



EXTRACT



Lifewatch Greece Portal

E-mail

Password

Sign In

[Forgot your password?](#)

[Register](#)



Literature Mining

<https://lm.portal.lifewatchgreece.eu>

Biodiversity literature and data constitute a vast public resource open to mining and knowledge extraction.

Associating organisms to key features of their life, for example the environment in which they live, the way they feed, their breeding habits, is cornerstone in explaining biodiversity patterns and informing ecological decisions.

The Literature Mining virtual Lab (LM-vLab) aims at both:

- the automatic extraction of species - traits associations from the literature
- augmenting [Lifewatch Greece](#) species related information based on the above

SPECIES

Identification of Taxonomic Mentions in Text

ENVIRONMENTS

Identification of Environment Descriptive Terms in Text

EXTRACT

Interactive Extraction of Metadata

JensenLab

Cellular Network Biology



The [EXTRACT](#) annotation tool, and the [ENVIRONMENTS](#) and [SPECIES/ORGANISMS](#) taggers are relevant LM-vLab tools to this end. All three are being employed for standard compliant term suggestion to describe the environmental context of metagenomic records, while [ENVIRONMENTS](#) is also being used for identification of [Environment Ontology](#) terms in text and the annotation of the [Encyclopedia of Life](#).

All tools are developed in collaboration and maintained at the group of Prof. Lars Juhl Jensen, Novo Nordisk Foundation Center for Protein Research, Copenhagen, Denmark.

Developed by HCMR and FORTH

EXTRACT

Interactive Extraction of Metadata

extract@hcmr.gr

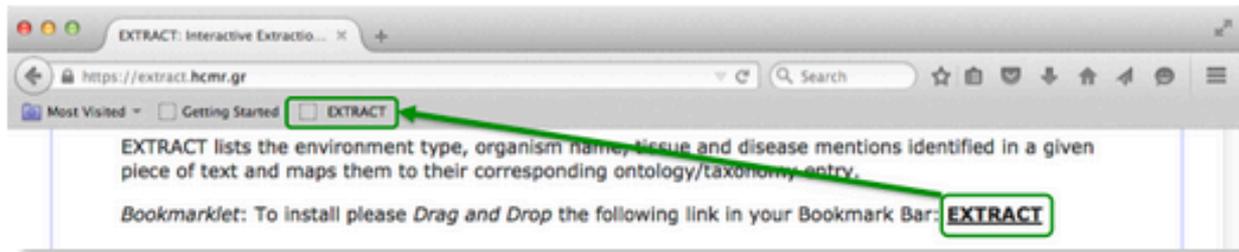
About

Demo

Help

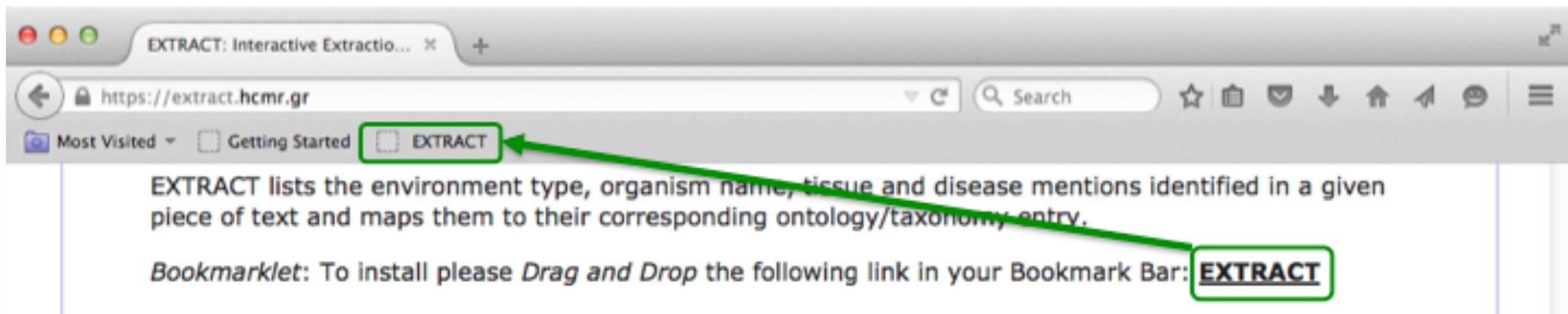
EXTRACT lists the environment type, organism name, tissue and disease mentions identified in a given piece of text and maps them to their corresponding ontology/taxonomy entries.

Bookmarklet: To install please *Drag and Drop* the following link in your Bookmark Bar: [EXTRACT](#)



Usage: a. select a piece of text of interest in a web page and then b. click on the bookmarklet. c. A pop-up such as the following will appear (supported browsers: [Chrome](#), [Firefox](#), [Safari](#)). By hovering the mouse cursor over the text tags or the table rows you can visually inspect which words have been identified as which entities.

EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation Pafilis E, Buttigieg PL, Ferrell B, et al.. (2016). **Bioinformatics**, 2016, baw005. doi:10.1093/bioinformatics/btv04



EXTRACT: Interactive Extractio... ✕ +

https://extract.hcmr.gr

Most Visited ▾ Getting Started EXTRACT

EXTRACT lists the environment type, organism name, tissue and disease mentions identified in a given piece of text and maps them to their corresponding ontology/taxonomy entry.

Bookmarklet: To install please *Drag and Drop* the following link in your Bookmark Bar: **EXTRACT**

Drag n' Drop Installation



EXTRACT: Interactive Extractio... ✕ +

https://extract.hcmr.gr

Most Visited ▾ Getting Started EXTRACT

EXTRACT lists the environment type, organism name, tissue and disease mentions identified in a given piece of text and maps them to their corresponding ontology/taxonomy entry.

Bookmarklet: To install please *Drag and Drop* the following link in your Bookmark Bar: **EXTRACT**

The screenshot shows a web browser window with the following elements:

- Browser Tab:** JGI GOLD | Study
- Address Bar:** <https://gold.jgi.doe.gov/studies?id=Gs005>
- Navigation:** Back, Forward, Home, Refresh, Search, and menu icons.
- Page Header:** Most Visited, Getting Started, and EXTRACT tabs.
- Left Sidebar:**
 - Sequencing Projects
 - Analysis Projects **56,349**
- Main Content Area:**

STUDY NAME	
GOLD Study ID	Gs0059071
Study Name	Marine Synechococcus communities from coastal surface water at La Jolla, California, USA
Other Names	Marine Synechococcus metagenome experiment
NCBI Umbrella Bioproject Name	
NCBI Umbrella Bioproject ID	
Legacy ER Study ID	14071
Legacy GOLD ID	Gm00146
Added By	Nikos Kyrpides on 2008-10-30
Last Modified By	Auto script update processes on 2014-06-19
STUDY DESCRIPTION	
PI	Brian Palaniappan
Description	From a sample of coastal California seawater, the marine cyanobacteria of the genus Synechococcus were enriched by flow cytometry-based sorting and the population metagenome was analyzed with 454 sequencing technology.
Relevance	
Study Information Link	
Study Information Link URL	
Study Information Visibility	Public

<https://gold.jgi.doe.gov/studies?id=Gs0059071>

The screenshot shows a web browser window with the URL <https://gold.jgi.doe.gov/studies?id=Gs005>. The page displays details for a study with ID Gs0059071. A blue circle highlights the 'EXTRACT' button in the top navigation bar. Another blue circle highlights the 'Description' field in the 'STUDY DESCRIPTION' section.

STUDY NAME	
GOLD Study ID	Gs0059071
Study Name	Marine Synechococcus communities from coastal surface water at La Jolla, California, USA
Other Names	Marine Synechococcus metagenome experiment
NCBI Umbrella Bioproject Name	
NCBI Umbrella Bioproject ID	
Legacy ER Study ID	14071
Legacy GOLD ID	Gm00146
Added By	Nikos Kyrpides on 2008-10-30
Last Modified By	Auto script update processes on 2014-06-19
STUDY DESCRIPTION	
PI	Brian Palaniappan
Description	From a sample of coastal California seawater, the marine cyanobacteria of the genus Synechococcus were enriched by flow cytometry-based sorting and the population metagenome was analyzed with 454 sequencing technology.
Relevance	
Study Information Link	
Study Information Link URL	
Study Information Visibility	Public

<https://gold.jgi.doe.gov/studies?id=Gs0059071>

[Show help page](#)
[Open popup in a new tab](#)
[Close](#)

EXTRACT ? ↑ ×

Selected text

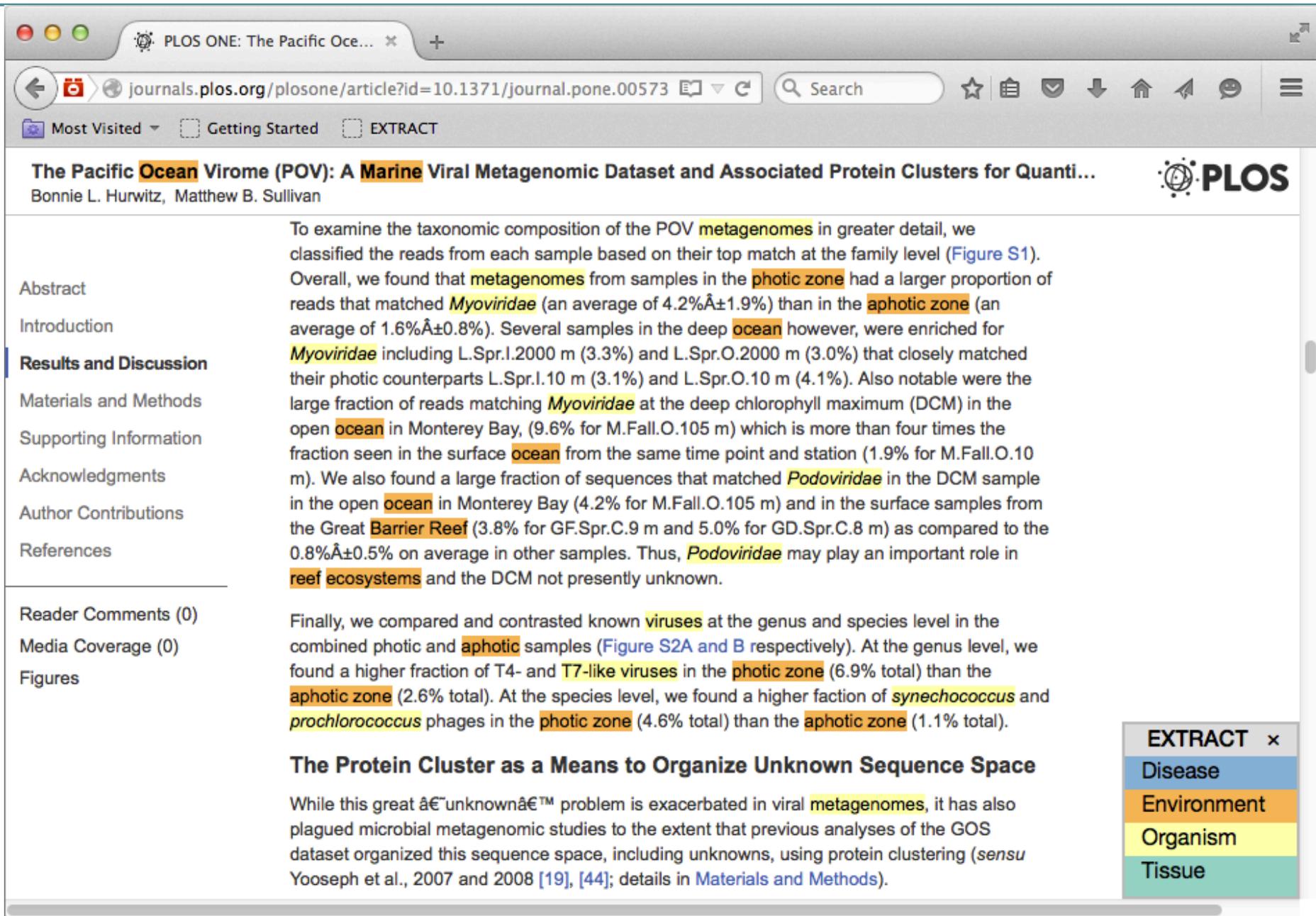
From a sample of **coastal** California **seawater**, the marine **cyanobacteria** of the genus **Synechococcus** were enriched by flow cytometry-based sorting and the population **metagenome** was analyzed with 454 sequencing technology.

Identified terms

Type	Name	Identifier
Environment	Coast	ENVO:00000303
Environment	Sea water	ENVO:00002149
Organism	Cyanobacteria	1117
Organism	Metagenomes	408169
Organism	Synechococcus	1129

Annotated user selected text
 Identified entity summary table
 Highlighted on mouse over: related tags and entities
 Additional information link

Copy to Clipboard and **Save** as tab separated values the list of extracted entities along with the selected text and the source page URL



The screenshot shows a web browser displaying a PLOS ONE article. The browser's address bar shows the URL: journals.plos.org/plosone/article?id=10.1371/journal.pone.00573. The article title is "The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quanti...". The authors listed are Bonnie L. Hurwitz and Matthew B. Sullivan. The article is categorized under "EXTRACT".

The article content is as follows:

Abstract

To examine the taxonomic composition of the POV metagenomes in greater detail, we classified the reads from each sample based on their top match at the family level (Figure S1). Overall, we found that metagenomes from samples in the photic zone had a larger proportion of reads that matched *Myoviridae* (an average of 4.2%±1.9%) than in the aphotic zone (an average of 1.6%±0.8%). Several samples in the deep ocean however, were enriched for *Myoviridae* including L.Spr.I.2000 m (3.3%) and L.Spr.O.2000 m (3.0%) that closely matched their photic counterparts L.Spr.I.10 m (3.1%) and L.Spr.O.10 m (4.1%). Also notable were the large fraction of reads matching *Myoviridae* at the deep chlorophyll maximum (DCM) in the open ocean in Monterey Bay, (9.6% for M.Fall.O.105 m) which is more than four times the fraction seen in the surface ocean from the same time point and station (1.9% for M.Fall.O.10 m). We also found a large fraction of sequences that matched *Podoviridae* in the DCM sample in the open ocean in Monterey Bay (4.2% for M.Fall.O.105 m) and in the surface samples from the Great Barrier Reef (3.8% for GF.Spr.C.9 m and 5.0% for GD.Spr.C.8 m) as compared to the 0.8%±0.5% on average in other samples. Thus, *Podoviridae* may play an important role in reef ecosystems and the DCM not presently unknown.

Introduction

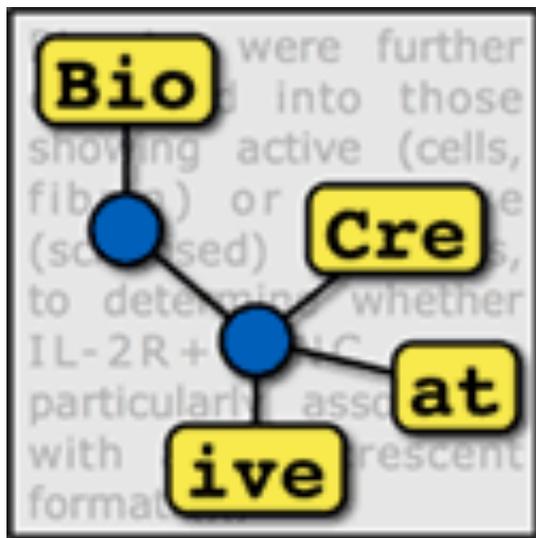
Finally, we compared and contrasted known viruses at the genus and species level in the combined photic and aphotic samples (Figure S2A and B respectively). At the genus level, we found a higher fraction of T4- and T7-like viruses in the photic zone (6.9% total) than the aphotic zone (2.6% total). At the species level, we found a higher fraction of *synechococcus* and *prochlorococcus* phages in the photic zone (4.6% total) than the aphotic zone (1.1% total).

The Protein Cluster as a Means to Organize Unknown Sequence Space

While this great "unknown" problem is exacerbated in viral metagenomes, it has also plagued microbial metagenomic studies to the extent that previous analyses of the GOS dataset organized this sequence space, including unknowns, using protein clustering (*sensu* Yooseph et al., 2007 and 2008 [19], [44]; details in Materials and Methods).

The sidebar on the right contains the following tags:

- EXTRACT x
- Disease
- Environment
- Organism
- Tissue



BioCreative V: Interactive Annotation Task (IAT) Dr. L. Hirschman, Dr. C. Arighi *et al.*
Challenge: March – August 2015
Presentation: September 2015, Sevilla, Spain
Metagenomics Record Annotation Session
 (Department of Energy [DE-SC0010838])

<http://www.biocreative2015.org>

Pafilis E, Buttigieg PL, Ferrell B, *et al.* (2015). Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 384–395.

http://www.biocreative.org/media/store/files/2015/IAT_extract_1.pdf

EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation Pafilis E, Buttigieg PL, Ferrell B, *et al.*. (2016). **Bioinformatics**, 2016, baw005. doi:10.1093/bioinformatics/btv04

- **Easy**: installation, tagging a web page, invoke the popup by processing selected text, saving results to a file
- **Adequate** NER accuracy (4/10 reported FN/FP, still two of them satisfied)
- **Speedup** in the range of 15–25%.
- Saving by time by avoiding looking up the ENVO identifier for every term
- **Manual** document inspection still needed
- EXTRACT helps **in finding terms** that would have been missed by the curators (e.g. due to non-familiarity with terminology)
- Average score 8.3 out of 10: they would **recommend EXTRACT**

EXTRACT

Interactive Extraction of Metadata

extract@hcmr.gr

About

Dem

Help

Please find below:

- Practical tips on how to use EXTRACT
- Curation assistance points
- Technical points and troubleshooting cases
- Using EXTRACT within other resources

Points in blue are a good starting point as they provide you with basic information about the EXTRACT bookmarklet, such as how to install and use the bookmarklet, and the EXTRACT popup description. Some points about record annotation with standardized metadata are listed afterwards (in green), followed by troubleshooting cases (in orange). Information on how to use EXTRACT within other resources can be found at the end (in purple).

[Show All](#) / [Hide All](#)

▶ How do I install the EXTRACT bookmarklet?

▶ How do I use EXTRACT?

▶ How can I use the EXTRACT popup for curation?

▶ Which types of entities can EXTRACT identify?

▶ How can I enlarge the EXTRACT summary popup?

▶ How can I use EXTRACT on documents that are not web pages?

▶ Why should I annotate samples with standards-compliant metadata?

▶ How should I annotate an outdoor sample with environment metadata?

▶ How can I annotate a host-associated/disease-related sample?

▶ Can EXTRACT suggest sections to study in a full-text article?

<https://extract.hcmr.gr>

EXTRACT

Interactive Extraction of Metadata

extract@hcmr.gr

About

Demo

Help

► How can I add the EXTRACT popup in my own web pages?

▼ Can I invoke the EXTRACT tagger programmatically?

In addition to the high-level `ExtractPopup` web method used in the previous section, EXTRACT offers a robust and fine-grained Application Programming Interface (API) to its named entity recognition engine. The core methods of this REST API are presented below:

GetEntities

`GetEntities` (<http://tagger.jensenlab.org/GetEntities>) returns the unique list of the entities identified in the document. The entities belong to the specified `entity_types` and the response follows the specified `format`.

Request:

```
http://tagger.jensenlab.org/GetEntities?
document=Both+samples+were+dominated+by+Zetaproteobacteria+Fe+oxidizers.+This+gro
up+was+most+abundant+at+Volcano+1,+where+sediments+were+richer+in+Fe+and+containe
d+more+crystalline+forms+of+Fe+oxides.&entity_types=-2+-25+-26+-27&format=tsv
```

Response:

```
Zetaproteobacteria    -2      580370
sediments             -27     ENVO:00002007
Volcano -27          ENVO:00000247
```

Parameter	Type	Content
document	required	the plain or html-formatted text to be tagged

<https://extract.hcmr.gr>, screenshot cropped for illustration purposes



Thank You

