World Wide Technology, Inc.

State of California Ocean Protection Council

Marine Protected Areas Infrastructure Management Assessment: Technical Design

March, 2017 (Revised April, 2017)

RESOURCES LEGACY FUND
Creative Solutions. Lasting Results.

CALIFORNIA
OCEAN PROTECTION COUNCIL

![World Wide Technology, Inc.]

Contents

Legal Disclaimer

The information contained herein is proprietary and confidential to World Wide Technology (WWT) and the specific client for which it was prepared. This document may not be reproduced or redistributed in any format, written or electronic, without express written consent of all parties involved.

WWT certifies the information in this document to be correct and true, to the best of its knowledge, at the time of its publication. All reasonable measures have been taken to ensure that the information provided is as accurate and up-to-date as possible at the time this document was completed.

Introduction

In 1999, California enacted the Marine Life Protection Act to protect California's marine natural heritage through establishing a statewide network of Marine Protected Areas (MPAs). MPAs are designed to preserve the diversity and abundance of California's marine life, the ecosystems they create, and the habitats they populate through a variety of programs ranging from conservation efforts to human activity restrictions.

Significant state-funded efforts have been focused on fishery, ecological, oceanic, and tribal research and scientific monitoring within the MPAs in order to benchmark baseline conditions and evaluate change over time through ongoing monitoring. The first phase of benchmarking occurred from 2007-2012 for the Central Coast MPA region, and subsequently for the other three MPA regions. Over 300 data packages have been created through state-funded research through this baseline benchmarking effort, and the size and complexity of the data will grow into the future as efforts transition to focus on long-term monitoring.

A key goal of the State is to multiply the value of the data collected by allowing it to be readily accessible and usable for anyone interested in the condition of MPAs. These stakeholders include decision-makers within the Department of Fish and Wildlife and Fish and Game Commission who are responsible for the success of the MPA Program, researchers and scientists involved in MPA research, as well as the general public who have a vested interest in the condition of the oceans. For this to happen, a robust data platform is required to support the upload, storage, discoverability, and download of these datasets.

An initial instance of the MPA data platform, OceanSpaces.org, was created in partnership with Ocean Science Trust in 2012. As the priorities and needs of the platform have evolved since the first deployment of the website, a more sophisticated infrastructure is required to support the needs of the State and platform user base into the future. World Wide Technology was engaged to assess the current infrastructure and provide a design for the platform to meet these revised priorities.

This document contrains a mid-level design that describes a proposed architecture to support the MPA data platform. The design was developed through eight weeks of assessment which included numerous interviews with a variety of stakeholders to understand user needs, as well as analysis of the environment in which the data platform currently resides and the new infrastructure being developed by the California Natural Resources Agency.

Objective

Develop a mid-level technical deployment solution for California's Marine Protected Area (MPA) Data Management Plan that will improve the discoverability, relevance and usability of the state's MPA monitoring data for Resource Managers, Decision Makers, Data Contributors and Users

Scope

MPA regions in California, which include the North Coast from the California/Oregon border to Alder Creek, North Central coast from Alder Creek near Point Arena to Pigeon Point, Central Coast from Pigeon Point to Point Conception, and South Coast, Point Conception to the California/Mexico border, including Channel Islands MPAs.

MPA Data Platform Requirements

The objectives of the MPA data platform are to encourage:

- Sustained use of MPA monitoring data by making it accessible, discoverable, durable, and appropriately described

- Science-informed decision making supported by results of MPA monitoring

- Interoperability of data to allow for higher-level synthesis

- Use of data by decision makers, resource managers, contributors, and the scientific community

To support these objectives, we have focused on developing infrastructure that will:

- Ingest data and store it in a secure manner

- Allow data to be easily discoverable and downloadable through robust search capabilities, potentially map-based

- Display data easily through understandable visualizations that capture key metrics governing MPA programs

- Integrate data and/or platform with complementary data sources

Current State

Through interviews with Ocean Science Trust, we understand the current system to function as following:
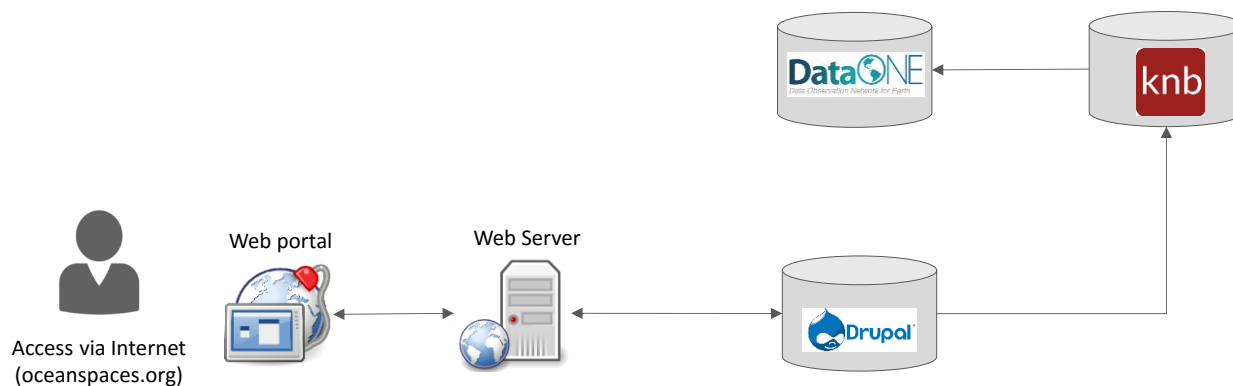
The core platform consists of the OceanSpaces Drupal instance, which provides community blogging, news publishing, and data cataloging. In addition, the OceanSpaces platform has an API connected to the ecological data repository KNB that will allow data to be stored and accessed on the KNB platform, which is in turn connected to the global scientific data repository DataONE.

The platform has valuable features that should be maintained going forward. These include:

Community user base and blogging: The OceanSpaces platform has a sizeable user base. Individuals who were interviewed as part of the user needs assessment most frequently mentioned using the

OceanSpaces platform for news update and community networking. Given the substantial effort already dedicated to building the platform and establishing a reputation in the community, the social aspect of the platform should be retained.

Metadata and upload standards: Ocean Science Trust has established metadata standards and data upload guidelines for contributors. These are appropriate for the present time but should be revised as data formats evolve.

KNB connectivity: OceanSpaces has an agreement with Knowledge Network for Biocomplexity (KNB), an international data repository that facilitates ecological and environmental research, to upload and store data on KNB's server. Data uploaded to OceanSpaces from March 2017 onwards will be stored on both OceanSpaces servers and the KNB server, and will be accessible from the KNB data portal. Moreover, KNB is also part of a larger data portal called DataONE, which allows OceanSpaces data to be distributed through additional channels. The connection to KNB and DataONE opens up OceanSpaces to a far wider audience and scientific community, and it is a priority of the State to make data accessible for the broader community.

Recommended Future State

While the current platform meets the needs of users today, it does not scale readily to accommodate the more sophisticated elements the platform requires going forward, such as effective search capabilities and large image and spatial data set storage.

The platform that the CNRA will deploy provides robust data cataloging and analytics capabilities that can accommodate the needs of the MPA Platform as it grows and expands. The CNRA platform, combined with some components of the existing system, can meet the needs of key user groups, state requirements, and scale into the future. In addition, a centralized state platform will facilitates cross-departmental analyses and connection with other state data sources.

Overall Architecture

We recommend an architecture that integrates Drupal as the content management system for community webpage functionalities and CNRA's ODP CKAN instance as the data cataloging system for data upload, download, and search. In addition, CKAN should be utilized for storage of data sets, both structured and unstructured. All data will be stored in the ODP platform and pushed to KNB from ODP.

External users will continue to access the data portal through OceanSpaces.org, while internal state users may access the ODP CKAN environment directly through a secured internal network. In addition, select external power users can be allowed direct access to the ODP if authorized.

 The ADP Hortonworks Hadoop data lake can be utilized for work-intensive analytics internal users may wish to do on MPA data. This platform will be valuable as MPA data increases in size and complexity. The ADP can also be used to perform analytics on MPA data in conjunction with other state data stored in the platform.
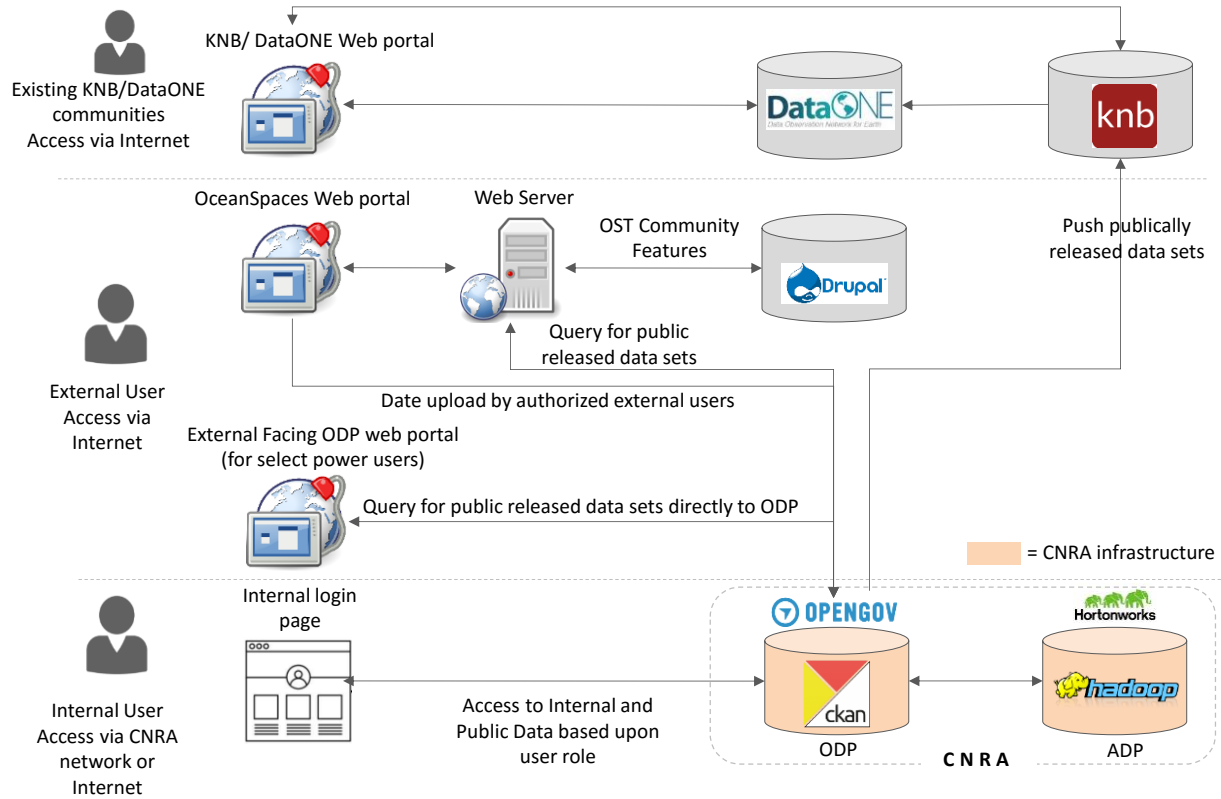
Figure 1: Recommended Future State Architecture

CKAN

CKAN is an open source content management system designed for managing, publishing, and searching data. It is extensively used to create open data websites by local, state, and national governments, research institutions, and scientific bodies. Both Drupal and CKAN contain features common to content management systems, such as user account registration, taxonomy, and page layout customization. However, Drupal's features are better utilized for community webpages and multi-user blogging, while CKAN specializes in building data hubs and data portals. CKAN has been used to build many government data portal websites; examples include the EU data portal www.europeandataportal.eu, the Australian federal data portal http://data.gov.au/ ,and the National Oceanic and Atmospheric Administration (NOAA) data platform https://data.noaa.gov/dataset. Many others can be found on https://ckan.org/instances/. The CNRA environment supports CKAN through implementation of OpenGov, a Software-as-a-Service product build on CKAN.

CKAN provides a framework supported by a large developer community with many pre-built functionalities relevant to open data platforms. Developers often rely on work created and shared by others in the profession, and having pre-made material such as packages of features or modules to use limits the need to create custom material and reduces effort in the long-term to maintain the platform. CKAN is appropriate for the MPA data platform due to a wide catalogue of pre-built functionalities, and a large community of developers that have created extensions and plugins

CKAN Capabilities

Many out-of-the-box capabilities will be relevant to the MPA data repository platform. These are described below.

Storage

CKAN can accommodate a variety of data sources such as tabular CSV or Excel spreadsheet, metadata in XML files, reports in PDF document and image files. The CNRA's OpenGov instance is hosted on Amazon Web Services.

Authorization

Different levels of access privileges within an organization can be given to users, including the ability to administer the website, upload data, edit or delete existing data, and view private data.

Search

CKAN has built in search capabilities through keyword and geospatial selection.

Users can search on dataset metadata, including titles, user-defined tags, formats, etc. In addition, CKAN can index text from within the document, enabling users to search for words or phrases within the document itself.

CKAN also is able to process co-ordinate geometries, which allows geospatial search capabilities through a spatial search plug-in. Users searching for data can filter results by geographic location by selecting an area on a map. See http://docs.ckan.org/projects/ckanext-spatial/en/release-v1.8/ and http://docs.ckan.org/projects/ckanext-spatial/en/latest/spatial-search.html#spatial-search-widget for more documentation.

Data Discovery

CKAN dataset information pages can contain a variety of metadata and data previews including the name, description, and other metadata. Visualization plugins can be installed to view data prior to download through tables, graphs, and maps. More details can be found under 'Data Visualization'.
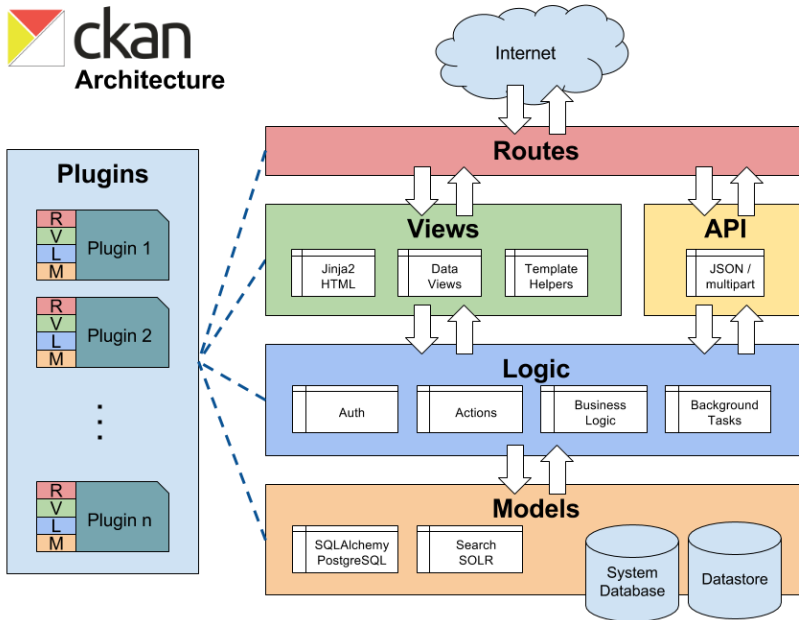
Usage Statistics

CKAN provides usage statistics such as number of dataset views. An HTTP header can be included to link to Google Analytics.

Flexibility

Much of the CKAN platform is extensible beyond the core feature set.

The CKAN architecture

Integrating Drupal and CKAN

Drupal and CKAN have APIs that allow flexible integration of the platforms. We recommend utilizing a "side-by-side" approach that has both systems operating alongside each other, rather than having CKAN sit behind Drupal. In this manner, CKAN receives requests directly from the webserver. It is possible to feed calls to CKAN from Drupal, however, this adds unnecessary complexity.
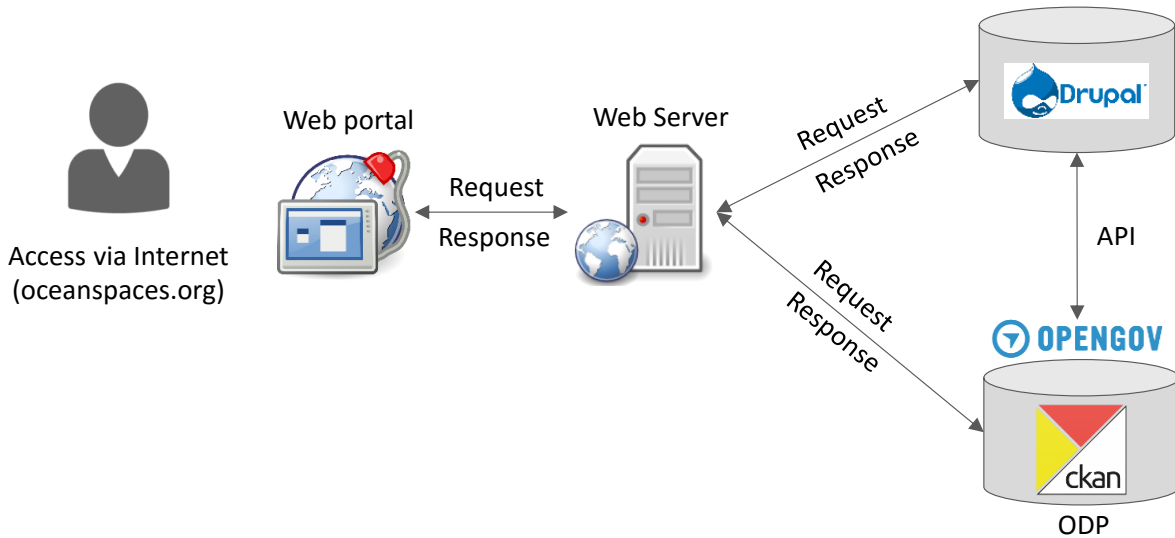
In this approach, Drupal and CKAN must share the same theme. However, theme-sharing is straightforward to implement.

Normal web API requests can be implemented through drupal_http_request calls or the python requests library.

Hortonworks

CNRA's Hortonworks Hadoop data lake can be leveraged for computationally intensive analytics and staging, storage, and processing of large data sets. A key goal of the state will be to perform internal analyses on MPA data, potentially in conjunction with data from other departments or confidential data that is not accessible through the public-facing OceanSpaces portal. The Hortonworks environment provides a secure environment to join data, run data analysis scripts in R or Python, and visualize data with tools such as Tableau. It can be an environment for teams to collaborate and share results and visualizations on both MPA data and cross-departmental data.

Migration Plan

The sub-sections below lay out a plan and guidelines for migrating to the CNRA CKAN environment.

Planning

An internal technical resource should be hired before migrating any data in order to ensure a timely, coordinated transition. He or she will need to act as the central point of contact for Ocean Science Trust, its vendors, any new vendors hired by the State, and the CNRA.

Next, OPC needs to formally define the terms of engagement with OST given that most users will be entering and landing the data portal through the Oceanspaces.org front-end. OST cooperation will be needed to ensure the migration proceeds in a smooth manner and that they continue to maintain the social aspect of OceanSpaces.

A CKAN managed service provider also should be engaged to build out the CKAN environment. Several are listed under 'Support Services'. OPC should review the migration plan with the vendor and ensure they are capable of providing the required services.

Once the CKAN managed service vendor has been selected, OPC can prepare the CKAN environment, build the required APIs to connect Drupal to CKAN and migrate data. The CNRA should be closely involved in the initial migration process to provide appropriate access to the ODP cluster and set up the environment.

Clear and firm timelines should be set with the CNRA for platform implementation. This should be done in consultation with the CNRA and in alignment with OPC priorities to determine a time frame that is reasonable. We expect a 90-120 day migration time frame barring no unforeseen hurdles and assuming availability of resources and data access. The migration process should start in August/September 2017 in order to meet OPC goals.

In the unlikely event of unforeseen delays in the buildout of the CNRA environment, OPC should begin building out the CKAN instance in the cloud through consultation with the vendor. The vendor may provide hosting services and have clusters already deployed at a cloud provider that can be readily utilized. The platform should easily port over to the CNRA environment

Implementation

OPC needs to define SLAs, roles and responsibilities, authorization, and security attributes with the CNRA to establish the CKAN environment. CNRA should be expected to offer support SLAs such as the ones described in 'Support Services'. These should be defined in the onboarding phase.

Current security protocols and authorization policies should be adopted for the new platform. The upload process should match the current process as closely as possible to minimize the learning curve for contributors.
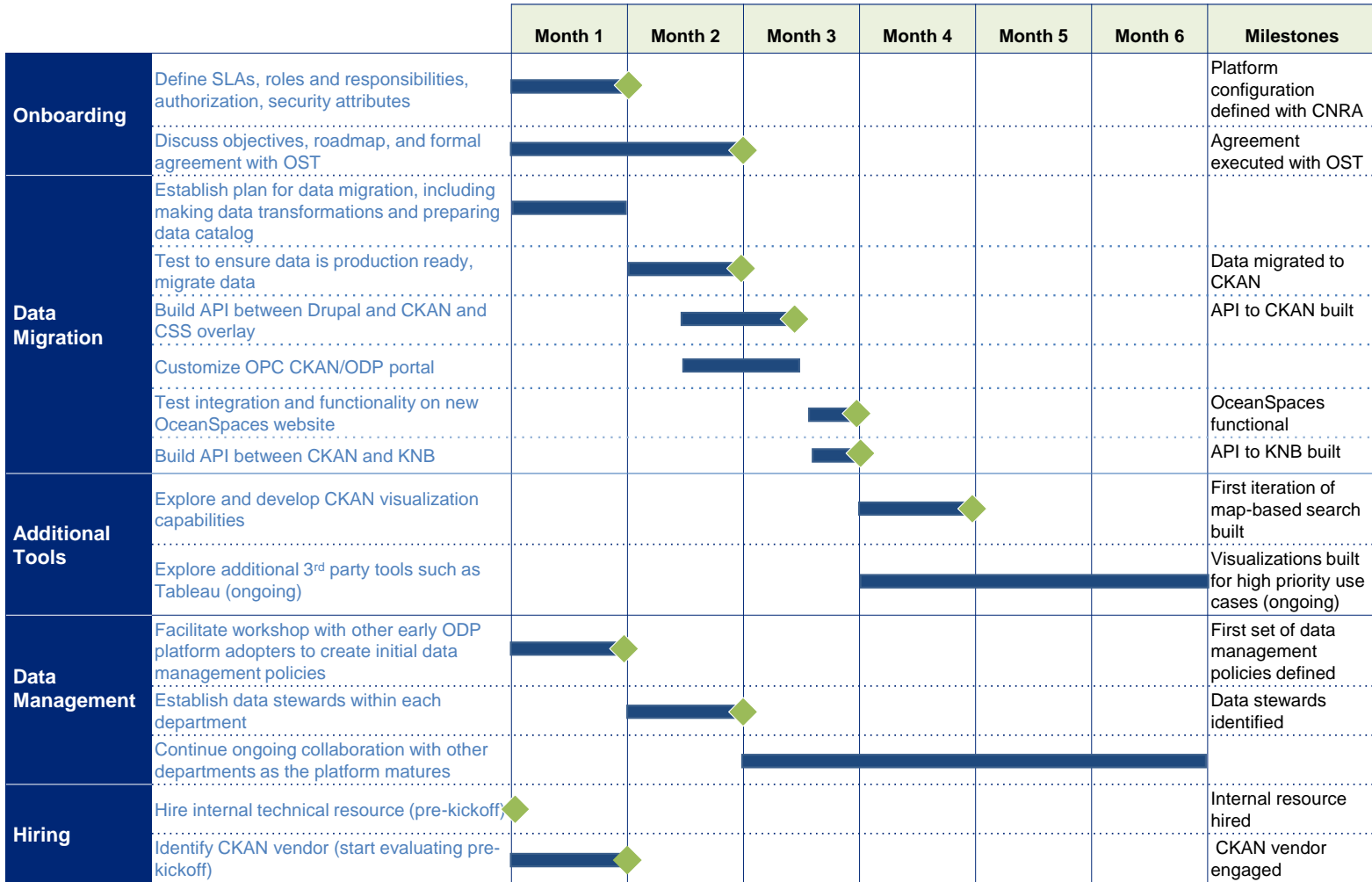
In the meantime, OPC needs to hold meetings with OST to discuss their plan for moving forward. OPC will need to gain cooperation from OST in order to stand up the CKAN environment under the OceanSpaces.org banner. It is unlikely that excessive OST involved will be required, however, OST will need to provide OPC with access to the current data storage environment and Drupal platform. They should be available for support when needed and join weekly/bi-weekly planning meetings.

After the initial configuration is complete, data migration can begin. The CKAN vendor should establishing the plan for data migration, including making data transformations and preparing the data catalog. Tests should be run to ensure the data is compatible with CKAN. After the data is ready, data should be copied over to the ODP.

The vendor will need to build an API and CSS overlay to connect Drupal with CKAN. In addition, an API will be needed to connect CKAN with KNB. KNB has an API guide at https://knb.ecoinformatics.org/#api

Some additional data governance standards may need to be implemented specifically for CKAN, such as the appropriate tagging of data.

The below diagram describes an estimated timeline for the primary migration activities. It assumes that the internal resource has been hired to manage the migration process and that no unforeseen hurdles are experienced.

		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Milestones
Onboarding	Define SLAs, roles and responsibilities, authorization, security attributes	■◆						Platform configuration defined with CNRA
	Discuss objectives, roadmap, and formal agreement with OST	■	◆					Agreement executed with OST
Data Migration	Establish plan for data migration, including making data transformations and preparing data catalog	■						
	Test to ensure data is production ready, migrate data		■◆					Data migrated to CKAN
	Build API between Drupal and CKAN and CSS overlay		■	◆				API to CKAN built
	Customize OPC CKAN/ODP portal		■					
	Test integration and functionality on new OceanSpaces website			■◆				OceanSpaces functional
	Build API between CKAN and KNB			■◆				API to KNB built
Additional Tools	Explore and develop CKAN visualization capabilities				■◆			First iteration of map-based search built
	Explore additional 3rd party tools such as Tableau (ongoing)				■		■	Visualizations built for high priority use cases (ongoing)
Data Management	Facilitate workshop with other early ODP platform adopters to create initial data management policies	■◆						First set of data management policies defined
	Establish data stewards within each department		■◆					Data stewards identified
	Continue ongoing collaboration with other departments as the platform matures			■			■	
Hiring	Hire internal technical resource (pre-kickoff)	◆						Internal resource hired
	Identify CKAN vendor (start evaluating pre-kickoff)	■◆						CKAN vendor engaged

1. Timeline assumes internal technical resource has been hired and no sever delays are encountered ◆ = milestone

Figure 2: Migration Timeline

Support Structure

Internal

Platform Management

We recommend that while OST retain primary responsibility for managing Drupal, the State should take responsibility for the majority of the platform, including buildout and maintenance of data upload, download, search, storage and visualizations. The first step is hiring an internal resource to serve as Program Manager for the MPA Data Platform and provide day-to-day oversight of OceanSpaces. A transition period will be needed to transfer responsibilities over from OST to the State, and close collaboration will be needed with OST to facilitate knowledge transfer. An incremental transfer of responsibilities should occur between OPC and OST across the various workstreams such as coordinating data collection from scientists, managing metadata and upload standards, managing technology vendors, and creating visualizations.

After the initial migration period, OPC may want to consider hiring an additional technical resource to augment the internal team responsible for managing and advancing the platform. A developer with similar programming skills to the job description below should be considered who will be responsible for building visualizations, imagery and video data, performing internal cross- departmental analyses. The individual will not need project management experience but should have ample hands-on experience manipulating and visualizing data.

Technical Resource Hiring

The State should hire an internal technical resource who will assist with the overall management of the OceanSpaces platform, both from an organizational and technical point of view. The individual should have a solid understanding of the technical components and tools required of the platform in order to be able to guide the State on decision-making and strategic direction of the platform. The individual will also be responsible for overseeing deployment of the platform, managing vendor relationships, and resolving issues. The individual will act as the key liaison between the Ocean Protection Council and other departments, such as the CNRA and DFW/FGC as required, and assist with tasks such as providing technical support to the internal user base, and helping to facilitate data analyses. The individual will also be responsible for developing the map-based search and visualizations, potentially with assistance from State resources or third party services.

Below is a sample job description for the role:

Responsibilities:

Provide day-to-day management of the MPA Data Platform and ensure platform meets the evolving needs of its user community. This includes providing advice, guidance, and support of the MPA platform infrastructure, development of new features and capabilities, installation of new applications, issue resolution and bug fixes, platform monitoring and troubleshooting. Provide platform capacity planning and optimization. Design, implement, and maintain platform security and data management. Manage vendors as required. Work in collaboration with the CNRA's IT group to migrate and stand up MPA Platform in ODP environment.

Qualifications
- Project management experience deploying platforms for internal projects or external facing customers
- Familiarity with Open Data Platforms, particularly CKAN environments
- At least 3 years of experience with web development including Python, Javascript, HTML, CSS
- Hands on experience with the Hadoop stack (MapReduce, Sqoop, Pig, Hive, Flume)
- Front end skills, particularly in data visualization
- Experience in managing complex, distributed systems
- Experience with Java, Pig, Hive, or other languages a plus
- Demonstrated ability to apply problem analysis and resolution techniques to complex system problems
- Drupal familiarity a plus
- Environmental or marine ocean protection interest a plus

CNRA

The OPC should define SLAs for expected levels of service including:

- Uptime – uptime should be at least 97%
- Issue Resolution – the CNRA should guarantee some level of response time when an issue is detected, e.g., a 1 hour response time and 24 hour resolution period
- Resource guarantee – CNRA should guarantee available technical support services for business hours and designate an individual who will act as point of contact for OPC.

The OPC should also clearly understand the services CNRA will provide regarding the initial data migration and ongoing support CNRA will provide through OpenGov such as:

- Number of datasets allowed
- Number of views and visualizations allowed
- Number of users allowed
- Support for geospatial data
- Graphic design and website UI buildout
- Administrative and user training
- API support
- Dashboards and reporting
- ArcGIS Connection
- Custom Data Exploration and Visualization capabilities
- Professional Services hours

Vendors

Several vendors provide services to build and manage open data platforms. OPC should utilize vendors that provide managed services on a CKAN platform; these include OpenGov, the vendor CNRA is utilizing to build out their ODP platform, and Viderum, the commercial subsidiary of Open Knowledge.

These vendors provide managed services to build and maintain open data platforms. Services include:

- Open data portal development
- Data storage
- API development
- Web interface development and design
- Data ETL and visualization
- Training and education
- Ongoing support and maintenance

Cost

Below describes a sample budget comparing current OST costs to revised costs based on an estimate of required resources to support the MPA platform under the recommendations provided in this document. These figures are estimates only and subject to actual pricing from the CNRA, Open Data Platform vendor, and state resource rates.

World Wide Technology, Inc.

	Current Costs - OST Budget[1]			Future State - Revised Estimate[2]		
	Year 1	Ongoing	Notes	Year 1	Ongoing	Notes
Infrastructure Costs						
Data & metadata storage	5,000	-		250,000	100,000	CNRA's cost - includes ODP and ADP
Imagery storage and access	-	100,000		-	-	
Infrastructure Subtotal	**5,000**	**100,000**		**250,000**	**100,000**	
Ocean Science Trust FTEs						
Program Manager, Technology & Information Systems	84,000	84,000	30% OST FTE	84,000	84,000	Budgets for one resource from OST to continue to oversee the Drupal platform
Program Manager, MPA Monitoring	42,000	42,000	15% OST FTE	-	-	
Communications Coordinator	27,500	27,500	10% OST FTE	-	-	
Data & Imagery Manager	128,000	256,000	100% OST FTE	-	-	
Associate Scientist	95,000	95,000	35% OST FTE	-	-	
OST Subtotal	**376,500**	**504,500**		**84,000**	**84,000**	
Ocean Protection Council FTEs						
Program Manager, Technology & Information Systems	-	-		120,000	120,000	OPC MPA Platform Manager[3]
Developer/Programmer	-	-		50,000	100,000	Begins 6 months into project[3]
OPC Subtotal				**170,000**	**220,000**	
Maintenance and Development Costs						
OceanSpaces Maintenance and other Operating Costs	90,000	30,000		15,000	15,000	Provide ongoing maintenance of Drupal
Data Management Architecture Development	35,000	-		-	-	
Map-based Data Discovery Tool	40,000	-		-	-	
Data Management System Development	-	50,000		50,000	30,000	3rd party CKAN vendor costs
Maintenance and Development Subtotal	**165,000**	**80,000**		**65,000**	**45,000**	
Grand Total	**546,500**	**684,500**		**569,000**	**449,000**	

1. Figures taken from Data Management Plan published by Ocean Science Trust
2. All figures listed are estimates only for discussion
3. Estimates subject to State rates

Data Management

The purpose of data management is to ensure that data is collected, analyzed, and stored in a way that enables an organization to use it securely, consistently, and accurately. In a heterogeneous data environment like the State of California which consists of data from multiple sources, a coordinated approach to data management is critical to ensure data is used to its fullest potential.

A thoughtful approach to data management can help the State of California achieve its mission to provide best-in-class services to its citizens:

- Ensure data resources across the department are properly utilize to achieve department goals through improving interoperability and integration of data systems
- Enable data to be accessible and accurate throughout the state to support informed decision-making
- Maximize data value and minimize data risk through securely protecting confidential citizen data

Data environments can be highly variable and fragmented. When divisions within an organization have individually taken their own approach to data management, it can be difficult to consolidate them into a cohesive strategy. A top-down Data Management Strategy will seek to provide a coordinated framework to leverage data to create better decision- and policy-making.

Getting Started

The first step to designing a data management strategy is to bring departments together and facilitate discussion on specific goals and objectives that the data management strategy should accomplish. The early adopters of the CNRA's platform, including the California Water Resources, and California Energy Commission, should each identify data owners within their organization who can form an initial working group and are prepared to voice each organization's needs regarding data management. Data owners should have intimate knowledge of how each department uses data, have familiarity with existing systems, and understand current data management efforts, if any.

The next step is agree on shared goals and objectives for using data. This includes understanding who within the organization uses data, why, and how. It also includes discussing what data exists, and how the agencies may benefit from interdepartmental sharing of data. It is important to prioritize data management needs; target specific goals and align initial policies and standards with accomplishing the most important goals.

The working group should then begin to define policies and standards that can help form the data management strategy to accomplish the highest priority goals.

These include:

- Data Quality – setting data quality standards, cleansing processes, automated checks
- Data Standards – defining naming standards, data modeling standards, rules for data usage and data definitions

- Access and Authorization – setting role-base access standards, such as privileges to sensitive or confidential data
- Privacy, Compliance, Security – security around protecting sensitive data and ensuring compliance with any regulations
- Data Architecture and Integration – defining where data will sit and how it will be logically or physically separated to protect sensitive areas or domains

When the departments are ready to begin sharing data and doing cross-departmental analyses, it will be important to ensure the data is easily to find and access. Creating a glossary of available data will ensure data is searchable and publishing naming standards for each organization will help users identify the data they require.

Data Governance Leadership

It is vital to have people in place who will be responsible for creating policies and observing progress as the platform matures. A leadership team, sometimes called a 'Data Governance Council', should be formed to establish policies and procedures, ensure policies are enforced, and be accountable for data governance within their department.

Data stewards should also be identified to act as the flag bearers for the data management policies. They are responsible for driving adoption and implementation of policies within a specific division or group within a department. Data stewards are responsible for ensuring effective local protocols are in place to guide the appropriate use of data and that data is clean and processed to maintain data integrity. Data stewards can be data owners of large data sets who are already familiar with current processes and workflow.

Data QC

Currently, some minimal data QC occurs within the data upload process itself (e.g., 5 metadata files must be provided if 5 data files are uploaded) but most QC occurs post-upload by manual checks. These include:

- Correct metadata provided
- # data tables, # look up tables, metadata match
- Correct names for packages
- Complete project information form filled
- Methods file present
- Packages open without error

Automated scripts can be written to reduce manual labor. The first script that should be created is a check for 0 byte files. Next, more complex scripts can include text parsing to check that appropriate content of files. Some manual spot checking should still occur.

Currently, there is little to no data quality checking to ensure the data is statistically accurate as data accuracy is the responsibility of the uploader. Tools such as Ataccama can automatically profile data for pattern analysis, business rules checks, and frequency of values. These should be explored as data grows in complexity.

Additional Tools Required

Visualization Tools

CKAN has several built in tools that allow users to preview and graph the data online before they download it. The CKAN page for a dataset can contain one or more visualizations of the data or file contents (a table, a bar chart, a map, etc).

Different view types are implemented via plugins. Plugins include:

Data Explorer: Displays data as a table; an example can be seen at http://catalogue.beta.data.wa.gov.au/dataset/australian-sea-lion-monitoring/resource/915df47f-0ae7-49e5-9d75-c43cd9286243

Figure 3: Screenshot of CKAN data preview capability

Graphs: Data that is provided in tabular format can be displayed as customizable graphs through user selected fields for X and Y axis; example can be seen here: http://catalogue.beta.data.wa.gov.au/dataset/2014-15-western-australia-budget-economic-and-fiscal-outlook-table-and-chart-data/resource/4571c3b6-1ece-487c-9ac6-9fd352efc8ed

Figure 4: Screenshot of CKAN graphing capability

Geospatial Capabilities

CKAN has many geospatial features that include map-based search, map-based data display and integration with other geospatial platforms.

As mentioned previously, CKAN has several geospatial-related plugins that add geospatial search capabilities to CKAN search interface. These include a geo-indexing plugin, which allows data to be queryable by location, and spatial search plugin, which allows users to limit search results by selecting a boundary on a map.

Spatial data can be displayed as a map through a Web Mapping Service (WMS) extension; example can be seen here: http://catalogue.beta.data.wa.gov.au/dataset/dpaw-managed-lands-and-waters/resource/195af62a-092c-309c-9f99-a14bf40e78c1?view_id=40bd6238-b599-4d5b-bfa8-5021de24ccfa
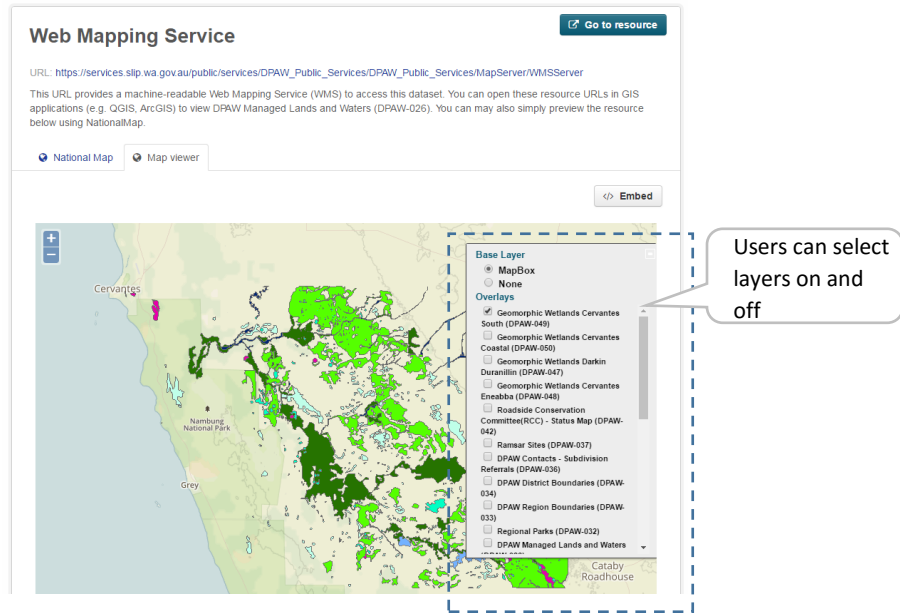
Figure 5: Screenshot of CKAN imbedded map capabilities

Additional open source frameworks exist for mapping data. CKAN has plugins to several, including OpenLayers (http://openlayers.org/) and Leaflet (http://leafletjs.com/), both of which have libraries of Javascript features for map-based data display. An example can be seen at: https://data.lacity.org/A-Prosperous-City/Map-of-Restaurants/ycz4-j47g
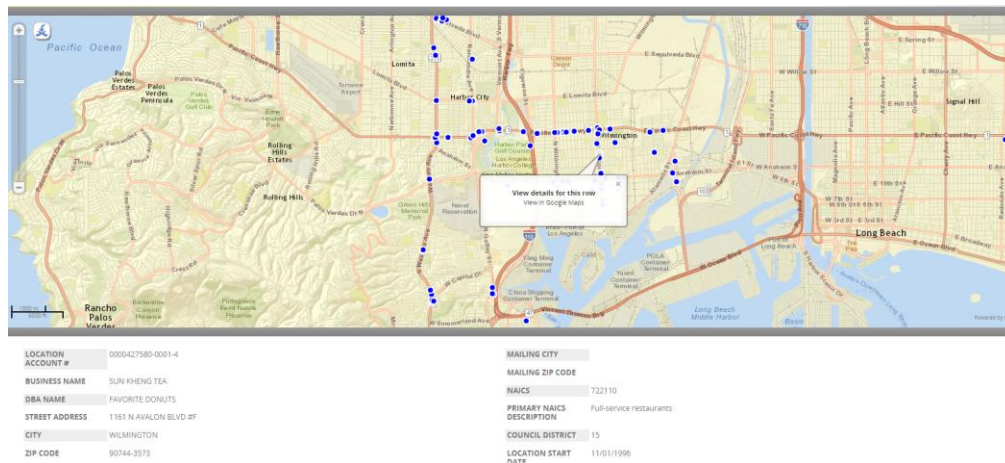

Figure 6 Screenshot of a map imbedded in CKAN displaying restaurants in the L.A. region

Moreover, CKAN can be integrated with other geospatial platforms and import metadata in a number of formats. CKAN can be federated with an ArcGIS portal and harvest data from an ArcGIS Open Data catalog; more documentation can be found here: http://doc.arcgis.com/en/open-data/provider/federating-with-ckan.htm.

Custom Dashboards

Dashboards to display data for a specific use case can be created in many visualization software tools that exist on the market.

CKAN provides some rudimentary dashboard capabilities (see https://github.com/ckan/ckanext-dashboard), but more sophisticated software exists.

The State has a Tableau license, which can be used to build non-editable public-facing dashboards that can display data in a wide variety of formats. Some learning is required to familiarize users with the tool, but it can allow the State to build a broad spectrum of visualizations.

In addition, power users can utilize Tableau for internal purposes, such as deeper analysis for MPA program assessments or to monitor important metrics. Multiple users across departments can collaborate and share analyses internally, as well as join data from multiple departments.

Google Maps provide a flexible and free API that can be used in conjunction with Tableau for dynamic mapping features. See https://developers.google.com/maps/documentation/javascript/examples/layer-fusiontables-query for a simple example of types of features that can be overlayed onto a Google Maps instance. Javascript libraries such as jQuery, DOJO, JSCharts can be used to create more visually appealing features.

Image Storage

The ODP and ADP can be utilized for image storage. Storage capabilities should be sufficient for large file sizes.

KNB vs DataONE

OceanSpaces should continue to push data to KNB until the CNRA platform is sufficiently large to warrant becoming a member node of DataONE. KNB is an appropriate choice for organizations who desire the support of an established platform as KNB provides stability, storage, data governance, security, replication, as well as a community of users who access KNB to search for data.

DataONE does not provide any storage capabilities beyond backup data replication services, and is a centralized catalog that contains aggregated metadata from all of its member nodes. Data is downloaded from the member's repository. This allows the platform to be scalable and keep data at the edge of the network.

As OceanSpaces grows in user base and its infrastructure is stable enough to stand on its own, it should explore becoming a member node of DataONE. Some implementation is required to pass the requirements to become a member; DataONE provides a four-stage roadmap which includes planning, developing, testing, and operating, before an organization can fully be integrated into the DataONE network. More detail can be found here: https://www.dataone.org/member-node-deployment-process.