

**Response to**

**National Geospatial Intelligence Agency**

**Disparate Data Challenge**

## Administrative Information

|  |  |
| --- | --- |
| **Company** | Envitia Inc. |
| **Address** | 11710 Plaza America Drive, Suite 2000, Reston, VA, 20190 |
| **Point of Contact** | Robin Parrish |
| **Phone Number** | 703-597-6261 |
| **Fax Number** | 703-871-5111 |
| **Email Address** | robin.parrish@envitia.com |
| **Web Page URL** | [www.envitia.com](http://www.envitia.com) |
| **Business Type** | Small Business |

## Background

Envitia Inc. is a wholly-owned subsidiary of Envitia Group PLC, employing US Nationals promoting Envitia Group’s software products and undertaking product and solutions development in the USA. Envitia Inc. undertook the development of the proof of concept described herein and will undertake any further work if this proof of concept is progressed with the support of NGA.

Envitia is a well-respected SME, selling software for over 15 years in the US. Envitia is well-known to NGA through work on Open Standards (OGC) and through collaboration with projects undertaken by Envitia UK for the UK Ministry of Defence and NATO. Envitia has considerable experience in data modelling, metadata, geospatial catalogues and registries and the deployment of web-based command and control/intelligence applications.

## Problem Summary

There are considerable issues in exploiting open data. Aside from licensing issues, there are general issues of effective discovery, characterization, and preparation for exploitation. Resources on a given website may need to be accessed via diverse interfaces and a mixture of web service endpoints and downloads. A plethora of web service types and file formats exist with little metadata (formal or informal) provided.

Consequently, there is considerable manual activity and data duplication needed to exploit data from a given site, for example, downloading large datasets which end up not being useful in the chosen context.

## Envitia’s Solution

Envitia’s approach to this challenge is to exploit a mixture of open source and Envitia COTS components to mediate data into an Open Standards-compliant Catalogue Service endpoint (compliant with CSW-ISO and CSW-ebRIM) and GeoPortal which provide metadata on resources within a selected website. This enables the first key element of exploitation, discovery. The architecture proposed also supports progression to full analysis exploitation. Envitia provides a searchable GeoRegistry with both constituent data resources, as well as a registered web processing services. By integrating this sophisticated data model into the GeoPortal, the web client becomes a single point of presence for all available data and analytics.

The architecture uses a two stage process:

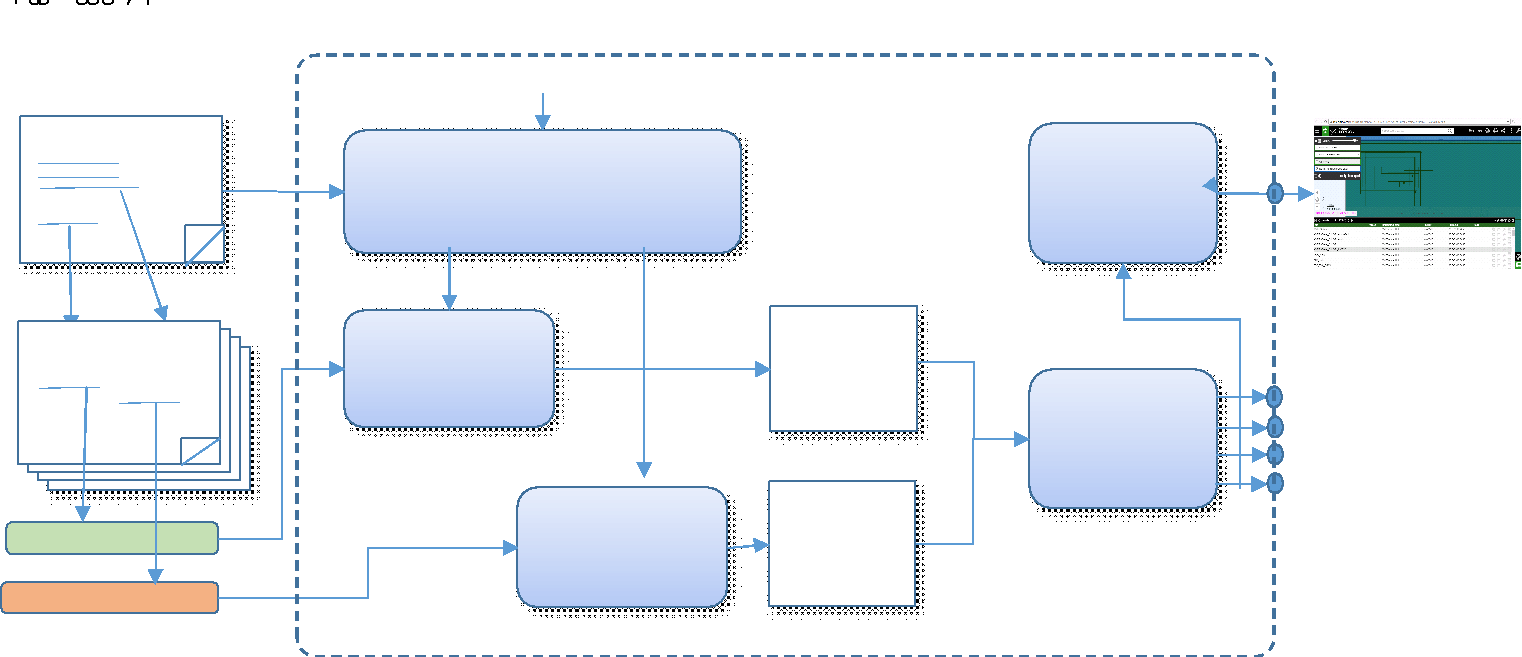
* Given a website, the first process is to identify harvestable resources. To do this, Envitia deployed a scanner which, given a base URL, will analyze all resources on the site (sub-pages, links, etc.).
* The scanner, having identified a resource (text block, URL, etc.) will analyze this using a range of potential processors, qualified by the resource discovered, e.g. a web service URL, a download link, etc. Appropriate analysis tools are then executed to fully characterize the resource.

The characterization of the resource uses ISO 19115 metadata, creating a pseudo-dataset object then a set of services (one or more) or download links which deliver the dataset.

For the purposes of demonstrating the concept, the overall framework has been deployed together with two specific processors.

* The OGC Web Service Processor takes a URL and executes a standard OGC request on the URL to obtain an OGC Capabilities document. At present the processor looks specifically for OGC WMS/WMTS, but it could evaluate any URL (covering cases where only the base URL is present) or OGC web service. If the processor receives a valid response, it will harvest this for metadata and add it to the catalogue service. Adding this immediately means the service can be discovered in the catalogue and portal (all bounding extents and other key metadata are shown). Access through the GeoPortal then allows GIS interactivity, including overlaying with other layers and querying.
* The geo-file processor is invoked when a URL is found which would initiate a download. Clearly it is undesirable to download all content in order to characterize it. To avoid this, the processor exploits virtual file querying (presently for a select number of file types, GeoTIFF and NITF) to request only the relevant geospatial metadata be transferred over the network. Suitable metadata is included in the catalogue, and again a pseudo data-set metadata item is created, this time with a download link referencing the full download.

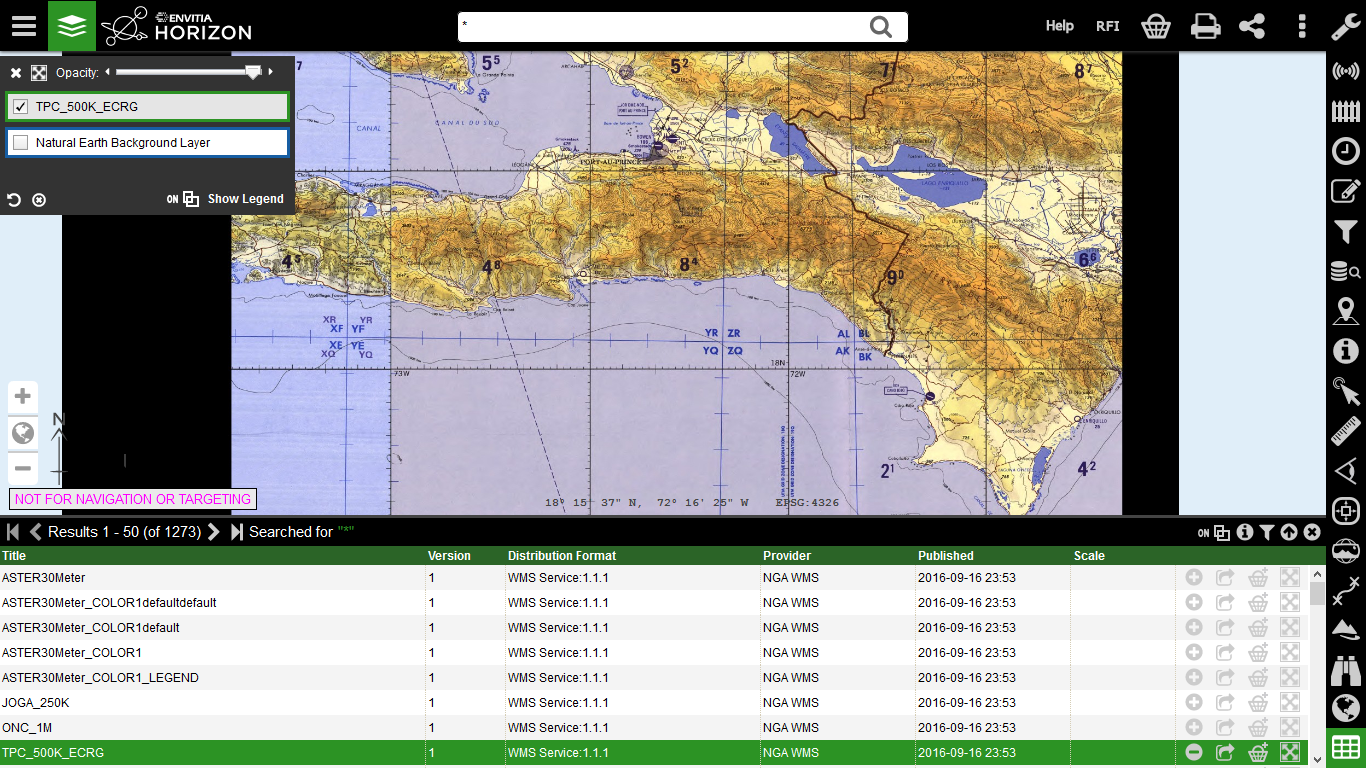
The framework and these two analysis processors have been deployed in an Amazon cloud together with the Envitia GeoRegistry and Envitia Horizon GeoPortal. The overall configuration has been deployed in an Amazon EC2 Environment (Figure 1).

.

**Figure 1 – Deployed Data Discovery and Exploitation Proof of Concept**

At present the process to invoke the analysis of a website is performed by a command line operation (taking the base URL of the website to exploit) on the server deployed on the cloud instance. Envitia Inc. is currently working to wrap the functionality of the overall scanner in a web service as a generic restful service, as a SOAP request or a OGC compliant WPS request, to allow NGA to experiment with scanning a range of other websites. However, we believe this demonstration shows the type of result available from the process, albeit in canonical form.

Having deployed this experimental configuration and run the scanner across the data links provided for the challenge, 1500 web map services and flat files were detected and catalogued. Figure 2 shows these represented in the Envitia Horizon GeoPortal.



**Figure 2 – Envitia GeoPortal showing a WMS visualization harvested from the sample data links.**

They can be searched for and their metadata queried; in the case of services they can be displayed and in the case of datasets they can be downloaded.

**Further Work**

This infrastructure offers significant opportunities for expansion:

* Allowing for example better characterization by specifically analyzing the text around a URL, to populate the metadata descriptions or keywords.
* Allowing automated execution of schema mapping based on the metadata characterization to allow the content to be standardized.
* Assembly of collections of information to be gathered and exploited. The infrastructure does integrate easily with the Web Object Server (WOS) development being undertaken by Envitia in the OGC Testbed 12 which is specifically sponsored by NGA. This tool would provide an excellent source of information to validate the WOS Concept.

Envitia looks forward to NGA’s response to the ideas presented in this paper.

## Appendix A to Envitia’s Disparate Data Submission

Visit <http://iot.envitia.com:9080/EnvitiaHorizon/viewer> using either Chrome of Firefox.

Please review the following video for a usage demonstration.

<https://www.youtube.com/watch?v=FAQ_nQnYJ-c>