Allow different perspectives to process and evaluate data differently
  – Clearance means something different to truck driver than it does to someone in CIA
  – A ‘Tank’ means something different to an infantry man than to a logistician.
Seamlessly Discover Information Across Structured and Unstructured Data

• One Box Example

- Google search for "detroit weather"
  - About 42,600,000 results (0.06 seconds)
  - Weather for Detroit, MI:
    - 50°F | 10°C
    - Current: Mostly Cloudy
    - Wind: N at 5 mph
    - Humidity: 37%
    - Forecast: Thu 51°F | 10°C, Fri 47°F | 8°F, Sat 63°F | 14°F, Sun 57°F | 14°F
  - Detailed forecast: The Weather Channel - Weather Underground - AccuWeather

- Search Other Collections
  - Search String: 
  - Max Results / Source: 
  - Sources:
    - USDA Forest Service
    - Aldo Leopold Research Institute
    - Minerals Management Service
    - US Government Printing Office
    - US Fish and Wildlife Service
    - Bureau of Land Management
    - NOAA's National Centers for Coastal Ocean Science
    - National Park Service
    - National SeaGrant Library
    - US Army Corps of Engineers - Institute for Water Resources
    - US Geological Survey

• Source Selection
Crowdsource to Improve Enrichment

- Enable people to rank the results
  - How accurate is the data
  - Were the right data elements collected
  - Is the Ontology Accurate
  - Is the translation correct
  - Manual Entity Tagging
  - Tag Finders – RSS Feed example of Machine Learning

- Use that Feedback to Improve the Collection and Enrichment Pipeline
Push OSINT Services to the People

On-Line, On-Demand OSINT Services Environment

• Web Services
• End User Environment Integrations
  – I2, Palantir, Thetus, ESRI, Visual Analytics, Inspire, MarkLogic etc.
• Application Access
  – Data validation, data collection, integration
• Federated Search
  – Internal, OSINT, Subscription, PKI etc.
• Browser Plugins
Summary

• We must not miss out on the internet as a source for intelligence

• Analysts must have an interface for discovering valuable content and that content must be tagged and delivered in a manner that supports the knowledge discovery process of the analyst.

• We must start today
Contacts

Booth 205

- Andy Lasko - Andrew.Lasko@KapowSoftware.com
- Brady Balls - Brady.Balls@KapowSoftware.com
- 703.489.1445