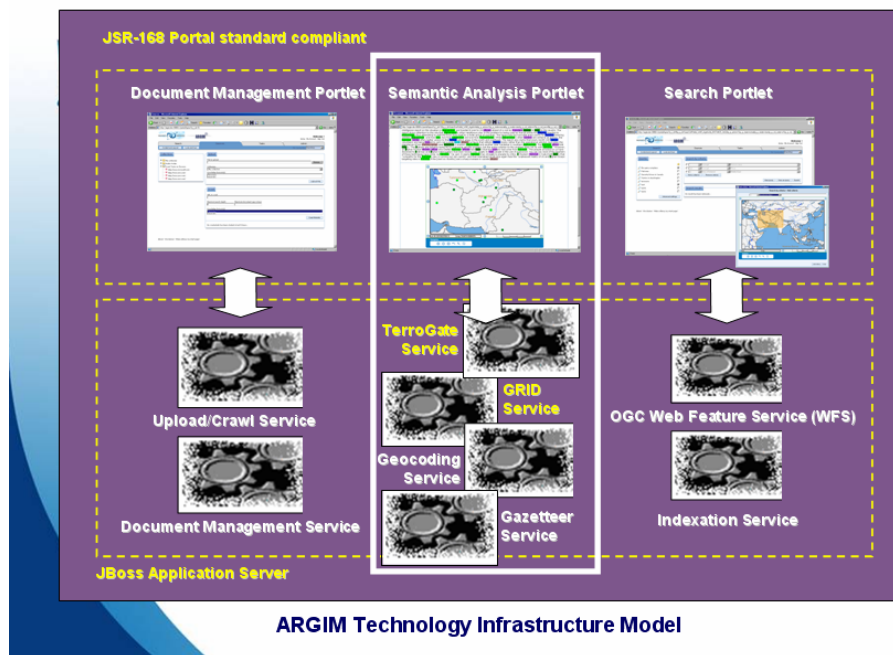


**Marc-André Morin**  
System Architect  
Defence R&D Canada

### Expression of Interest in the topics of the Meeting

Defence R&D Canada (DRDC) is an agency of the Canadian Department of National Defence. Its mission is to improve Canada's defence capabilities, through research and development. My R&D area of expertise is the Knowledge and Information Management, specifically Geographic Information Systems (GIS). I am involved in a major applied research project: ARGIM, a GEOINT solution integrating knowledge exploitation and geospatial technologies.

ARGIM stands for Applied Research for Geospatial Information Management. The Project Leader is Dr Alain Auger, a Defence Scientist with expertise in computational linguistics. The ARGIM project leverages on both natural language processing and GIS expertise in order to develop a new GEOINT framework, or a wide toolbox, based on free and open source software to rapidly integrate, deploy and evaluate new Information & Knowledge Management concepts and technologies.



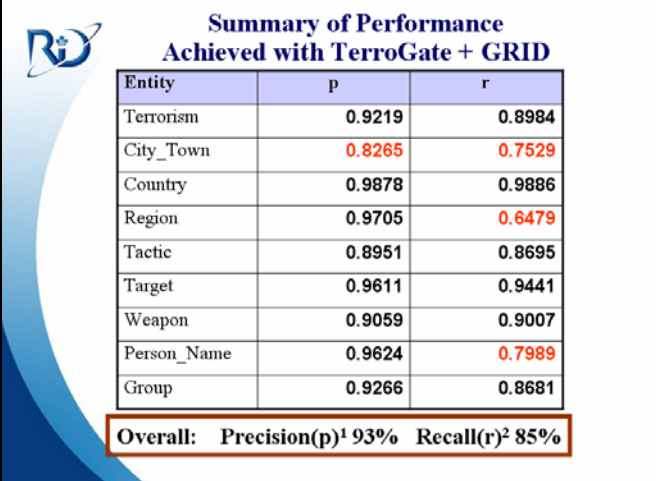
Designed and implemented by our team, the current prototype relies on Service Oriented Architecture (SOA). The portal includes:

1. Document Management Service for grabbing, structuring and sharing sources of information;
2. *Natural Language Processing (NLP)* services for semantic-based text search and analysis, including automatic annotation of geographically-related entities in unstructured documents;
3. GIS capabilities for indexation, retrieval, visualization and analysis of spatio-temporal information. Notice that specifications from the Open Geospatial Consortium (OGC) are taken into considerations.

NLP services include:

1. TerroGate service, a new information retrieval system dedicated to the terrorism domain recognition (tactics, weapons, targets, groups, etc.);
2. GRID service, a geoparser for geographic pattern-based recognition;
3. Geocoder for assigning geometries to representative geographic location terms contained in texts;
4. Finally, gazetteers for feeding all services described previously in their specific activities.

TerroGate and GRID have both been developed by DRDC Valcartier. They are both named entities extraction technology based on the free and open source software called GATE (<http://gate.ac.uk/>), a human language processing system to develop linguistic-based technologies. In fact, the generic “Named Entities Extractor” module of GATE has been modified, specialized and trained in order to increase its performance over electronic texts related to the terrorism and the geospatial domains.



The image shows a slide titled "Summary of Performance Achieved with TerroGate + GRID". It features a table with columns for Entity, p (Precision), and r (Recall). The overall performance is summarized as Precision(p)<sup>1</sup> 93% and Recall(r)<sup>2</sup> 85%.

Entity	p	r
Terrorism	0.9219	0.8984
City_Town	0.8265	0.7529
Country	0.9878	0.9886
Region	0.9705	0.6479
Tactic	0.8951	0.8695
Target	0.9611	0.9441
Weapon	0.9059	0.9007
Person_Name	0.9624	0.7989
Group	0.9266	0.8681
<b>Overall: Precision(p)<sup>1</sup> 93% Recall(r)<sup>2</sup> 85%</b>		

As shown in the figure above, poorest results in terms of precision<sup>1</sup> and recall<sup>2</sup> are mostly related to geographic named entities. Precision for extraction of cities and towns within text documents is relatively low essentially because of several partial matches. For example, “Washington” could be tagged where it should have been “Washington D.C.” Therefore, precision could be increased by improving pattern matching rules, grammar and algorithms. However, to increase recall relative to cities, towns and regions – where a lower percentage is linked to many locations or place names that are missing in our gazetteer – we have to focus on richer gazetteers and domain ontologies and that will help us to complete gaps within our in-house gazetteer. To achieve that goal, the Alexandria Project is probably the best source of information. Thus, the experimentation of the ADL Feature Type Thesaurus and the ADL Gazetteer are probably a good start to improve and redesign our geospatial knowledge domain.

Proper identification and georeferencing of information with a geographical context is a big challenge, and digital gazetteers can be part of the solution, and maybe, part of the problem... Loading the whole gazetteer in memory to conduct natural language analysis is unfeasible and useless. To do so, we must categorize location-related pattern matching rules and gazetteer’s features according to a scale of visibility. For instance, it is totally inadequate to process a grammar rule responsible for detecting zip/postal codes and load in memory all street names related to a city when a document is related to world wide topics, such as earthquakes. Thus, our geoparsing application needs to support and integrate more than one gazetteer, where each one will have its own specificity. It is obvious that digital gazetteers play a key role within our infrastructure, not only for georeferencing place names but also to feed our parsing applications like TerroGate and GRID.

Finally, geosemantic ambiguities are a real dilemma for georeferencing or geocoding place names. If we do a search for “Paris” by using the ADL Gazetteer Client, we get 57 matches. How can the machine choose the right one? Also, how can we georeference assertions such as: “20 miles north of Santa Barbara”? This is what ARGIM and also The Alexandria Textual Geospatial Integration project (TGI) try to address. It will be very interesting to put this subject on the table for an in-depth discussion in order to identify opportunities for collaborative efforts.

<sup>1</sup> **Precision** ratio is the proportion relevant named entities among the retrieved ones

<sup>2</sup> **Recall** ratio is the proportion of retrieved named entities among all relevant ones