EUROPEAN COMMISSION

Community research

Climate Information Portal

CLIPC DELIVERABLE (D -N°: 3.2)

Vocabulary Discovery Services

File name: CLIPC_32final.docx

Dissemination level: PU (public)

Author(s): ***Rob Thomas (BODC)***
Alexandra Kokkinaki (BODC)

Reviewer(s): ***Victoria Bennett (STFC)***
Jan-Willem Noteboom (KNMI)

Reporting period: *01/12/2013 – 31/12/2015*

Release date for review: ***25/02/2016***

Final date of issue: ***XX/XX/2016***

Revision table			
Version	**Date**	**Name**	**Comments**
1	25/02/2016	First version	Circulated for discussion and internal review
2	04/03/2016	Second version	Incorporated comments from internal review.

Abstract

The CLIPC project is building a data service layer across different environmental datasets. Vocabularies and mappings are integral in allowing a data service layer to function. As well as using established vocabularies, CLIPC is creating vocabularies which need to be well governed both in content and technical support. This document describes improved functionality delivered by the NVS for CLIPC to allow easier creation and continued content governance of CLIPC vocabularies and mappings, and the search functionality to identify appropriate vocabularies and their content.

Project co-funded by the European Commission's Seventh Framework Programme (FP7; 2007-2013) under the grant agreement n°607418

Table of contents

Executive Summary

The aim of this deliverable is to describe the improved functionality available from the NERC Vocabulary Server (NVS) in support of the CLIPC project. The CLIPC portal aims to deliver a data service layer across a range of datasets, which are produced by different science communities. These different science communities mark up their data using discipline specific vocabularies and different communities are at varying levels of maturity in the metadata models and the vocabularies they use to populate the required metadata fields. The NVS already supports interoperability across a number of environmental communities; the marine community with the SeaDataNet (SDN) and Marine Environmental Data Information Network (MEDIN) vocabularies, and also the Climate and Forecast community by hosting the Climate and Forecast Standard Names. CLIPC Deliverable 5.2 has identified metadata and controlled vocabularies for data, quality control and uncertainties in Earth Observation data. These vocabularies and mappings need to be reliably served and available to the community and the NVS is to be used to take on this responsibility for the CLIPC project. CLIPC has funded improvements to the NVS Editor allowing authorized users to maintain vocabularies (insert, modify or deprecate terms) through a web form, either on a term by term basis or through a bulk upload feature. Before the recent development activity the NVS allowed for machine to machine interoperability through 3 mechanisms: RDF restful API, SOAP API and a SPARQL endpoint. A web based search tool (NVS Search) has been built to facilitate direct searches by users from their web browser. This has been built on the NVS SPARQL endpoint. Previously search results were returned as XML. The search results are now delivered with content negotiation, as RDF/XML in the case of machine to machine interactions or as html to a user's web browser. The html that is returned allows the user to follow mappings between vocabularies as hyperlinks within the web page or follow hierarchies within concept schemes.

Introduction

CLIPC Deliverable 3.1 included an outline of how vocabulary services and Knowledge Organization Systems could be used by the CLIPC project. Deliverable 5.2 has identified the controlled vocabularies that will be used for data, quality control and uncertainty information within CLIPC. Since not all the metadata fields required to markup data for CLIPC have appropriate vocabularies established, where gaps existed domain experts were identified to determine the concepts within the vocabulary and their definitions. This has established the content governance of these vocabularies. The NERC Vocabulary Server (NVS) is to be used to serve these vocabularies and maintain the technical governance of providing these vocabularies alongside already established environmental vocabularies. CLIPC is aiming to assimilate data from a number of different environmental disciplines that may use vocabularies that overlap or are very similar in scope or concepts. Rather than markup all data with one system of vocabularies it is more practical to determine mappings between the major vocabularies used across these disciplines. Mappings can be used to indicate if terms are equivalent ("rain (in English)" is the same as "rein (in German)"), similar ("rain" is related to "snow"), broader ("precipitation" is broader than "rain") or narrower ("snow" is narrower than "precipitation") in meaning than another term. As part of the CLIPC project the NERC Vocabulary Server (NVS) web interface has been overhauled and enhanced to improve the ability to create, maintain and search these hosted vocabularies and the associated mappings.

As reported in D3.1, previously the most efficient way to create a new vocabulary was to email spreadsheets to BODC for loading internally. This could lead to short delays in publication. The NVS Editor has now been redesigned and through the revised system authorized users (those with governance rights) for a vocabulary can add, modify or deprecate concepts both on a term by term basis or for the vocabulary as a whole through a bulk upload function. In addition to mappings between concepts in vocabularies both hosted on the NVS, new functionality allows users to add mappings between NVS concepts and concept URLs that are external to the NVS. Details of how to use the editor are provided in section 1. All this functionality is available to the user from their web browser (https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_editor/).

Prior to the CLIPC project the NVS content could be searched through a SPARQL endpoint, RESTFUL or SOAP API. These all required some level of technical knowledge of web services to interrogate the NVS. The results were returned in a web browser as XML, which was not presented in a human user friendly way. While the ability to search the NVS via machine to machine technologies was well established a more human friendly search and results presentation was required. The search functionality developed allows a user to search for a vocabulary or specific concepts directly from a web browser (https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search/). The search results are

then presented using content negotiation: human readable style sheets. Details of the different search functionality that has been added to the NVS are provided in section 2.

1. Using the NVS Editor to manage vocabularies and mappings

The NERC Vocabulary Server (NVS) editor can be located from the BODC homepage as shown or by using the following link https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_editor/.

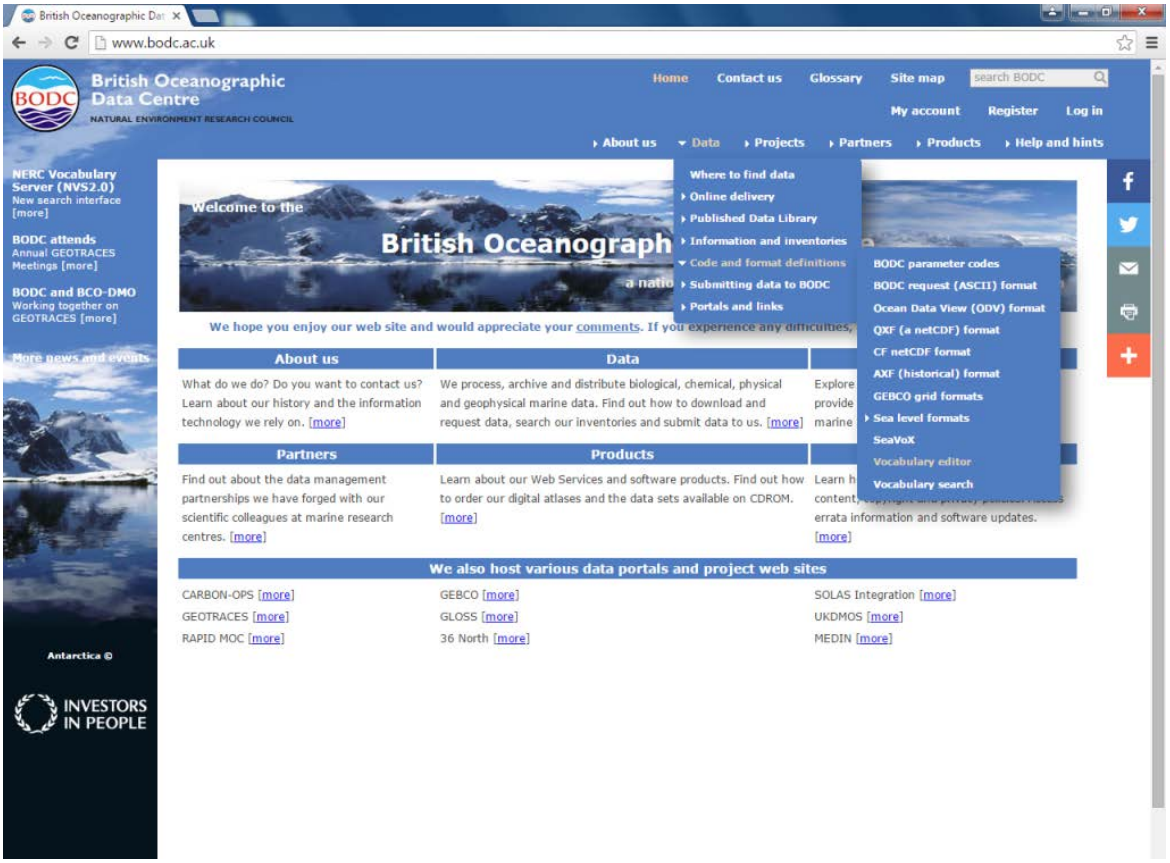

Figure 1.1: Accessing the NVS Editor from the BODC homepage.

The editor is intended to be used for updating existing vocabularies. If a new vocabulary needs to be set up then details of the vocabulary (Name, Description and Governance), as described in Deliverable 3.1, should be emailed to BODC (enquiries@bodc.ac.uk). Users should search the NVS to determine if appropriate vocabularies already exist and can seek advice from the BODC Vocabularies Management Group. It may be that an existing vocabulary can be extended. Once the “container” for the newly requested vocabulary is in place, the vocabulary can be populated through the NVS Editor.

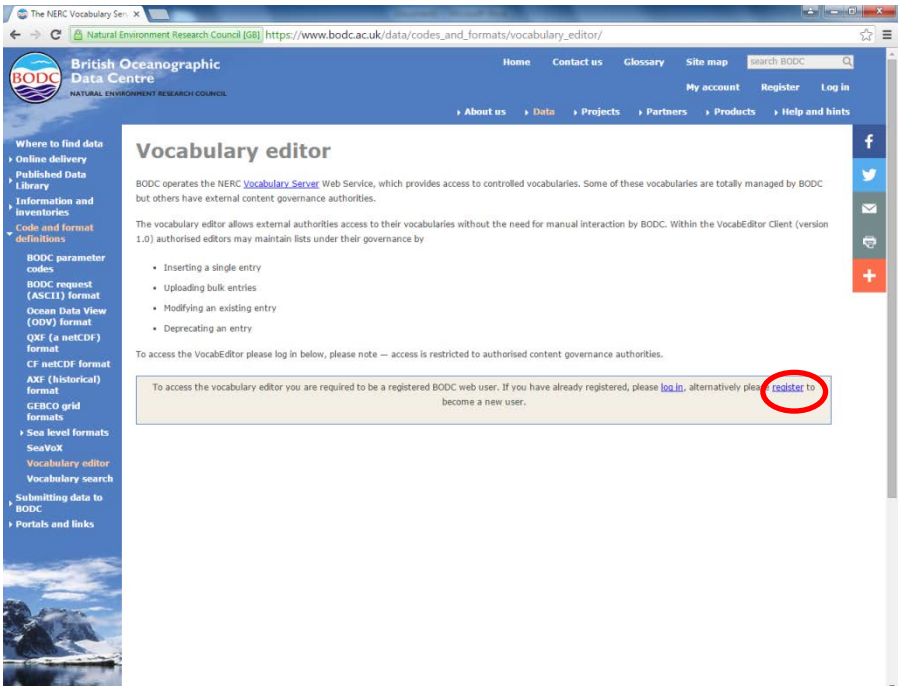

Figure 1.2: The editor requires a user to be registered before use.

Registration is through a simple web form. The details of which vocabularies a user wishes to have update permissions on should be supplied in the free text field. BODC will confirm with the appropriate governance group that the user should be granted these permissions.

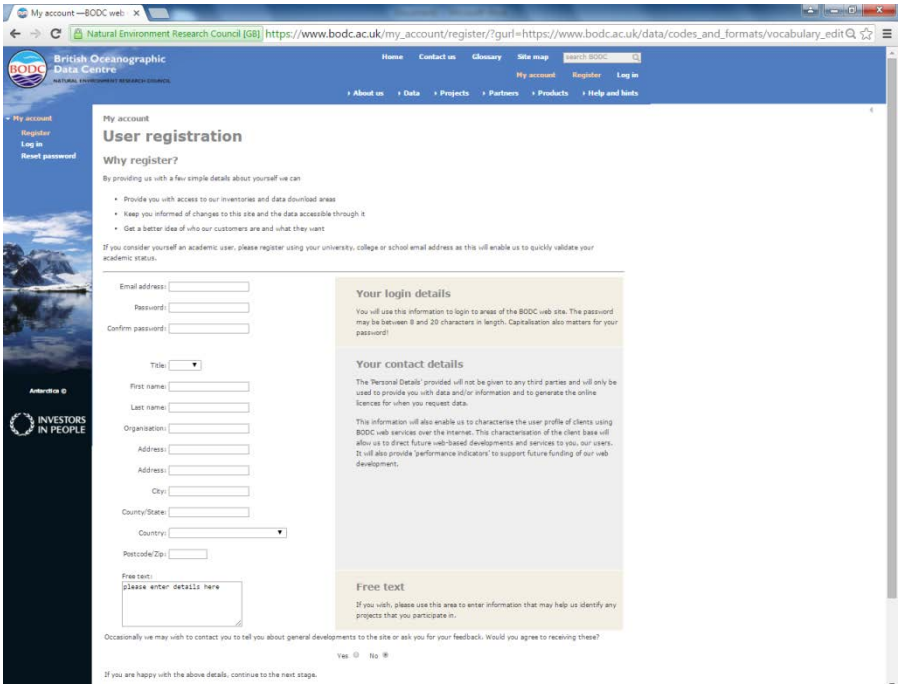

Figure 1.3: BODC website user registration page.

If a user is already registered and needs new permissions for a vocabulary they should email enquiries@bodc.ac.uk.

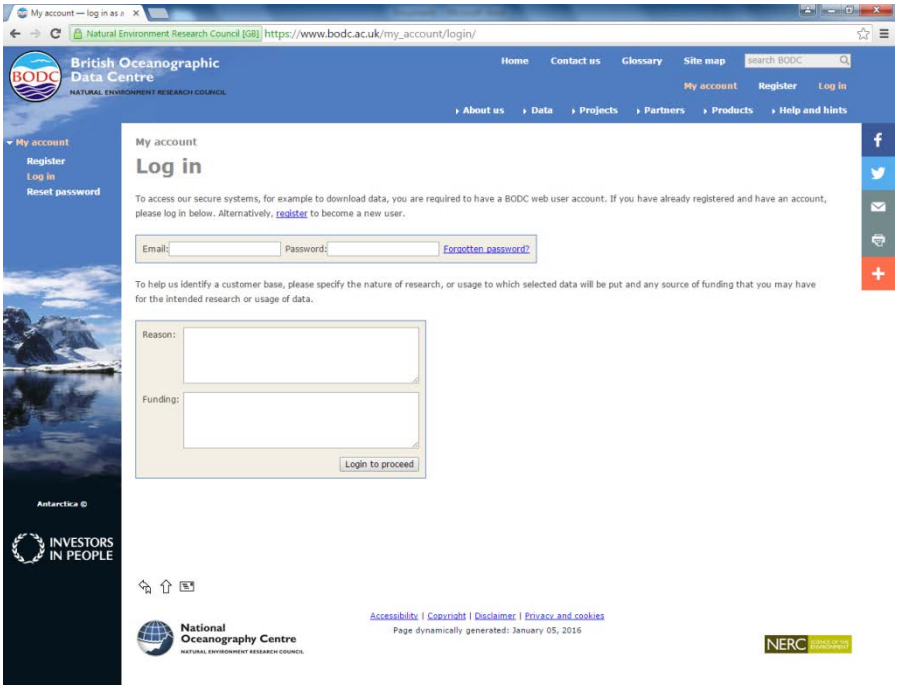

Figure 1.4: BODC website login page.

After logging in the user will be presented with a list of the vocabularies that they have permission to edit. Clicking on the hyperlinked List ID will take you to the vocabulary as displayed on the NVS2.

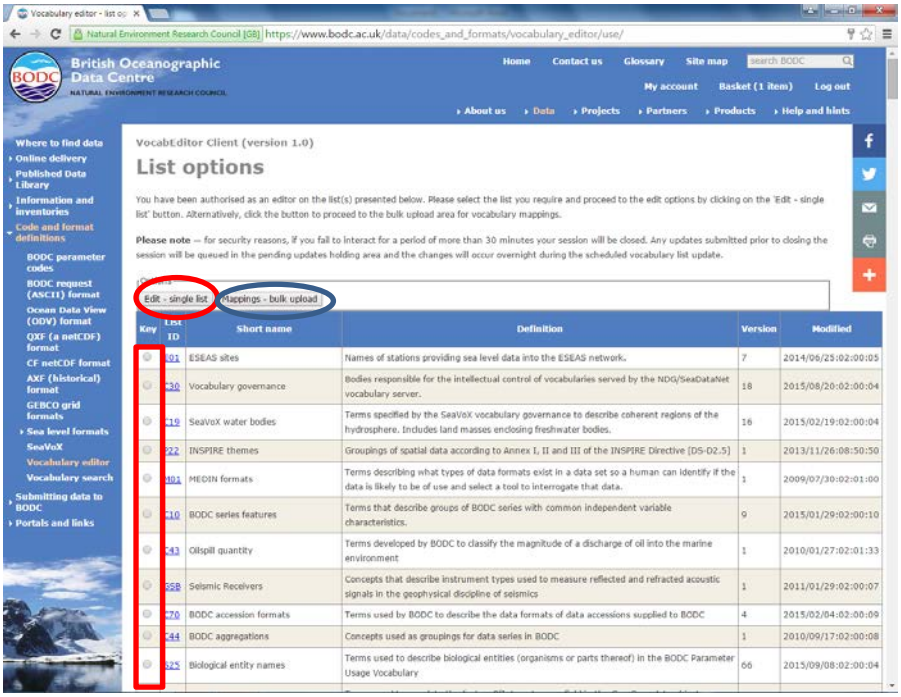

Figure 1.5: NVS Editor page showing vocabularies available for editing by the user.

Select the appropriate vocabulary using the radio button next to the vocabulary List ID and choose the "Edit – single list" option highlighted in red (see next page).
There is an option to add mappings for terms by using the "Mappings – bulk upload" option highlighted in blue.

1.1 Editing a vocabulary list

There are online help details available when using the editor.

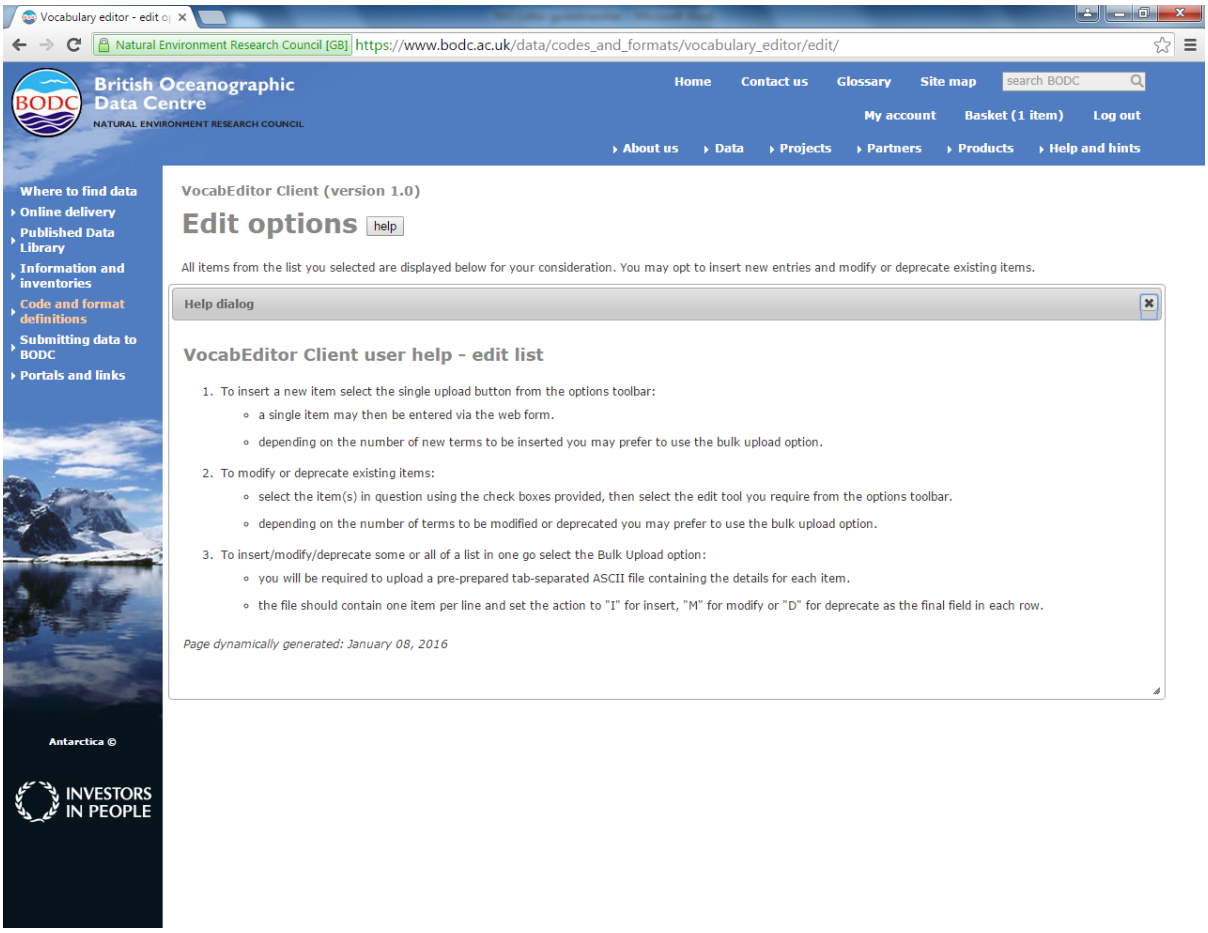

Figure 1.6: NVS Vocab Editor user help for making edits to a vocabulary.

Having selected the vocabulary to edit, the user will be presented with the terms already within the list and four options for editing (Single insert/Bulk update/Modify/Deprecate).

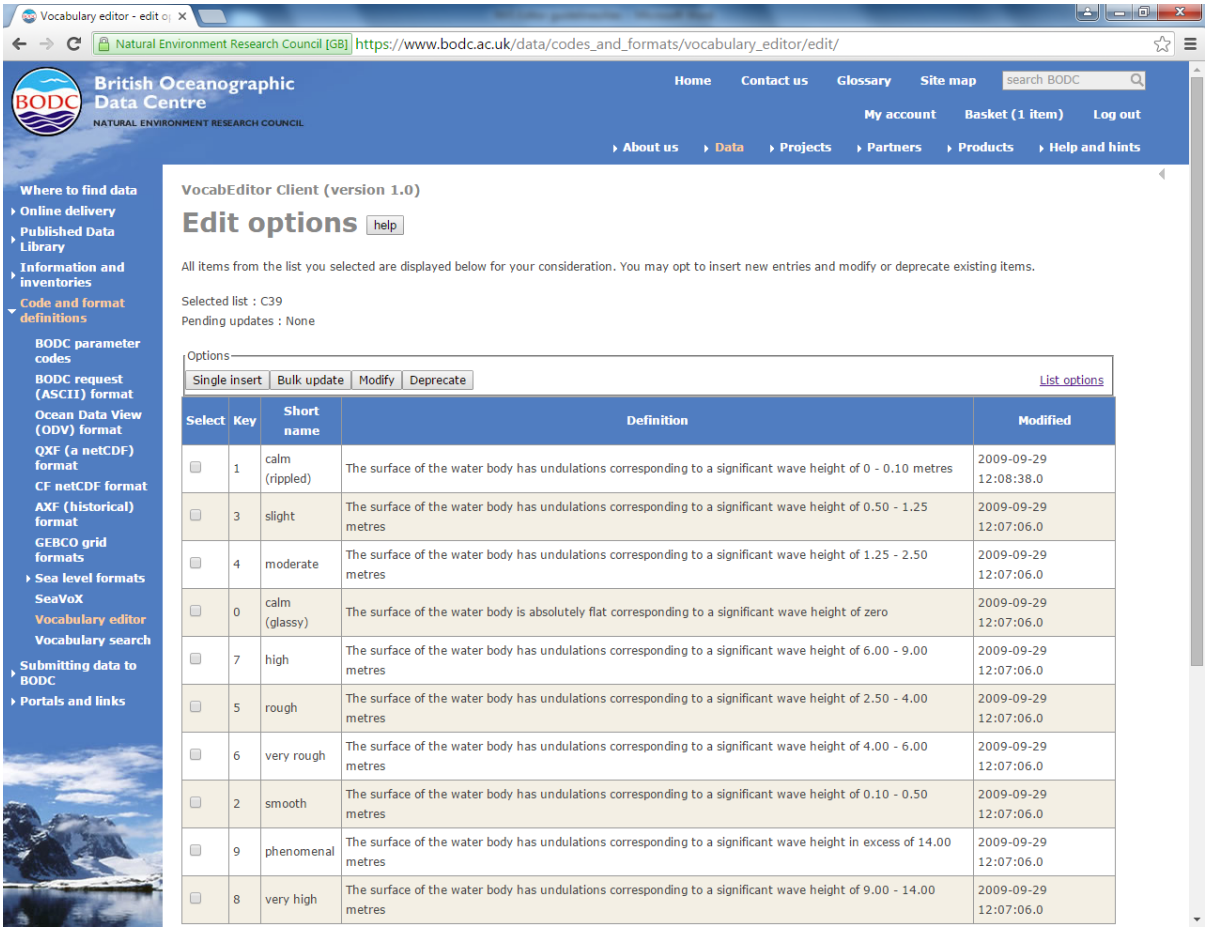

Figure 1.7: Example of how a selected vocabulary's terms are listed for editing.

a) Single insert

The user manually types text into each of the required fields.

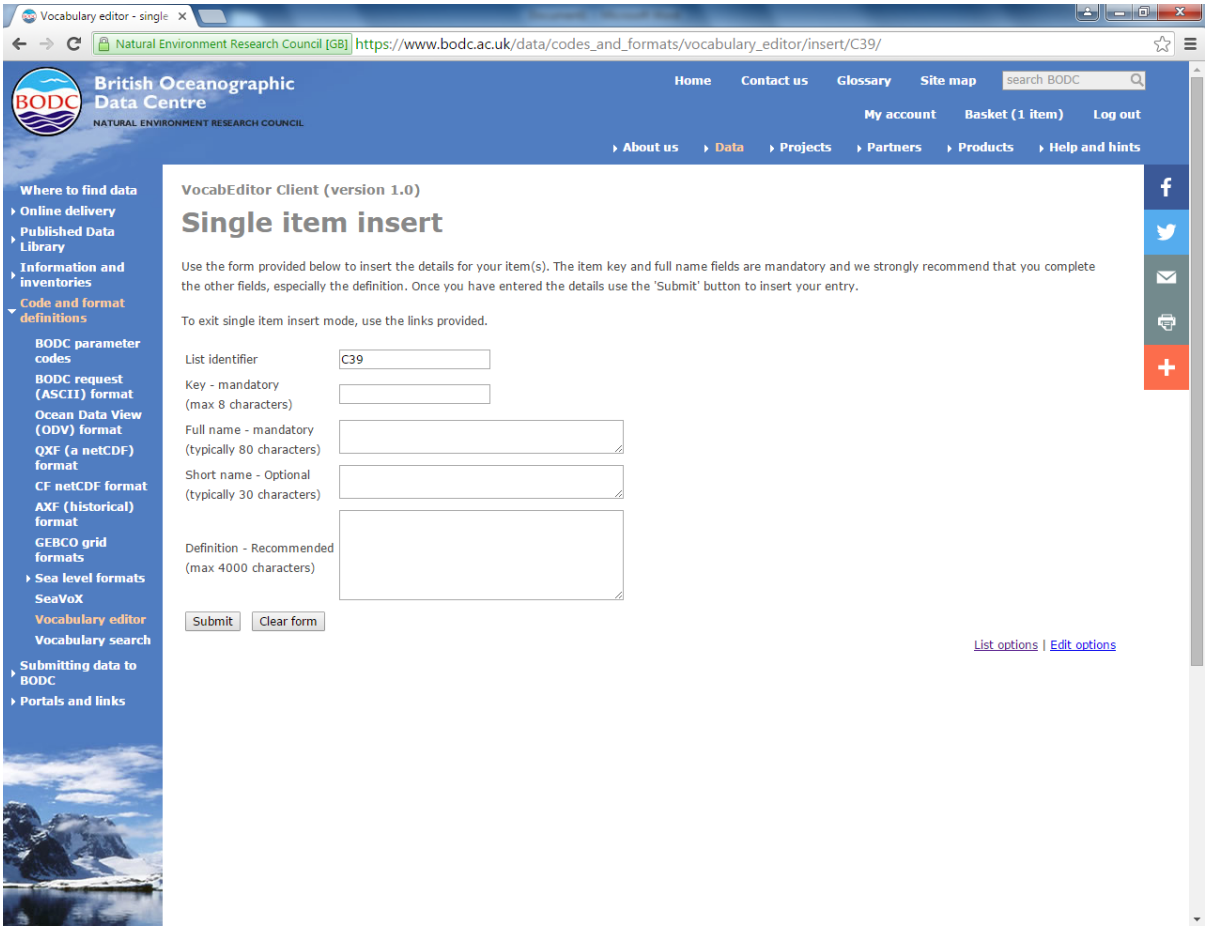

Figure 1.8: Web form for inserting a single item into a vocabulary.

b) Bulk update

The upload file must be TAB delimited with five elements per row as detailed below.

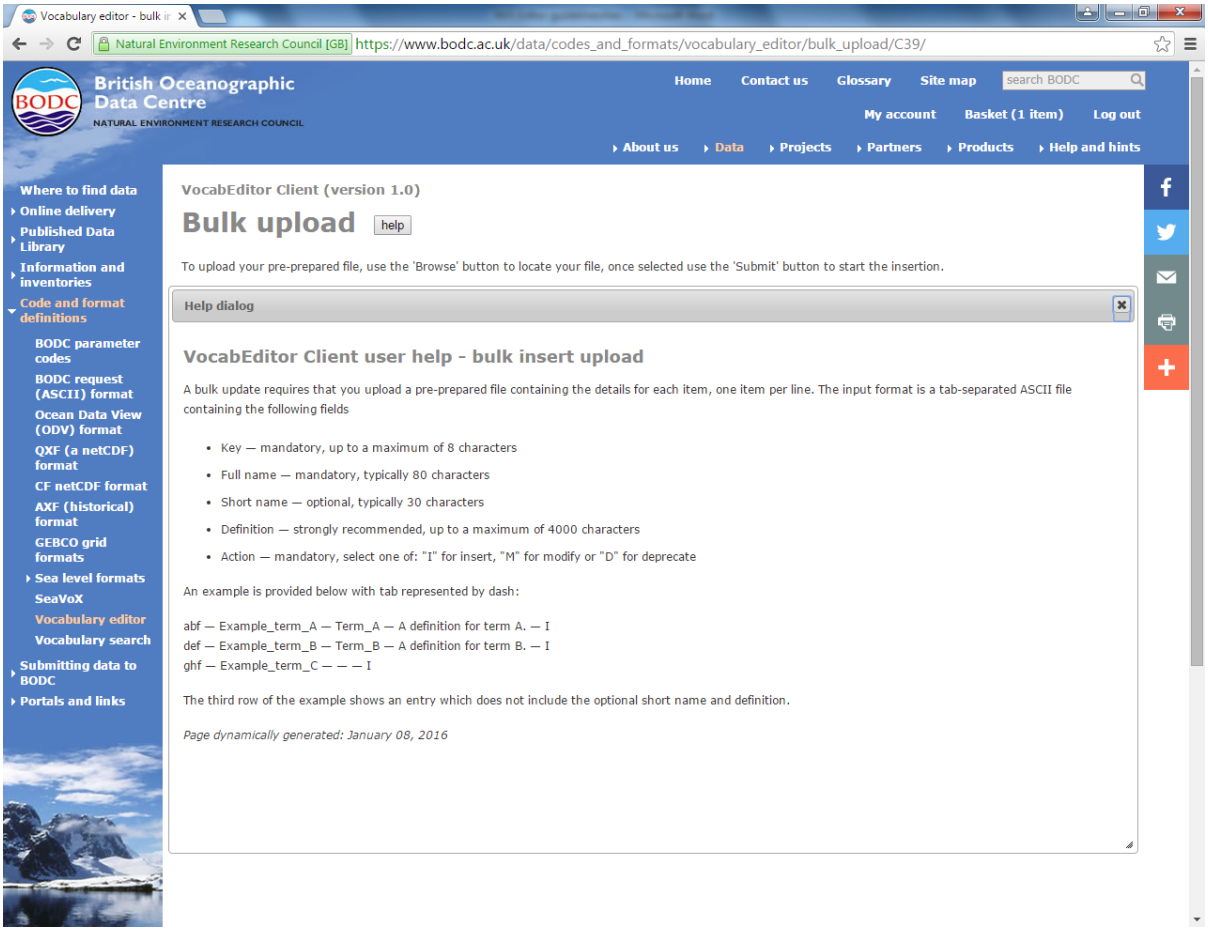

Figure 1.9: NVS Vocab Editor user help detailing how to set up a file for the bulk update of a vocabulary.

The content of the vocabulary bulk upload file should be laid out as tab delimited text as follows; containing Key, Full name, Short name, Definition and Action in that sequence.

abf	Example_term_A	Term_A A definition for term A.	D
def	Example_term_B	Term_B A definition for term B which is being modified.	M
ghf	Example_term_C	Term_C A definition for term C.	I

On loading, each row in the file will be reported with a status code from the following list.

VocabEditor Client user help - List terms upload- status codes

Status code	Meaning
200	Success - item inserted - currently queued in the holding area.
400	Specified insert is already member of the list or is currently queued in the holding area.
401	User not authorised - user session has elapsed please login again.
403	User not authorised - insufficient permissions for list specified.
404	Incorrect format. The list identifier not a valid list.
500	Oracle error - please try again (please contact enquiries@bodc.ac.uk if problem persists).

The rows that are unsuccessful should be corrected accordingly and reloaded or the user contact BODC to resolve any issues over permissions.

c) Modify

Terms to be modified are selected, then all fields except the Key can be modified. The action will be applied by clicking the “Modify” button.

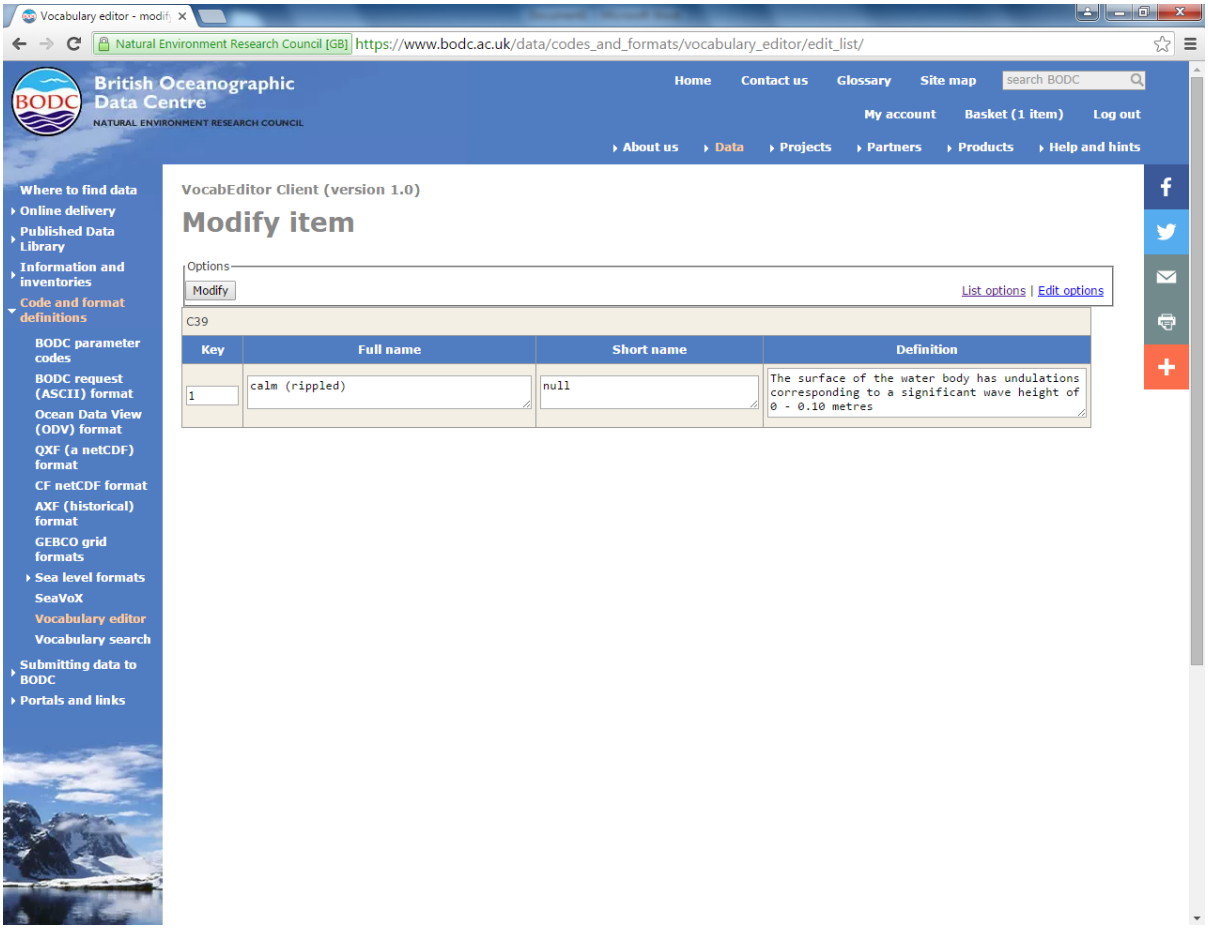

Figure 1.10: NVS Vocab Editor web form for modifying an existing item.

The key field cannot be changed. The full name, short name and definition should only be changed to clarify details of the concept NOT to refer to a different concept. If a term is to be replaced then the deprecate function should be used and a new concept added to the list.

d) Deprecate

This requires terms to be selected prior to clicking the "Deprecate" button. The user is then given a second chance to confirm the term(s) to be deprecated.

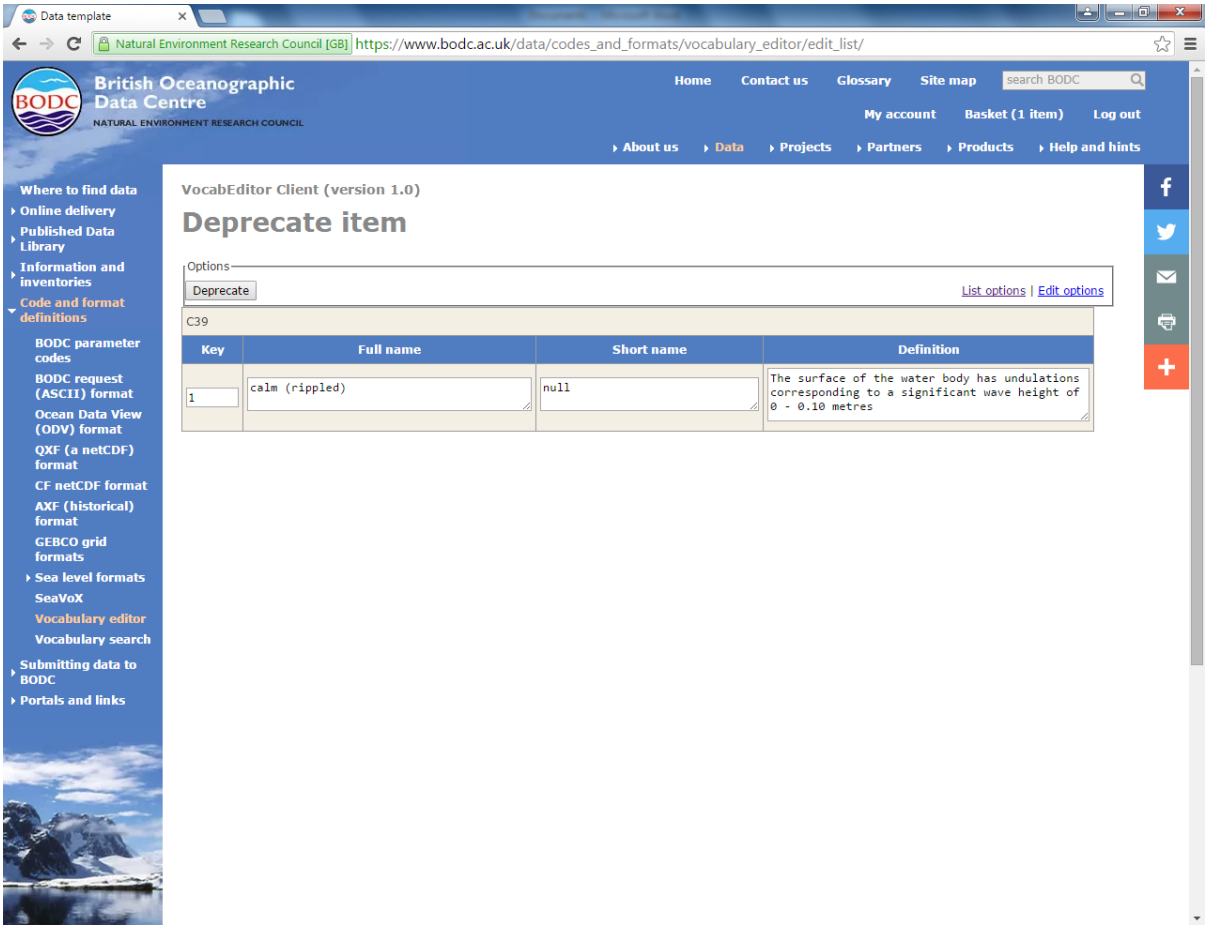

Figure 1.11: NVS Vocab Editor web form for deprecating an existing item.

1.2 Mappings

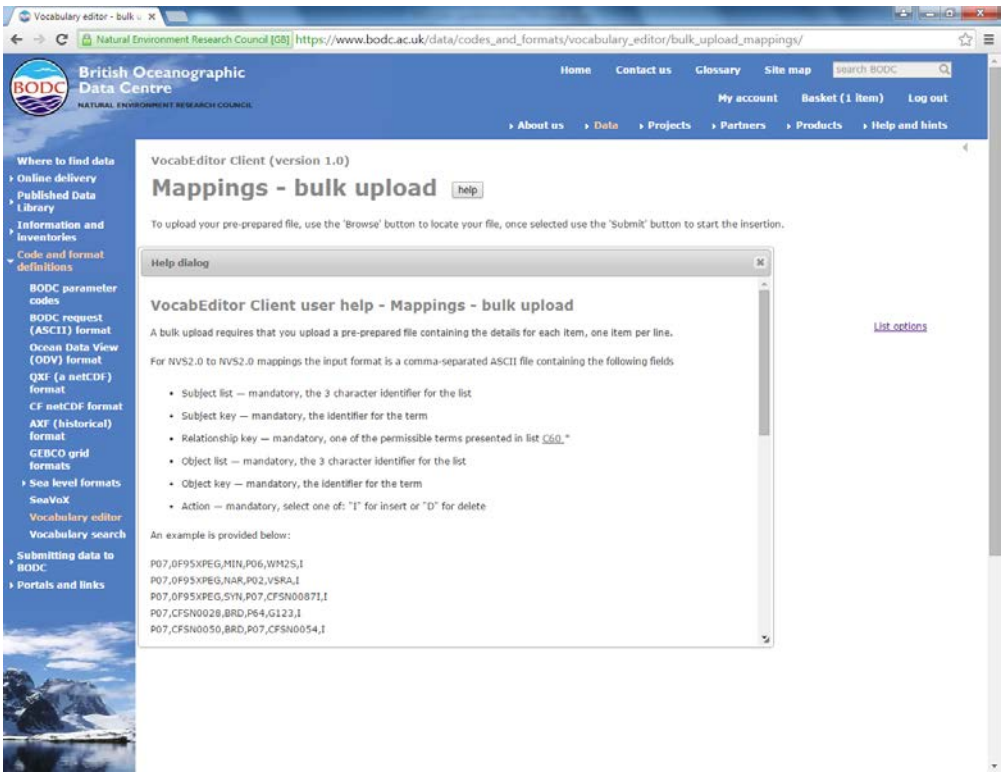

Figure 1.12: The above screen print provides the example layout for inserting mappings between two NVS vocabulary lists.

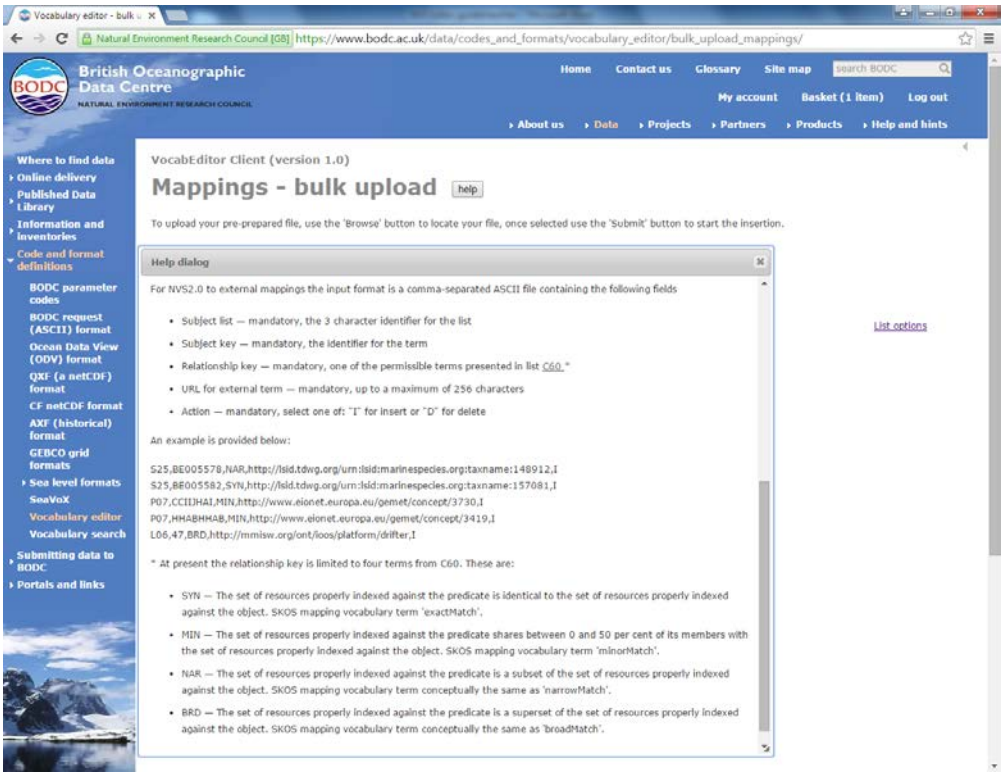

Figure 1.13: The following screen print provides the example layout for inserting mappings between an NVS vocabulary list and a non-NVS served concept URL.

The content of the mappings bulk upload file should be laid out as comma separate value text as follows; containing Subject List, Subject Key, Relationship Key, URL for external term and Action in that sequence.

```
S25,BE005578,NAR,http://lsid.tdwg.org/urn:lsid:marinespecies.org:taxname:148912,I
S25,BE005582,SYN,http://lsid.tdwg.org/urn:lsid:marinespecies.org:taxname:157081,I
P07,CCIIJHAI,MIN,http://www.eionet.europa.eu/gemet/concept/3730,I
P07,HHABHHAB,MIN,http://www.eionet.europa.eu/gemet/concept/3419,I
L06,47,BRD,http://mmisw.org/ont/ioos/platform/drifter,I
```

After loading the mappings each row will be returned with a status code from the following list.

VocabEditor Client user help - Mappings - status codes

Status code	Meaning
200	Mapping successfully uploaded.
400	Mapping specified already pending.
401	User has insufficient permissions to generate a mapping for this list.
409	An empty or null mapping, i.e. a blank line in your input file.
410	Incorrect format. Expected mapping is: Subject list, Subject key, Relationship key, Object list, Object key.
412	Incorrect format. Invalid option as a mapping term cannot be mapped to itself.
414	Incorrect format. List C60 (the relationship option) is not a permitted value for the Subject and/or Object list.
415	Incorrect format. One or more of the list or key terms entered were not found.
416	Incorrect format. Expected mapping is Subject list, Subject key, Relationship key and URL.
417	Invaild URL. A mapping to a NVS2.0 URL (i.e. vocab.nerc.ac.uk) is not permitted. Please use the NVS2.0 to NVS2.0 option for such mappings.
418	Invalid URL. An invalid response was returned from the external URL. Please check that it exists.
500	Oracle error. Please try again (please contact enquiries@bodc.ac.uk if the problem persists).

The rows that are unsuccessful should be corrected accordingly and reloaded.

2. Using the NVS search to discover vocabularies and concepts

A web-based Vocabulary Search has been built as part of the CLIPC project.

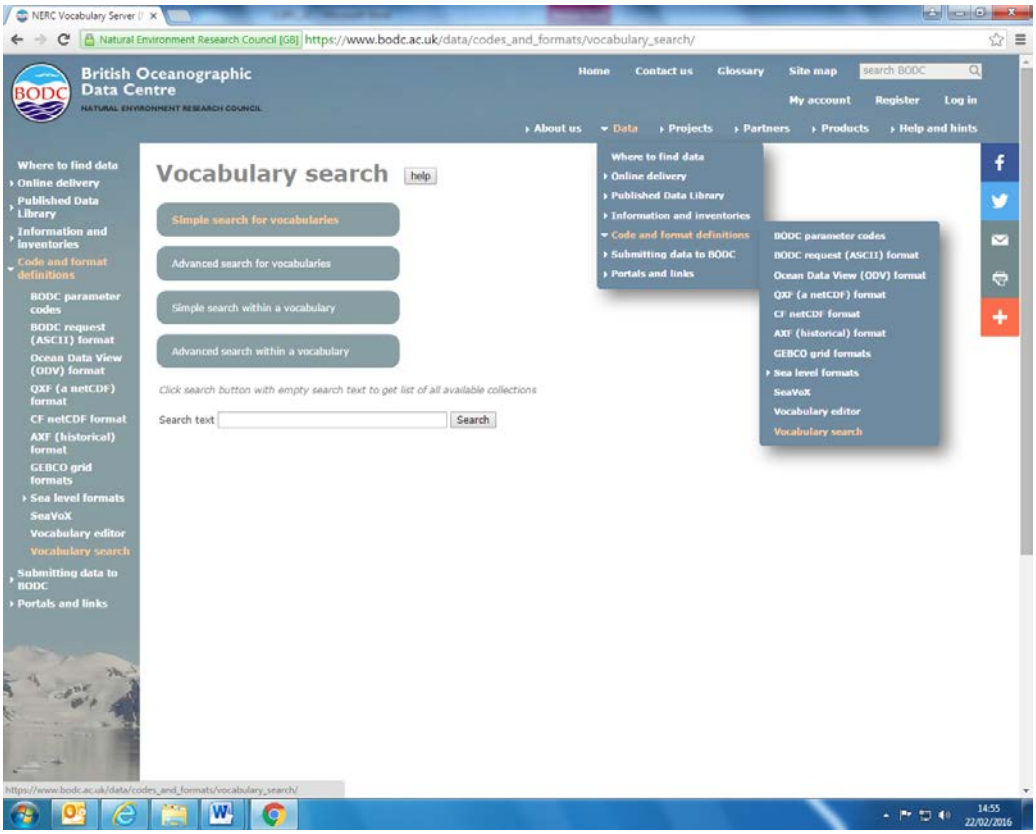

Figure 2.2: Locating the NVS Vocabulary search from the BODC web site.

The search tool has four different options offering two types of searches both with either simple or advanced behaviour:

- The first search type searches for controlled vocabularies (options 1 and 2 on the web page) and then allows searching of the located vocabularies.
- The second search type (options 3 and 4 on the web page) searches a vocabulary known to the user by the identifier

The two types of search behaviour that can be applied are:

- Simple search behaviour (options 1 and 3) returns every possible hit, including near matches and matches with fields other than titles like definition, alternative label and identifier.
- Advanced search behaviour (options 2 and 4) by default adds case sensitivity to the search text and searches only within the identifier, title and alternative label. Controls are provided that allow the advanced search to be broadened if required

Search strings for both kinds of behaviour may include one or more '%' characters as wildcards. The ordering of text within the search strings is important. For example when searching in vocabulary P01 the measurement type must come before what was measured, this in turn must come before the matrix in which the measurement was made.

The advanced search controls are:

- Case sensitive - determines whether case is taken into account during the search.

- Exact match - determines whether the search string is automatically wrapped in wildcard characters before the search. If checked the characters are not added and the search string must match the target string and not just be a part of it.
- Target inclusion - allows the fields of the vocabulary catalogue or vocabulary entry searched to be specified.
- Type selection - controls whether a search for vocabularies is based on the vocabulary catalogue or the text of the vocabulary entries.
- Deprecation status - controls whether deprecated vocabulary entries are searched (searches within vocabularies only).
- Additional filter - allows an inclusive (e.g. lead or mercury) or exclusive (e.g. lead not liver) additional search strings to be specified (searches within vocabularies only).
- Results per page - controls output pagination.

Each search function is described in more detail and with example screenshots in the following sections.

2.1 Simple search for vocabularies

This is the default search when arriving at the search tool, as shown in Figure 1. The user adds the text to be searched in the search field. The results shown in Figure 2 show a list of Vocabularies containing the searched term, ranked by frequency of appearance in each vocabulary.

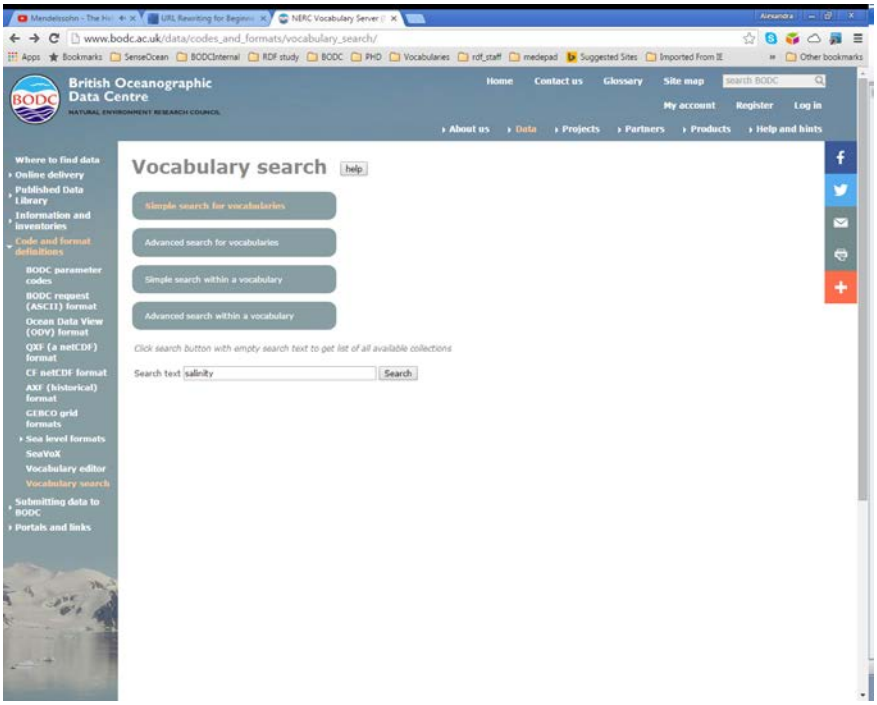

Figure 2.2: Simple search for Vocabularies.

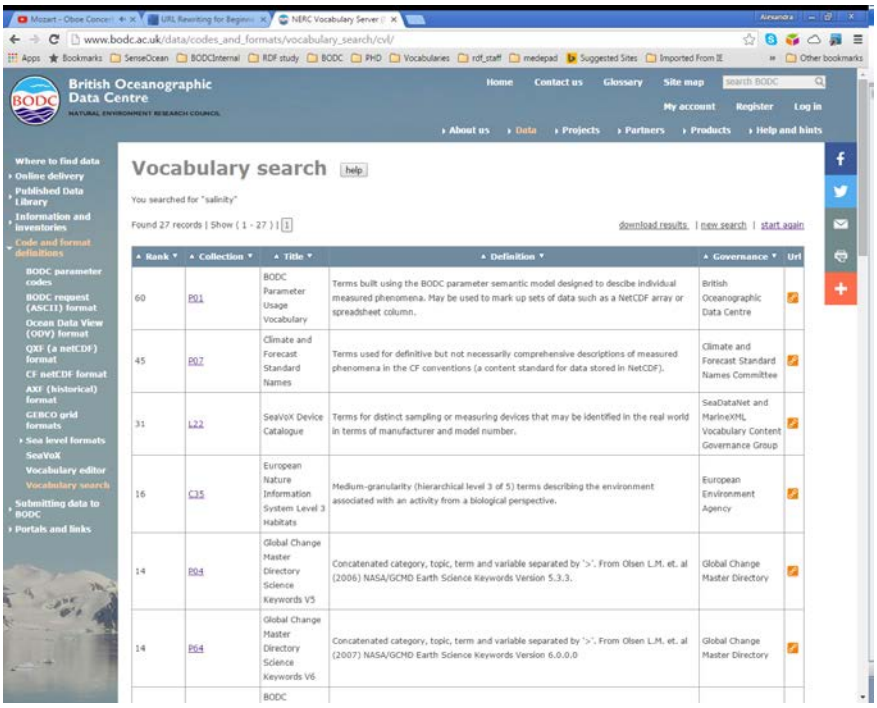

Figure 2.3: Simple Search for Vocabularies Results.

2.2 Advanced search for vocabularies

The advanced search for vocabularies by default adds case sensitivity to the search. The following example searches in the catalogue descriptions, to find vocabularies described as GCMD, as shown in Figure 3. The results are shown in Figure 4.

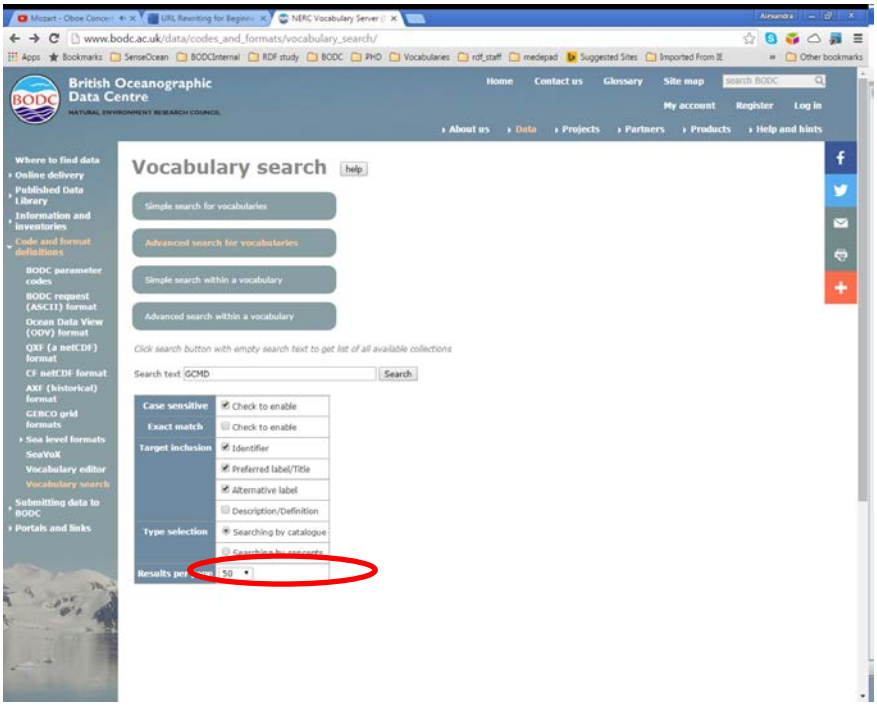

Figure 2.4: Advanced search for vocabularies

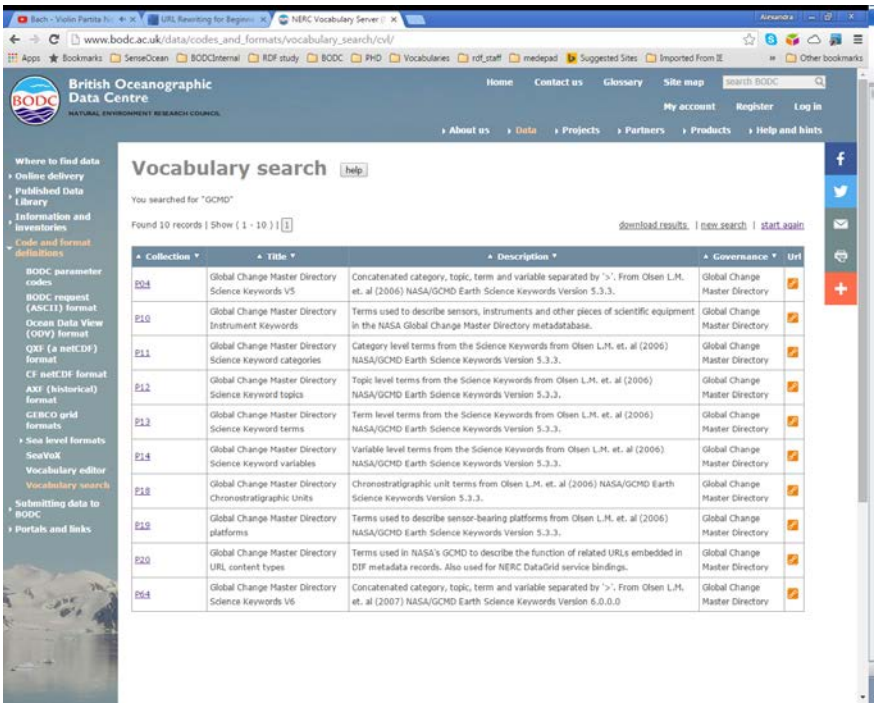

Figure 2.5: Advanced search for vocabularies results.

2.3 Simple search within a vocabulary

In this example the user selects one of the available Vocabularies shown in Figure 5, and adds a term to be searched only within the selected vocabulary. Hovering over each Vocabulary code reveals the title of the Vocabulary. The result (Figure 6) is a list of concepts containing the search term or string with links to the NVS concepts.

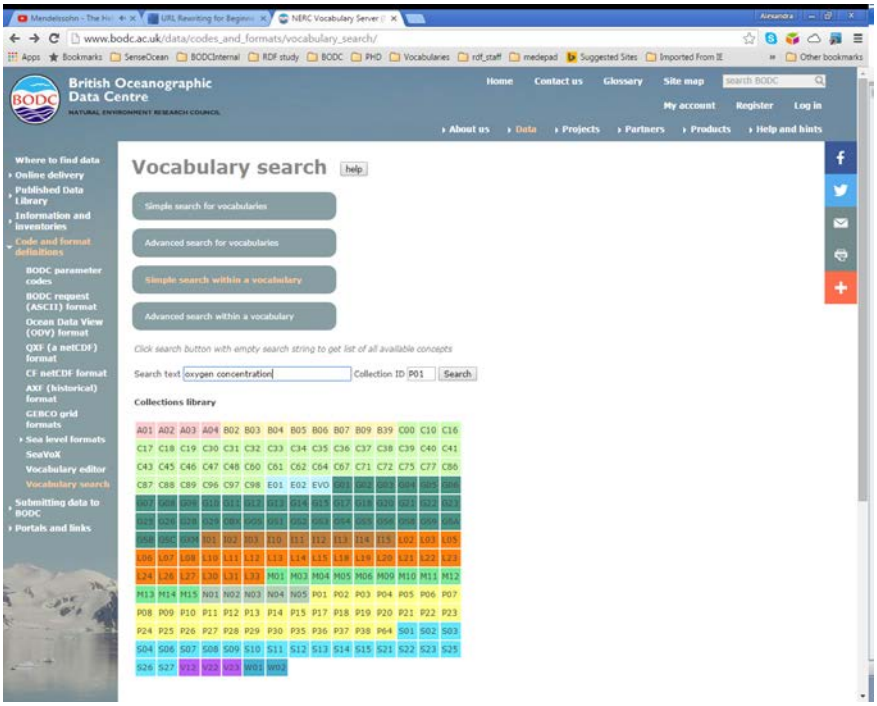

Figure 2.6: Simple search within a vocabulary.

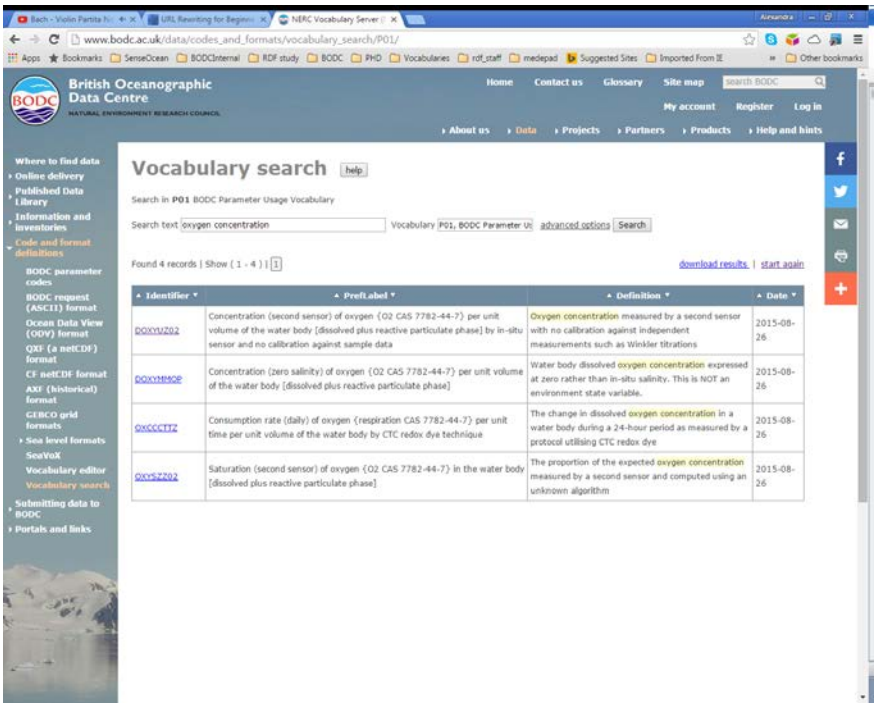

Figure 2.7: Simple search within a vocabulary results.

2.4 Advanced search within a vocabulary

Advanced users can select the Vocabulary to be searched in advanced mode. In the following example the user searches for “Mytilus but not edulis” as shown in Figure 7. Figure 8 shows the result concepts satisfying the query.

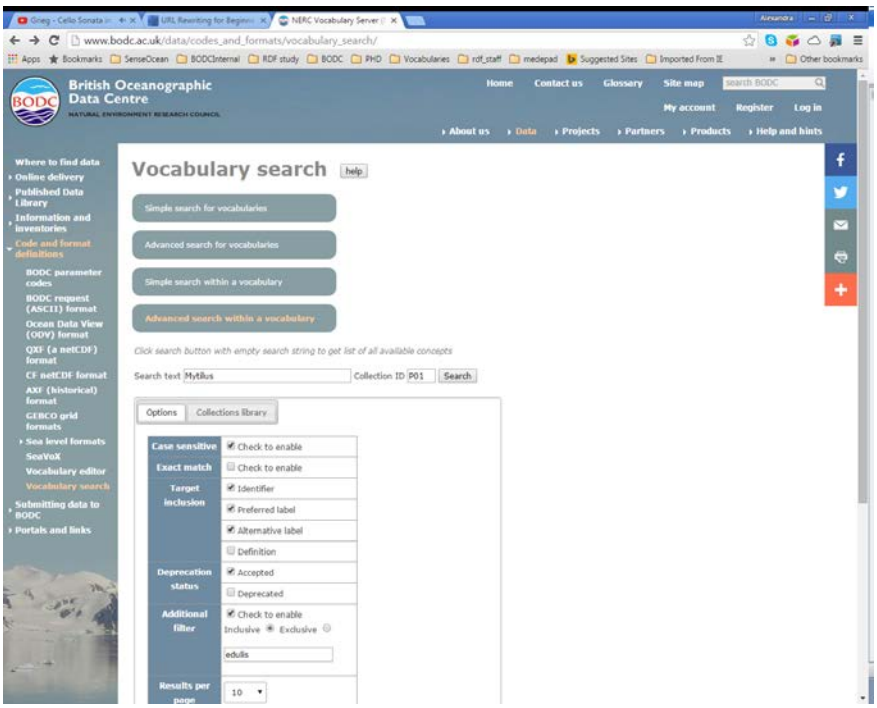

Figure 2.8: Advanced search within a vocabulary.

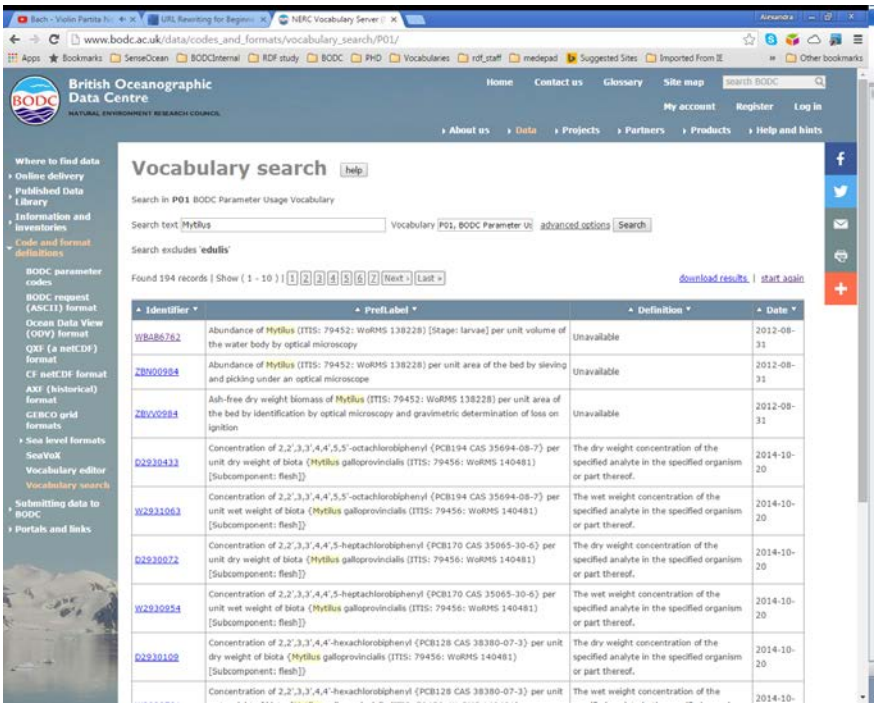

Figure 2.9: Advanced search within a vocabulary results.

3. Presentation of search results

The NVS now displays results using content negotiation through 2 formats:

- RDF
- HTML

Content negotiation is a mechanism defined in the Hyper Text Transfer Protocol (HTTP) specification that makes it possible to serve different versions of a document from the same URI. In this way the client can specify which version they would like to receive from the server when it sends a request. So when an HTTP client attempts to dereference a URI, the request can specify which types of content the client would prefer to receive in response. In the request for the URI the value of the Accept header gives format types corresponding to the preferred content types. When the server receives the request, it can use the value of the Accept field to select the most appropriate response from those available, attempting to meet the preference of the client as closely as possible. This is achieved by including an Accept field in the header of the request message from the client to the NVS. In our case if a human accesses an NVS page, the browser's Accept field will most probably equal text/html and the server returns html to the clients browser. If a machine accesses NVS then the Accept header will most probably be application/rdf+xml. This enables the appropriate display as either machine to machine or as a human readable browser display.

4. Future work

The NVS is built on the Simple Knowledge Organization System (SKOS), which employs a semi-formal organization structure through mappings (e.g. "broader than", "narrower than"). Overlaying SKOS with OWL allows the creation of formal ontologies with concepts as classes and allows for richer queries and inferencing.

The NVS currently allows only 4 mapping options between concepts (narrower, broader, same, related). Therefore we are looking to increase the number of predicates available for use in mappings. Deliverable 5.2 has identified the predicates that would be of use to the CLIPC project.

There are currently 2258 mappings between SDN Parameter Discovery Groups (P02) and the Climate Forecast (CF) Standard Names (P07). Of the 3060 CF Standard Names there are 2123 currently mapped to P02 concepts. The remaining CF Standard Names will be mapped to P02 concepts as appropriate with the help of the CF Standard Names co-ordinator, Alison Pamment at STFC.