# Interoperability in Scientific Thesauri

Justine Withers

San Jose State University

Abstract

Thesauri can make scientific domains clear at the same time they create barriers between different domains. Evaluating the overlap of two related thesauri reveals gaps that hamper information sharing. Coastal and Marine Ecological Classification Standard (CMECS) serves a specific need for classifying federal research data sets but lacks flexibility and links to other systems. The United Nations Food and Agriculture Organization (FAO) Aquatic Sciences & Fisheries Abstracts (ASFA) Thesaurus covers a broader range of more access points. However, it fails to link many related terms.

Interoperability in Scientific Thesauri

Imagine a marine ecologist in a remote part of the western coast of the United States. He is an employee of the National Marine Fisheries Service as well as an adjunct professor at a state university. As a research scientist, he collects and disseminates data, publishes papers, and consumes the research of others. How can he find relevant information and share his findings efficiently? For instance, the metadata for his research data sets must follow federal guidelines: the Coastal and Marine Ecological Classification Standard (CMECS) and U.S. Integrated Ocean Observing System (IOOS®). If that data is used in a research paper, it is indexed under whatever terminology is used by the abstracting service.

Comparing the terminology available in two common marine science thesauri reveals an interoperability gap that could hamper sharing of information.

## CMECS for federal data sets

Developed by the United States National Oceanic and Atmospheric Administration (NOAA), Environmental Protection Agency (EPA), and Geological Survey (USGS), among other organizations, the Coastal and Marine Ecological Classification Standard (CMECS) classifies "physical, biological, and chemical data" to "organize information about coasts and oceans and their living systems." (Digital Coast, n.d.)

The Federal Geographic Data Committee (FGDC) oversees data collection for the United States government, and for marine sciences, they approved the use of CMECS in conjunction

with Integrated Ocean Observing System (IOOS®), which standardizes description of physical data, including "water temperature, water level, currents, winds, and waves" (FGDC, 2013).

Basically, IOOS defines the units and methods of measurement and CMECS puts the results into the greater aquatic context.

## ASFA for aquatic science articles

If our scientist writes a paper based on a recent sea cruise, it might be abstracted in the Oceanic Abstracts database, available from ProQuest. Oceanic Abstracts uses two thesauri: *Aquatic Sciences & Fisheries Abstracts (ASFA)*, for subjects, and *Taxonomic Terms,* which covers the Latin names of organisms.

ASFA was developed by the Fisheries and Aquaculture Department of the Food and Agriculture Organization (FAO) of the United Nations and covers "the world's literature on the science, technology, management, and conservation of marine, brackish water, and freshwater resources and environments, including their socio-economic and legal aspects." (FAO, 2013)

## Study design

A set of keywords relevant to marine science were entered in the CMECS and ASFA thesauri to ascertain their availability, overlap, and extent of related terms. They were then entered into the Oceanic Abstracts database to measure resource retrieval.

### Test set

Keywords were collected from the titles of articles appearing in the 1987 and 2007 runs of Pacific Science,"an international, multidisciplinary journal reporting research on the biological and physical sciences of the Pacific basin." (University of Hawai'i Press, 2011) Titles not rel-

evant to marine science were removed. The remaining titles were then parsed to identify all significant words.

The set contains 62 keywords or phrases.

**Criteria**

Shiri (2012, p. 245) summarizes many studies of thesauri and their measures of usability. The most frequently occurring measures that could be measured within the constraints of this study were satisfaction, success/failure, preference, user understanding, and relevance. I combined these with the other measures to identify four general and objective criteria—interoperability, retrieval rate, query formulation and reformulation, and visible context of terms—and one fuzzy, subjective measure—perceived ease of use or satisfaction.

## Study results

CMECS is available from a website (http://cmecscatalogue.org) as well as a PDF. On the website, one can search keywords, browse several tree structures—geographic regions, marine regions, and various modifiers—and drill down into various ecosystem components.

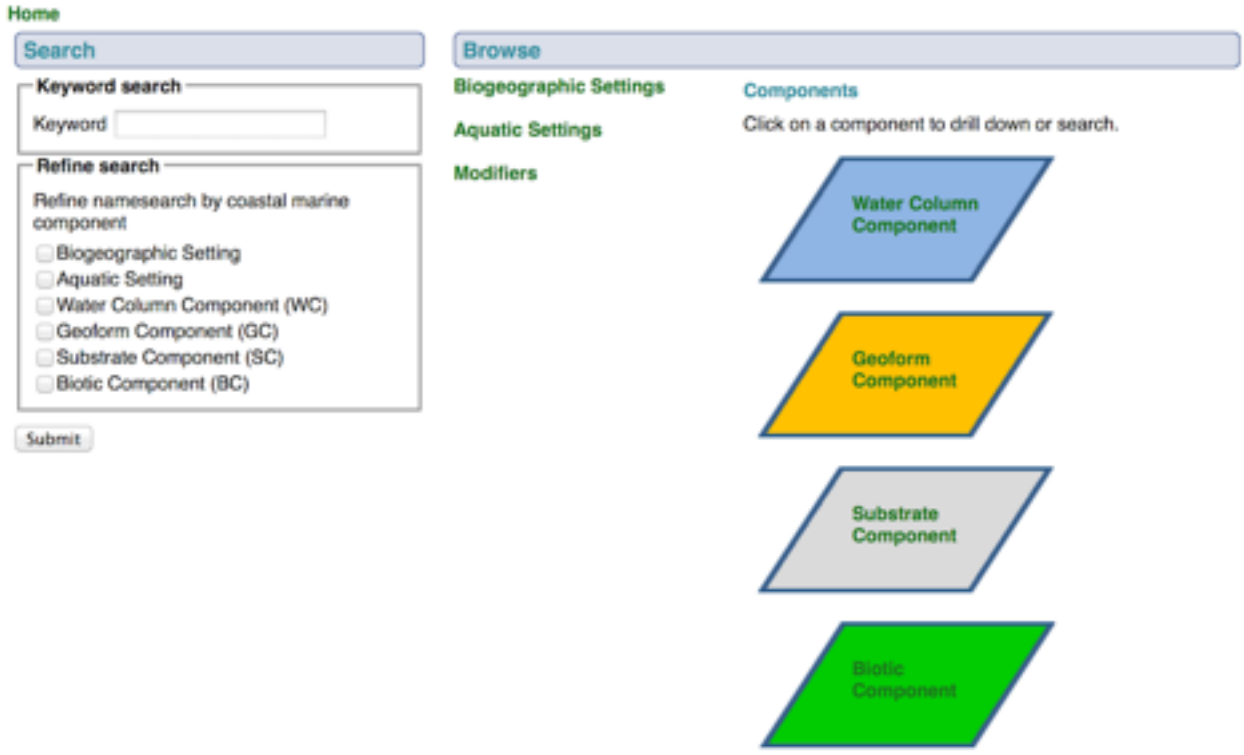


*Figure 1.* CMECS home page. Users can search specific keywords or browse by geography, marine regions, and types of marine life.

CMECS is designed to only classify resources and is not connected to any retrieval system.

The ASFA thesaurus is available from its own website (http://www4.fao.org/asfa/asfa.htm) as well as within the ProQuest database interface.

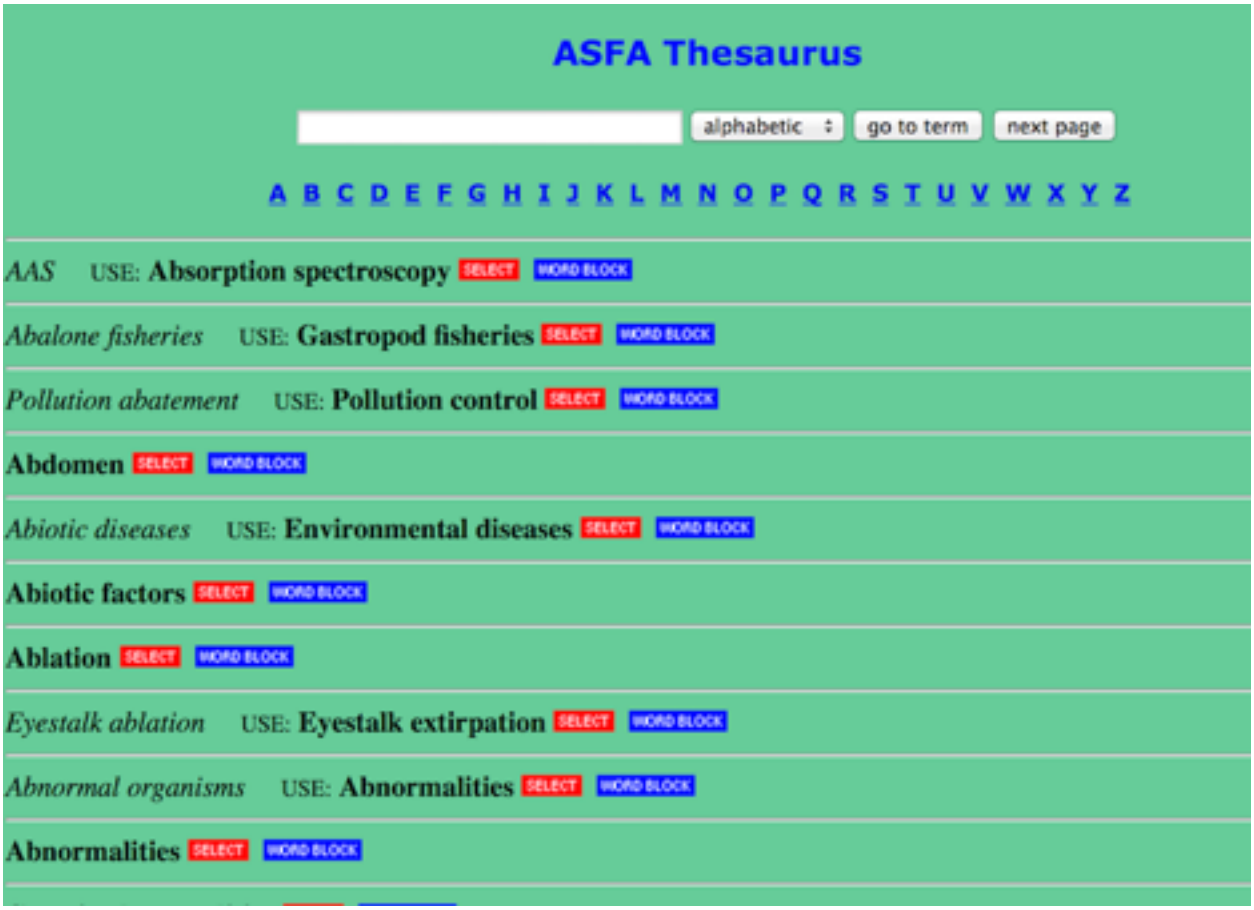


*Figure 2.* ASFA Thesaurus interface. Selected terms appear in detail in a pane to the right.


*Figure 3.* ASFA Thesaurus interface through ProQuest.

**Query formulation**

CMECS offers only one way to formulate queries: entering a term in the Keyword Search field and narrowing it down to one or more of the basic categories for browsing.

A successful query offers a list of relevant headings.

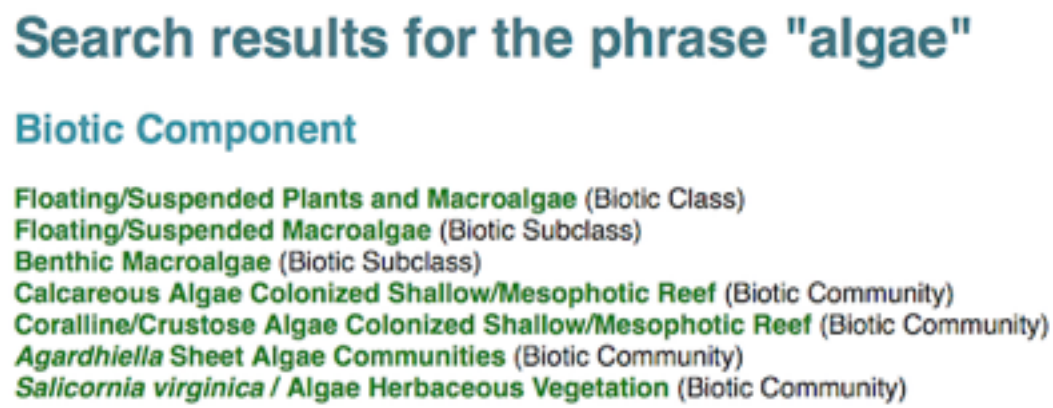


*Figure 4.* Results from a keyword search in CMECS.

An unsuccessful query shows only a blank area: the user is out of luck with no suggestions or error corrections.


*Figure 5.* Unsuccessful keyword query in CMECS.

Interactive query formulation is considered a benefit to users (Shiri, 2012, p. 272) and CMECS is lacking. It seems to operate under the assumption that users are already experts, familiar with the appropriate terminology and its context.

Compare the CMECS results with that of ASFA. The scope note and related terms are typical of a thesaurus. Little interaction is available, other than being able to access the search field directly.

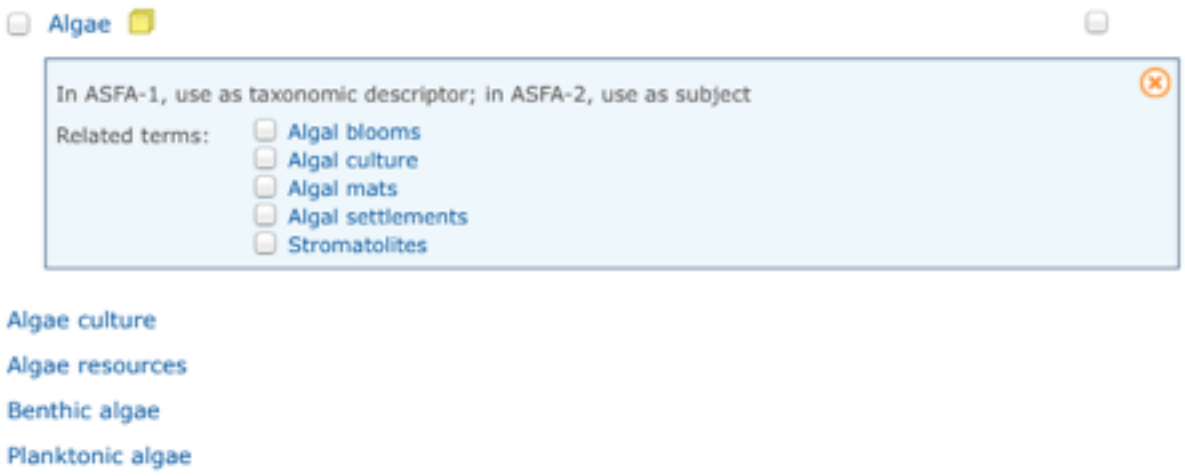


*Figure 6.* Results from a keyword search in ASFA.

Within the Oceanic Abstracts database, query formulation is more flexible. The suggested Boolean searches help narrow results and identify appropriate index terms.

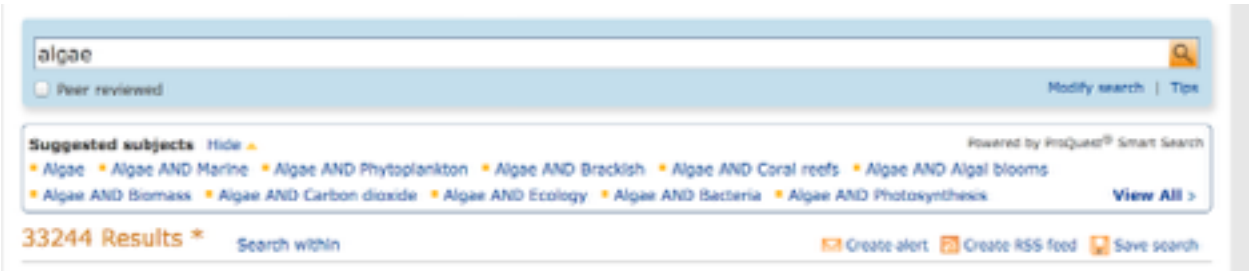


*Figure 7.* Suggested search terms in Oceanic Abstracts.

However, the Modify Search link returns the user to the database home page and offers little guidance.


*Figure 8.* Modify Search link in Oceanic Abstracts. Options are limited and unintuitive.

**Visible context**

In its two browsing interfaces, CMECS offers the most help to the user in visualizing the context of a term in the larger ecosystem.

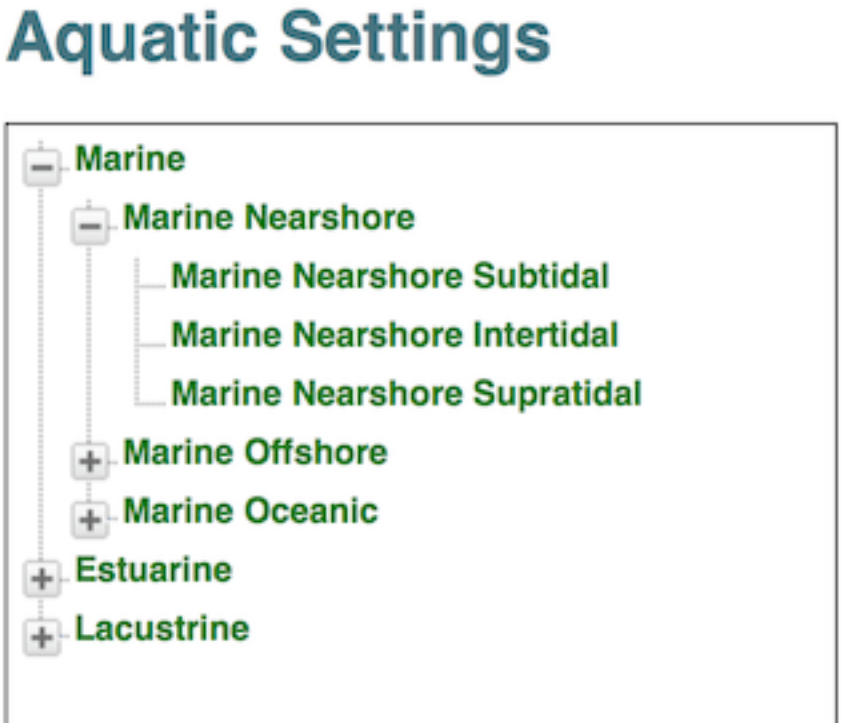


*Figure 9.* Tree structure in CMECS. Users can drill down the subclasses of environments and species.

Again, the user must already be an expert to understand the difference between the terms.

ASFA shows context in the typical thesaurus conventions of related terms and alternate headings.

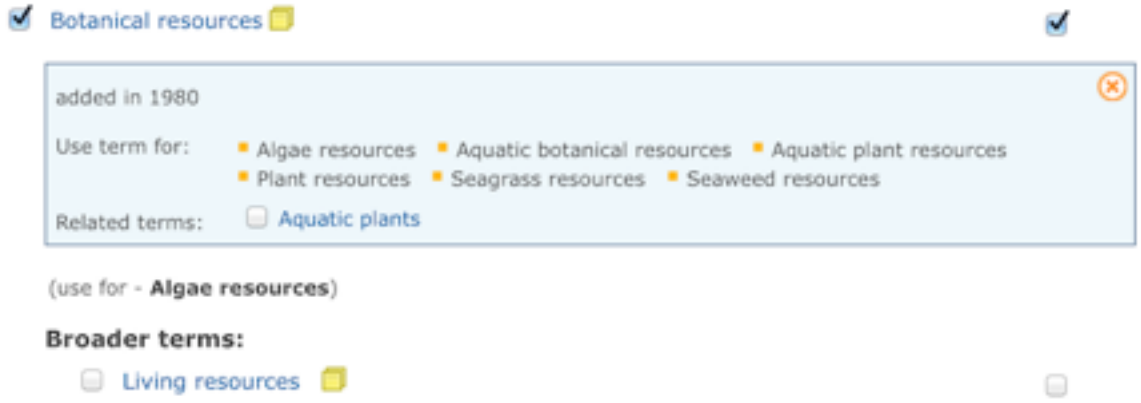


*Figure 10.* Related terms in ASFA.

ASFA is also more forgiving in error recovery, allowing the user to enter a layman's term and retrieving the domain-specific one.

**Interoperability**

Because CMECS and ASFA do not as yet interact systematically, in this study interoperability is measured by the appearance of a term in both thesauri. Nine out of 62 terms appear in both thesauri. Even this low number must be qualified by noting the unavailability of many of the terms in CMECS. The terms that are shared are quite broad and common, such as algae, deep-water (a shared concept, although appearing under different names), and coral reef.

In addition, the availability of linked data increases the interoperability of a classification system (Wittenburg & Broeder, 2002, Mendez & Greenberg, 2012). CMECS does not yet offer RDF triples for its classification. ASFA has been converted to triples as part of the AGROVOC

Linked Open Data set (LOD). (AIMS, 2012) The linked data is not available through ProQuest, however.

**Retrieval rate**

From the keyword set, twelve appear in the CMECS thesaurus, either in the original form or a synonym found through browsing. CMECS is limited to a very specific range of topics.

Thirty keywords appear in ASFA, along with additional related terms. A high level of experience with the subject is sometimes needed to choose the appropriate heading. For instance, if a user searches on *trophic group distribution,* six possible headings are offered: *Trophic levels, Trophic relationships, Trophic status, Trophic structure, Trophic zonality, and Trophodynamic cycle.* ASFA offers no scope notes for these so the user needs to already know what the mean or have a good textbook nearby.

Another factor affecting retrieval is the use of a separate thesaurus for taxonomic names. Twenty-nine of the keywords are specific names of species, either common or scientific. Taking this into consideration, only three keywords are unavailable from the Oceanic Abstracts database.

A more general retrieval rate of articles from the OA database was satisfactory. Only one term, diversity, was too broad to be useful. The broad term of *water column structure* retrieved 476,422 articles and the OA interface offered several options to narrow the results: AND sediments, AND marine, AND phytoplankton, AND lakes, AND freshwater, as well as other options.

Names of specific creatures, such as *Linckia multiform* and *Metograpsus oceanicus* retrieved fewer than twenty articles. Scientific names revealed an anomaly and weakness in OA's use of thesauri. Searching on *Pacific Pygmy Octopus* retrieves 215 articles; its scientific name,

*Octopus digueti,* retrieves only 44. Similar results occurred with other common and scientific name pairs, suggesting that the two terms are not mapped, reducing user success.

Another anomaly in OA is its treatment or, rather, non-treatment of spelling errors. When *photsynthesis* was accidentally entered, OA returned 76 articles, each of which included the typographical error. With the keyword spelled correctly, 124,965 articles come back, easily narrowed with suggested phrases.

In other situations, OA offered alternate searches that returned no results, seemingly wanting to correct a typo that did not exist.

**Satisfaction**

In the more touchy-feely realm of "user satisfaction," the tree structure navigation of CMECS was instantly intuitive, both in its operation and in providing context to terms. However, its lack of feedback was frustrating.

The ASFA Thesaurus in its original mint green and hyperlink blue is awful to look at. However, its interface is clearer than that in ProQuest. Scope notes and related terms appear in a separate pane so that the user can see details without losing the context of the original term. In ProQuest, all other terms disappear when linking to another term and one cannot see details of a term without clicking another button. Additionally, the original ASFA interface offers both alphabetical and Keyword in Context displays.

As far as retrieving articles within the OA database, the thesaurus seems to operate relatively well behind the scenes. Articles containing related terms are retrieved and related and narrower terms are offered to narrow results. The visible thesaurus is not directly linked to the data-

base, however, so after a search term is identified, it must be entered manually into the Search page.

## Conclusions

Both thesauri show the prevalent siloing of scientific classification systems. Even the obvious mapping of common to scientific names is not reliable. Gaps such as create challenges for novices and experts alike: laypeople cannot use their own terminology to find scientific concepts; domain experts cannot share their potentially relevant research with other fields because of differences in vocabulary.

The challenge is especially obvious in CMECS. Because it covers such a specific range of concepts, but must serve many federal departments, it should actively map to other classification systems.

References

Agricultural Information Management Standards (AIMS). (March 5, 2012). *ASFA Thesaurus linked to AGROVOC LOD.* Retrieved from http://aims.fao.org/news/aquatic-sciences-and-fisheries-abstracts-asfa-thesaurus-linked-agrovoc-lod

Digital Coast. (n.d.) *Coastal and Marine Ecological Standard.* Retrieved from http://www.c-sc.noaa.gov/digitalcoast/publications/cmecs

FAO Fisheries and Aquaculture Department. (2013). *Aquatic Sciences and Fisheries Abstracts (ASFA).* Retrieved from http://www.fao.org/fishery/asfa/en

Federal Geographic Data Committee (FGDC). (2013.) *Marine and Coastal Spatial Data Sub-committee: status update and discussion.* Retrieved from http://www.fgdc.gov/participation/coordination-group/meeting-minutes/2013/september/mcds-report-cg-20130910.ppt/view

Méndez, E. and Greenberg, J. (2012). Linked data for open vocabularies and HIVE's global framework. *El Profesional de la Información, 21*(3) 236–244.

University of Hawai'i Press. (2011.) *Pacific Science.* Retrieved from http://www.uhpress.hawai-i.edu/t-pacific-science.aspx

Wittenburg, P., & Broeder, D. (2002). Metadata overview and the semantic web. In P. Austin, H. Dry, & P. Wittenburg (Eds.), *Proceedings of the international LREC workshop on re-sources and tools in field linguistics.* Paris: European Language Resources Association.