

	<MMI /> Case Study
Title	WHOI Cruise Data Harvesting
Website	http://gdc.ucsd.edu:8080/digarch http://tango.whoi.edu
Contact	Dru Clark; Stephen Miller
Organization	Scripps Institution of Oceanography
Audience	Academic researchers; WHOI staff; students; teachers
Discipline	Marine Geology and Geophysics, also supports Physical, Biological and Chemical Oceanography, Acoustics
Problem	No coherent long-term access to WHOI cruise data
Solution	Develop auto-harvest approach similar to previous SIO efforts, but making use of new approach with greater use of Instrumentation Block and associated controlled vocabularies
Input	Digital online and physical media scattered throughout institution with minimal metadata
Output	Digital Library of data, images, and documents, organized by cruise leg
Steps in approach	<p>Preparation</p> <ul style="list-style-type: none"> • Implement a working digital library system at WHOI, patterned after SIO approach • Create framework for “A Multi-Institutional Approach to Digital Archiving” document to record our progress, including motivation, obstacles, and achievements, adding content as available in chapters for Cruise, Submergence and next generation SIOExplorer. • Create framework for “User Manual for WHOI Cruise Archiving” • Finalize list of available WHOI digital object types, with descriptions (mb, Athena, calliope, IMET, ctd, adcp, nav, synopsis, other?) and update controlled vocabularies • If needed, modify SIO navabs software for producing abstracted WHOI navigation files • Establish staging area on tango.whoi.edu system, with directories as needed • Define approach for loading available objects into staging • Create scripts for auto-harvesting .metadata from objects in staging • Define approach for creating ado, mif pairs with modified standard SIO adoCreator tool • Define approach for publishing in WHOI digital library <p>Operation</p> <ul style="list-style-type: none"> • For each cruise leg, copy selected files to staging directory • Extract metadata from files and other resources • Run adoCreator to create (ado,mif) pairs • Publish in WHOI digital library <p>Further usage of results</p> <ul style="list-style-type: none"> • Apply similar approach for submergence vehicles
Results to date	Manual prototype tested; automatic procedures work-in-progress
Future goals	Turn over to WHOI for routine use; extend to Alvin and ROV harvesting
Lessons learned	It is challenging to create multi-institution progress by committee, need designated small groups to lead the way and create prototypes for others to evaluate.
Re-usability	This capability can easily be adapted to any project that uses the mtf and cv approach, as in collections at SIO, WHOI, Oregon State. Adapting to a non-mtf project would be straightforward, but would require development. Technical skills required are: database, xml, metadata
Presentations	http://gdc.ucsd.edu:8080/digarch/about-project/presentations/
Reports	NSF Annual Report (http://gdc.ucsd.edu:8080/digarch/about-project/presentations/DIGARCH_0455998_Annual_Report.pdf) Helly, “Scalable models of data sharing in Earth sciences,” http://www.sdsc.edu/~hellyj/papers/helly2002GC000318-1.pdf
Data products	WHOI cruise synopses websites Knorr (http://www.whoi.edu/marops/research_vessels/knorr/synopses_main.html) Atlantis (http://www.whoi.edu/marops/research_vessels/atlantis/synopses_main.html) WHOI multibeam catalog (http://mbdata.whoi.edu/) WHOI data media inventory (http://gdc.ucsd.edu:8080/digarch/Content-harvesting/inventory/) WHOI digital library prototype (http://tango.whoi.edu) SIO operational digital library (http://SIOExplorer.ucsd.edu) SIO digital library with new metadata approach (http://siox.sdsc.edu)
Tools	Tools used or developed, including links adoCreator for WHOI
Profile or template	http://gdc.ucsd.edu:8080/digarch/technical-development/metadata/MTF/
Standards	SIO/SDSC mtf approach with canonical and native blocks
Format	ado, mtf, mif

Dictionaries	http://gdc.ucsd.edu:8080/digarch/technical-development/Dictionary/
Controlled Vocabularies	http://gdc.ucsd.edu:8080/digarch/technical-development/Controlled%20Vocabulary/
Ontologies	Not yet
Other products	
Funding sources	"DIGARCH" NSF/Library of Congress CISE IIS 0455998 Multi-institution Testbed for Scalable Digital Archiving
Start date	2005-06-01
End date	2007-05-30
Sustainability plan	TBD, WHOI responsibility for ongoing effort
Case Study contributor	Stephen Miller
Case Study date	2006-10-13