

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

Final Report

Johnathan Kool, Geoscience Australia

16 January 2017, RPv2





Australian Government Geoscience Australia

Enquiries should be addressed to: Johnathan Kool johnathan.kool@ga.gov.au

Preferred Citation

Kool, J. 2017. NESP Project B3 - Enhancing access to relevant marine information –developing a service for searching, aggregating and filtering collections of linked open marine data: Final Report on Project Milestones. Report to the National Environmental Science Program, Marine Biodiversity Hub. Geoscience Australia.

Copyright

This report is licensed by the University of Tasmania for use under a Creative Commons Attribution 4.0 Australia Licence. For licence conditions, see <u>https://creativecommons.org/licenses/by/4.0/</u>

Acknowledgement

This work was undertaken for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Science Programme (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

Exec	utive Summary	1
1.	INTRODUCTION	1
2.	APPROACH	2
3.	RESULTS	3
4.	FUTURE DIRECTIONS	6
Арре	endix A – Sample GBRMPA VOCABULARY TERMS	7



List of Figures

Figure 1- Demonstration web site start screen	3
Figure 2 - Browsing using vocabulary terms	4
Figure 3 - Free text search example	4
Figure 4 - Combining search techniques	5
Figure 5 - Images tagged with 'shark'	5



EXECUTIVE SUMMARY

This is a summary of outcomes from NESP Project B3, *Enhancing access to relevant marine information –developing a service for searching, aggregating and filtering collections of linked open marine data.* As part of the project, software was developed capable of tagging unstructured collections of documents (including Word documents, PDFs, spreadsheets, presentations, metadata and images among others) using an arbitrary vocabulary (term list), and then being able to structure, search, and deliver those documents through a web service. A demonstration web page interface was developed to highlight the capabilities of the service, and the results have been presented to NESP partner organisations, including representatives of: the Department of the Environment and Energy (DoEE), Parks Australia (under DoEE), IMOS/AODN, the Australian Institute of Marine Science (AIMS), and the Great Barrier Reef Marine Park Authority (GBRMPA). The capabilities of the software can be replicated and integrated into partner websites (e.g. the North West Atlas, and the NESP website).

1. INTRODUCTION

The capability to efficiently and effectively search for information is essential for scientific research, and for developing effective policy. Evidence-based decision-making lies at the heart of both, and the ability to survey access the full range of available information is critical to this process. Even if information addressing specific problems of interest already exist, if their existence or means of access is unknown, then they cannot be leveraged effectively. Across many organizations, there are often extensive collections of documents, including reports, data, spreadsheets and metadata representing a significant collective investment of time and resources. These collections may be organised in file systems, however their complete structure, contents and organisation may only be known to a few select individuals, or it is also possible that no individual will have a full understanding of the data, owing to their size and complexity. Different individuals and groups are also likely have divergent interests regarding uses of the information. For example, researchers may wish to structure documents according to data type or discipline, whereas managers and policy experts may have a greater interest in structuring documents according to associated legislation or policy outcome. Ideally, some sort of organisational system should be available that is capable of traversing large unstructured document collections, and allow for identification and extraction of relevant content that can then be organised according to an individual user's needs.

Data storage and retrieval is often handled through storing entries in a database or indexing system, which can then be queried using a range of attributes. Given the size and diversity of the information collections, it will not be feasible to organise them manually, and therefore some kind of automatic method for harvesting entries is needed. Additionally, it would be advantageous for users to be able to find additional items of linked interest. One way of achieving this is through tagging information with URLs (hyperlinks), describing its content in a machine-readable manner that can be parsed into a human-readable format. This 'Linked Data' paradigm is an emerging approach in online data management that provides a means of describing data collections and their attributes such that they can be searched and queried in a distributed manner. Linked Data has been promoted by Sir Tim Berners-Lee as the next step in the evolution of the World Wide



Web¹. The advantage of using this approach is that it provides a means of searching and aggregating distributed sources of information, and standards and software tools have already been developed by organisations such as the W3C and Apache Software Foundation.

To summarize, to improve access to the content of large collections of unstructured document collections, a system is needed that is able to:

- Sort and filter large unstructured collections of information and returning results in a structured manner.
- Permute the structure of search results, depending on topics of interest (for example, structure results according to named geographic location, data type, and author or alternatively by applicable legislation, impact type, and named geographic location).
- Use custom topic/keyword lists (as simple or as complex as desired).
- Perform in an automated manner to ensure ease of use (i.e. minimal manual effort).
- The system should take advantage of emerging Linked Data concepts, standards and tools.

To address this, we have developed prototype software called '*Hydroid*' that performs these functions, as well as a prototype web interface to demonstrate the capabilities of the software to deliver unstructured document collections in a structured manner in an interactive way.

2. APPROACH

Hydroid takes advantage of a number of existing software technologies (e.g. Apache Solr, Stanbol, Jena) and orchestrates them together to automatically tag and structure document collections. The software can be run on a single machine, or on a collection of machines and services. The software operates by traversing a collection of documents stored in a file folder or storage service (e.g. Amazon S3 Simple Store), and matching them against a list of terms with defined meanings (a 'vocabulary'). Copying the files into a storage location (through copying files into a directory or uploading to an S3 location) is the only back-end effort required. Indexing of the document occurs automatically on a periodic basis, and does not require shutting down the system. The vocabulary can be simple or complex, with or without defined relationships among terms. Through linking documents with vocabulary terms, documents also become associated (tagged) with the properties and relationships defined by the vocabulary. In addition to tagging, the full content of each document is also indexed and stored for guerying. For demonstration purposes, a vocabulary was created using headings harvested from GBRMPA's 25-year Strategic Assessment and management plan (http://bit.ly/1sWUH2a), although other vocabularies could be used instead. Example terms include topics such as coral trout (under Marine communities and species \rightarrow fishes \rightarrow bony fishes), humpback whales (under Marine communities and species \rightarrow mammals \rightarrow whales) and coral reefs (under Marine habitats; see Appendix 1 for the sample vocabulary structure).

The software uses Apache Tika as a means of identifying and reading the broad range of document types, and is capable of extracting information from over a thousand formats (see



Page 2

¹ https://www.ted.com/talks/tim_berners_lee_on_the_next_web

National Environmental Science Programme

<u>https://tika.apache.org/1.14/formats.html</u> for the list), and content is analysed using OpenNLP for natural language processing - giving machines the ability to understand word context, associations, and to distinguish different word uses.

The software can be queried using a URL string, which returns machine-readable content which can then be formatted in a human readable manner, in the same manner that HTML can be rendered as a viewable web page. Documents can be grouped according to the matched vocabulary terms, or searched for their contents using a URL query.

The software also has the capability to process images (e.g. .jpg, .tiff, .png). The software is able to parse metadata associated with the image (e.g. author, camera information etc.), but also image content through the Google Vision image recognition service.

The complete software code is available at <u>https://github.com/GeoscienceAustralia/hydroid</u> and <u>https://github.com/GeoscienceAustralia/hydroid-client</u> respectively, and instructions for setting up a new instance are available at: https://github.com/GeoscienceAustralia/docker_hydroid_demo.

3. RESULTS

To provide a tangible example of a user interface, a web interface was created to showcase how the service could be used to deliver collections of documents in a structured manner. Sample content from the NERP/NESP Marine Biodiversity Website and material provided by Parks Australia were copied into the storage location, and the GBRMPA sample vocabulary (described above) was used for tagging. The sample web interface is available at http://hydroid-dev-web-lb-1763223935.ap-southeast-2.elb.amazonaws.com.

•	i hydroid-dev-web-lb	-1763223935.ap-so	outheast-2.elb.amazonaws.com/#/	▼ 80%	C Q File indexing		→ ☆ 自		↓ 1		
🔊 Mo	📓 Most Visited 🛞 Getting Started 🚺 Web Slice Gallery 🛞 Suggested Sites 🝕 Views AGDC Content										
	And	stralian Government Geoscience Australia	Hydroid: Semantically-enabled search for m	marine information							
		What is yo	our interest ?			Search					
	Historic heritage		Welcome to Hydroid!								
	Indigenous heritage										
	Marine chemical proces	sses >									
	Marine commerce										
	Marine communities and	d species 🗲									
	Marine Ecological Proce	esses >									
	Marine habitats										
	Marine impacts										
	Marine invasives										
	Marine physical proces	ises >									
	National heritage										
	Terrestrial habitats sup Marine systems	porting									

Figure 1- Demonstration web site start screen

National Environmental Science Programme

Page 3 Marine Biodiversity Hub

Project B3 - Enhancing access to relevant marine information – Final Report

From the entry page, users can search in one of two ways. They can browse the collection in a structured manner using the vocabulary tree on the left:

sited 🐨 Getting Started	Web Slice Gal	lery 🛞 Suggested Sites 💐 Views AGDC Content							
Australia Greek	in Government ner Australia 1	lydroid: Semantically-enabled search for marine information A About - Research - People -							
	What is you	ar interest ? Search							
Marine communities and species	236	Search Results							
birds	127	Reports and articles							
seabirds	127	NWSJEMS_TR2.pdf	😑 🕹						
shorebirds.	65	Created on: Nov 17, 2008							
fahea	136	The distribution of seagrasses in Shark Bay, Western Australia, with The mangroves of the Lowendal Islands and Montebello							
bony fishes	6	Pretensary revestigations involving davy deging of Internet of the Incine Views their platforked,San turties and interagement of marine sellind programs. The response of Internet Mickagenta Neuwenglaer) to offshure selsmic. Photosynthesis and photoprotection in mangroves under field conditions in and Western A							
black teatfish									
conal trout	8								
rays	75	Christmas Is resarch and managment - Workshop Report FINAL JULY 2013.pdf	≥ 1						
sharks	136	Authors: a02172							
invertebrates	238	Climate change - increase sea temperatures, storm intensity and frequency, impacts of adjacent oceanic ecosystems (e.g. on							
ascidiana		seabirds)coraits are the major habitat forming organisms inWhite Syndrome coral disease outbreak during 2000 had a major impact on coral communities by selectively killing Acroports plate coraits, up to 90% mortality of these coraits at soResident orcen turtles originate							
corais	236	from areas Beyond North Keeling							
crown-of-thorna starfish	3								
sponges	61	Assessing the Effectiveness of Tramp Ant Projects to Reduce Impacts on Biodiversity	🗢 🕹						
macroalgae	6	Created on: Apr 23, 2013 Authors: Lori Lach, The University of Western Australia							
Halmeda									
marimala	54	recos on man, insects, crustacceans, simal birros, reprises and trops, and shall mammalis To our knowledge, there has to date been no investigation into non-target impacts of fiproni batting on Christmas Island mammalisThe Action Plan for Australian birdsThus our							
dolphine	67	sample estimates ~vrm or birds in the bioregion would be affected by yellowWe assessed a total of 49 species for potential risk to yellow crezy ants on Christmas Island, comprising 21 birds, 3 mammalis, 10 reptiles, and 15 crabs							
dugongs	60								

Figure 2 - Browsing using vocabulary terms

Or using the free text search box at the top:



Figure 3 - Free text search example

National Environmental Science Programme





Project B3 - Enhancing access to relevant marine information - Final Report

It is also possible to combine approaches by searching on a term and filtering on the basis of vocabulary terms. For example, performing a free text search on 'Kool' and filtering on 'Connectivity'.

(hydroid-dev-web-l	b-1763223935.ap-soutl	heast-2.elb.	amazonaws.co	om/#/?query=Koo	ol&facet=connectivity	/	80% C ⁱ	Q prescript	ve		→ ☆ 自		+	î	≡
😕 Most Visited 🛞 Getting Started 🚺 Web Slice Gallery 🛞 Suggested Sites 考 Views AGDC Content															
	Australian Government	Hydroid	d: Semanti	cally-enabled	search for mari	ne information									_
	Geoscience Australia	ft	About +	Research +	People -										
	Kool								Search						
Marine Ecologica	6	Sea	arch Result	s											
connectivity	5		Reports and	d articles											
Deart 5		Ма	rine Biodive	ersity Hub								=	Ł	₩0	
Reset F	iler	Crea	ated on: Oct 12 hors: Bax. Nic ()	, 2012 CMAR, Hobart)											
															=
		Ма	rine Biodive	ersity Hub								=	Ŧ	₩0	
		Crea	ated on: Oct 12 hors: Bax, Nic (, 2012 CMAR, Hobart)											
		Ge Crea	ohab2014p ated on: Apr 25	rogram.pdf								2	*	₩Đ	
		CE	RF Hub fina	al report 2010 v	veb.pdf							=	Ŧ	₩0	
		Crea	ared on: Jun 24	2011											
															. L
		CE	RF_Hub_fir	nal_report_201	0_web.pdf							=	Ł	₩0	
		Crea	ated on: Mar 24	, 2011											
L															

Figure 4 - Combining search techniques

The service also makes it possible to search images on the basis of their metadata, as well as their content. Images are processed using Google Vision to perform image recognition.

Australian Govern	Hy	droid: Semantically-e	nabled search for mar	ine information			
Geoscience Austra	alia 🔒	About - Resea	rch - People -				
humpback whales	16	0					
mangroves	35	C Images					
microalgae	3				Access of the subscription of the second		
microbes	1	T.			Sawfishes and River Sharks		
plankton	33				11111		
reptiles	52					11 steado	
crocodiles	14	Add to cart	Add to cart	Add to cart	Add to cart	Add to cart	Add to ca
sea snakes	16						
turtles	11						
loggerhead turtles	11	tren	-	and the second second	1 - 1 - 12	Treatenis for Surveying and Tagging Sawdohos and River Sharks	
seagrasses	23		And			man a	1
Marine Ecological 38 Processes	5					The best Markey and Markey and Prices And	- 7
compatition	42	Add to cart	Add to cart	Add to cart	Add to cart	Add to cart	Add to ca

Figure 5 - Images tagged with 'shark'

It is also possible to save selected search results similar to a 'shopping cart' found on commercial web sites. The demonstration website provides the capability to download all of the selected files simultaneously, however the page scripting could be modified to forward tagged results to a web



service. For example, items tagged as being 'mappable' could be forwarded to a map service for plotting.

It is important to distinguish between the interface and the underlying service. The software service is what provides the search capability, and retrieval of the information as a stream of data. The interface was designed as a means of presenting the search results in a human-friendly fashion, but the results could be used in other websites and web services without needing to use the web page design or layout shown here. The data can be used as a stream into a range of outlets, and could be reused in a broad range of ways with minimal effort.

4. FUTURE DIRECTIONS

Geoscience Australia has plans to continue developing the technology, specifically through improving the natural language processing capabilities of the software, and training the software to recognize a richer collection of entities for data extraction. Additional development work may also take place under ongoing NESP Project D1 – integrating some of the collection delivery capabilities into the online North West Atlas. This would provide additional search options for retrieving data, documents and other content types relating to marine areas of north west Australia and their biota. This activity would be performed in coordination with Dr. Karen Miller of AIMS (K.Miller @aims.gov.au).

Fundamentally, the project delivers new capabilities for organising, sorting, filtering and delivering large collections of (previously) unstructured document collections. The software provides access to these documents as a service that can be leveraged by a broad range of web-based interfaces. The software can be repurposed and re-branded by Hub partners for their own needs - for example, to help deliver collections of images, or manage large collections of reports and grey literature. AIMS have discussed possibilities for integrating Hydroid's capabilities with their Drupal-based Content Management System (CMS). This would help with managing their image collection, and providing new means of accessing the content of their web site. The contact for this activity would be Dr. Eric Lawrey (e.lawrey@aims.gov.au). Ongoing work with Parks Australia to help manage collections of information regarding Commonwealth Marine Reserves possibility, is another with the point of contact being Amanda Parr (Amanda.Parr@environment.gov.au). Jeanette Corbitt (jeanette.Corbitt@environment.gov.au), the director of the Information Strategy and Major Projects within ERIN also expressed interest in the work, particularly with regards to adding additional functionality to a whole-of-NESP information discovery portal. There was also some discussion regarding whether the approach could also be used to integrate with SPIRE, the Department of the Environment and Energy's records management system. The software can be refined with regards to identifying duplicates, ranking documents according to their relevance, and identifying documents that similar users have found useful. These capabilities can be incorporated into websites or applications by partner agencies and organizations, or also offered as a service or translated into products for uptake by external developers and users.

Hydroid provides a number of opportunities to help improve information search and retrieval by Hub partners and the public in general. We look forward to the opportunity to continue development of this software, and furthering its application.

Marine Biodiversity Hub

Page 6

APPENDIX A – SAMPLE GBRMPA VOCABULARY TERMS

•	Historic h	eritage	•	Marine i	nvasives
	0	World War II		0	Crown-of-thorns outbreaks
	0	Lightstations		0	Disease outbreaks
		-			
•	Indigeno	us heritage	•	Marine p	physical processes
	0	Sacred Sites		0	Currents
				0	Cyclones
•	Marine c	hemical processes		0	Freshwater inflow
	0	Nutrient cycling		0	Light
	0	Salinity		0	Sea level
				0	Sea temperature
•	Marine c	ommerce		0	Sedimentation
	0	Fisheries		0	Wind
	0	Marine recreation			
	0	Marine research	•	National	heritage
	0	Marine tourism		0	National beauty
	0	Ports		-	
	0	Shipping	•	Terrestri	al habitats supporting marine ecosystems
				0	Forests
•	iviarine c	Direde			Kalillorest Freshwater wetlands
	0	Birus • Sophirds		0	Freshwater wetianus
		 Seabilius Shorobirds 			
	0	Fishes			
	0	 Bony fishes 			
		Black teatfish			
		Coral trout			
		 Bays 			
		 Sharks 			
	0	Invertebrates			
		 Ascidians 			
		 Corals 			
		 Crown-of-thorns starfish 			
		 Sponges 			
	0	Macroalgae			
		 Halimeda 			
	0	Mammals			
		 Dolphins 			
		 Dugongs 			
		 Whales 			
		Humpback whales			
	0	Mangroves			
	0	Microalgae			
	0	Microbes			
		Pidriktori Poptilos			
	0				
		 Crocouries See Snaker 			
		 Jea Shakes Turtles 			
		 Loggerhead turtles 			
	0	Seagrasses			

Page 7 Marine Biodiversity Hub



www.nespmarine.edu.au

Contact: Johnathan Kool Geoscience Australia

Cnr Jerrabomberra Ave and Hindmarsh Dr, Symonston, ACT 2609 johnathan.kool@ga.gov.au | tel +61 02 6249 5842