A Framework for the Systematic Collection of Open Source Intelligence

Line C. Pouchard, Jonathan M. Dobson, Joseph P. Trien

Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37934
pouchardlc@ornl.gov, dobsonjd@ornl.gov, trienjp@ornl.gov

Abstract

Following legislative directions, the Intelligence Community has been mandated to make greater use of Open Source Intelligence (OSINT). Efforts are underway to increase the use of OSINT but there are many obstacles. One of these obstacles is the lack of tools helping to manage the volume of available data and ascertain its credibility. We propose a unique system for selecting, collecting and storing Open Source data from the Web and the Open Source Center. Some data management tasks are automated, document source is retained, and metadata containing geographical coordinates are added to the documents. Analysts are thus empowered to search, view, store, and analyze Web data within a single tool. We present ORCAT I and ORCAT II, two implementations of the system.

Introduction

The 109th Congress (2006) defines Open-Source Intelligence (OSINT) as "intelligence that is produced from publicly available information and is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement."[1] Starting with the 9/11 Commission, OSINT, as part of all-source intelligence, has been named as one of the resources that the Intelligence Community (IC) needed to better exploit to succeed in the post-cold war era. It has also been singled out by the Intelligence Reform and Terrorism Prevention Act of 2004 (Public Law 108-458) and the Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction, in its report to the President released on March 31, 2005. Open Source has gained importance, advocacy, access to financial resources, credibility and a number of coordinating and disseminating efforts within the Department of Defense, the Central Intelligence Agency (CIA), the State Department, Department of Homeland Security (DHS) and the Office of the National Intelligence (ODNI). At the September 2008 Open Source Conference, Michael V. Hayden, CIA director, emphasized that "open source intelligence is widely seen as both an essential capability and a formal asset in our national security infrastructure". The DNI's strategic plan states that "no aspect of collection requires greater consideration or holds more promise than open source." Open Source intelligence has always been available, but considered at best a minor asset when compared to classified intelligence, and at worst, a wasteful, unreliable, untrue, or outright deceitful source of information in the IC.

OSINT is the content provided by published literature, news coverage in all media, online content, multi-media repositories, and other public and commercial sources. Some important benefits of Open Source Intelligence exist. OSINT can provide contextual information about larger intelligence subject matters. Local news coverage, which is available world-wide through the Internet as soon as it is published on the Web, is quicker to emerge during real-time crises. Measuring credibility, establishing provenance, and bias discovery context become crucial. Another benefit of OSINT over classified information is providing better access to socio-cultural practices. When compiled with the right tools, heterogeneous media and content that have incompatible data formats can be integrated.

However, working with OSINT is a challenge. The ever increasing volume and diversity of sources can be overwhelming. The amount of available data and the heterogeneous data formats create obstacles to efficient use. According to a 2003 study of the Joint Forces Intelligence Command, seventy percent of an analyst's time is spent in data management tasks (Andes, 2006). OSINT's minor role has resulted in the lack of mature technologies and practices to collect, present, analyze and disseminate open source data. Finally, with its value and volume increasing, the collection and exploitation of OSINT on a large scale has increased concerns about

privacy and the right of the US government to collect information about its citizens, without their knowledge (the latter falls outside the scope of this research).

The purpose of this research is to develop an approach for collecting and analyzing OSINT that harnesses the power of information technology while supporting analysts' judgment calls and interpretation tasks. We developed a method supported by computing tools for identifying and retrieving online information, building, accessing and storing collections tailored to specific intelligence requirements, and for presenting results. This paper focuses on how to make Web news content and content from the Open Source Center more valuable to the Intelligence Community by creating automated, stored, and customized collections searchable across time and enhanced by rich metadata. The Open Source Center (OSC) was created in 2005 by the Office of the Director of National Intelligence in response to the findings of 9/11 Commission that OSINT needed to be better exploited. It is an enterprise Web portal for the Intelligence Community.

Approach and Architecture

We consider OSINT in digital form, including text and multimedia that analysts may access for the purpose of producing Intelligence Reports and Estimates. Web data is often badly formatted for reasons such as late standardization of the HTML specifications and the extreme flexibility of browsers. The content itself includes no measure of accuracy or quality. Data available from OSC undergoes vetting, and tends to contain in-depth and summary material that increases its value. The OSC portal supports keyword searches, saved searches, browsing, and daily email of subscription results. The analyst may download and store single articles at a time. Bulk download of data overnight using ftp is an option. In the case of data transferred using ftp, decompressing, transporting, viewing and loading into an application or a browser require additional manual operations. For both Web data and OSC data, storing articles in a database is needed to ensure persistence of personnel collections of data over time.

Data collection proceeds in a familiar workflow:

- Task 1: a user enters keywords in a search engine.
- Task 2: a list of links is returned.
- Task 3: the user selects a few items of interest.
- Task 4: desired articles are saved into a file one at a time.
- Task 5: articles are stripped from HTML code if processing by an application other than a Web browser is desired.

All except 2 are manual tasks. With OSC data, task 1 may be replaced by automatic search subscription.

Our approach aims at optimizing this process by automating certain tasks while utilizing user input at critical junctions. This approach provides the advantage of reducing tedious manual work while allowing the user to retain control of the articles selected for storage. Because we use Web standards for data models and connectivity in our tools, such as XML, RDF, and RSS, systems built on this approach embed interoperability. Granted the caveat that automatic building of a document collection is only the first step in an analyst's workflow, our pre-processing tools are still useful to prepare and deliver HTML documents for more sophisticated analysis.

Automatically building customized collections of Web data is demonstrated with two implementations. With both implementations, the analyst can work on local copies of the data:

1) One system, called ORCAT I (ORNL Collection Analysis Tool) combines RSS (defined below) and database technologies to support download, storage, and saved queries for Web data, and
2) ORCAT II, the second system, uses batch uploads instead of RSS, Sesame, an RDF repository, and sophisticated queries.

RSS stands for Real Simple Syndication. It is a Web 2.0 protocol by which content providers, for instance media distributors, such as Google News, Yahoo News, or an enterprise network, publish a list of Web links, each referring to a news item. The provided lists are XML-based and contain some metadata. A user can subscribe to this content using an RSS feed reader. Many are available on the Web and may be integrated in a browser. However, RSS feed readers do not typically include the capability of saving a local copy of the content, as ORCAT I does.

In ORCAT I, the user subscribes to RSS feeds through a Web interface. The system downloads and stores the results returned by the content provider in the ORCAT I database. The architecture of ORCAT I is presented in Figure 1. Web data is automatically cleaned up (HTML processing) and compressed (Data Compression). The source of the data, its date of publication and date of access are preserved. ORCAT I stores articles in mySQL, a common database implementation. Users access and view Web content through the Graphical User Interface. A Command Line Interface and an Application Programmer's Interface (API) are also available.

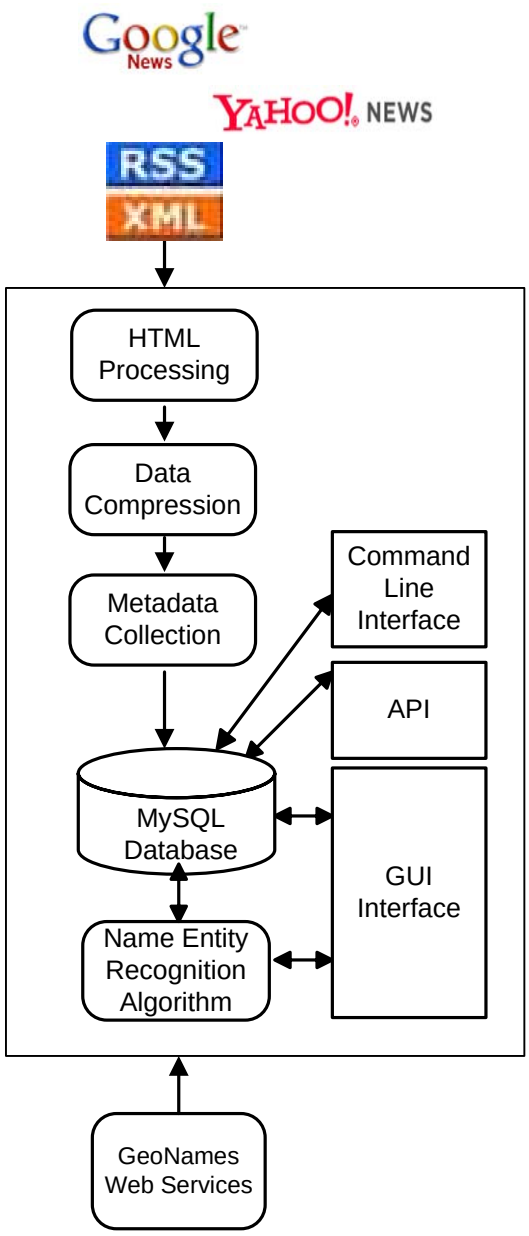

Figure 1: The architecture of ORCAT I.

In order to add geo-referencing capabilities, ORCAT I uses a Named Entity Recognition (NER) component from the Freeling 2.0 framework (Atserias, et al). The choice of Freeling was dictated by the existence of libraries, a GNU GPL license, a modular design, and performance results. Freeling is based on the AdaBoost algorithm. Carreras and Wu, using this algorithm, ranked 8 and 16 respectively in the CoNLL-2003 language independent Named Entity Recognition shared task competition (Sang and Meulder, Carreras et al., Wu et al. 2003). The NER function allows ORCAT I to extract city names from each article. Geographical coordinates are imported by Web services from GeoNames.org. This capability currently allows geo-referencing the names of cities quoted in an article. The ORCAT I interface displays these references as a place mark on a Google map displayed and accessible from within the interface. Under the current implementation of ORCAT.1, the NER pipeline needs improvements.

ORCAT II stores data in an RDF repository. RDF is the Resource Description Framework, a format and Web standard from the World Wide Web Consortium (Beckett, 2004). Built on XML technology, RDF describes data items in a statement of the form (subject-property-object). Using an analogy from the database world, each statement (or triplet) would correspond to a row with only three columns. The entire data set is described by triplets. RDF allows querying data based on every item in the triplet. Queries on the properties of a subject can yield all objects related by the same property to a subject. For instance, a query on the property "visit" of the subject "Osama" would yield every instance that is related to Osama via the "visit" property.

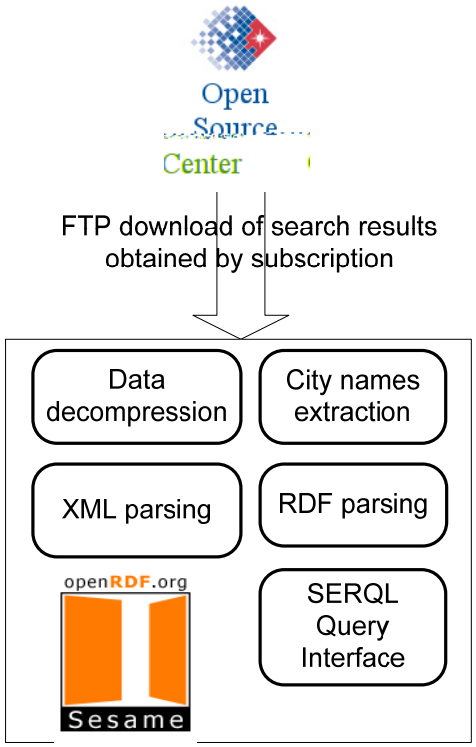

Figure 2: The architecture of ORCAT II

With ORCAT II, OSC data is OSC data is obtained by subscription, processed, and uploaded in an RDF repository (figure 2). OSC data delivered by ftp is already cleaned up and structured using an XML schema and annotated with rich metadata such as Article Location, Topic, Topic Location, etc. The architecture of ORCAT II is shown in Figure 2.

Results

ORCAT I is implemented in the C++ and C# languages. ORCAT I exploits the capabilities of existing search engines, the availability of news media in digitized format, and the content produced daily on the Web by blogging communities. The ORCAT I main window is illustrated in Figure 3.

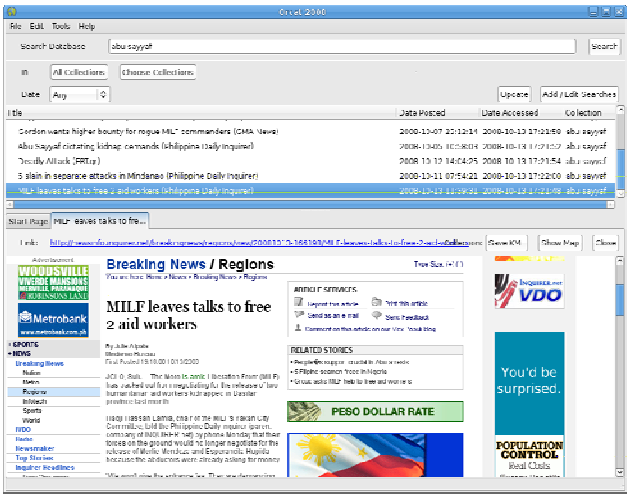

Figure 3: ORCAT displays an article from the Philippine Daily.

ORCAT I can display any content accessible through a Web browser including streaming audio and video files. The browser window at the bottom of the main panel includes links and the ability to find keywords within a page. The top panel displays a search box, a list of content items, the collections they belong to, the dates published and accessed. The source URL and the number of articles in the collection are also displayed. Updating the collection is automatically performed with content available as soon as it is published. Desired content persists in the analyst's database. The use of ORCAT I would automate tasks 3, 4, and 5 outlined in the approach. Task 1 (query input) is executed manually once for each new search. Launching the automatic update may be done manually or triggered by another application.

ORCAT I enables cross-collections searches. The articles stored in the ORCAT I database and resulting from Web-wide news searches can further be manipulated and searched. Queries constrained by source, dates and collection can be performed solely against ORCAT I data. Because search topics change over time and old results can be searched along new ones, the cross-searching function allows manual correlation of data. Access to material no longer existing on the Web may take importance for the in-depth analysis of an event or a person. The persistence of Web data in ORCAT I enables temporal searches so that an analyst can build a picture of the historical context.

Figure 4 shows the ORCAT I window for adding a new search. The user can build a collection based on a

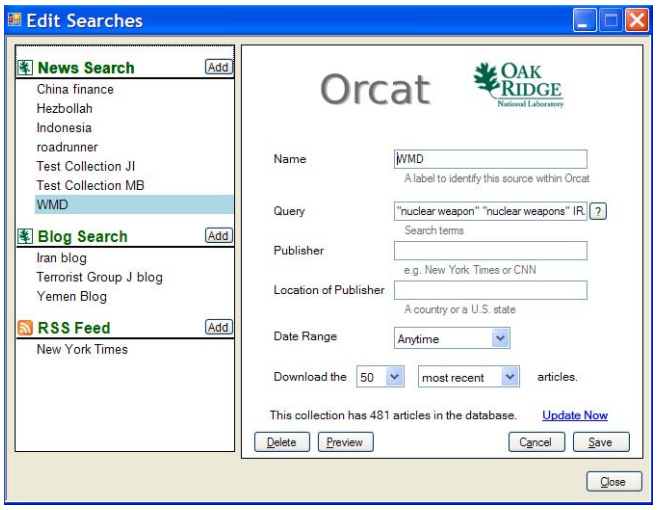

Figure 4: The ORCAT I search interface.

keyword search of a news aggregator distributing content by RSS (e.g. Yahoo news), or build a collection with everything provided by a certain provider (e. g. New York Times). In the case of a keyword search, the query parameters are directly passed from ORCAT I to the news search engine.

Figures 5 and 6 illustrate the functionality of the Show Map button, displaying place marks within the article viewing window. The different levels of map resolution and the toggling between map and satellite view afforded by Google Maps are available within ORCAT I.

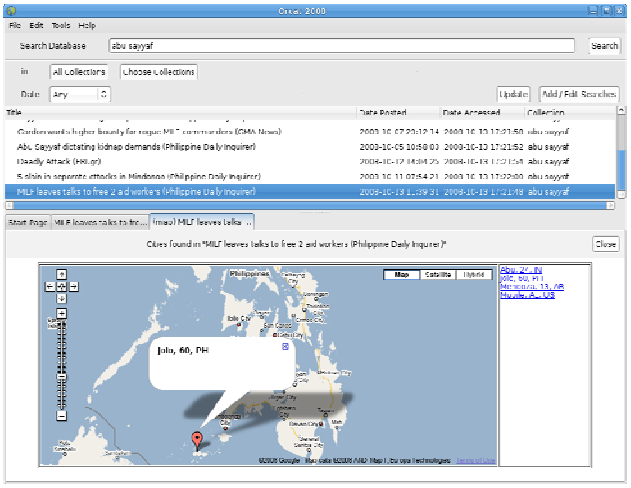

Figure 5: Google Maps displays a city named in the article of interest.

ORCAT II takes advantage of the search subscription and FTP download capabilities of OSC. Based on active subscriptions, a local server receives zipped content from OSC over night. As illustrated in Figure 3, the ORCAT II pipeline processes this content, unzipping it, transforming XML data into RDF data, and loading it into an RDF repository (Sesame). An item of interest (news, blog,

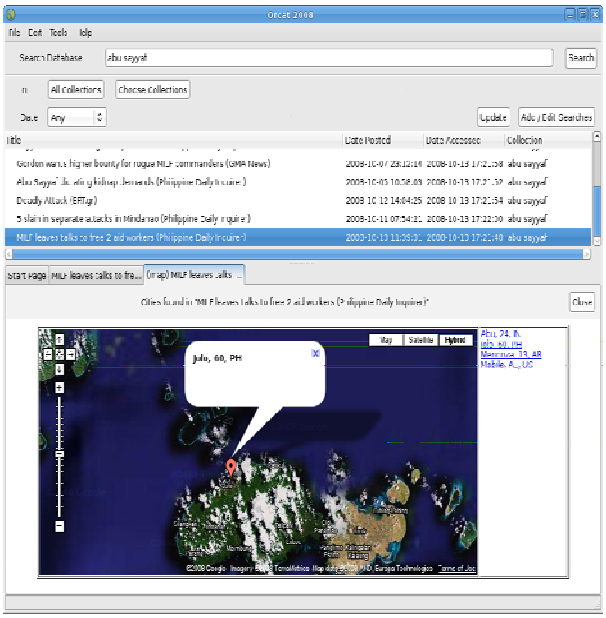

Figure 6: A satellite view of Jolo, Philippines, displayed within ORCAT.

transcripts, etc.) is annotated with information concerning its source, topic, level of classification, and many other OSC annotations. OSC provides several types of location. Source and article topic location by country are distinguished, as a news source reports about events around the world. The ORCAT II architecture adds a component that extracts named-entities from the article text and correlates them with the topic country indicated in the OSC metadata. Geo-coordinates are then uploaded from the GeoNames Web service, similarly as in ORCAT I. Thus, articles in ORCAT II are annotated with OSC metadata, augmented with city location and coordinates, and can be displayed on a Google-type of map.

The SeRQL query language used in conjunction with the RDF repository allows precise control over searches (Broekstra J. and Kampman, A., 2003). SeRQL obtains results that would be difficult to get in a database. For instance, an analyst interested in knowing which cities in Asia suffer from terrorism activities would be able to get a list in one query with ORCAT II. Furthermore, as ORCAT II uses OSC as its source, the results are vetted. By contrast, a Google search would return over twelve million hits for the query "terrorism in Asia". A search in OSC data alone returns 9, 566 items at the time of this writing. With the added technology of ORCAT II the manageability of results is greatly enhanced.

Discussion

The ORCAT I and II architectures are modular and the systems are available with a command-line interface so that more sophisticated analysis tools, such as modeling and profiling tools, may be added in a plug-and play manner. Both use OSINT data in digital form. OSINT is viewed as less costly and less risky than sensitive and human sources and must be collected systematically for added-value (Best and Cumming, 2007). ORCAT I collects raw Web data while ORCAT II collects vetted OSC data.

ORCAT I integrates and displays in a smooth interface the functionality of several "freely" available tools. Content is downloaded from the Web and processed immediately. During crises, the ability to obtain material as soon as it is posted, rather than obtaining vetted but late or limited material may be crucial despite the caveat of quality. ORCAT I and II's local storage capabilities ensure that content items are still available after removal from the Web. Items are displayed in the context of their source so that an analyst may visit the source Web site, for instance, in order to quickly form an opinion about the source.

The geo-referencing function enables discovery of news items by proximity, something that ordinary databases cannot do. For instance, a location of interest can be enlarged to include geographical areas not covered by the keyword search. New articles flagged by a place mark in proximity will be shown, even if the topic is not related.

The reasoning and inference capabilities of ORCAT II are based on those of RDF implemented in the Sesame server. Our implementation is currently rudimentary. In particular, it lacks an easy-to-use interface that would hide the complexity of SERQL queries.

Related Work

Commercial content distributors such as Dialog and Lexis-Nexis operate search subscription services that deliver metadata-rich content to a user's desktop. As they are primarily focused on serving business needs, their costs are extremely high. Articles are delivered for a fee, with a complete citation plus full-text document reaching up to $ 10 per article. This is an estimate since fee schedules are not public and depend on the database, the number of keywords in a search, whether the entire article, a title or a summary are included. Most services deliver content in bulk overnight, and thus are not up-to-date on breaking news. In addition, the data management system needed to efficiently retain and process this data is often delivered by third party software at an additional cost.

An approach for developing a relational database for storing information regarding extremist activities and international terrorism has been described by Hale (2006). Metadata regarding specific details about the event, geographical location, individual involvement and other variables were collected. Hale focuses on a specific category of events and not on general Web data. Although geographical location is mentioned, there is no attempt at mapping documents.

Many articles propose methods for collecting OSINT, analyzing it and successfully utilizing it, but they focus on the process (Thibault, G. Gareau, M. and Le May F. 2007), the value (Steele, 2006), or the uses of OSINT (Bass, 2006), and not technologies that would make it easier to use.

Conclusion and Future Work

We presented an approach to building customizable desktop collections of documents from two very different sources, the general Web and the ODNI Open Source Center, and two implementations. These lightweight desktop tools combine many desirable but commonplace technologies in a useful package. ORCAT I implements automatic storing of real-time Web data in a local database, cross-searching, geo-referencing, and display of locations on Google Maps. ORCAT I presents a smooth, easy-to-use Graphical User Interface where analysts can perform all information retrieval, storage, and search-related tasks within a single tool. ORCAT II stores and processes vetted OSC data, adds topic city and geographical coordinates, and uses the SeRQL query language to resolve open-ended queries.

Future work includes improvements to the Named-Entity Recognition function and the visual display of information. Functionality such as query expansion and the ability to record personal notes may also be added.

References

109th Congress National Defense Authorization Act For Fiscal Year 2006, Public Law 109-163, Section 931.

Andes, James. Oak Ridge National Laboratory. Conversation with the author (November 2006).

Asterias, J., Casas, B., Comelles, E., González M., Padró L. and M. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, ELRA. Genoa, Italy. May, 2006.

Bass, S. The Challenges of Information Management in the Networked Battlespace: Unmanned Aircraft Systems, Raw Data and the Warfighter. Master' s thesis, June 2006. Air Force Institute of Technology Report A389354.

Beckett, D. et al. 2004. Resource Description Framework (RDF). World Wide Web Consortium.

Best, R. and Cumming, A. Open Source Intelligence (OSINT): Issues for Congress. Congressional Research Service RL 34270, December 5, 2007.

Broekstra, J.; Kampman, A.; 2003. SeRQL: A Second generation RDF Query Language. In *Proceedings of the SWAD Europe Workshop on Semantic Web Storage and Retrieval*, Vrije Universitet, Amsterdam, NL, 13014 November 2003.

Carreras X., Màrquez, L. and Padró, L. A simple named entity extractor using AdaBoost. In *Proceedings of the seventh Conference on Natural Language Learning at HLT-NAACL 2003, Volume 4*, 152 – 155. Edmonton, Canada.

Hale, W. 2006. Information versus intelligence: construction and analysis of an open source relational database of worldwide extremist activity. International *Journal of Emergency Management* 3:280-297.

Open Source Center (OSC) is available on the Web at http://www.opensource.gov.

Steele, R. Information Operations: Putting the 'I' Back Into Dime. Storming Media Pentagon Reports A046444.

Thibault, G. Gareau, M. and Le May F. Intelligence collation in asymmetric conflict: A Canadian armed forces perspective. *Proceedings of the 10th International Conference on Information Fusion*: 1-8.

Tjong Kim Sang, E.; De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the seventh Conference on Natural Language Learning at HLT-NAACL 2003, Volume 4*, 142 – 147. Edmonton, Canada.

Wu, D. Ngai, G, and Carpuat, M. A stacked, voted, stacked model for named entity recognition. In *Proceedings of the seventh Conference on Natural Language Learning at HLT-NAACL 2003, Volume 4*, 202-203. Edmonton, Canada.

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.