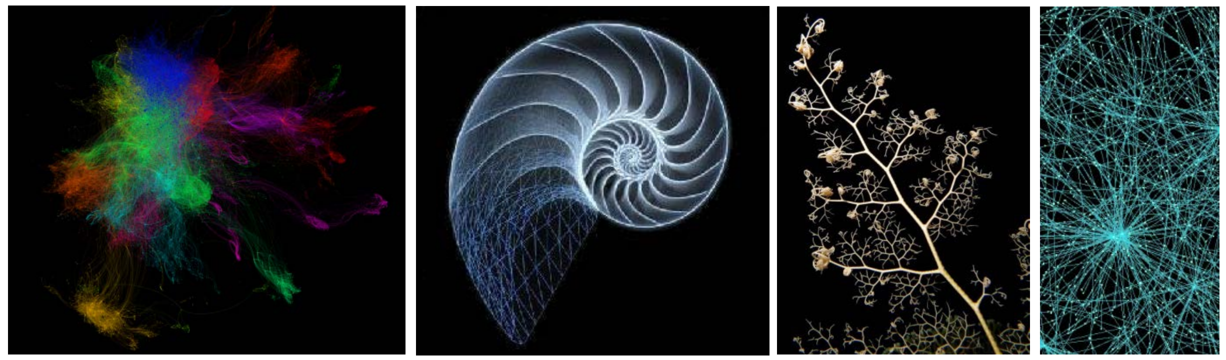
Marine Biodiversity Hub

National Environmental Science Programme

Project B3 - Enhancing access to relevant marine information: developing a service for searching, aggregating and filtering collections of linked open marine data – Scoping study

Johnathan Kool, Geoscience Australia

Milestone report
25 November 2015



www.nespmarine.edu.au

Enquiries should be addressed to:
Johnathan Kool
Geoscience Australia
Cnr Jerrabomberra Ave and Hindmarsh Drive
Symonston, ACT 2609

Tel: 02 6249 5842
e-mail: johnathan.kool@ga.gov.au

Copyright

Acknowledgement

This work was undertaken for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Research Program (NESP). NESP Marine Biodiversity Hub partners include the University of Tasmania; CSIRO, Geoscience Australia, Australian Institute of Marine Science, Museum Victoria, Charles Darwin University, the University of Western Australia, Integrated Marine Observing System, NSW Office of Environment and Heritage, NSW Department of Primary Industries.

Important Disclaimer

The NESP Marine Biodiversity Hub advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, the NESP Marine Biodiversity Hub (including its host organisation, employees, partners and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

This project seeks to improve the searchability, discoverability and delivery of marine information through the development of an online service capable of searching, filtering and organizing linked open marine data. The service would also provide the capability of forwarding the collections of discovered data to web services for subsequent processing into products of higher utility. This work will improve access to existing data collections, and will facilitate the development of new applications by acting as an aggregator of links to sources of marine information. The work will benefit managers (i.e. Department of the Environment staff) by providing fast and simple access to a wide range of relevant marine information products, and offering a means of quickly synthesizing and aggregating information from multiple sources.

1. INTRODUCTION

An emerging priority in information management is building smarter information search engines that are tailored to specific types of end users. These end-user-focused systems can provide highly targeted and relevant results immediately, in contrast to the common experience of sifting through extensive collections of potentially irrelevant items.

Currently, Australia has a broad array of scientific information products related to the marine environment. Examples not only include geospatial (GIS) information, but also products such as documents, websites, tables, images and video. Furthermore, different user groups will have different approaches when looking for information. For example, one user group may have a geographic focus (e.g. Western Australia, Great Barrier Reef), another may have a policy focus (water quality management, conservation planning), or alternatively, there may be particular species of interest (e.g. sharks, crown-of-thorns starfish). It is also possible that users may wish to identify materials that combine the different factors (e.g. find information pertaining to sharks and conservation planning, but do not include whale sharks, items related to the Great Barrier Reef, or video). Traditional methods of information delivery (e.g. portal data lists) present limited options for sifting through large data collections in a manner flexible enough to accommodate a variety of search strategies and entry points.

To advance our ability to utilise marine information, this report sets out a plan of work to establish an online service that can:

- Search arbitrary collections of linked data using a search string (a ‘Google’ search)
- Bin/Classify the returned items into categories of interest (faceting)
- Return items in order of relevance with respect to the search string
- Filter/subset the collection of returned items
- Provide the ability to forward the information subset on to other services for further processing

We intend to develop a flexible interface focused on the information requirements of the Department of the Environment to search, filter and deliver connections to linked open marine data (this tool could also be applied to data within the Department). This will help provide efficient access to a wide range of information sources using a Google-type search interface that will have an intuitive feel for non-specialist users. The user experience should be simple, familiar, and straightforward – no more complicated than interacting with shopping interfaces such as Amazon or eBay, yet capable of letting users quickly and easily find and access collections of information that are relevant to their needs. The interface will accept a single search string, and will return an ordered/sorted list of results. The resulting collections of data resources can then be forwarded to other web services for plotting, ingestion into modelling tools and virtual laboratories, or saving as a report. This contrasts with previous efforts to generate map ‘portals’ by providing a targeted subset of information tailored to individual needs that can be updated dynamically (akin to an Amazon search for marine information as opposed to books and household goods).

By enhancing the accessibility and utility of a raft of marine information the proposed project aligns with Marine Biodiversity Research Priorities to:

- Develop and trial decision making tools that will support managers to define and prioritise activities.
- Provide meaningful and accessible information on the status and trends of key social and economic values associated with the marine environment.
- Improve our knowledge of key marine species and ecosystems to underpin their better management and protection
- Enhance the role of citizen science in the management of marine biodiversity.

The enhanced search capability of the system will be enabled through the development and use of semantically-enabled data. Semantically-enabled data is information that is structured in such a way that it is possible for machines to understand its meaning without human intervention. Using semantically-enabled data, new opportunities emerge for analysing and delivering information resources stemming from the ability for computer systems to aggregate and filter items in an automated manner, and to identify connections across a network of linked concepts. For example, linking projects, organizations, research subjects, spatial and temporal information and information delivery capabilities as a network of relationships. The better-described the information, the greater the improvement in terms of its capabilities for application to specific information needs.

2. REVIEWING EXISTING LINKED OPEN DATA CAPABILITY

Although there are some ongoing linked open data initiatives within the Australian Government, the present use of linked open data is not widespread. The W3C, the principal international standards organization for the World Wide Web, has developed the following assessment criteria for linked data:

★	On the web – available on the Web, whatever format, under an open license
★★	Machine-readable – structured data, e.g. Excel instead of image scan of a table
★★★	Non-proprietary formats – e.g., CSV instead of Excel
★★★★	RDF standards – RDF & SPARQL using HTTP URIs
★★★★★	Linked RDF – RDF linked to other data to provide context

Data.gov.au maintains catalogue metadata in RDF form (e.g. Water Quality Zones), however this information is metadata only, and is not linked to other data objects, meaning that presently, the collection only conforms to the four-star standard. Geofabric, a BOM project for representing key inland hydrological features has fully linked RDF, as do the People, Datasets and Licenses collections of the Bioregional Assessments program (a collaboration among the

Department of the Environment, CSIRO, the Bureau of Meteorology and Geoscience Australia). Note that these projects have no direct relation with the marine environment, however they do demonstrate that there is both interest and uptake of the linked open data concept by Australian Government agencies for the purpose of better managing and utilising environmental information. While some work has been done to incorporate linked open data as part of the eReefs program (Car, 2013), within the various agencies and organizations responsible for collecting and curating marine data the move towards taking advantage of linked open data is still in its early stages. However, already in this project we are able to demonstrate a workflow that is capable of taking existing metadata and documents, and enhancing them in an automated manner, and have executed this workflow on existing Marine Biodiversity Hub resources.

3. CONCEPTUAL DEVELOPMENT

The linked open data delivery system will be split into two components: (i) back-end software accessed through an application programming interface (API), and; (ii) a front-end web-based graphic user interface (GUI). The term ‘API’ refers to the interface between different software programs, facilitating their interaction, similar to the way the user interface facilitates interaction between humans and computers. The API provides a mechanism for communicating and interacting with software through a prescribed set of programming commands. The web-based GUI is designed to take human input, and translate it into those same programming commands. This way, the capabilities of the system can be leveraged by humans in an interactive manner, or by machines in an automated manner.

4. BACK-END APPLICATION PROGRAMMING INTERFACE (API)

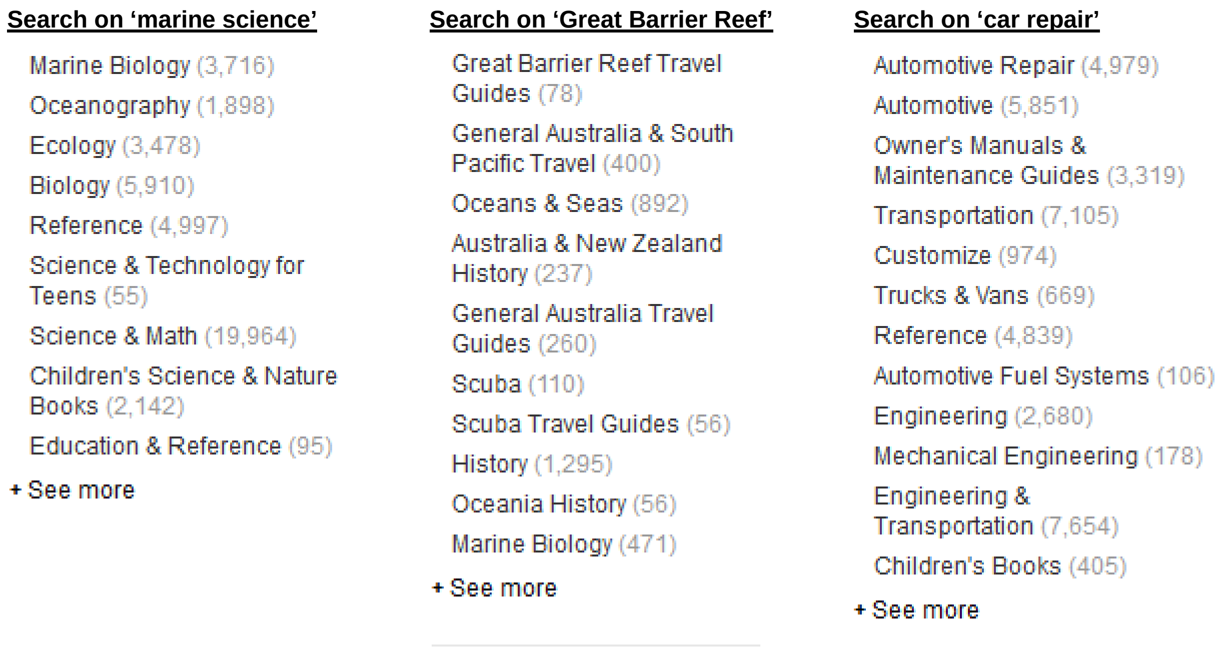
4.1 Sources of information

The fundamental operating unit of the linked open data delivery system is the *document*. Documents can consist of full text (e.g. a traditional text document), or tagged information (e.g. web page content, or metadata), and the format of the document may vary (e.g. Text, XML, JSON, JSON-LD etc.). The distinguishing feature of a document is that it contains information that can be represented as text, as opposed to binary or encoded data. The text-based nature of documents ensures that irrespective of their internal structure, their content can be searched and parsed in a consistent manner; documents, regardless of their internal structure, will always be composed of strings of characters.

4.2 Document enhancement with semantic tagging

Although documents (whether they contain metadata or text) do contain a great deal of information, the information that they contain is typically not structured in such a way that makes it possible to dynamically organize collections of documents based on their content. By ‘dynamic organization’ we mean that the resources can be partitioned and grouped in a permutable manner. That is , it is possible to organize a collection of resources on the basis of geography (e.g. state), sample gear type and taxa sampled (in that order) or on the basis of sampling organization, taxa sampled, measurement type and geography (in that order). By way of example, Amazon.com returns dynamic groups (also referred to as *dynamic facets*) when searching its product catalogue (Figure 1). Thus, searching Amazon's book list for the topic 'marine science' generates subcategories such as: *Marine Biology*, *Oceanography*, *Ecology* and *Science & Math*. A search on 'Great Barrier Reef' splits into *Great Barrier Reef Travel Guides*, *Oceans and Seas*, *Oceania History*, *Marine Biology* and *Science & Math* (among others). In contrast, a search on 'car repair' returns classes such as *Automotive Repair*, *Trucks & Vans* and *Engineering*. The key point here is that a fixed classification system that would be meaningful when searching for books on car repair would be neither useful nor appropriate for classifying and grouping information resources related to the Great Barrier Reef. A way of categorizing the information dynamically at search time is needed. Through the searches on 'marine science' and 'Great Barrier Reef', it also becomes clear that there are multiple avenues of connecting to the same topic – for instance, the concept of 'Marine Biology', accessed via a subject-based search and a location-based search respectively.

Figure 1 – Examples of dynamic faceting from Amazon.com. The first search uses the search term 'marine', the second uses 'Great Barrier Reef', and the third uses 'car repair'



Search on 'marine science'

- Marine Biology (3,716)
- Oceanography (1,898)
- Ecology (3,478)
- Biology (5,910)
- Reference (4,997)
- Science & Technology for Teens (55)
- Science & Math (19,964)
- Children's Science & Nature Books (2,142)
- Education & Reference (95)

+ See more

Search on 'Great Barrier Reef'

- Great Barrier Reef Travel Guides (78)
- General Australia & South Pacific Travel (400)
- Oceans & Seas (892)
- Australia & New Zealand History (237)
- General Australia Travel Guides (260)
- Scuba (110)
- Scuba Travel Guides (56)
- History (1,295)
- Oceania History (56)
- Marine Biology (471)

+ See more

Search on 'car repair'

- Automotive Repair (4,979)
- Automotive (5,851)
- Owner's Manuals & Maintenance Guides (3,319)
- Transportation (7,105)
- Customize (974)
- Trucks & Vans (669)
- Reference (4,839)
- Automotive Fuel Systems (106)
- Engineering (2,680)
- Mechanical Engineering (178)
- Engineering & Transportation (7,654)
- Children's Books (405)

+ See more

This capacity for dynamic and flexible sub-setting is important, because it gives users the ability to find, group and filter collections of information via searching in an unstructured manner, and letting the computer system perform the tedious, but helpful work of sorting the collection of information into natural subgroups within the context of the search chain.

Dynamic faceting is possible if terms within documents have been marked up with tags that distinguish the different forms of content within the document. Examples can include common tags such as 'Title', 'Abstract' and 'Authors', but any word or phrase can be categorised using an arbitrary number of classifier tags.

4.3 Automating the enhancement process

In order to take full advantage of document enhancement and semantic tagging, it is essential to have a means of automatically marking up documents and metadata. Manually tagging documents would require prohibitive amounts of (unnecessary) effort on the part of data providers, and would certainly be unsustainable in the long term as new data sets continually arise. A more practical approach is to take advantage of existing software packages that are capable of automatically extracting key terms from structured documents (Named Entity Recognition – NER), and then linking those terms with established term lists (vocabularies). Existing software with this capability include OpenNLP (Java) and the Natural Language Toolkit (Python).

Vocabularies are used to provide a controlled list of terms that are able to unambiguously identify different objects, concepts or actions in a form that lends itself to automation and interpretation using computer systems. The controlled nature of vocabularies ensures that there is a defined structure that machines can use to organize the information, while the modular nature of vocabularies also ensures that they can be modified, updated and replaced as necessary, allowing the system to evolve over time.

The following are examples of existing vocabulary services that could be used as part of this project:

Dbpedia	A translation of Wikipedia entries into linked open data form
Geonames	Over 8 million placenames as linked open data
CSIRO Linked Data Registry	Australian scientific vocabulary registry
NERC Vocabulary Server	Lists of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic and wider community.
CF Standard Names	Standard names when working with NetCDF and climate data
IGSN	A 9-digit alphanumeric code that uniquely identifies samples from the natural environment and related sampling features
ORCID	Provides a persistent digital identifier for researchers
QUDT	Quantities, units, dimensions and data types ontology.
Dublin Core	Terms used to describe web resources, physical resources and objects.
SKOS	A common data model for sharing and linking knowledge organization systems
OWL	Provides a formal way of describing taxonomies and classification networks.
FOAF	A machine-readable ontology describing persons, their activities and their relations to other people and objects
Vcard/Organization Ontology	Specifications for describing people and organisations

It is also possible to generate new vocabularies to specify terms and concepts that have not been captured elsewhere, but it is usually best to take advantage of previous work whenever possible to avoid duplicating effort, and to build user consensus around the term lists. Additional formal vocabularies that could be developed and incorporated into this work could include Australian government entities (agencies, organizations and units); key policies, laws and regulations (e.g. linking to ComLaw), and defining key marine biophysical features of interest.

4.4 Linking data

A critical aspect of vocabularies is that they are not simply word lists and classifiers, but can also link to documents that provide a formal description of their subject, with their own links to related resources (identified using Uniform Resource Indicators – URIs – e.g. http://xxx.yyy). Presently, the commonly used format for these objects is RDF (Resource Description Framework). RDF provides a specification for structuring documents, based on the idea of making statements about resources in the form of subject-predicate-object expressions known as *triples* (e.g. Fig 2). A *subject* is the resource being described, and a *predicate* defines the nature of the relationship between the *subject* and the *object*. For example, a document on Geoscience Australia's Oceanic Shoals Survey might represent the concept that "the National Environmental Science Program is a program under the Department of the Environment" as *subject*: "National Environmental Science Program"; *predicate*: "is a program of"; and *object*: "Department of the Environment". Note that the terms "National Environmental Science Program" and "Department of the Environment" may contain links to other vocabulary terms which in turn refer to RDF documents describing these two entities. The result is a network of linked documents that are machine readable, and the connections among them are machine-traversable (Fig. 3). This is the fundamental concept underpinning the notion of linked open data and the 'semantic web'. Note that triple elements can also be prefaced with a vocabulary reference, meaning that it is possible to mix, match and swap the vocabularies being used for term referencing.

4.5 Information delivery as a web service

Although the idea of being able to search, organize, filter and traverse a collection of linked information products is a powerful one, it is essential that these capabilities be accessible in a manner that lends itself to interoperability, scalability, and ease of use. As previously indicated, the interface should be no more complicated than navigating a commercial website such as Amazon.com, and by the same token, should be able to easily plug into and interact with standard web components and services – RESTful APIs offer a means of accomplishing this. REST (Representational State Transfer) is a style of software architecture that operates on the basis of constraining the manner in which different components, connectors, and data elements interact. This is a consistent, whole-of-software industry-recognised framework by which web services operate and communicate with one another – effectively conforming to an operational standard. This approach ensures that developing interfaces that interact with the web service are simple, that individual components can be easily changed and replaced, that individual components are portable, and that the manner in which various components communicate can be clearly interpreted. Reducing the complexity of communication also offers benefits in terms of system-level performance and scalability. RDF output formats include JSON-LD, RDF/XML, RDF/JSON, Turtle format and N-Triples. In particular, XML and JSON are widely recognised and used, and a broad array of tools and software packages exist to work with these formats. For example, XML can be easily filtered and rearranged (e.g. using XSLT), and visualised in a flexible manner through the use of stylesheets (e.g. cascading style sheets - CSS).

5. SOFTWARE ARCHITECTURE OPTIONS

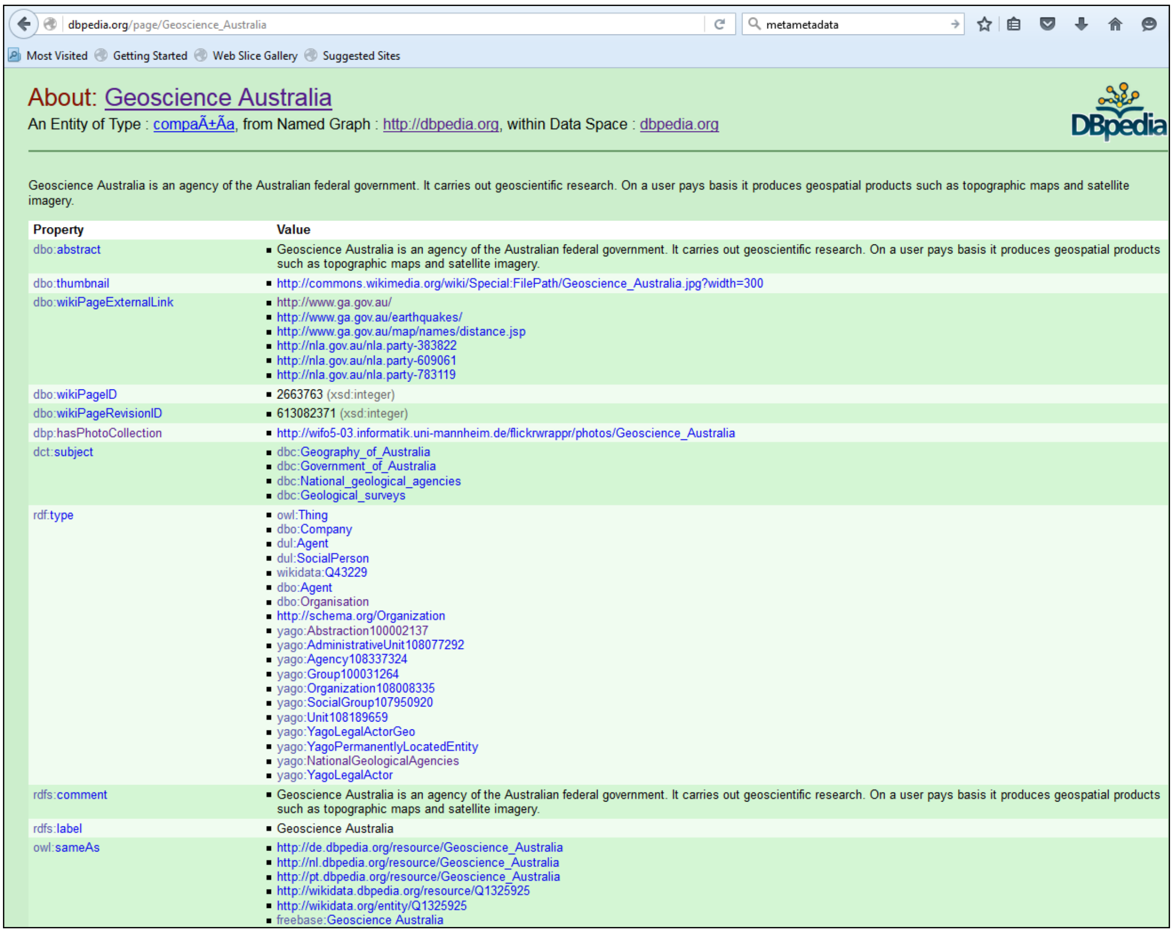
Despite the complexities in delivering the features described above, several software packages have been developed that are capable of delivering these capabilities, including: PoolParty, Topquadrant, CKAN and Apache Stanbol (Table 1).

Table 1 – Cross-comparison of linked open data engine features

	PoolParty	Topquadrant	CKAN	Apache Stanbol
Privately owned?	Yes	Yes	No	No
Open source?	No	No	Yes	Yes
Purchase cost	US$6588/y	$3450+/license/y	Free	Free
Language	Java	Flex	Python	Java
Linked Data	Yes	Yes	Yes	Yes
Tag Vocabularies	Yes	Yes	Yes	Yes
Web management interface	Yes	Yes	Yes	Yes
Natural Language Processing module	Yes	Yes	No	Yes

Stanbol was selected as our test development environment of choice owing to the free, open source nature of the software, its use of Java as a development language, and its natural language processing capabilities.

Figure 2 – Example rdf for 'Geoscience Australia' from DBPedia. DBPedia is a project that is automatically translating Wikipedia entries into linked RDF data.



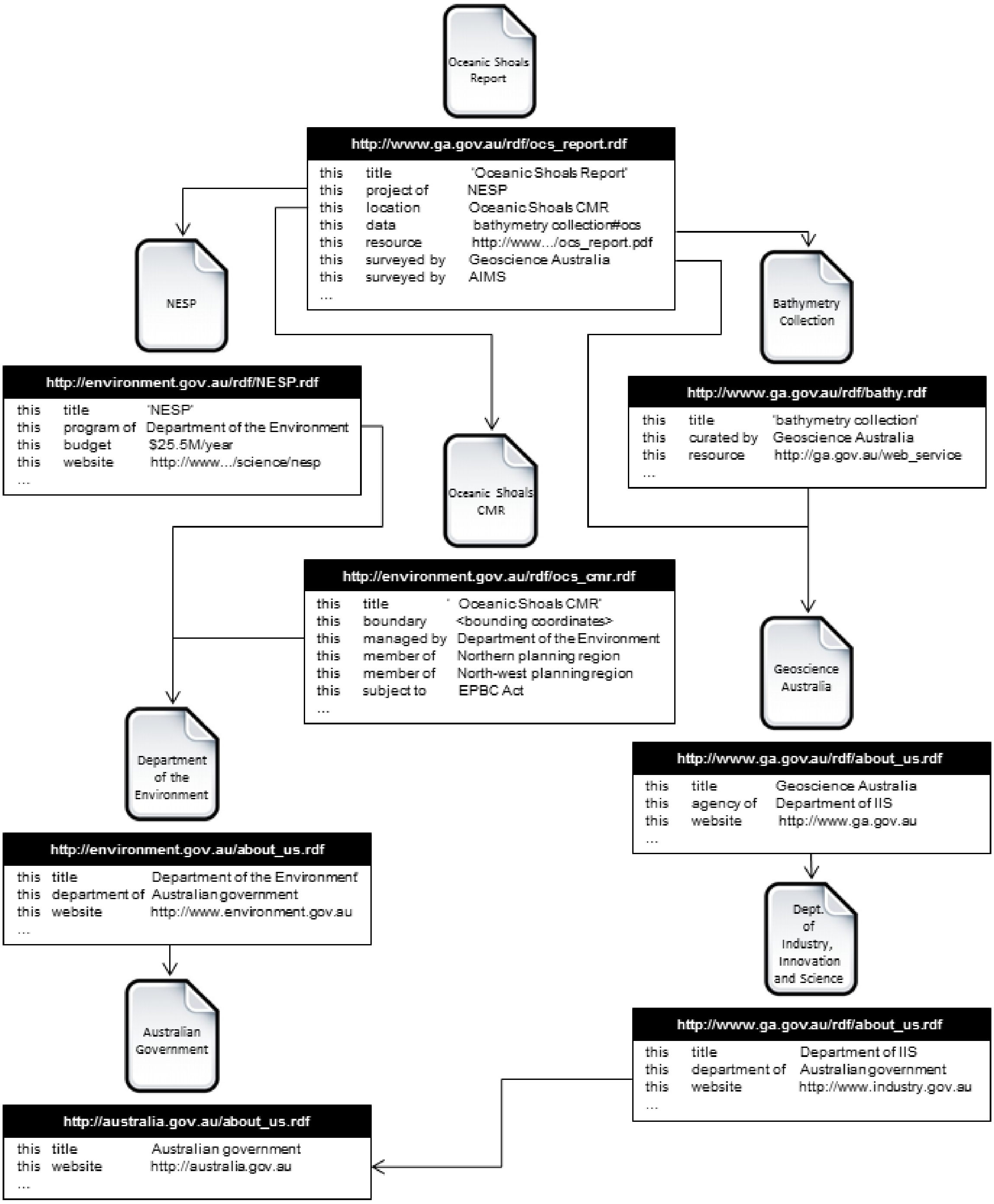
About: Geoscience Australia

An Entity of Type : compaÃ±Ãa, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Geoscience Australia is an agency of the Australian federal government. It carries out geoscientific research. On a user pays basis it produces geospatial products such as topographic maps and satellite imagery.

Property	Value
dbo:abstract	▪ Geoscience Australia is an agency of the Australian federal government. It carries out geoscientific research. On a user pays basis it produces geospatial products such as topographic maps and satellite imagery.
dbo:thumbnail	▪ http://commons.wikimedia.org/wiki/Special:FilePath/Geoscience_Australia.jpg?width=300
dbo:wikiPageExternalLink	▪ http://www.ga.gov.au/ ▪ http://www.ga.gov.au/earthquakes/ ▪ http://www.ga.gov.au/map/names/distance.jsp ▪ http://nla.gov.au/nla.party-383822 ▪ http://nla.gov.au/nla.party-609061 ▪ http://nla.gov.au/nla.party-783119
dbo:wikiPageID	▪ 2663763 (xsd:integer)
dbo:wikiPageRevisionID	▪ 613082371 (xsd:integer)
dbp:hasPhotoCollection	▪ http://wifo5-03.informatik.uni-mannheim.de/flickrwrappr/photos/Geoscience_Australia
dct:subject	▪ dbc:Geography_of_Australia ▪ dbc:Government_of_Australia ▪ dbc:National_geological_agencies ▪ dbc:Geological_surveys
rdf:type	▪ owl:Thing ▪ dbo:Company ▪ dul:Agent ▪ dul:SocialPerson ▪ wikidata:Q43229 ▪ dbo:Agent ▪ dbo:Organisation ▪ http://schema.org/Organization ▪ yago:Abstraction100002137 ▪ yago:AdministrativeUnit108077292 ▪ yago:Agency108337324 ▪ yago:Group100031264 ▪ yago:Organization108008335 ▪ yago:SocialGroup107950920 ▪ yago:Unit108189659 ▪ yago:YagoLegalActorGeo ▪ yago:YagoPermanentlyLocatedEntity ▪ yago:NationalGeologicalAgencies ▪ yago:YagoLegalActor
rdfs:comment	▪ Geoscience Australia is an agency of the Australian federal government. It carries out geoscientific research. On a user pays basis it produces geospatial products such as topographic maps and satellite imagery.
rdfs:label	▪ Geoscience Australia
owl:sameAs	▪ http://de.dbpedia.org/resource/Geoscience_Australia ▪ http://nl.dbpedia.org/resource/Geoscience_Australia ▪ http://pt.dbpedia.org/resource/Geoscience_Australia ▪ http://wikidata.dbpedia.org/resource/Q1325925 ▪ http://wikidata.org/entity/Q1325925 ▪ freebase:Geoscience Australia

Figure 3 - Example of linked RDF documents. Content within the oceanic shoals report RDF links to items like NERP and the Oceanic Shoals CMR, which in turn link to additional documents.



http://www.ga.gov.au/rdf/ocs_report.rdf

this	title	'Oceanic Shoals Report'
this	project of	NESP
this	location	Oceanic Shoals CMR
this	data	bathymetry collection#ocs
this	resource	http://www.../ocs_report.pdf
this	surveyed by	Geoscience Australia
this	surveyed by	AIMS
…		

http://environment.gov.au/rdf/NESP.rdf

this	title	'NESP'
this	program of	Department of the Environment
this	budget	$25.5M/year
this	website	http://www.../science/nesp
…		

http://www.ga.gov.au/rdf/bathy.rdf

this	title	'bathymetry collection'
this	curated by	Geoscience Australia
this	resource	http://ga.gov.au/web_service
…		

http://environment.gov.au/rdf/ocs_cmr.rdf

this	title	' Oceanic Shoals CMR'
this	boundary	<bounding coordinates>
this	managed by	Department of the Environment
this	member of	Northern planning region
this	member of	North-west planning region
this	subject to	EPBC Act
…		

http://www.ga.gov.au/rdf/about_us.rdf

this	title	Geoscience Australia
this	agency of	Department of IIS
this	website	http://www.ga.gov.au
…		

http://environment.gov.au/about_us.rdf

this	title	Department of the Environment
this	department of	Australian government
this	website	http://www.environment.gov.au
…		

http://www.ga.gov.au/rdf/about_us.rdf

this	title	Department of IIS
this	department of	Australian government
this	website	http://www.industry.gov.au
…		

http://australia.gov.au/about_us.rdf

this	title	Australian government
this	website	http://australia.gov.au
…		

6. PROOF OF CONCEPT – A WORKING EXAMPLE

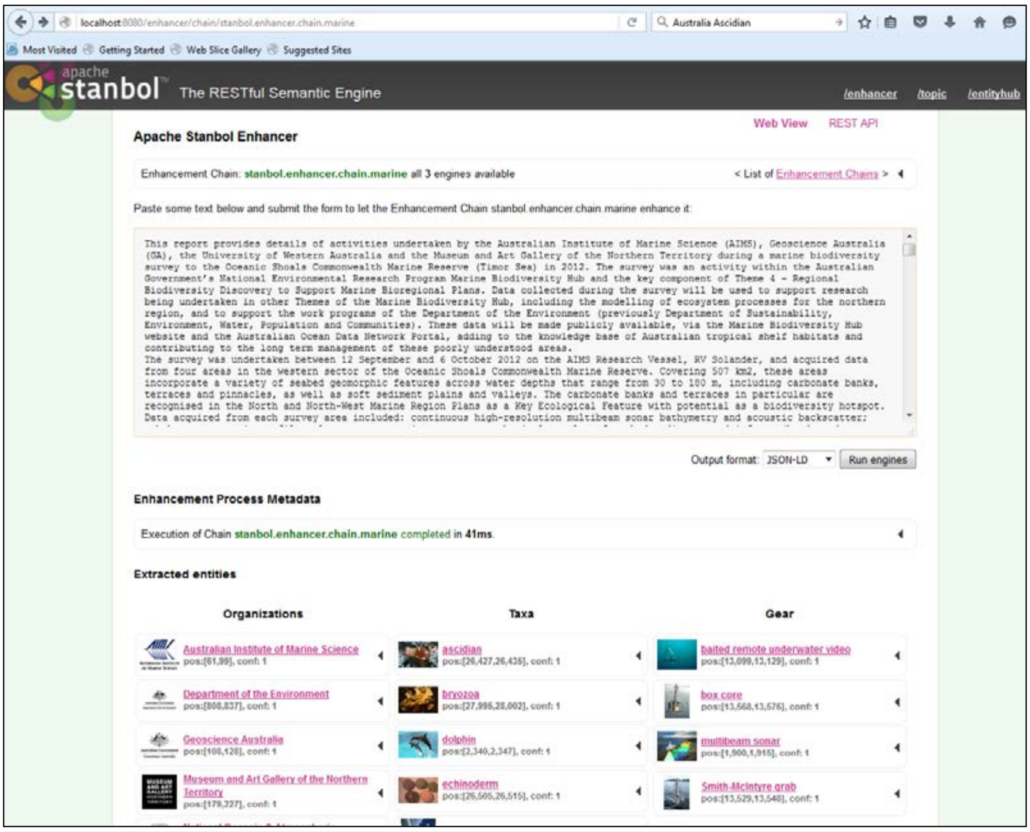
Using Stanbol as a test environment, we demonstrate the dramatic increase in depth and range of information that can be obtained from a typical marine science product through the process of semantically-marking up a document. In this example, we use the metadata record for the *Oceanic Shoals post-survey report* (available at http://www.ga.gov.au/metadata-gateway/metadata/record/gcat_e091142b-a617-0d43-e044-00144fdd4fa6/xml).

Within Stanbol, identifying the document type is first handled by Apache Tika. Tika is a toolkit that is capable of detecting and extracting metadata and text from over a thousand different file types, including Microsoft Office files (.doc, .docx, .xls, .ppt), PDF files, and XML, among others. Next, the document object is passed to OpenNLP so that the content can be parsed into logical units – for example: identifying organizations, proper names, times and dates, places or concepts. This parsing is highly customisable – using anything from simple word list matching to more sophisticated language modelling. Stanbol has been designed to be able to incorporate language models trained using the OpenNLP package.

Once the generic word types have been identified within the document, the terms need to be associated with specific terms in the vocabularies. Within Stanbol, vocabularies can be hosted locally (as a 'Managed Site'), or copied from a remote source (a 'Referenced Site' - a remote Linked Data server). Examples of enhancement using Stanbol's interactive web interface are shown in Fig. 4, with a complete list of extracted terms and associated vocabularies in Fig. 5. Metrics generated as part of the process of identifying and linking terms can also be used to filter the results. For example, metrics on the frequency of word occurrence or confidence in identifying terms can be used to prevent terms with low relevance from being captured in the enhanced document.

In our example, this process is used to enhance/mark up a single document, however, the enhancement process can be repeated on any number of documents to build up a collection. The collection can then be searched using the typical means of keyword searching, or also on the basis of markup type/vocabulary of interest (Fig. 7) Stanbol stores enhanced document information as RDF (Fig .8) within an instance of Solr/Lucene (Solr layers additional functionality on top of Lucene's core document search capabilities. In terms of function, Lucene can be considered as an open-source Google-style search engine).

Figure 4 – Document enhancement by Stanbol via the interactive user interface. The first example is an enhancement of the Oceanic Shoals Survey Report, and the second is an enhancement of the Flinders Survey abstract.



Apache Stanbol Enhancer

Enhancement Chain: stanbol.enhancer.chain.marine all 3 engines available   < List of Enhancement Chains >

Paste some text below and submit the form to let the Enhancement Chain stanbol.enhancer.chain.marine enhance it:

```
This report provides details of activities undertaken by the Australian Institute of Marine Science (AIMS), Geoscience Australia
(GA), the University of Western Australia and the Museum and Art Gallery of the Northern Territory during a marine biodiversity
survey to the Oceanic Shoals Commonwealth Marine Reserve (Timor Sea) in 2012. The survey was an activity within the Australian
Government's National Environmental Research Program Marine Biodiversity Hub and the key component of Theme 4 - Regional
Biodiversity Discovery to Support Marine Bioregional Plans. Data collected during the survey will be used to support research
being undertaken in other Themes of the Marine Biodiversity Hub, including the modelling of ecosystem processes for the northern
region, and to support the work programs of the Department of the Environment (previously Department of Sustainability,
Environment, Water, Population and Communities). These data will be made publicly available, via the Marine Biodiversity Hub
website and the Australian Ocean Data Network Portal, adding to the knowledge base of Australian tropical shelf habitats and
contributing to the long term management of these poorly understood areas.
The survey was undertaken between 12 September and 6 October 2012 on the AIMS Research Vessel, RV Solander, and acquired data
from four areas in the western sector of the Oceanic Shoals Commonwealth Marine Reserve. Covering 507 km2, these areas
incorporate a variety of seabed geomorphic features across water depths that range from 30 to 180 m, including carbonate banks,
terraces and pinnacles, as well as soft sediment plains and valleys. The carbonate banks and terraces in particular are
recognised in the North and North-West Marine Region Plans as a Key Ecological Feature with potential as a biodiversity hotspot.
Data acquired from each survey area included: continuous high-resolution multibeam sonar bathymetry and acoustic backscatter;
```

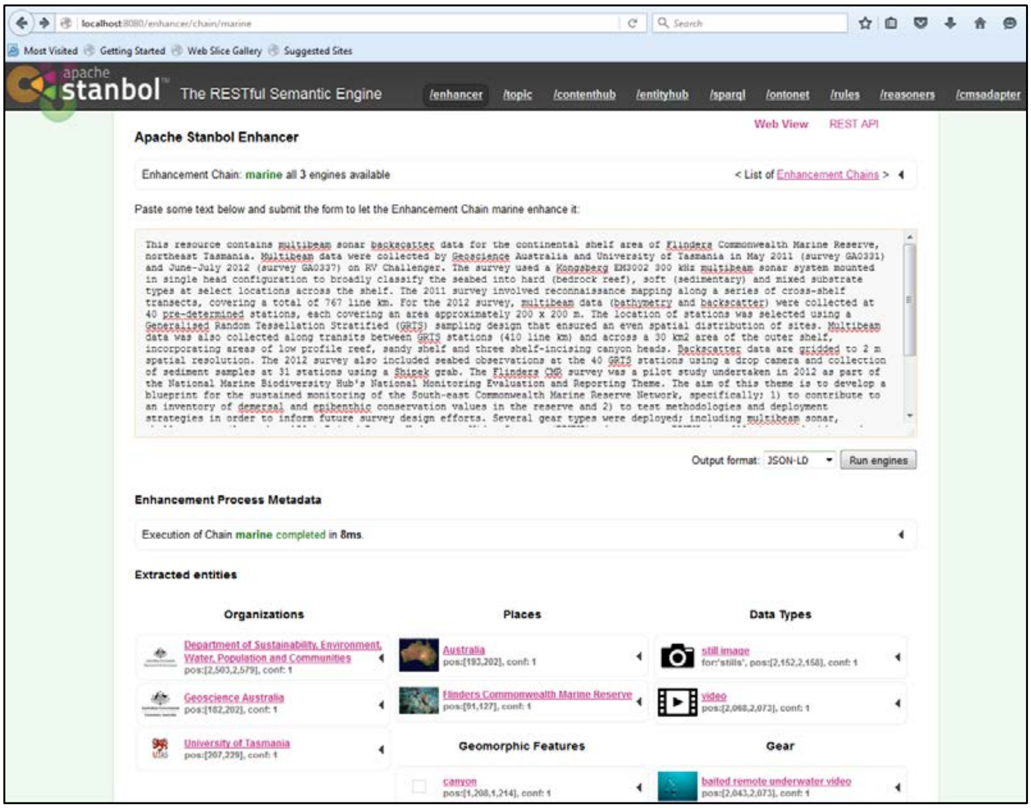
Output format: JSON-LD ▾ Run engines

Enhancement Process Metadata

Execution of Chain stanbol.enhancer.chain.marine completed in 41ms.

Extracted entities

Organizations	Taxa	Gear
Australian Institute of Marine Science pos:[61,99], conf: 1	ascidian pos:[26,427,26,435], conf: 1	baited remote underwater video pos:[13,099,13,129], conf: 1
Department of the Environment pos:[808,837], conf: 1	bryozoa pos:[27,995,28,002], conf: 1	box core pos:[13,568,13,576], conf: 1
Geoscience Australia pos:[108,128], conf: 1	dolphin pos:[2,340,2,347], conf: 1	multibeam sonar pos:[1,900,1,915], conf: 1
Museum and Art Gallery of the Northern Territory pos:[179,227], conf: 1	echinoderm pos:[26,506,26,515], conf: 1	Smith-McIntyre grab pos:[13,529,13,548], conf: 1



Apache Stanbol Enhancer

Enhancement Chain: marine all 3 engines available   < List of Enhancement Chains >

Paste some text below and submit the form to let the Enhancement Chain marine enhance it:

```
This resource contains multibeam sonar backscatter data for the continental shelf area of Flinders Commonwealth Marine Reserve,
northeast Tasmania. Multibeam data were collected by Geoscience Australia and University of Tasmania in May 2011 (survey GA0331)
and June-July 2012 (survey GA0337) on RV Challenger. The survey used a Kongsberg EM3002 300 kHz multibeam sonar system mounted
in single head configuration to broadly classify the seabed into hard (bedrock reef), soft (sedimentary) and mixed substrate
types at select locations across the shelf. The 2011 survey involved reconnaissance mapping along a series of cross-shelf
transects, covering a total of 767 line km. For the 2012 survey, multibeam data (bathymetry and backscatter) were collected at
40 pre-determined stations, each covering an area approximately 200 x 200 m. The location of stations was selected using a
Generalised Random Tessellation Stratified (GRTS) sampling design that ensured an even spatial distribution of sites. Multibeam
data was also collected along transits between GRTS stations (410 line km) and across a 30 km2 area of the outer shelf,
incorporating areas of low profile reef, sandy shelf and three shelf-incising canyon heads. Backscatter data are gridded to 2 m
spatial resolution. The 2012 survey also included seabed observations at the 40 GRTS stations using a drop camera and collection
of sediment samples at 31 stations using a Shipek grab. The Flinders CMR survey was a pilot study undertaken in 2012 as part of
the National Marine Biodiversity Hub's National Monitoring Evaluation and Reporting Theme. The aim of this theme is to develop a
blueprint for the sustained monitoring of the South-east Commonwealth Marine Reserve Network, specifically: 1) to contribute to
an inventory of demersal and epibenthic conservation values in the reserve and 2) to test methodologies and deployment
strategies in order to inform future survey design efforts. Several gear types were deployed; including multibeam sonar,
```

Output format: JSON-LD ▾ Run engines

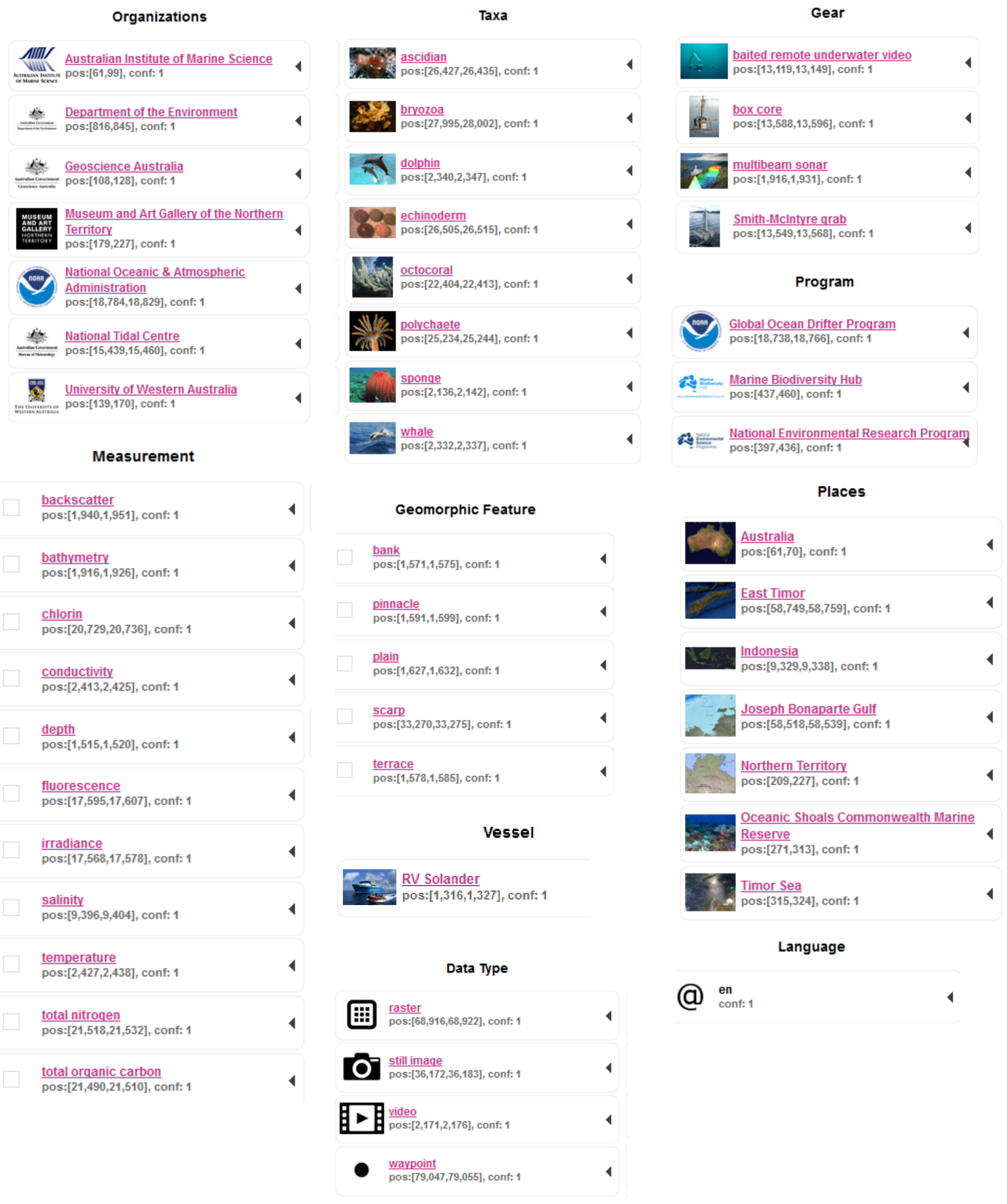
Enhancement Process Metadata

Execution of Chain marine completed in 8ms.

Extracted entities

Organizations	Places	Data Types
Department of Sustainability, Environment, Water, Population and Communities pos:[2,503,2,579], conf: 1	Australia pos:[193,202], conf: 1	still image for:'stills', pos:[2,152,2,158], conf: 1
Geoscience Australia pos:[182,202], conf: 1	Flinders Commonwealth Marine Reserve pos:[91,127], conf: 1	video pos:[2,068,2,073], conf: 1
University of Tasmania pos:[207,229], conf: 1	**Geomorphic Features**	**Gear**
	canyon pos:[1,208,1,214], conf: 1	baited remote underwater video pos:[2,043,2,073], conf: 1

Figure 5 - Examples of terms extracted from the NERP Oceanic Shoals Survey Report (categorised by reference vocabulary). The extracted terms are also linked with online URLs, which can be other RDF nodes, websites or literal values. Photos and logos are retrieved as part of this process. This is representative of the markup process for a single document.



Organizations

Australian Institute of Marine Science
pos:[61,99], conf: 1

Department of the Environment
pos:[816,845], conf: 1

Geoscience Australia
pos:[108,128], conf: 1

Museum and Art Gallery of the Northern Territory
pos:[179,227], conf: 1

National Oceanic & Atmospheric Administration
pos:[18,784,18,829], conf: 1

National Tidal Centre
pos:[15,439,15,460], conf: 1

University of Western Australia
pos:[139,170], conf: 1

Measurement

□ backscatter
pos:[1,940,1,951], conf: 1

□ bathymetry
pos:[1,916,1,926], conf: 1

□ chlorin
pos:[20,729,20,736], conf: 1

□ conductivity
pos:[2,413,2,425], conf: 1

□ depth
pos:[1,515,1,520], conf: 1

□ fluorescence
pos:[17,595,17,607], conf: 1

□ irradiance
pos:[17,568,17,578], conf: 1

□ salinity
pos:[9,396,9,404], conf: 1

□ temperature
pos:[2,427,2,438], conf: 1

□ total nitrogen
pos:[21,518,21,532], conf: 1

□ total organic carbon
pos:[21,490,21,510], conf: 1

Taxa

ascidian
pos:[26,427,26,435], conf: 1

bryozoa
pos:[27,995,28,002], conf: 1

dolphin
pos:[2,340,2,347], conf: 1

echinoderm
pos:[26,505,26,515], conf: 1

octocoral
pos:[22,404,22,413], conf: 1

polychaete
pos:[25,234,25,244], conf: 1

sponge
pos:[2,136,2,142], conf: 1

whale
pos:[2,332,2,337], conf: 1

Geomorphic Feature

□ bank
pos:[1,571,1,575], conf: 1

□ pinnacle
pos:[1,591,1,599], conf: 1

□ plain
pos:[1,627,1,632], conf: 1

□ scarp
pos:[33,270,33,275], conf: 1

□ terrace
pos:[1,578,1,585], conf: 1

Vessel

RV Solander
pos:[1,316,1,327], conf: 1

Data Type

raster
pos:[68,916,68,922], conf: 1

still image
pos:[36,172,36,183], conf: 1

video
pos:[2,171,2,176], conf: 1

● waypoint
pos:[79,047,79,055], conf: 1

Gear

baited remote underwater video
pos:[13,119,13,149], conf: 1

box core
pos:[13,588,13,596], conf: 1

multibeam sonar
pos:[1,916,1,931], conf: 1

Smith-McIntyre grab
pos:[13,549,13,568], conf: 1

Program

Global Ocean Drifter Program
pos:[18,738,18,766], conf: 1

Marine Biodiversity Hub
pos:[437,460], conf: 1

National Environmental Research Program
pos:[397,436], conf: 1

Places

Australia
pos:[61,70], conf: 1

East Timor
pos:[58,749,58,759], conf: 1

Indonesia
pos:[9,329,9,338], conf: 1

Joseph Bonaparte Gulf
pos:[58,518,58,539], conf: 1

Northern Territory
pos:[209,227], conf: 1

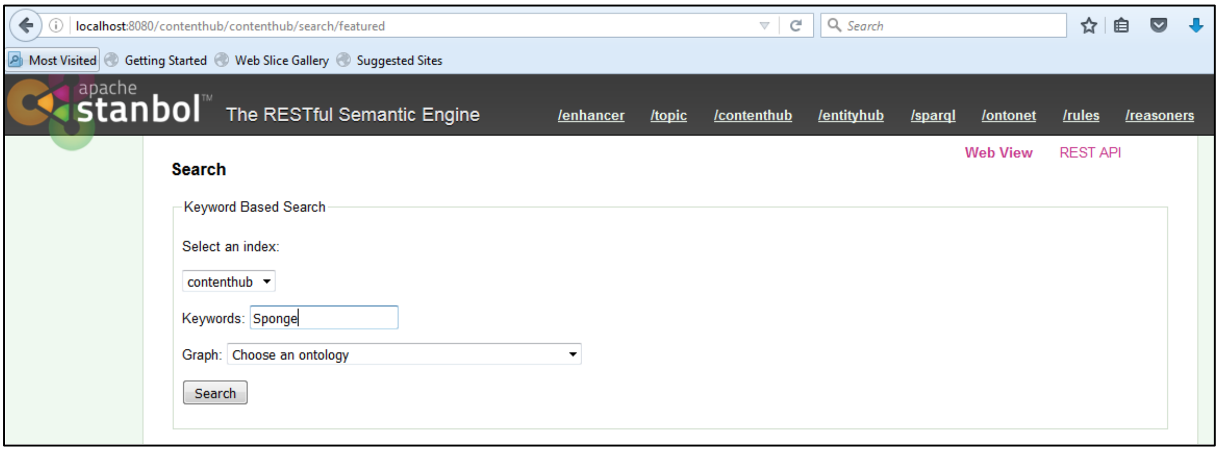
Oceanic Shoals Commonwealth Marine Reserve
pos:[271,313], conf: 1

Timor Sea
pos:[315,324], conf: 1

Language

@ en
conf: 1

Figure 6 – Example search on tagged data (example term: Sponge) using Stanbol's built-in search capabilities.



Search

Keyword Based Search

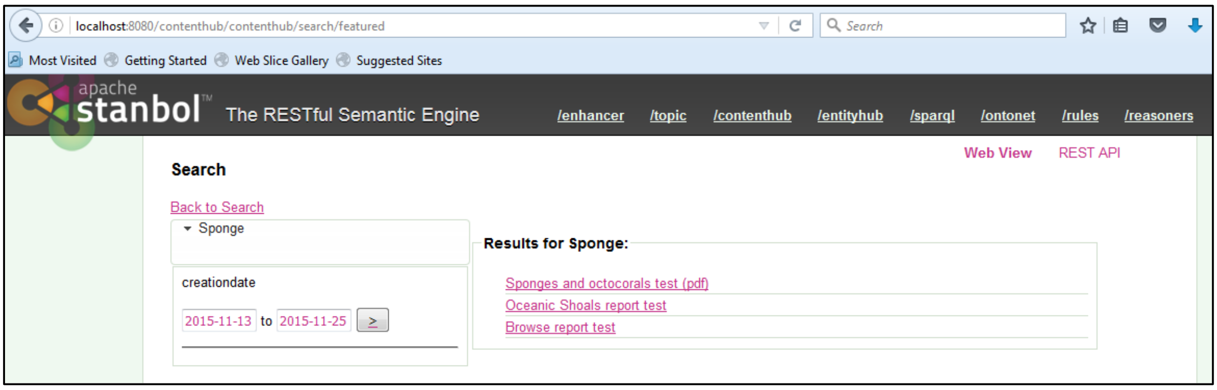
Select an index:

contenthub

Keywords: Sponge

Graph: Choose an ontology

Search



Search

Back to Search

Sponge

creationdate

2015-11-13 to 2015-11-25 >

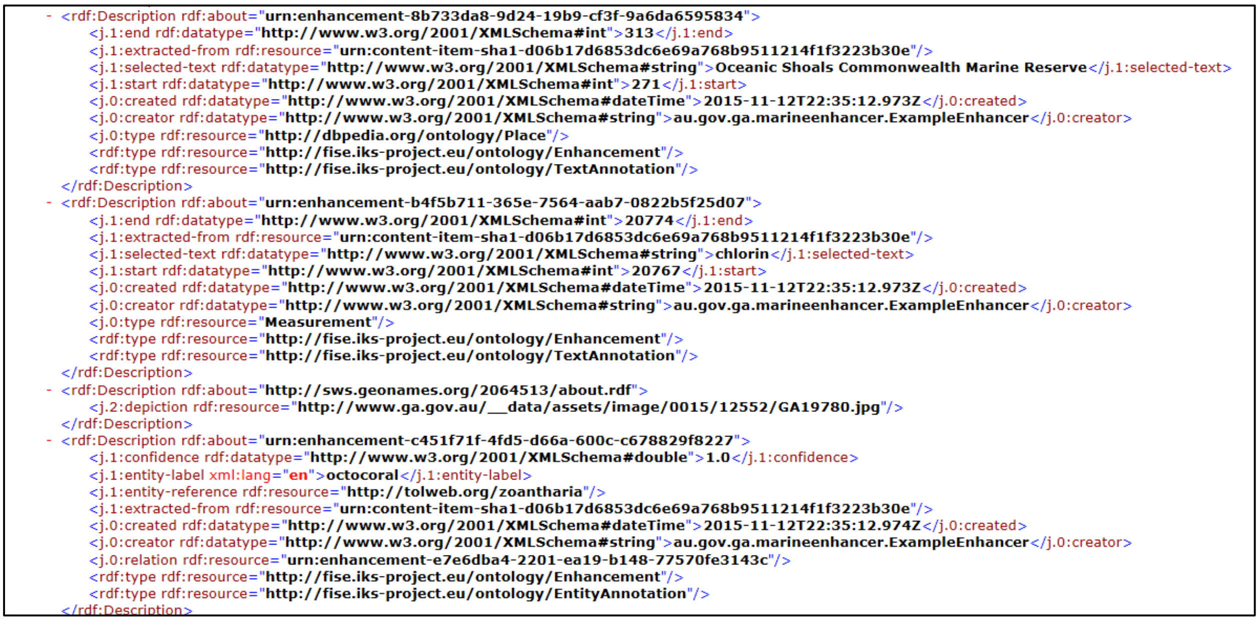
Results for Sponge:

Sponges and octocorals test (pdf)

Oceanic Shoals report test

Browse report test

Figure 7 – Example of stored RDF information (formatted as RDF/XML). Note: full file content is much longer than the sample shown here.

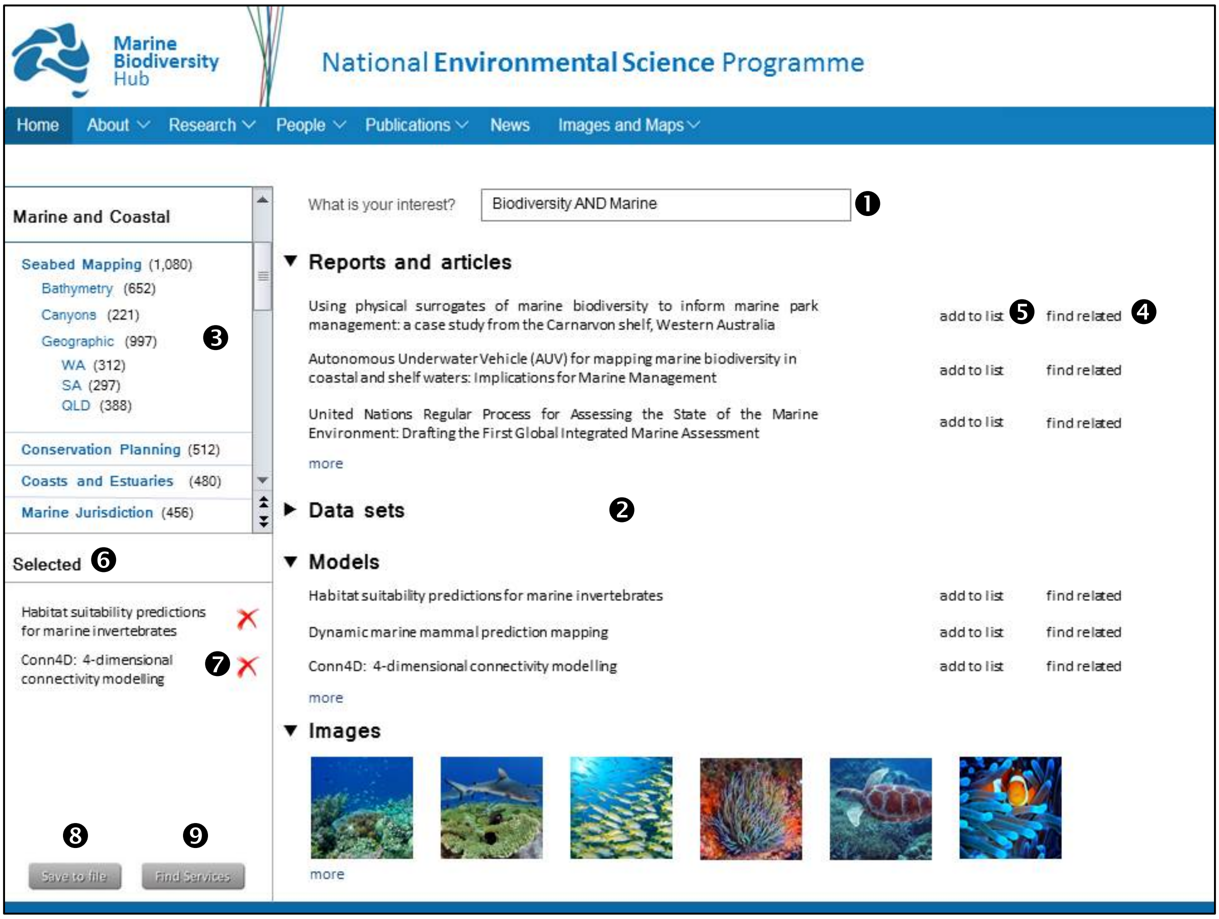


```
- <rdf:Description rdf:about="urn:enhancement-8b733da8-9d24-19b9-cf3f-9a6da6595834">
      <j.1:end rdf:datatype="http://www.w3.org/2001/XMLSchema#int">313</j.1:end>
      <j.1:extracted-from rdf:resource="urn:content-item-sha1-d06b17d6853dc6e69a768b9511214f1f3223b30e"/>
      <j.1:selected-text rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Oceanic Shoals Commonwealth Marine Reserve</j.1:selected-text>
      <j.1:start rdf:datatype="http://www.w3.org/2001/XMLSchema#int">271</j.1:start>
      <j.0:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2015-11-12T22:35:12.973Z</j.0:created>
      <j.0:creator rdf:datatype="http://www.w3.org/2001/XMLSchema#string">au.gov.ga.marineenhancer.ExampleEnhancer</j.0:creator>
      <j.0:type rdf:resource="http://dbpedia.org/ontology/Place"/>
      <rdf:type rdf:resource="http://fise.iks-project.eu/ontology/Enhancement"/>
      <rdf:type rdf:resource="http://fise.iks-project.eu/ontology/TextAnnotation"/>
  </rdf:Description>
- <rdf:Description rdf:about="urn:enhancement-b4f5b711-365e-7564-aab7-0822b5f25d07">
      <j.1:end rdf:datatype="http://www.w3.org/2001/XMLSchema#int">20774</j.1:end>
      <j.1:extracted-from rdf:resource="urn:content-item-sha1-d06b17d6853dc6e69a768b9511214f1f3223b30e"/>
      <j.1:selected-text rdf:datatype="http://www.w3.org/2001/XMLSchema#string">chlorin</j.1:selected-text>
      <j.1:start rdf:datatype="http://www.w3.org/2001/XMLSchema#int">20767</j.1:start>
      <j.0:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2015-11-12T22:35:12.973Z</j.0:created>
      <j.0:creator rdf:datatype="http://www.w3.org/2001/XMLSchema#string">au.gov.ga.marineenhancer.ExampleEnhancer</j.0:creator>
      <j.0:type rdf:resource="Measurement"/>
      <rdf:type rdf:resource="http://fise.iks-project.eu/ontology/Enhancement"/>
      <rdf:type rdf:resource="http://fise.iks-project.eu/ontology/TextAnnotation"/>
  </rdf:Description>
- <rdf:Description rdf:about="http://sws.geonames.org/2064513/about.rdf">
      <j.2:depiction rdf:resource="http://www.ga.gov.au/__data/assets/image/0015/12552/GA19780.jpg"/>
  </rdf:Description>
- <rdf:Description rdf:about="urn:enhancement-c451f71f-4fd5-d66a-600c-c678829f8227">
      <j.1:confidence rdf:datatype="http://www.w3.org/2001/XMLSchema#double">1.0</j.1:confidence>
      <j.1:entity-label xml:lang="en">octocoral</j.1:entity-label>
      <j.1:entity-reference rdf:resource="http://tolweb.org/zoantharia"/>
      <j.1:extracted-from rdf:resource="urn:content-item-sha1-d06b17d6853dc6e69a768b9511214f1f3223b30e"/>
      <j.0:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2015-11-12T22:35:12.974Z</j.0:created>
      <j.0:creator rdf:datatype="http://www.w3.org/2001/XMLSchema#string">au.gov.ga.marineenhancer.ExampleEnhancer</j.0:creator>
      <j.0:relation rdf:resource="urn:enhancement-e7e6dba4-2201-ea19-b148-77570fe3143c"/>
      <rdf:type rdf:resource="http://fise.iks-project.eu/ontology/Enhancement"/>
      <rdf:type rdf:resource="http://fise.iks-project.eu/ontology/EntityAnnotation"/>
  </rdf:Description>
```

7. FRONT-END USER INTERFACE DESIGN & FEATURES

Although Stanbol provides an 'out-of-the-box' web interface for displaying and interacting with enhanced data, and for performing searches on the data collection, the interface should be customised to streamline and simplify the process of searching for and retrieving results. As previously discussed, the user interaction should be no more complicated than browsing products on any number of commercial websites (e.g. Amazon, eBay). Although the final design of the user interface will evolve depending on the features that are included, design requirements, and the results of user feedback, the user interface might look something like the following:

Figure 8 – Mock-up of the front-end graphic user interface design.



What is your interest? Biodiversity AND Marine

▼ Reports and articles

- Using physical surrogates of marine biodiversity to inform marine park management: a case study from the Carnarvon shelf, Western Australia — add to list — find related
- Autonomous Underwater Vehicle (AUV) for mapping marine biodiversity in coastal and shelf waters: Implications for Marine Management — add to list — find related
- United Nations Regular Process for Assessing the State of the Marine Environment: Drafting the First Global Integrated Marine Assessment — add to list — find related

more

▶ Data sets

▼ Models

- Habitat suitability predictions for marine invertebrates — add to list — find related
- Dynamic marine mammal prediction mapping — add to list — find related
- Conn4D: 4-dimensional connectivity modelling — add to list — find related

more

▼ Images

more

Marine and Coastal

Seabed Mapping	(1,080)
Bathymetry	(652)
Canyons	(221)
Geographic	(997)
WA	(312)
SA	(297)
QLD	(388)
Conservation Planning	(512)
Coasts and Estuaries	(480)
Marine Jurisdiction	(456)

Selected

- Habitat suitability predictions for marine invertebrates ✗
- Conn4D: 4-dimensional connectivity modelling ✗

The interface should be able to be embedded within web page designs (for example, here within the page design for the Marine Biodiversity Hub). Users will initially search using a Google-style search box (1), which will return a collection of information resources organised by product type (2). On the left is the faceting menu which will be used to dynamically partition the results by vocabulary terms. Clicking on individual items (3) should link to the actual associated resource, and related items should be discoverable by clicking on the 'find related' buttons (4). Clicking on 'add to list' (5) will save the corresponding item to the 'Selected' area (6), similar to saving selections to a shopping cart on a commercial website. Items can be removed from the selected box by clicking on 'remove' icons (7). With the selected list, the results can be saved to a document file (e.g. RDF/XML, JSON-LD, Turtle, N-Triples, YAML) (8), or the search service can be re-invoked to look for services that would complement/enhance the list of selected items (e.g. downloading, mapping etc.) (9). Note that this interface is designed for human interaction, and the number of selected items is expected to be limited. Working with large selection lists would be better handled using the API.

8. WHAT CAPABILITIES WILL THE SYSTEM DELIVER (WHAT IS IN SCOPE)?

We have been successful in developing a prototype workflow using Stanbol capable of enhancing documents with semantic markup, and storing the information for a search. Continued implementation of the project (to be carried out during calendar year 2016) will include the development of a functional linked open data delivery service, including the enhancement of existing CERF and NERP information. With respect to the capabilities of the service, it will deliver:

- Automated tagging of documents using vocabularies and associated terms
- Querying of information resources via a RESTful API
- Development of a user-friendly GUI, similar to a commercial online shopping interface, including the ability to hyperlink to retrieved information resources
- The ability to save a retrieved list of resources as a file which could then be forwarded on to other services (e.g. for customised views, automated downloading, mapping etc.)

We will also engage with stakeholders and partner organizations, to align the development of the tool to complement their needs through user testing and acceptance.

9. WHAT CAPABILITIES WON'T THE SYSTEM DELIVER (WHAT IS OUT OF SCOPE)?

The tool is also intended as a means of accessing sources of information, and forwarding metadata on to other services. However, developing a range of processing services will take place as a separate project, once the initial capabilities of the discovery system have been proven.

Although we will design the system such that it can be upgraded and enhanced, a software system that incorporates any and all sources of marine information from both inside and outside of the Marine Biodiversity Hub is clearly outside the scope and budget of the project. Additionally, while we anticipate that the tool will harvest documents and metadata, it will not be designed to ingest raw data or data streams.

The use of controlled vocabularies and RDF creates new opportunities to enable search and access of information products in other languages. For example, RDF documents may contain links to corresponding terms in other languages, opening up new search options for discovering related information (which could then potentially be automatically translated). The significance of this is that it would significantly broaden the audience for Australian data resources, and would facilitate the exchange of information in international collaborations. Developing this functionality will be deferred to future projects.

10. PERSONNEL

Project lead or **project manager**
Johnathan Kool, Geoscience Australia,
Tel: 02 5842 6249 E-mail: johnathan.kool@gmail.com

Senior scientific advisor
Brendan Brooke, Geoscience Australia

Development team
Liam O'Brien – Solution Architect, Geoscience Australia
Erin Zimmer – Java Team Lead, Geoscience Australia
Nicholas Car – Data Architect, Geoscience Australia

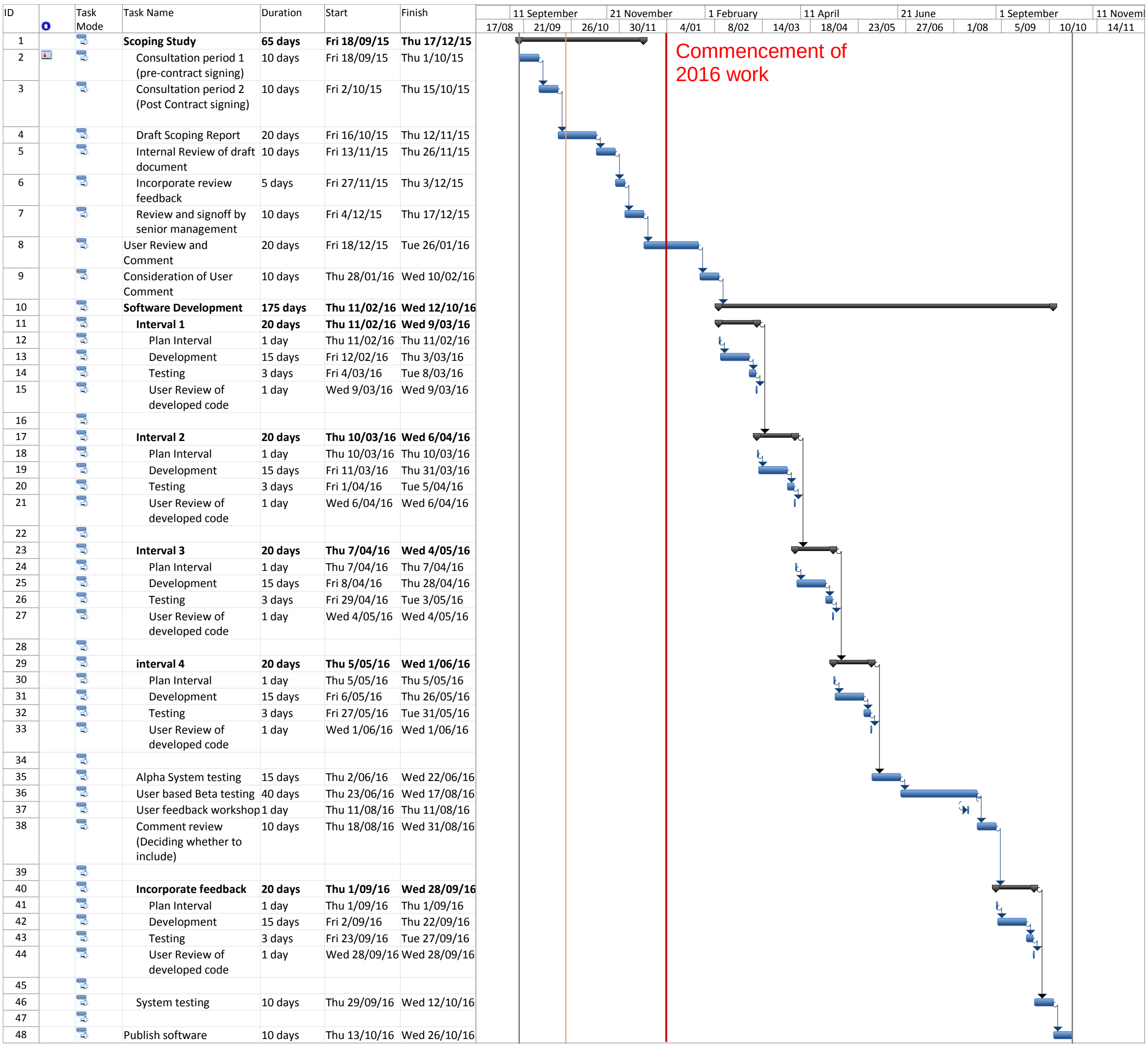
Darren Reid – Developer, Geoscience Australia

Stakeholder organisations

Parks Australia
Department of the Environment
ERIN
IMOS/AODN
AIMS
CSIRO
Great Barrier Reef Marine Park Authority
Bureau of Meteorology

11. PROJECT TIMELINE

Figure 9 – Proposed project timeline. A user testing workshop is scheduled for August 2016.



ID		Task Mode	Task Name	Duration	Start	Finish
1			**Scoping Study**	**65 days**	**Fri 18/09/15**	**Thu 17/12/15**
2			Consultation period 1 (pre-contract signing)	10 days	Fri 18/09/15	Thu 1/10/15
3			Consultation period 2 (Post Contract signing)	10 days	Fri 2/10/15	Thu 15/10/15
4			Draft Scoping Report	20 days	Fri 16/10/15	Thu 12/11/15
5			Internal Review of draft document	10 days	Fri 13/11/15	Thu 26/11/15
6			Incorporate review feedback	5 days	Fri 27/11/15	Thu 3/12/15
7			Review and signoff by senior management	10 days	Fri 4/12/15	Thu 17/12/15
8			User Review and Comment	20 days	Fri 18/12/15	Tue 26/01/16
9			Consideration of User Comment	10 days	Thu 28/01/16	Wed 10/02/16
10			**Software Development**	**175 days**	**Thu 11/02/16**	**Wed 12/10/16**
11			**Interval 1**	**20 days**	**Thu 11/02/16**	**Wed 9/03/16**
12			Plan Interval	1 day	Thu 11/02/16	Thu 11/02/16
13			Development	15 days	Fri 12/02/16	Thu 3/03/16
14			Testing	3 days	Fri 4/03/16	Tue 8/03/16
15			User Review of developed code	1 day	Wed 9/03/16	Wed 9/03/16
16						
17			**Interval 2**	**20 days**	**Thu 10/03/16**	**Wed 6/04/16**
18			Plan Interval	1 day	Thu 10/03/16	Thu 10/03/16
19			Development	15 days	Fri 11/03/16	Thu 31/03/16
20			Testing	3 days	Fri 1/04/16	Tue 5/04/16
21			User Review of developed code	1 day	Wed 6/04/16	Wed 6/04/16
22						
23			**Interval 3**	**20 days**	**Thu 7/04/16**	**Wed 4/05/16**
24			Plan Interval	1 day	Thu 7/04/16	Thu 7/04/16
25			Development	15 days	Fri 8/04/16	Thu 28/04/16
26			Testing	3 days	Fri 29/04/16	Tue 3/05/16
27			User Review of developed code	1 day	Wed 4/05/16	Wed 4/05/16
28						
29			**interval 4**	**20 days**	**Thu 5/05/16**	**Wed 1/06/16**
30			Plan Interval	1 day	Thu 5/05/16	Thu 5/05/16
31			Development	15 days	Fri 6/05/16	Thu 26/05/16
32			Testing	3 days	Fri 27/05/16	Tue 31/05/16
33			User Review of developed code	1 day	Wed 1/06/16	Wed 1/06/16
34						
35			Alpha System testing	15 days	Thu 2/06/16	Wed 22/06/16
36			User based Beta testing	40 days	Thu 23/06/16	Wed 17/08/16
37			User feedback workshop	1 day	Thu 11/08/16	Thu 11/08/16
38			Comment review (Deciding whether to include)	10 days	Thu 18/08/16	Wed 31/08/16
39						
40			**Incorporate feedback**	**20 days**	**Thu 1/09/16**	**Wed 28/09/16**
41			Plan Interval	1 day	Thu 1/09/16	Thu 1/09/16
42			Development	15 days	Fri 2/09/16	Thu 22/09/16
43			Testing	3 days	Fri 23/09/16	Tue 27/09/16
44			User Review of developed code	1 day	Wed 28/09/16	Wed 28/09/16
45						
46			System testing	10 days	Thu 29/09/16	Wed 12/10/16
47						
48			Publish software	10 days	Thu 13/10/16	Wed 26/10/16

Note: timelines are indicative, and represent duration intervals, not effort (i.e. 7 days to work on a task, not 7 days of work on a task)

Australian Government
National Environmental Science Programme

Australian Government
Australian Institute of Marine Science

CSIRO

Australian Government
Geoscience Australia

IMOS
Integrated Marine Observing System

Charles Darwin University

University of Tasmania

The University of Western Australia

NSW Government

Museum Victoria

www.nespmarine.edu.au

Contact:
Johnathan Kool
Geoscience Australia

Cnr Jerrabomberra Ave and Hindmarsh Drive, Symonston, ACT, 2609
johnathan.kool@ga.gov.au | tel +61 2 6249 5842