

	<MMI /> Case Study
Title	Controlled Vocabularies for Metadata Harvesting
Website	http://gdccoll.ucsd.edu/
Contact	Dru Clark; Stephen Miller
Organization	Scripps Institution of Oceanography
Audience	Digital archivists and developers
Discipline	Marine Geology and Geophysics, also supports Physical, Biological and Chemical Oceanography, Acoustics; approach is relevant to any discipline
Problem	Legacy holdings suffer from missing or erroneous critical supporting information (metadata). Repairing the information in the original source documents would be painful and cost-prohibitive.
Solution	Use database technology in a staging area to perform quality control before final publishing in a digital library. Extract metadata from existing files as available and place them in a database. Analyze variance in parameter values across the collection. Distill a controlled vocabulary for each parameter. Finally, repair the database of metadata values with the help of the controlled vocabulary
Input resources	mgd77 underway geophysical files for about 800 cruises, each with sensor information in header. CCDS staging directories for about 800 cruises, with 100,000 data files as loaded in previous version of SIOExplorer digital library
Output products	Corrected database of metadata values for about 800 cruises
Steps in approach	<p>Preparation</p> <ul style="list-style-type: none"> Define the scope of the collection to be built (data from all SIO expeditions, organized by each cruise leg) Define the metadata profile (mtf) to identify the required parameters Create an initial controlled vocabulary from a priori knowledge, with values for each parameter in the mtf, as appropriate. Build a PostgreSQL database to hold the results. (Dru, further details) <p>Operation</p> <ul style="list-style-type: none"> Extract metadata from existing files and other resources as available and populate a plain ascii intermediate (.metadata) file for each cruise leg Populate a database with the values from the intermediate files for all cruise legs For each parameter, analyze the range of values across the collection of cruises Create a new version of the controlled vocabulary, adding certified acceptable values from the previous analysis Determine mappings of existing values to acceptable values. This includes synonyms, abbreviations, spellings, cases. Repair the database contents by applying mapping. Repair the database by correcting any remaining errors. Keep audit trail of all changes, available as SQL commands from the repair phase. After experience with this process is gained at the Cruise block level of metadata, apply the same approach to native blocks: Documentation, Instrumentation, Products and Samples. Instrumentation will require the greatest effort. <p>Further usage of results</p> <ul style="list-style-type: none"> Use the resulting database values to seed the metadata for each object to be published in the digital library, incorporating specific values such as lat/lon fore each object.
Results to date	Manual prototype tested; automatic procedures work-in-progress
Future goals	Test for entire cruise loadings of varying complexity and age. When satisfied, apply to approximately 800 cruise legs.
Lessons learned	A second generation approach to a problem can convert a morass of semi-valid information into something much more reliable.
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/ Presentation for Workshop on Harvesting Data and Metadata (work-in-progress) Poster for Dec 2006 AGU (work-in-progress)
Reports	Helly, "Scalable models of data sharing in Earth sciences," http://www.sdsc.edu/~hellyj/papers/helly2002GC000318-1.pdf
Tools	Tools used or developed, including links (Dru...)
Metadata specification	http://siox.sdsc.edu/metadata/mtf/current/MTF_current
Standards	SIO/SDSC mtf approach with canonical collection, ado and native blocks

Format	ado, mtf, mif
Data products	Coming soon: revised .metadata files for the entire SIOExplorer collection, ready for publishing as ado, mif pairs in the next version of the digital library, http://siox.sdsc.edu
Dictionaries	http://siox.sdsc.edu/metadata/dictionary/current/
Controlled Vocabularies	http://siox.sdsc.edu/metadata/CV/current/CV_current
Ontologies	Not yet
Protocols	
Schemas	
Web services	none
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	Geological Data Center, SIO responsibility for ongoing effort
Case Study contributor	Stephen Miller
Case Study date	2006-11-01