WCGA RDF Data Registry Design Assessment

Report by Ecotrust to the WCGA Regional Data Framework Action Coordination Team

In partial fulfillment of NOAA Award # NA12NOS4730004

Authors:
Tim Welch
Ryan Hodges
Tanya Haddad

Jan. 17th, 2013

For questions or comments, please contact Tim Welch, Senior Developer, at Ecotrust, 721 NW 9th Avenue, Suite 200 Portland, OR 97209; twelch@ecotrust.org, 503-467-0786
TABLE OF CONTENTS

1. THE CHALLENGE ............................................................................................................................................. 2

2. INITIAL ASSESSMENT ..................................................................................................................................... 2
   2.1. Assessment of WCGA ACT Priority Resources ......................................................................................... 2
   2.2. Identify Expertise and Educational Resources ............................................................................................ 3
   2.3. Identify Resource Providers .......................................................................................................................... 5
   2.4. Data Registry Requirements .......................................................................................................................... 6
   2.5. Define Data Registry Requirements ............................................................................................................... 6
   2.6. Identify Additional Recommendations .......................................................................................................... 7
   2.7. Define Publishing Workflow ......................................................................................................................... 8

3. DATA CATALOG ASSESSMENT ......................................................................................................................... 9
   3.1. Identify Potential Catalog Products ............................................................................................................. 9
   3.2. Catalog Comparison ....................................................................................................................................... 14
   3.3. Integration Testing .......................................................................................................................................... 16
   3.4. Final Conclusions ......................................................................................................................................... 18


LIST OF TABLES

Table 1 - Domain experts engaged for this assessment ......................................................................................... 3
Table 2 - Metadata and catalog service educational resources ............................................................................. 4
Table 3 - Glossary ................................................................................................................................................... 4
Table 4 - Examples of Relevant West Coast Data Catalogs .................................................................................... 5
Table 5 - Priority resource providers not publishing metadata ........................................................................... 5
Table 6 - Data Catalog Comparison Matrix ........................................................................................................... 16
Table 7 - Test Results ............................................................................................................................................... 17

LIST OF FIGURES

Figure 1 - WCGA priority resource list sample ..................................................................................................... 3
Figure 2 - WCGA Data Registry workflow diagram ............................................................................................... 9
1. THE CHALLENGE

The stated purpose of the WCGA Data Registry is:

"Development of a prototype West Coast Data Catalog to build the human and technological components of the WCGA Regional Data Framework to support and inform regional ecosystem-based management, coastal marine spatial planning, and regional ocean management issues as described by the WCGA"

The challenge to this is figuring out how to coordinate and combine select regional information gathering efforts in an efficient and coherent way. The goals of this data registry project can be summarized as the following:

- Combine WCGA priority resources from different sources across the West Coast and share them with many users and applications; where priority resources are defined to be information products and services, both spatial and non-spatial, that are seen as critical to meeting the needs of the WCGA.
- Collect metadata once and keep it where it can be maintained most effectively.
- Make it easy to find resources, understand how they can be used to meet a particular need, and how they can be acquired and used.
- Provide access to these resources in a way that is conducive to discovery.
- Encourage use of current metadata standards and publishing solutions.
- Overall, support good planning and governance at multiple levels and scales.

What follows is an assessment of the functional and technical requirements for the registry, a feasible workflow for aggregating and publishing resources from regional providers through the data registry, and finally thoroughly researched options for the overall architectural design of the data registry. Each step was guided by the decisions made previously, beginning with the stated purpose and goals above.

2. INITIAL ASSESSMENT

2.1. Assessment of WCGA ACT Priority Resources

In Task 1 of this project, Ecotrust began by engaging the 10 WCGA Action Coordination Teams as well as core members of the RDF Data Working Group to identify an initial list of resources that should be published through the registry. This list is currently being amended, prioritized, and categorized by the RDF Data Working Group through further discussions with stakeholders and will be finalized in early 2013.
2.2. Identify Expertise and Educational Resources

During this assessment, a number of domain experts have provided insight, expertise, and assistance:

<table>
<thead>
<tr>
<th>Name</th>
<th>Expertise/Domain</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tanya Haddad</td>
<td>WCGA RDF member, Oregon Marine Data Network, Metadata standards</td>
<td>Data gathering and communication with West Coast groups, creation of sample set of ISO metadata records for testing, troubleshooting GeoNetwork and Geoportal</td>
</tr>
<tr>
<td>Robby Wilson</td>
<td>NOAA CMSP Data Registry</td>
<td>Experience developing the NOAA CMSP data registry</td>
</tr>
<tr>
<td>Marten Hogeweg</td>
<td>Geoportal lead developer, ESRI</td>
<td>Assistance configuring and testing ESRI Geoportal</td>
</tr>
<tr>
<td>Scott Toews</td>
<td>California Geoportal</td>
<td>Information on California Geoportal and harvesting strategies with RDF Registry</td>
</tr>
<tr>
<td>Marc Rempel</td>
<td>Geoportal - Oregon Spatial Data Library (OSDL)</td>
<td>Assistance configuring ESRI Geoportal</td>
</tr>
<tr>
<td>Douglas Nebert</td>
<td>Data.gov technical lead. Senior Advisor for Geospatial Technology, System-of-Systems Architect, FGDC Secretariat</td>
<td>Insight into the GSA’s redesign of data.gov to use CKAN. Details on federal transition to ISO and support for legacy FGDC</td>
</tr>
<tr>
<td>Tom Kralidis</td>
<td>GeoNetwork and CKAN core developer, creator of OWSLib and PyCSW</td>
<td>Assistance testing CKAN and harvesting from GeoNetwork. Insight into PyCSW development</td>
</tr>
<tr>
<td>Adria Mercader</td>
<td>CKAN core developer with Open Knowledge Foundation, UK</td>
<td>Assistance configuring and testing CKAN, insight into CKAN enhancements for data.gov re-release</td>
</tr>
<tr>
<td>Craig Risien</td>
<td>NANOOS data management and user products</td>
<td>Discussion on what ocean observation products would be useful to publish</td>
</tr>
<tr>
<td>Brent Wood</td>
<td>GeoNetwork core developer</td>
<td>Assistance configuring and testing GeoNetwork</td>
</tr>
<tr>
<td>Bruce Crevensten</td>
<td>GeoNetwork and CKAN testing for</td>
<td>Insight into potential to create a hybrid</td>
</tr>
</tbody>
</table>

Figure 1 - WCGA priority resource list sample
A few high-level educational resources were identified that were useful for understanding some of the subtleties of FGDC/ISO metadata and metadata catalog services. See the documentation and manuals for each respective software project to learn more about their implementation of these standards.

Table 2 - Metadata and catalog service educational resources

<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalog Services (CSW) at a glance, with example requests</td>
<td><a href="http://dev.geonode.org/trac/wiki/CSW_at_a_glance">http://dev.geonode.org/trac/wiki/CSW_at_a_glance</a></td>
</tr>
<tr>
<td>Querying CSW servers with Python</td>
<td><a href="http://geopython.github.com/OWSLib/#csw">http://geopython.github.com/OWSLib/#csw</a></td>
</tr>
</tbody>
</table>

Table 3 - Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Registry</td>
<td>A website that aggregates and republishes multiple external datasets for a specific purpose (e.g. WCGA Data Registry, NOAA CMSP Data Registry). Sometimes referred to as a data portal, data clearinghouse, or data catalog.</td>
</tr>
<tr>
<td>Data Catalog</td>
<td>A data catalog is generally speaking a repository for datasets, hosting them directly. It may offer metadata features and the ability to link to datasets rather than storing them directly. In this way a data catalog can also be a metadata catalog.</td>
</tr>
<tr>
<td>Metadata Catalog</td>
<td>Database that makes data discoverable by publishing metadata records which describe and provide links to acquire the data in different forms. The actual dataset are stored elsewhere.</td>
</tr>
<tr>
<td>Metadata Profile</td>
<td>A profile shortens or extends a metadata standard and changes conditionality (e.g. ISO North American Profile of the ISO 19115 standard)</td>
</tr>
<tr>
<td>Metadata Extension</td>
<td>An extension adds to a metadata standard in order to document additional information (e.g. Remote Sensing Extension of the FGDC CSDGM standard)</td>
</tr>
<tr>
<td>ISO 19139:2007</td>
<td>2007 version of the XML metadata schema for geographic information published by the ISO. Learn more in the NOAA courses</td>
</tr>
<tr>
<td>JSON</td>
<td>Javascript Object Notation - simple schema for structured objects. Commonly used for data between a browser client and server in a web application</td>
</tr>
<tr>
<td>REST or RESTful API</td>
<td>Use of this term in this document does not refer to the ESRI REST API for ArcGIS Server. REST is not a standard, it's a set of principles or design patterns for creating a web interface. Another term for this is RESTful (<a href="http://tomayko.com/writings/rest-to-my-wife">http://tomayko.com/writings/rest-to-my-wife</a>)</td>
</tr>
<tr>
<td>CSW</td>
<td>CSW is an OGC standard for accessing a catalog of geospatial records over the web. (<a href="http://en.wikipedia.org/wiki/Catalog_Service_for_the_Web">http://en.wikipedia.org/wiki/Catalog_Service_for_the_Web</a>)</td>
</tr>
<tr>
<td>Web accessible folder (WAF)</td>
<td>A folder on a web server that is accessible to the public</td>
</tr>
</tbody>
</table>
2.3. Identify Resource Providers

Metadata for priority WCGA resources are expected to be, and in some cases already are, published online. The WCGA Data Registry must be able to harvest a subset of metadata records from these repositories.

<table>
<thead>
<tr>
<th>Catalog</th>
<th>Contact</th>
<th>Catalog Product</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data.gov</td>
<td>Doug Nebert</td>
<td>CKAN</td>
<td>CKAN-based release early 2013</td>
</tr>
<tr>
<td>Oregon Coastal Atlas</td>
<td>Tanya Haddad, Oregon Coastal Management program</td>
<td>GeoNetwork</td>
<td>Under development - <a href="http://www.coastalatlas.net/geonetwork">http://www.coastalatlas.net/geonetwork</a></td>
</tr>
<tr>
<td>NOAA National Geophysical Data Center</td>
<td>N/A</td>
<td>ESRI Geoportal</td>
<td><a href="http://www.ngdc.noaa.gov/geoportal/catalog/search/browse.page">http://www.ngdc.noaa.gov/geoportal/catalog/search/browse.page</a></td>
</tr>
<tr>
<td>USGS</td>
<td>N/A</td>
<td>Mercury</td>
<td><a href="http://mercury.ornl.gov/clearinghouse">http://mercury.ornl.gov/clearinghouse</a></td>
</tr>
</tbody>
</table>

Many of the resources that are a priority to the WCGA are not currently published through an online metadata repository of some kind. Assistance will be provided as possible to help make this happen.

<table>
<thead>
<tr>
<th>Provider</th>
<th>Contact</th>
<th>Resources</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oregon Nearshore Ecological Data Atlas</td>
<td>Jon Bowers</td>
<td>Nearshore ecological data products</td>
<td>TBD</td>
</tr>
<tr>
<td>Pacific Coast Ocean Observing System (PACOOS)</td>
<td>Chris Romsos - OSU Goldfinger Lab</td>
<td>Ocean observation, datasets, services, and products</td>
<td><a href="http://pacoos.org/Datamgt.htm">http://pacoos.org/Datamgt.htm</a></td>
</tr>
<tr>
<td>Central and Northern California Ocean Observing System (CeNCOOS)</td>
<td>Fred Bahr, Jen Patterson</td>
<td>Ocean observation tools, datasets, services, and products</td>
<td><a href="http://www.cencoos.org/sections/products/">http://www.cencoos.org/sections/products/</a></td>
</tr>
<tr>
<td>Southern California Coastal Ocean Observing System (SCCOOS)</td>
<td>Darren Wright</td>
<td>Ocean observation tools, datasets, services, and products</td>
<td><a href="http://www.sccoos.org/interactive-map/">http://www.sccoos.org/interactive-map/</a></td>
</tr>
<tr>
<td>Ecotrust</td>
<td>Charles Steinback</td>
<td>Consumptive and non-consumptive human use</td>
<td><a href="http://www.ecotrust.org/monitoring/">http://www.ecotrust.org/monitoring/</a></td>
</tr>
</tbody>
</table>
2.4. Data Registry Requirements

Metadata standards, data catalogs, and catalog services are beginning to mature and converge. Rather than designing a data registry that is all encompassing in its metadata support and features, the WCGA RDF ACT sees an opportunity to develop a data registry that is focused on providing quick aggregation, search and access to priority resources using current metadata and catalog standards. The willingness of the RDF ACT to provide assistance and funding where possible to resource providers to make this happen is seen as a clear requirement.

2.5. Define Data Registry Requirements

In order to assess technology options and narrow down to a feasible data registry design we’ll now make some declarations on what the WCGA data registry will, may, and will not provide during Task 1-3 of this project.

The WCGA Data Registry will:

1. Use out-of-the-box data catalog software to provide majority of registry features
2. Provide administrative access using SSL and password-based authentication
3. Use harvesting as the primary strategy for loading resources into the catalog
4. Allow new metadata to undergo a review and approval process before being made publicly available
5. Support harvest of metadata using the Dublin Core schema for non-spatial resources
6. Support harvest of metadata using ISO 19139/19115/19119 schemas for spatial resources
   a. Recommendation - require this for providers that offer web service links or one or more alternative links
   b. Recommendation - recommend this for resource providers that are creating new metadata.
7. Support harvest of metadata using the FGDC CSDGM schema for legacy spatial resources that do not need to link to web services.
   a. Recommendation - allow for resource providers that already have FGDC metadata and do not need to publish web service links.
8. Support harvest of metadata from partner catalogs using CSW 2.0.2 and the APISO profile
   a. Recommendation - CSW is a standardized catalog service specification and current resource providers’ data catalogs already support it
9. Allow administrators to associate controlled vocabulary/taxonomy terms (e.g. category, sub-category, theme, sub-theme) being created by the RDF Data Working Group during Task 1 of this project.
a. Recommendation - these attributes should be assigned after harvest, and not to the metadata document itself. The reason is this is WCGA-specific information that the resource provider should not have to be responsible for maintaining.

b. Recommendation - look for ways to incorporate the vocabulary/tags into faceted search capabilities to provide faster and more relevant search results to users. Allow users to search available data records across all of the metadata attributes (title, description, keywords, etc.)

10. Be usable on modern tablet computers

The WCGA Data Registry **may**:

11. Support harvest of web-accessible folders (WAF’s) as a simple way for resource providers to publish metadata
   a. Recommendation - This is a simple way to publish metadata but it has known shortcomings due to lack of universal identifiers for each record

12. Support harvesting ISO 19139 metadata documents from GIS servers using alternate non-standardized methods
   a. ArcServer using the REST API

13. Provide users with a preview image of spatial datasets or web services
14. Allow users to share resources with others via email and social media

The WCGA Data Registry **will not**:

15. Incorporate or use any formal ontologies to improve searching
   a. Recommendation - this could be a future possibility
16. Support non-standardized metadata records
17. Provide account or group-based access to private data layers. Assumed that all published resources are for public access
18. Store datasets directly, only provide access through linking in metadata documents.
   a. Recommendation - consider not trying to do too much with the data registry. Data and web service hosting is something that can be done separately.
19. Support ESRI XML or EML metadata records.
   b. Recommendation - if priority resource providers are using it and there’s future strong demand then EML profiles could be added at a later date.

**2.6. Identify Additional Recommendations**

To support these requirements the WCGA RDF should consider:

- Providing a basic guide, training, and/or ongoing assisting in creating metadata records
• Providing resource providers with minimum model compliant templates for ISO 19139 and Dublin Core.
• Providing a DTD or schema for validation of the documents
• Providing controlled vocabularies or ontologies for key metadata elements (recommended by Rex Sanders)

2.7. Define Publishing Workflow

Based on our assessment of the project requirements, we identified one possible workflow for information to flow from resource providers, through the WCGA Data Registry, and to the audience that needs it. Each of the numbered steps below is visualized in the diagram that follows.

1. Resource provider creates metadata documents (if they don’t already exist)
   a. Metadata document created from model template with at least minimal set of required fields
   b. Each metadata document includes Uniform Resource Identifiers (URIs aka URLs) for download links, web service endpoints, and other alternative ways to access the resource.

2. Resource provider publishes metadata records on a CSW-capable data catalog or in a web accessible folder (WAF)

3. The WCGA admin registers the resource providers data catalog or web accessible folder in the WCGA Data Registry for harvest of priority resource metadata records
   a. Harvest is done periodically (e.g. daily)
   b. Only priority resources identified in Task 1 are harvested. This can be described as filtering or 'cherry picking'
   c. If new records are found they are added, removed records are deleted, and changed records are updated. This keeps the two repositories synchronized.
   d. The WCGA admin verifies proper harvesting the first time and resolves any errors. This might involve making adjustments/corrections to the resource provider's metadata document.
   e. Harvested documents are stored in the internal database.
   f. Document attributes including web service links are extracted.

4. The WCGA data registry admin assigns additional attributes to resources
   a. These attributes are terms from the WCGA-specific controlled vocabulary/taxonomy
   b. These attributes are maintained in the WCGA data registry database and associated with the attributes from the original metadata document.

5. The WCGA data registry publishes resources through its public user interface (UI)
   a. Users discover WCGA priority resources through full-text search (faceted), tag/hierarchical search (category, sub-category, theme, sub-theme, state, ACT), etc.

6. Users access priority resources through links (URI's) provided in the registry (direct file download, web service endpoints, etc.)

7. The WCGA data registry re-publishes resources through its CSW service
   a. Other catalogs are able to query and harvest from the WCGA data registry, thus forming a federated network of data catalogs.
3. DATA CATALOG ASSESSMENT

3.1. Identify Potential Catalog Products

Four data catalog products were identified that could provide a significant number of the features required by the WCGA data registry out-of-the-box.

- PyCSW
- CKAN
- ESRI Geoportal
- GeoNetwork

Other data catalog products were identified, but were not considered due to not meeting one or more major requirements of the WCGA Data Registry.

- MetaCat - http://knb.ecoinformatics.org/knb/docs/
  - Strong support for EML/Dublin Core metadata schemas but not ISO geospatial schemas. Strong support for the DataOne service, which is not supported by any of the other catalog products currently. Does not support the CSW standard.
  - Developed in partnership by NASA, USGS, and others. Mercury is a robust data catalog that has broad support for metadata standards including FGDC,
Dublin-Core, Darwin-Core, EML, and ISO-19115. However, Mercury does not support harvesting or publishing using CSW. This may change with the interest in integrating ESRI Geoportal with DataOne, a product built on Mercury. (https://redmine.dataone.org/issues/3229).

**PyCSW**
http://pycsw.org/

The main focus of PyCSW is providing a very lightweight Python CSW server solution. Another goal is to allowing you to quickly publish your metadata repository and make your resources discoverable. A number of data catalog projects including CKAN, have begun using PyCSW to provide their CSW harvesting and serving capabilities.

**Links**
- PyCSW - OSGeo project spotlight
- Inside Idaho - Idaho's state geospatial clearinghouse, CSW capabilities provided by PyCSW. Learn about its use
- CKAN - data catalog behind data.gov.uk and soon data.gov. CSW capabilities provided by PyCSW.
- Open Data Catalog - data catalog behind Open Data Philly. CSW capabilities provided by PyCSW. Learn about its use of PycSW

**Pros**
- PyCSW was just certified OGC-compliant, and is in the process of becoming an official OSGeo project. This is meant to ensure longevity and stability of the project going forward.
- Supports multiple profiles for ISO 19139, Dublin Core, and FGDC CSDGM
- Full CSW harvesting and serving capabilities. Great for resource providers that want to stand up a simple catalog (without a user interface) that other catalogs can harvest from
- If you want to build a lightweight web interface on top of PyCSW, the CSW and RESTFUL API's provide that ability

**Cons**
- No graphical user interface
- No user account, groups, or authentication out of the box
- No custom categories outside of extending metadata profiles
- No faceted search

**Initial Conclusions**
- By itself, PyCSW is not a good option for the WCGA Data Registry. It functions best as a simple 'headless' CSW server or in combination with another data catalog product like CKAN.
CKAN
http://ckan.org/

CKAN was created by the Open Knowledge Foundation (OKFN) in the United Kingdom and is the data catalog platform behind data.gov.uk. CKAN is now beginning to catch on in the US. It was recently chosen to become the new data catalog behind data.gov and geo.data.gov, which will be unveiled sometime in early 2013. CKAN will replace the DMS data catalog platform behind data.gov and the ESRI Geoportal platform behind geo.data.gov.

Some of the reasons given for choosing CKAN over others for data.gov include:

- Modern, well-designed user interface out-of-the-box
- Robust faceted search with plans for more robust spatial search algorithms
- Responsive and enthusiastic support from the Open Knowledge Foundation
- Designed for publishing any type of resource, not just geospatial
- Flexible interface, easier to extend than Geoportal
- Searches are quicker and provide more relevant results out of the box than Geoportal

CKAN features currently under development for the 2.0 release and funded by the data.gov transition:

- Support for ISO and FGDC metadata harvesting
- Improved CSW support with the integration of PyCSW
- Support for harvesting from Geoportal using CSW
- User group and moderation capabilities

Other Pros

- Focus on user experience; getting people to the information they need quickly. Doesn't overwhelm the user with features.
- Very robust web API allowing you to access and edit any piece of information in CKAN, even build a whole new front-end web interface.

Cons

- Harvesting support for version 1.8 is limited to ISO metadata, and only from GeoNetwork. This will improve in early 2013 with the integration of PyCSW as the CSW server for CKAN.
- OKFN supports CKAN running on Linux but not Windows through their support services, although CKAN can run on Windows successfully.

Initial Conclusions
• Great user interface and user experience out-of-the-box with use of modern HTML/CSS. This would provide a solid starting point for the WCGA Data Registry.
• The 1.8 release of CKAN is lacking support for FGDC and Dublin Core metadata and the CSW publishing and harvesting capabilities are incomplete. At the current pace of development of the new data.gov, most of these missing features are expected to be available in 1-2 months for the 2.0 release, but this is not guaranteed.

ESRI Geoportal
http://www.esri.com/software/arcgis/geoportal

ESRI Geoportal is a mature data catalog product created by ESRI, and is now open source. It is in widespread use by state and federal agencies that have traditional GIS departments, but not for example by the oceanographic and ocean observation community.

ESRI Geoportal is used by many of the resource providers identified for the WCGA Data Registry.

Links
• Oregon Spatial Data Library
• geo.data.gov
• National Geophysical Data Center Geoportal

Pros
• Technical support and training available from ESRI
• Full CSW support and a number of metadata schemas and profiles
• Support for harvesting ArcGIS Server
• RESTFUL API, for searching metadata records only. Has support for JSON, which would be an option for creating a new lightweight user interface for the WCGA Data Registry and minimizing the need to alter server-side code. Note, that CSW would still be used for harvesting.
• Geoportal server is likely to ensure support is provided for ESRI products going forward including ArcGIS Server, ArcGIS Online, etc.
• A lot of development to date has been funded by NOAA, who has setup a large number of Geoportal instances.

Cons
• The default user interface is functional but the user experience overall is lacking. From the perspective of someone just approaching the site:
It can be difficult to find what you're looking for. The navigation and discover of resources using the search and tree views is not the most intuitive.

The map viewer features with their toolbars are designed for people with GIS experience.

Overall the interface appears to be designed more for power users, which can be overwhelming to the average user.

- The server-side Java framework, Spring, has been around since 2003 which reflects well on the projects maturity. Software framework design has also advanced a lot since then and Spring is considered by many to be perfectly usable but not the go-to framework for creating web applications in Java. It's worth considering how much of an investment should be made if custom development is needed to create the WCGA Data Registry.

- The core developer Marten is quite responsive but it's not necessarily clear where the project is headed or who is working on it. As a developer if you're going to work with a catalog and customize it, you need that transparency if you want to stay compatible with the changes.

**Initial Conclusions**

- Very robust data catalog features and CSW harvesting of resource providers should work well. Straightforward to use from an administrative standpoint. The project is mature and stable.

- The default user experience isn't well executed. 5 years ago the experience would be considered great. Since then there have been a lot of advances in user interface design and HTML/CSS standards that can and should be leveraged. If Geoportal were selected for this project, we would recommend that a new user interface be developed, built on the Geoportal RESTful API.

**GeoNetwork**  
[http://www.geonetwork-opensource.org](http://www.geonetwork-opensource.org)

GeoNetwork is a mature data catalog product and is one of the flagship projects of OSGeo.

GeoNetwork is popular outside the US and has excellent support for the EU INSPIRE initiative. GeoNetwork is increasingly developed in lock-step with other open source GIS projects including GeoServer, GeoWebCache, and GeoNode. Not surprisingly because the OpenGeo organizations employs core developers on all of these projects.

**Links**
• Dutch National Registry

Pros
• Technical support and training available from OpenGeo
• Full CSW support and a number of metadata schemas and profiles
• Standardized CSW interface
• RESTFUL API, for searching metadata records only. Has support for JSON which makes it possible to create at least basic custom web clients without needing an additional server-side component.

Cons
• The default interface is overly complex
• The default user experience is lacking

Initial Conclusions
• Very robust data catalog features and CSW harvesting of resource providers should work well. Managed to pass all of the CSW tests, even with slightly imperfect test records.
• The default user experience for GeoNetwork is functional but underwhelming. The use of GeoExt in the new upcoming version isn't well executed. The XML-based API is powerful but complex. If GeoNetwork were selected for this project, we would recommend developing a new user interface, built on top of the CSW interface.
• Bottom line. GeoNetwork is mature software product like Geoportal. It's used by a lot of groups. The user interface is functional but clunky and not always intuitive. There's a lot of features offered out of the box that the WCGA data registry doesn't need.

3.2. Catalog Comparison

The table below is a direct comparison of the 4 data catalog products. Key missing features are in red, areas of concern are in yellow. Developers of each of these software products had the opportunity to provide feedback.

* Note that the upcoming CKAN 2.0 release will include integration with PyCSW but the work is currently not finished. We decided to go ahead and include a 5th option in the table below for comparison called ‘CKAN 2.0 (anticipated)’. This is a best guess of what CKAN will be able to do in a few months based on what PyCSW can currently do now.
<table>
<thead>
<tr>
<th>Operating System</th>
<th>PostgreSQL, MySQL databases (SQL Server, PostgreSQL, MySQL, etc.)</th>
<th>Windows, Linux, OSX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata storage/access strategy</td>
<td>Uses an internal database model. Attributes are extracted from metadata documents into the model on harvest. This keeps the complexity of metadata schemas, profiles, and XSLT transforms out of the main codebase.</td>
<td>Uses an internal database model. Attributes are extracted from metadata documents into the model on harvest. This keeps the complexity of metadata schemas, profiles, and XSLT transforms out of the main codebase.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REQUIRED FEATURES</th>
<th>PyCSW</th>
<th>CKAN</th>
<th>CKAN 2.0 (anticipated)</th>
<th>ESRI GeoPortal</th>
<th>GeoNetwork</th>
</tr>
</thead>
<tbody>
<tr>
<td>User accounts (registered/admin)</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Secure authentication</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Custom categories (separate from metadata)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Search engine-like indexing</td>
<td>No</td>
<td>Yes. SOLR</td>
<td>Yes. SOLR</td>
<td>Yes. Lucene</td>
<td>Yes. Lucene</td>
</tr>
<tr>
<td>Faceted search</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Supports ISO 19139/19115/19119</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supports FGDC CSDGM</td>
<td>Yes</td>
<td>No. (expected v2.0)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Supports Dublin Core (ISO 15836)</td>
<td>Yes</td>
<td>No. (expected v2.0)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Harvesting feature</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Harvest filtering (cherry picking)</td>
<td>Yes, with custom profiles</td>
<td>Yes, with custom harvesters or profiles</td>
<td>Yes, with custom harvesters or profiles</td>
<td>Yes, with custom profiles</td>
<td>Yes</td>
</tr>
<tr>
<td>CSW server feature</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RESTful API feature</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-- output types supported</td>
<td>JSON, XML</td>
<td>JSON</td>
<td>Yes</td>
<td>JSON, XML</td>
<td>XML</td>
</tr>
</tbody>
</table>
3.3. Integration Testing
All four data catalogs were setup and configured in a development environment. They were then tested for their ability to harvest ISO and FGDC metadata documents from a web accessible folder, and from each other using the provided CSW interface. This testing was critical because the WCGA Data Registry will need be able to harvest from, at the very least, other Geoportal and GeoNetwork catalogs. By testing harvesting between each of these catalogs, it could be shown how robust their CSW publishing and harvesting capabilities are.

Test Suite
The metadata used for testing was a set of 18 sample metadata records developed or acquired by Tanya Haddad with Oregon DLCD. These were minimal records and 16 of them were ISO, 1 was FGDC CSDGM, and 1 was Dublin Core. FGDC records from the California Geoportal were also included in some cases. The sample metadata was loaded in turn into each of the catalogs and then a harvest was attempted from each of the other catalogs.

Table 6 - Data Catalog Comparison Matrix

<table>
<thead>
<tr>
<th>NON-REQUIRED FEATURES</th>
<th>PyCSW</th>
<th>CKAN</th>
<th>CKAN 2.0 (anticipated)</th>
<th>ESRI GeoPortal</th>
<th>GeoNetwork</th>
</tr>
</thead>
<tbody>
<tr>
<td>User groups</td>
<td>No</td>
<td>No. (expected v2.0)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Support for private resources (privileged access)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data upload</td>
<td>No. DIY</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Harvest alternatives</td>
<td>WMS, JSON</td>
<td>URL, Google spreadsheet, ArcGIS Server REST API (planned for 2013 with data.gov work)</td>
<td>URL, Google spreadsheet, ArcGIS Server REST API (planned for 2013 with data.gov work)</td>
<td>THREDDS, ArcGIS Server REST/SOAP, ESRI Metadata Services, URL</td>
<td>WebDAV, OGDC Web Services, Thredds, ArcSDE Server</td>
</tr>
<tr>
<td>Supports harvesting non-standard attributes in metadata?</td>
<td>Yes, custom profile</td>
<td>Yes, custom profile</td>
<td>Yes, custom profile</td>
<td>Yes, custom profile</td>
<td>Yes, custom profile</td>
</tr>
</tbody>
</table>
# Test Results

<table>
<thead>
<tr>
<th>CATALOG INTEGRATION TESTING</th>
<th>PyCSW</th>
<th>CKAN</th>
<th>CKAN 2.0 (anticipated)</th>
<th>ESRI GeoPortal</th>
<th>GeoNetwork</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO Harvest from PyCSW source using CSW</td>
<td>Success</td>
<td>CKAN Failed (expected v2.0)</td>
<td>Success</td>
<td>Failed</td>
<td>Success</td>
</tr>
<tr>
<td>ISO Harvest from CKAN source using CSW</td>
<td>CKAN Failed (expected v2.0)</td>
<td>CKAN Failed (expected v2.0)</td>
<td>Success</td>
<td>Failed</td>
<td>Success</td>
</tr>
<tr>
<td>ISO Harvest from CKAN 2.0 source using CSW (anticipated)</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
</tr>
<tr>
<td>ISO Harvest from GeoNetwork source using CSW</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
</tr>
<tr>
<td>ISO Harvest from Geoportal using CSW</td>
<td>Success</td>
<td>Failed (expected CKAN v2.0)</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
</tr>
<tr>
<td>FGDC Harvest from Geoportal source using CSW</td>
<td>Success</td>
<td>Failed (expected CKAN v2.0)</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
</tr>
<tr>
<td>Dublin Core Harvest from Geoportal source using CSW</td>
<td>Success</td>
<td>Failed (expected CKAN v2.0)</td>
<td>Success</td>
<td>Success</td>
<td>Failed</td>
</tr>
<tr>
<td>ISO/FGDC/DC Mixed Harvest from Geoportal source using CSW</td>
<td>Success</td>
<td>Not tested</td>
<td>Success</td>
<td>Success</td>
<td>Not tested</td>
</tr>
<tr>
<td>ISO Harvest from Web Accessible Folder (WAF)</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
</tr>
<tr>
<td>FGDC Harvest from Web Accessible Folder (WAF)</td>
<td>Success</td>
<td>Failed</td>
<td>Success</td>
<td>Success</td>
<td>Success</td>
</tr>
<tr>
<td>Dublin Core Harvest from Web Accessible Folder (WAF)</td>
<td>Failed (expected Jan)</td>
<td>Failed</td>
<td>Failed (expected Jan)</td>
<td>Success</td>
<td>Failed</td>
</tr>
<tr>
<td>ISO/FGDC/DC Mixed Harvest from Web Accessible Folder (WAF)</td>
<td>Partial Success (DC expected Jan)</td>
<td>Partial Success (ISO only)</td>
<td>Partial Success (DC expected Jan)</td>
<td>Success</td>
<td>Partial Success (ISO/FGDC only)</td>
</tr>
</tbody>
</table>

The ISO harvest testing conclusively showed that PyCSW, ESRI Geoportal, and GeoNetwork provide excellent support for harvesting via CSW and of web accessible folder.
We therefore conclude that CKAN 2.0 with PyCSW integration will also have excellent support.

Test harvesting of FGDC CSDGM and Dublin Core metadata records was done for PyCSW and Geoportal. Both were found to be able to harvest the single record we had of each using CSW and from a web accessible folder.

In addition to testing harvesting capabilities on a small scale we also verified that both PyCSW and CKAN could successfully harvest records from larger external catalogs out there (1000+ records) like the California Geoportal, Oregon Coastal Atlas, and ESRI Geoportal Demo Server.

3.4. Final Conclusions

Geoportal vs GeoNetwork
Both are mature Java-based software products, and have a similar feature set including good support for CSW serving and harvesting. They also share a shortcoming in their user interfaces. Where Geoportal stands out is its support for JSON with its RESTful API, the fact that ESRI is accessible as a vendor right here on the West Coast, and the fact that so many of the WCGA resource providers are already using the Geoportal product. For these reasons the WCGA IT Working Group recommended that ESRI Geoportal be selected over GeoNetwork and that the final decision be between CKAN and ESRI Geoportal.

Geoportal Conclusion

Mid 2012, the data.gov team chose to use CKAN over Geoportal and we took that decision very seriously. The main Geoportal shortcomings they expressed included a complex user interface, difficulty in managing hundreds of thousands of metadata records, slow speed of searching those records, and lack of sophistication of the search algorithm to provide relevant search results with so many similar records.

The data.gov team thinks that some of these issues of catalog size and search have been solved better by CKAN and that may be true. We think that these issues are less of a concern for the WCGA Data Registry, at least for the time being, because there is expected to be only a few hundred datasets. These resources are also expected to be hand-picked by the WCGA RDF to be the best products for West Coast ocean planning. The creation of a West Coast taxonomy for categorizing these resources will also allow us to provide alternative ways for people to discover these resources besides a keyword search, however sophisticated it may be. In our testing of Geoportal, we found the speed of searches to be sufficient for our needs.

CKAN Conclusion

CKAN integration of PyCSW is beginning this month (Jan. 2013) as part of the data.gov work. This will provide full harvesting and CSW capabilities for ISO, FGDC, and Dublin Core. This work is expected to take 1-2 months and culminate in the release of CKAN 2.0. This timeline appears to be within the timeline of this WCGA data registry project. However, there is always a risk that this work won't be completed on time. The code is also under transition and not in a useable state at this time.
CKAN vs. Geoportal

- Both CKAN and Geoportal are viable options for this project but both will need customization to provide the necessary features.
  - Selective harvesting. Both have limited support for selective harvesting of metadata records from another data catalog. (~4 work-weeks to develop)
  - Taxonomy. Both do not support creation of a taxonomy and search of records by taxonomy term, as it is currently being envisioned by the WCGA RDF. (~4 work-weeks to develop). Similar but insufficient features include CKAN - tags, tag vocabularies and Geoportal - collections, assertions
- Both have excellent harvesting capabilities of ISO, FGDC, and Dublin Core using both CSW and web accessible folders.
- Both have a robust administrative interface providing the necessary tools for managing harvesting and adding additional information to each record.
- CKAN's user interface out-of-the-box is closer to what's desired for the WCGA Data Registry but that doesn't mean it will take less time to develop the vision that comes out of the upcoming UI design phase. We think that if either catalog is able to expose all of the necessary information through its RESTful API to drive the user interface, then we can very quickly develop a polished and powerful custom user interface that is completely separate from these data catalog products.
  - ~ 1-2 work-months
- Geoportal is arguably a more complex code base. We think it would take less time for us to get up to speed with CKAN and its internal architecture and begin developing with it productively (~1.5 weeks vs. ~3 weeks with Geoportal)
- CKAN and PyCSW have a more active development team and a more transparent development process on GitHub.
- Geoportal has more users on the West Coast that together form an informal support network and pool of resources. However CKAN has better communication channels than Geoportal.
- Geoportal currently has better support for metadata harvesting through non-standard interfaces including the THREDDS API and ArcGIS Server API.
  - You can expect that Geoportal will continue to implement and keep up to date with supporting access to ESRI flagship products and services.
- Open Knowledge Foundation provides enterprise support for CKAN but only on the Linux operating system at this time.

Catalog Recommendation

As good of a fit as CKAN might be, the key updates underway for it won't be ready for at least a couple months. We recommend that we start the process of designing mockups for the registry user interface. At the same time we can proceed with using Geoportal, at least for the time being, and prove its capabilities and perceived shortcomings. Specifically, in the next 1.5 months (by the end of February 2013) we can:

- Create design mockups for the data registry based on the agreed upon requirements. Get agreement from the group on what the user experience will be including what different features will look like and be used at a high-level.
Setup the latest development version of Geoportal on a development server that is accessible by others (Linux).

Continue to build a rapport with the Geoportal dev team and other West Coast Geoportal users.

Prove the existing mechanism for selectively harvesting records with Geoportal, specifically from the California Geoportal.

Prove the ability to create a user interface separate from Geoportal using the Geoportal REST API by building a quick straw-man registry interface.

Assess the backend Geoportal development that needs to be done.
  o Includes modules that need to be created (taxonomy) or additional information that needs to be exposed through the RESTful API in order to drive the expected user interface features.

At the end of February we will summarize our successes as well as any shortcomings of Geoportal and its RESTful API as it pertains to the WCGA Data Registry and workflows it needs to support. At that time, if there are issues that are insurmountable with the available time and budget for this project we should consider our options. This might include requesting help from the ESRI core team or in the extreme case recommending a switch from Geoportal. The progress of CKAN might be further assessed at this point. If this is necessary, all would not be lost as the time spent designing the registry UI, developing metadata templates, and working with partners to develop metadata documents and harvesting workflows would be transferable.

User Interface Recommendation

The preference expressed by working group members seems to be to develop an entirely new public user interface for the WCGA Data Registry rather than starting with the Geoportal or CKAN interface. And because all of the administration of the registry is expected to be able to be done through the data catalog software, we believe this separate public user interface can be done without creating any additional backend systems or databases.

Our recommendation is to build this user interface using the REST API provided by both CKAN and Geoportal. This design is expressed in Figure 2, and decouples the data registry user interface from the data catalog software providing the following advantages:
  • Protects us from subsequent Geoportal revisions, particularly to the UI, that can render a lot of our work obsolete.
  • Allows us to switch to a different data catalog at a later date if needed, with a minimal amount of code changes. For example, CKAN also offers a RESTful interface for accessing all of its catalog functionality (http://docs.ckan.org/en/latest/api.html)

- Email from Doug Nebert (DOI), interim technical lead for data.gov taking over for Chris Musilek and response to NOAA metadata managers on status of data/geo.data.gov

GSA data.gov program has selected CKAN as the basis for the next generation of the data.gov infrastructure, replacing the DMS (metadata entry system and approval engine) and the catalog itself. Requirements are being developed that will lead to an imminent contract for CKAN enhancements that will allow it to also be a replacement for the geo.data.gov catalog. This will be used by the Geospatial Platform as the sole catalog, fed by harvest of agency metadata - in ISO or FGDC format or even from agency Portal for ArcGIS/AGOL instances. There are a lot of moving parts and we need to be sure that we get what we need and have a better integrated and easier-to-use environment. We look forward in the next two weeks to engage the agency stakeholders to develop/refine the requirements for this rapid development cycle, and look forward to NOAA's technical support and guidance.

The CKAN developers (OKFN) developed the prototype in less than one month, based on what has been done in Europe for their data/geodata community in the UK. As you know, they require ISO metadata for the INSPIRE initiative there, so we will be requiring a robust metadata handling and validation capability. We will still be able to ingest FGDC metadata and will translate it to ISO for presentation and indexing, but are recommending storing and presenting the original form on request in the UI.

Most agencies handcraft their geospatial metadata using GIS tools and metadata editors 'at home' and then submit them for harvest or expose them in a catalog service. data.gov is seeking to move away from the push model - enter and edit metadata from a form, store the metadata only at GSA - and towards the pull or harvest model from agency metadata stores for both data.gov "raw data" and geospatial metadata. We will continue to manage the virtual catalog through registered collections of metadata with approved content. In this environment, the need for a central metadata editor at data.gov is greatly reduced. Even small collections of metadata records would benefit from this 'closest to source' harvest model, giving control to the authors and avoiding stale repository models.

Also supporting a push model and editor for the nextgen data.gov would increase the complexity and cost of the system. If, however, a CKAN-based catalog were to be deployed as an organizational metadata management solution, then the ability to enter, update, and delete different types of metadata records in the agency catalog would be important. Given the spiral or incremental development approach we will take, we might see such capabilities added in the future based on future stakeholder requirements and funding if there was interest in supporting and developing agency catalogs (that we would continue to federate).

Please send me your thoughts and opinions on this approach. We look forward to your insight and thoughts moving forward.

Doug.
CKAN and Data.gov / Geo.data.gov Questions and Answers:

1. **Question:** CKAN data.gov will be able to handle the FGDC CSDGM format. This is good. Not everyone will be off CSDGM and completely transitioned to ISO for a few years. That will take some time. It will be interesting to see how the planned transform would handle the complexity of doing CSDGM to ISO. I have done a few transforms and even though the information in the CSDGM gets moved over, there is a certain amount of hand editing that I have needed to do.

   • **Answer:** Since CKAN already has parsers for ISO (and not for FGDC) we will use that mechanism to ‘feed the index’ - to put the relevant metadata content into the searchable ‘fields’ (facets in Lucene) for rapid retrieval. Fortunately, the relevant populated fields of interest are easily mapped between FGDC and ISO. Once the content is made searchable, then it is a matter of giving the user a choice of inspecting it in either ISO or native format. We can work on finesseing the transformation as we get more experience - we will be using Jacqui Mize’s (NOAA) transforms.

2. **Question:** Will agency metadata administrators eventually have the capability and permissions to make their own metadata visible? Currently there is still the issue with getting metadata visible on the current Geoportal Server due to conflicting documents between Data.gov and the whole old GOS concept. For example I want the entire NOS metadata inventory (or NOS WAF’s) made visible in geo.data.gov. Although I requested all to be shared and made visible, it seems they must be asked for individually and cannot happen automatically for my account as a federal account holder.

   • **Answer:** The simplification of the entire registration and approval process is a major goal of this consolidation. We want to still support differing views or filters on the data that would allow one to discern quickly what are ‘Platform’ nationally-significant A-16 data, federal data.gov data, and non-federal state/local/academic/tribal data. My thinking is that the geo.data.gov site will actually become just the geo publisher and user Community for data.gov. Search would be linked through the UI in data.gov or Platform, but in new and cleaner ways through the same API. We will also be mining and using the "metadata" created in registered agency ArcGISPortal instances and making those webmap mashups discoverable and actionable as resources, though they will continue to have light metadata. The data.gov "raw" catalog will be merged in also, with its smaller metadata schema.

3. **Question:** Currently, we cannot see which records are shared from the account, and I must do individual searches on the Data.gov (Geo.Data.gov) Geoportal to see if a record is visible (i.e. no way to do a bulk search for “NOS” inventory. So, will there be a way to identify which records are made visible for Data.gov? If not, will there be an easier way to get them made visible to the public?
• **Answer:** Absolutely. These are logical design criterion.

4. **Question:** Will CKAN be able to take advantage of ISO XML attributes? Currently simple ISO editors validate for a data, one cannot use a nilReason attribute to document Unknown or Inapplicable.

   • **Answer:** Certainly we could tune the XSL styling to display XML attributes.

5. **Question:** Will CKAN take advantage of elements like Browse Graphics for display?

   • **Answer:** Yes, as we used to do in GOS, to display a thumbnail with metadata results. Unfortunately, we lost a bunch of functionality over the past two years that we need to regain and go well beyond. This is our opportunity.