A Framework for the Systematic Collection of Open Source Intelligence

Dr. Line C. Pouchard
Computing and Computational Sciences
Oak Ridge National Laboratory

Jonathan D. Dobson
Software Development
Ingenium Professional Services

Joseph P. Trien
Computing and Computational Sciences
Oak Ridge National Laboratory

1. What is Open Source Intelligence?

"Intelligence that is produced from publicly available information, is collected, exploited, disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement." [109th Congress, 2006]

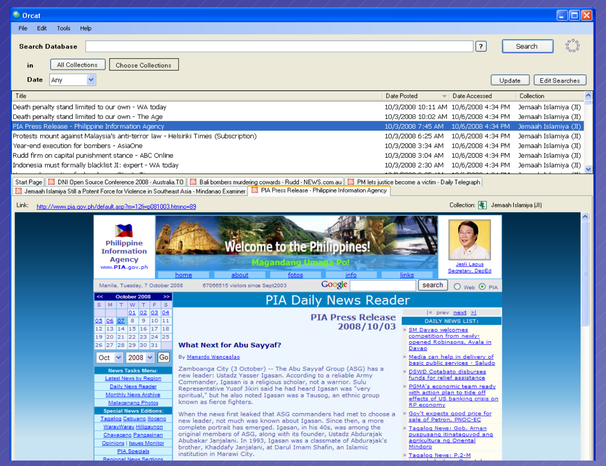
OSINT is content produced by published literature, local and global news coverage, online media, repositories, databases, commercial and non-commercial sources.

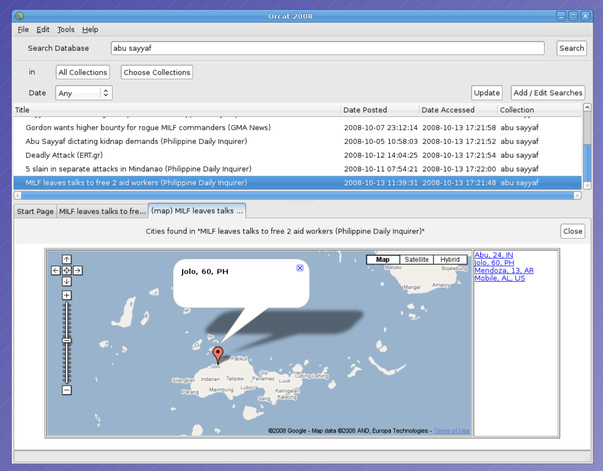
Difficult to use efficiently

2. ORNL Content Analysis Tools (ORCAT)

ORCAT is a lightweight, desktop tool that allows an intelligence analyst to build and organize their own customized collections of Web content.

Inputs come in as RSS feeds, then are processed and stored into the ORCAT database (MySQL).

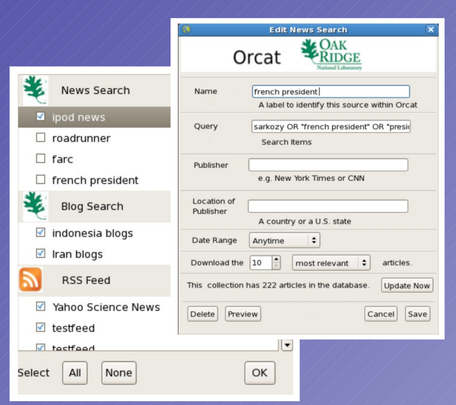




ORCAT can start processing content as soon as it is available. Preferred sources, including intranet or blog URIs can be provided by the user.

Keyword searches on the Web may be performed to build a feed. Searches on subsets from the ORCAT db are also possible.

3. Building Customized Collections with Metadata



A search interface to content providers is integrated into ORCAT.

Data is automatically downloaded, cleaned-up, stored and available for immediate or later use until deleted by the user.

Text is saved with timestamps, source, and title.

ORCAT provides some of the benefits of a news subscription service at the fraction of the cost.

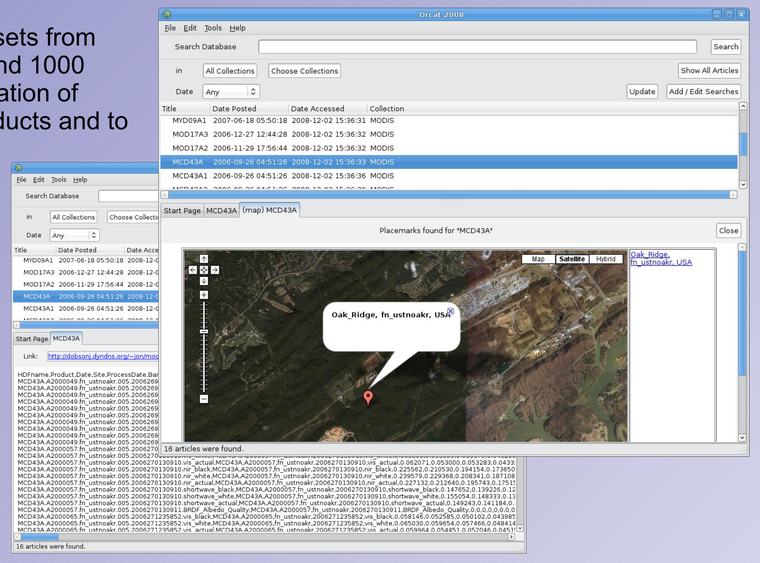
Geo-referencing capability.

4. Ingesting Physical Sensor Data: MODIS

MODIS is a collection of climate datasets from Field sites and Flux towers from around 1000 sites around the world. Used for validation of models and remote-sensing data products and to characterize field sites.

MODIS datasets consists of pixel readings for vegetation indices, surface reflectance, leaf area index, temperature, net photosynthesis, at various granularity level: 250m, 16-day average, 500m, 8-day average, yearly, etc...

(MODIS data courtesy of NASA)



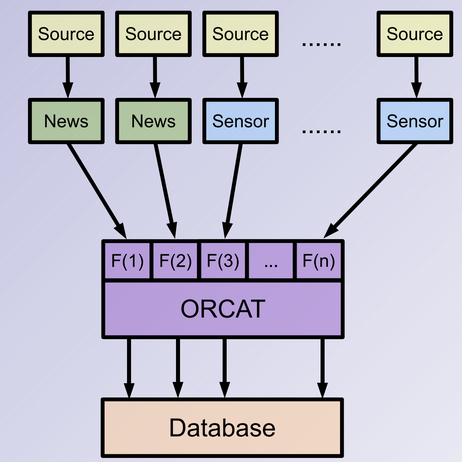
5. Improving Scalability

Target Platform
Large cluster of Symmetric Multi-processing machines.

Structuring patterns
Overall computation involves working with many independent sets of data.

Each feed is downloaded and processed through a loop, which can easily be broken into a set of independent threads.

Concurrency
A collection of feeds can be broken down into a set of independent tasks.
Functions performed on each feed must be done in the same order.



6. Future Work

- Improve the Named-Entity Recognition function.
- Improve the visual display of data.
- Implement query expansion functionality.
- Enable the ability to record personal notes.
- Now supported by the Extreme Scale Systems Center (ESSC).

7. Acknowledgements

The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Special thanks to David Labissoniere who developed an early version of ORCAT while completing an undergraduate degree at East Tennessee State University.