Standards Promoting and Uptake Activities at the Hellenic Centre for Marine Research

Evangelos Pafilis

Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC)

Hellenic Centre for Marine Research (HCMR), Heraklion Crete, Greece

pafilis@hcmr.gr, http://epafilis.info

DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–10
doi: 10.1093/database/bav126
Original article

Original article

Value, but high costs in post-deposition data curation

Petra ten Hoopen[1,*], Clara Amid[1], Pier Luigi Buttigieg[2], Evangelos Pafilis[3], Panos Bravakos[3], Ana M. Cerdeño-Tárraga[1], Richard Gibson[1], Tim Kahlke[4], Aglaia Legaki[3], Kada Narayana Murthy[5], Gabriella Papastefanou[3], Emiliano Pereira[6], Marc Rossello[1], Ana Luisa Toribio[1] and Guy Cochrane[1]

Environmental context metadata annotation

- Interactive
- Lightweight
- Term look up assistant
- Standards-compliant term suggestions
- Metagenomics records

EXTRACT

http://www.environmentontology.org/

Buttigieg PL, *et al*. 2013, J Biomed Semant.4:43.

Bioinformatics, 2015, 1–3
doi: 10.1093/bioinformatics/btv045
Advance Access Publication Date: 24 January 2015
Applications Note

Data and text mining

ENVIRONMENTS and EOL: identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life

Evangelos Pafilis[1,*], Sune P. Frankild[2], Julia Schnetzer[3,4], Lucia Fanini[1], Sarah Faulwetter[1], Christina Pavloudi[1], Katerina Vasileiadou[1], Patrick Leary[5], Jennifer Hammock[6], Katja Schulz[6], Cynthia Sims Parr[6,†], Christos Arvanitidis[1] and Lars Juhl Jensen[2,*]

http://environments.hcmr.gr,
http://environments.jensenlab.org

OPEN ACCESS Freely available online
PLOS ONE

The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text

Citation: Pafilis E, Frankild SP, Fanini L, Faulwetter S, Pavloudi C, et al. (2013) The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. PLoS ONE 8(6): e65390. doi:10.1371/journal.pone.0065390

http://species.hcmr.gr, http://species.jensenlab.org

http://www.ncbi.nlm.nih.gov/Taxonomy

Benson DA, *et al*. 2009, NAR

IMBBC
Institute of Marine Biology
Biotechnology and Aquaculture

http://extract.hcmr.gr

hcmr
ΕΛΚΕΘΕ

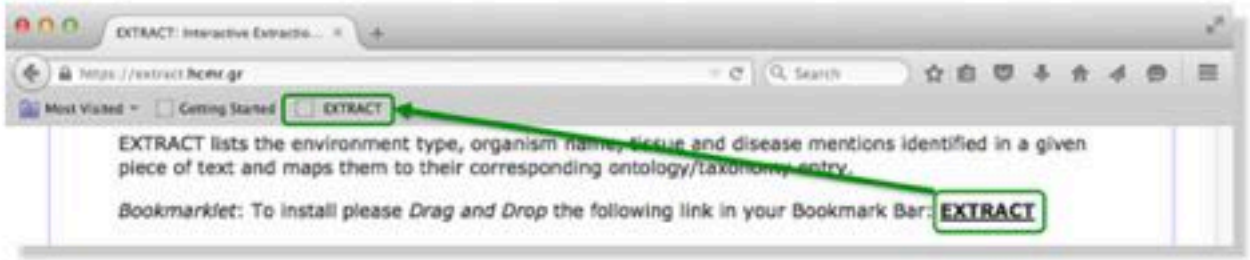
EXTRACT
Interactive Extraction of Metadata

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

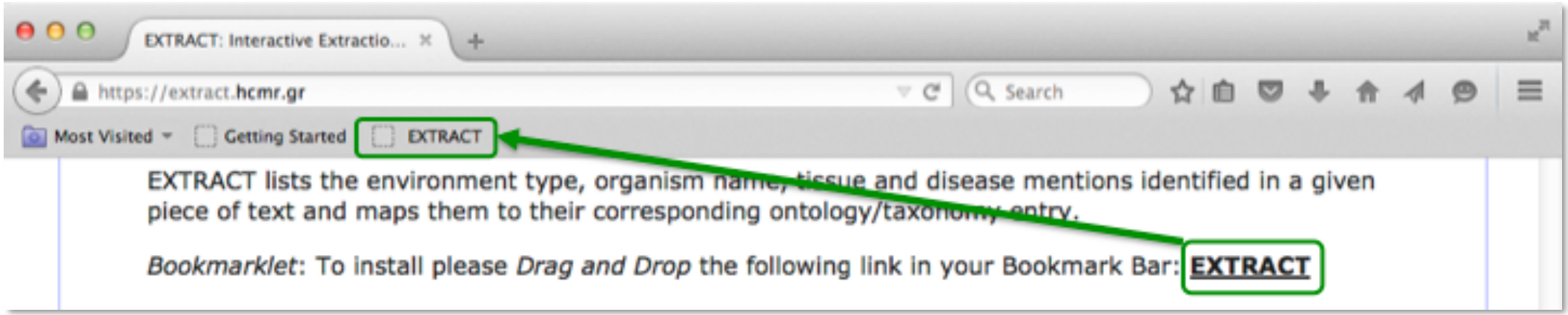


Usage: a. *select* a piece of text of interest in a web page and then b. *click on the bookmarklet*. c. A *pop-up* such as the following will appear (supported browsers: Chrome, Firefox, Safari). By *hovering* the mouse cursor over the text tags or the table rows you can visually inspect which words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

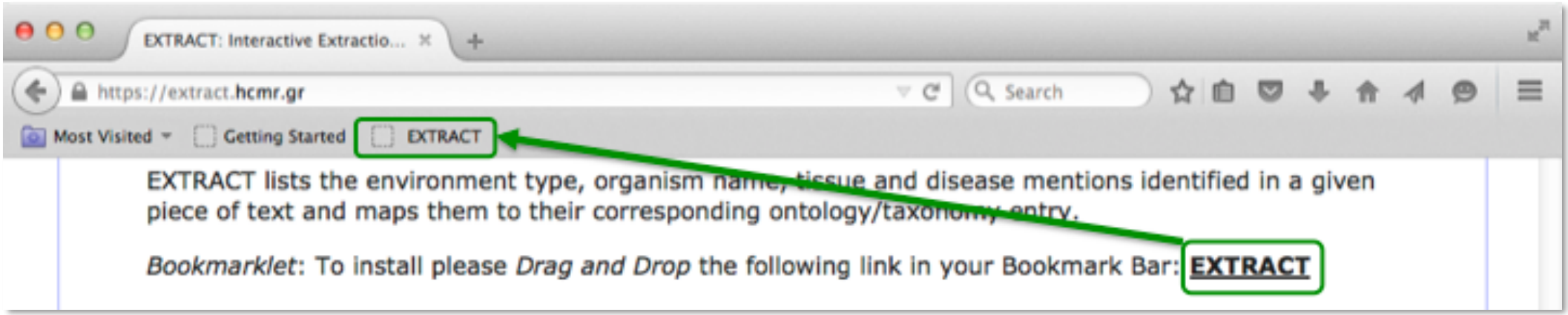
EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation Pafilis E, Buttigieg PL, Ferrell B, et al.. (2016). **Bioinformatics, 2016**, baw005. doi:10.1093/bioinformatics/btv04

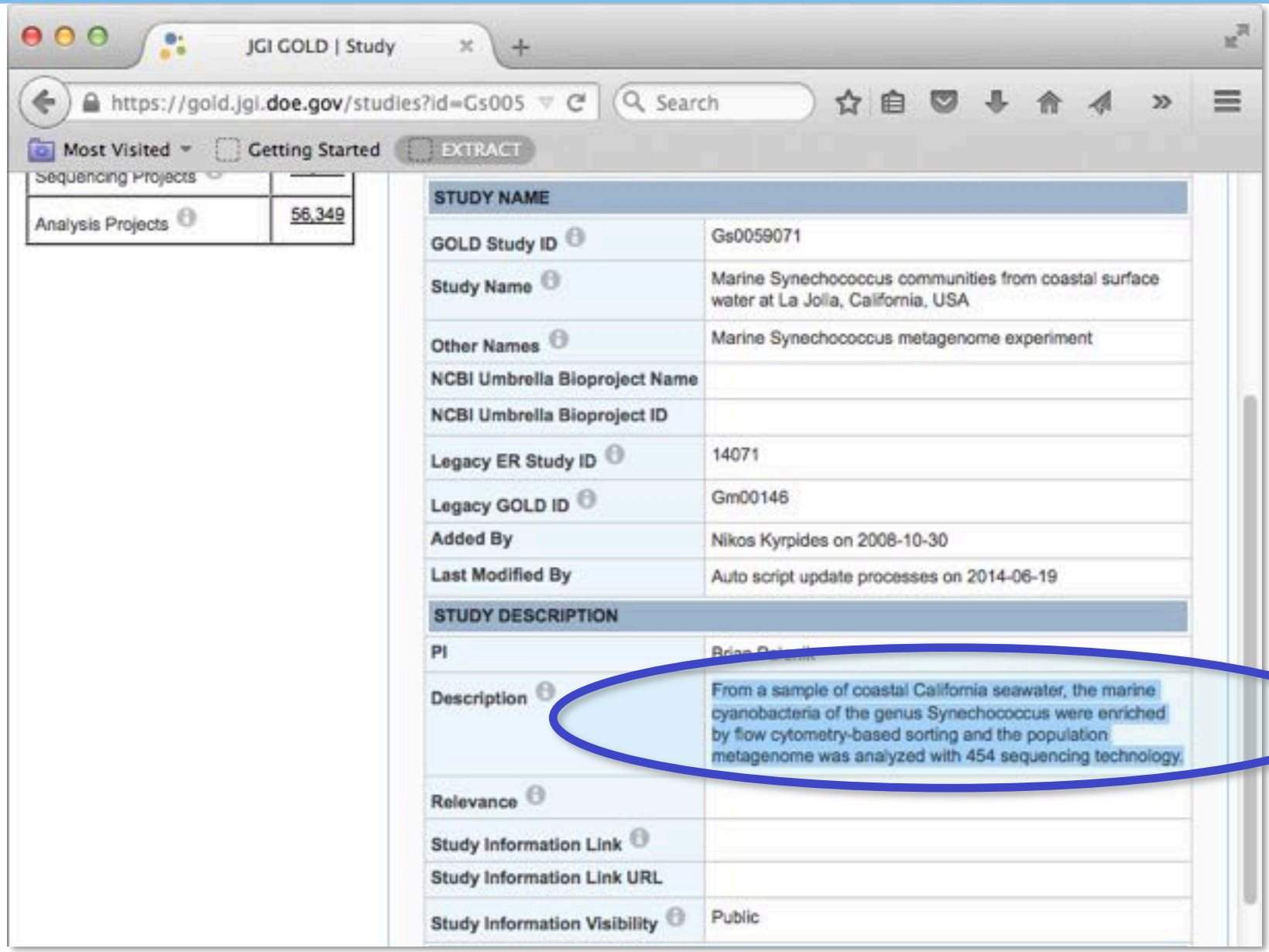
LIFE WATCH

Drag n' Drop Installation



Drag n' Drop Installation



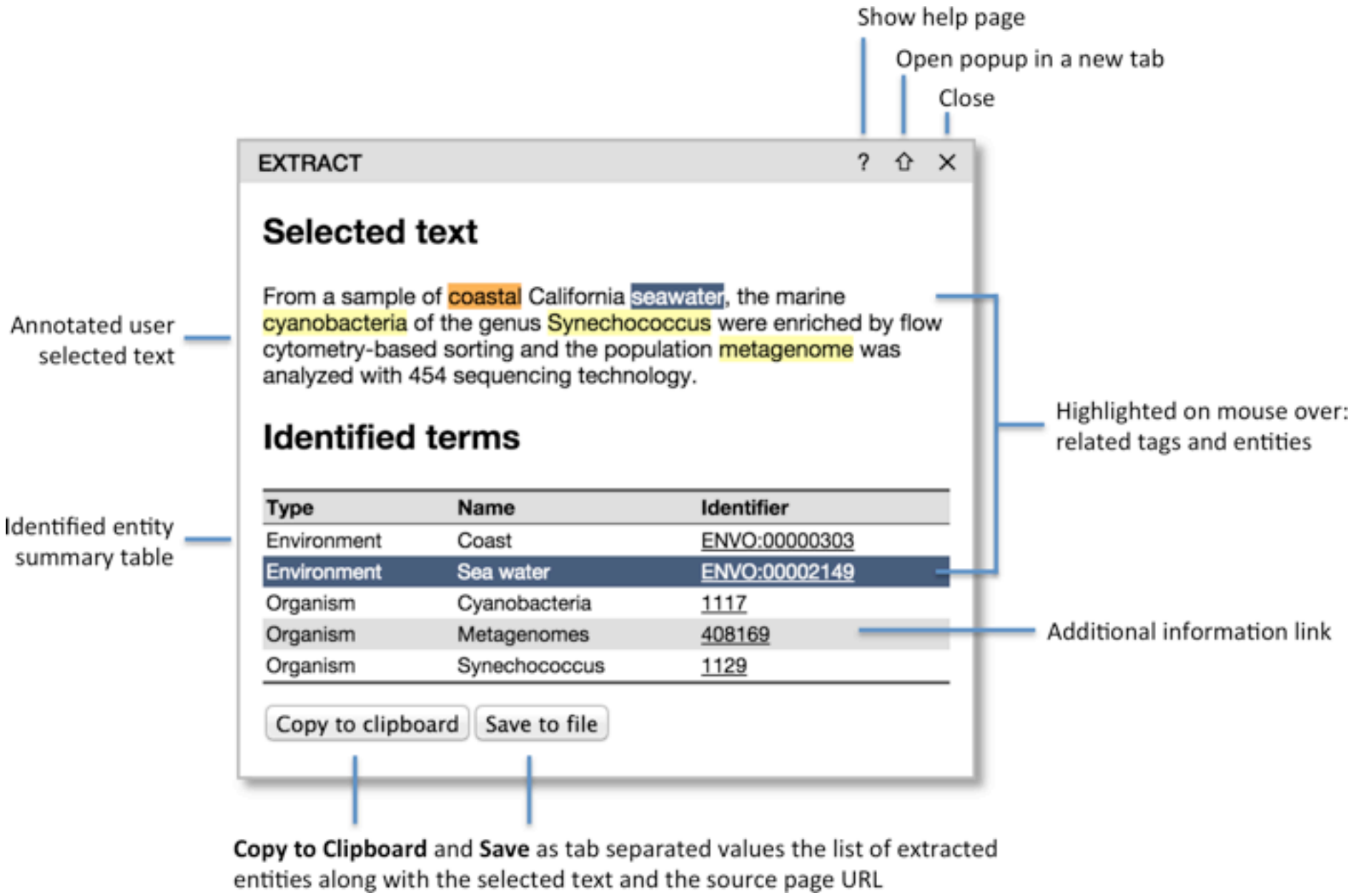


https://gold.jgi.doe.gov/studies?id=Gs0059071

IMBBC
Institute of Marine Biology
Biotechnology and Aquaculture

Popup – Multiple Entities

hcmr
ΕΛΚΕΘΕ

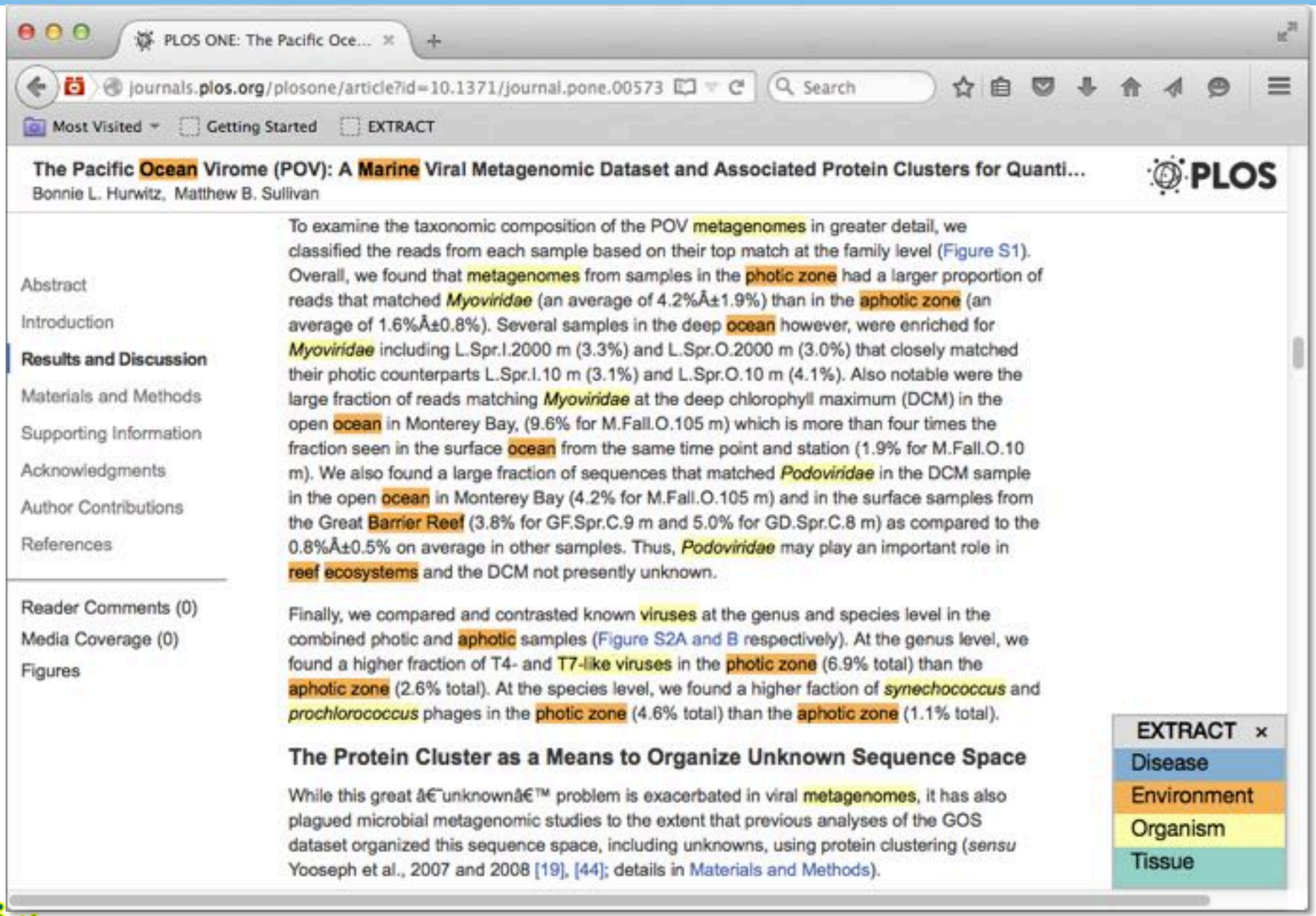


LIFE WATCH

IMBBC
Institute of Marine Biology
Biotechnology and Aquaculture

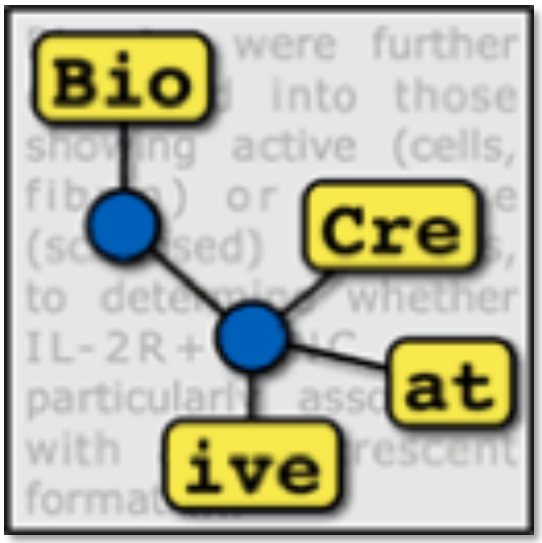
Full Page Tagging

hcmr
ΕΛΚΕΘΕ



Hurwitz & Sullivan, PONE, 2013, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0057355#s2

LIFE WATCH



http://www.biocreative2015.org

BioCreative V: Interactive Annotation Task (IAT) Dr. L. Hirschman, Dr. C. Arighi *et al.*

Challenge: **March – August 2015**

Presentation**: September 2015**, Sevilla, Spain

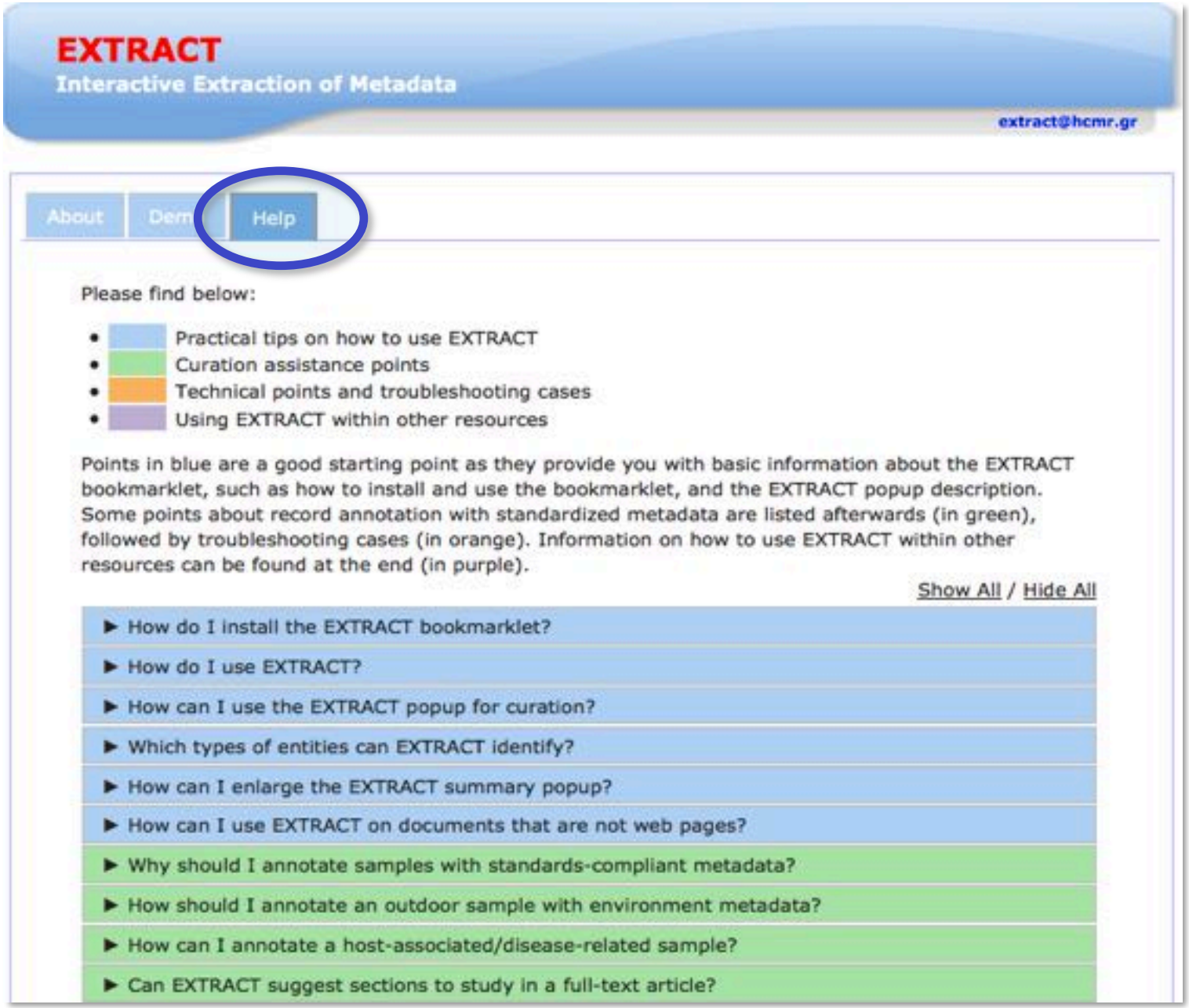
Metagenomics Record Annotation Session

(Department of Energy [DE-SC0010838])

Metagenomics Expert Evaluation

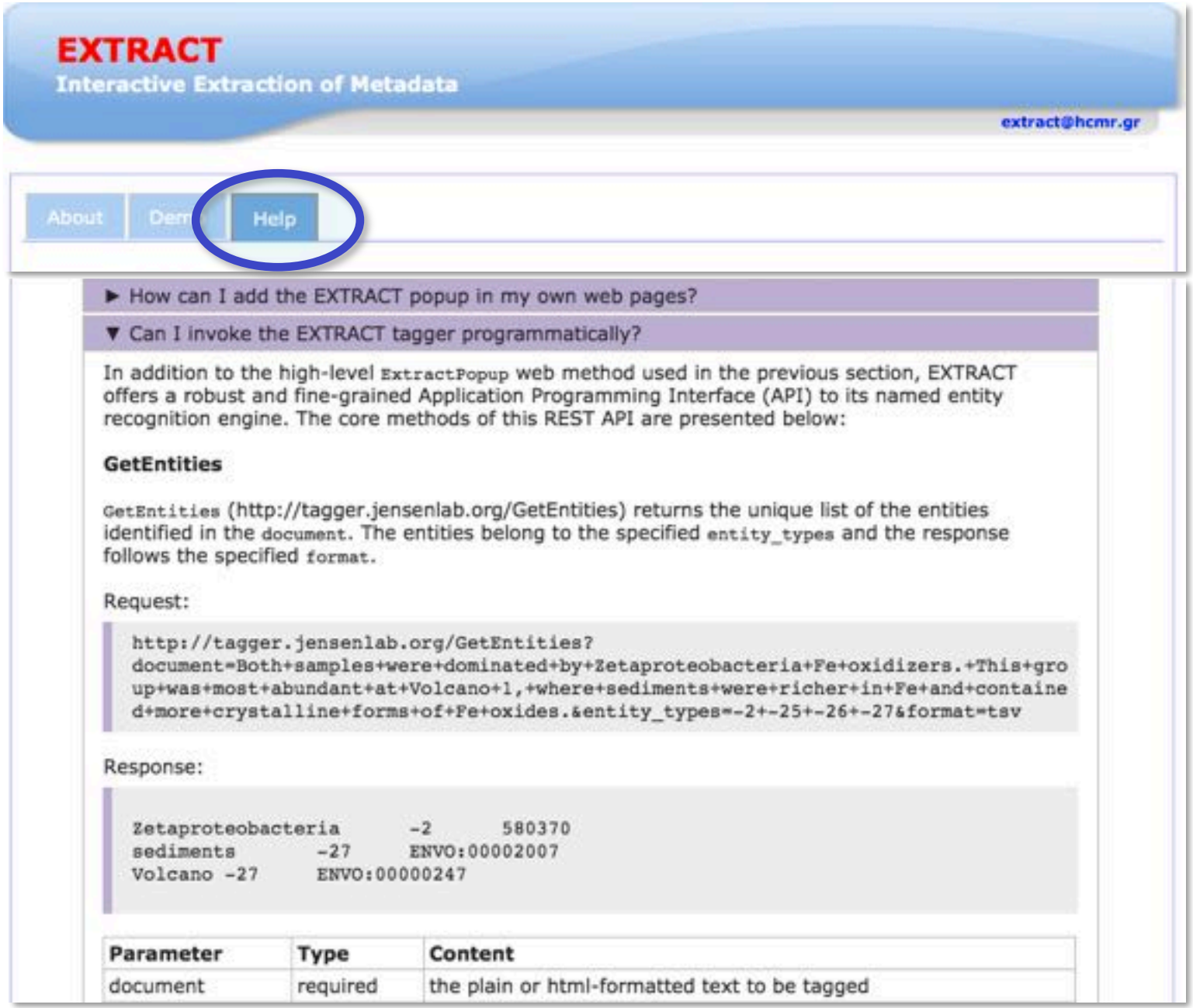
- Easy: installation, tagging a web page, invoke the popup by processing selected text, saving results to a file
- Adequate NER accuracy (4/10 reported FN/FP, still two of them satisfied)
- Speedup in the range of 15–25%.
- Saving by time by avoiding looking up the ENVO identifier for every term
- Manual 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

Documentation



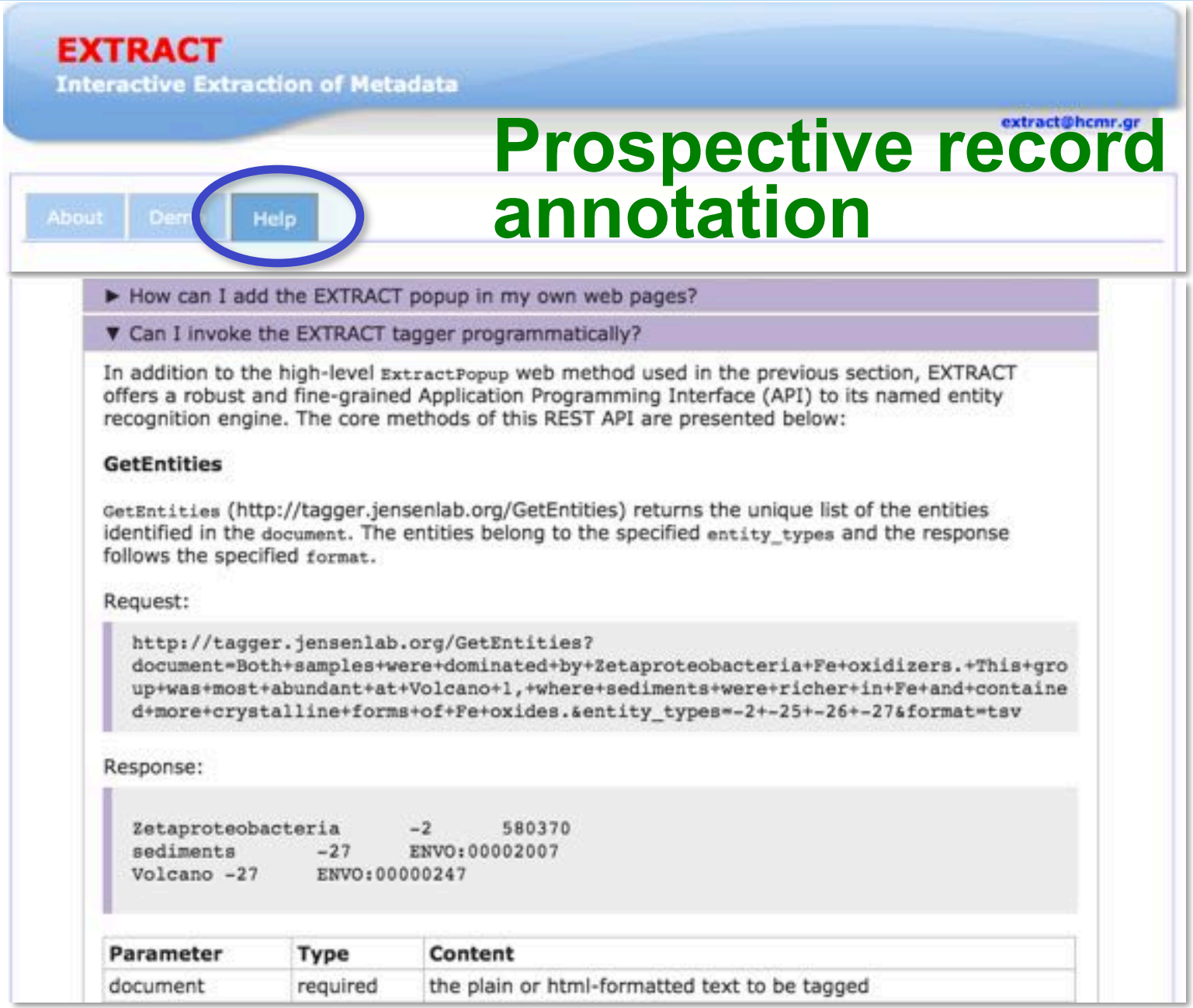
https://extract.hcmr.gr

API Documentation



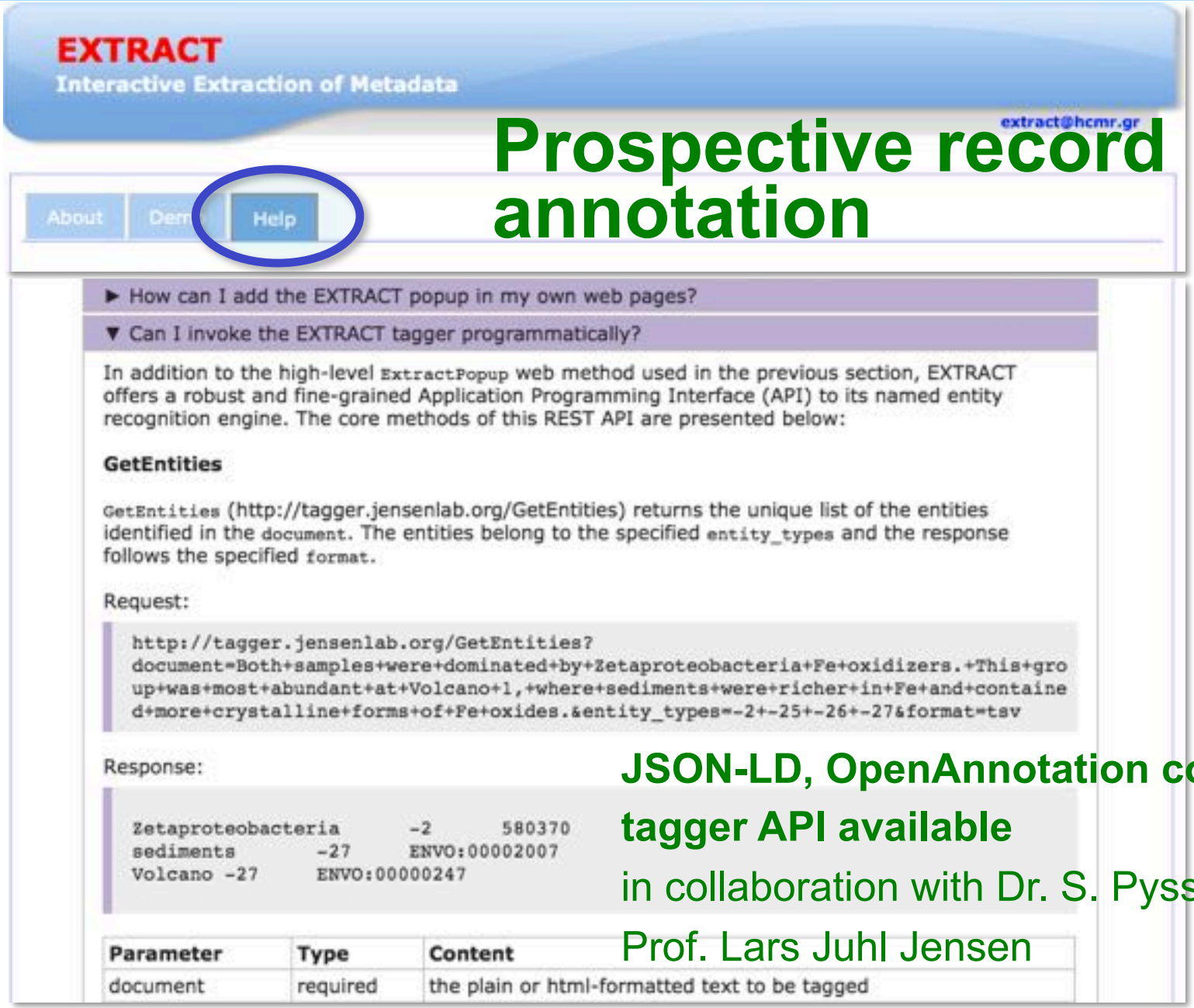
https://extract.hcmr.gr, screenshot cropped for illustration purposes

API Documentation



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

API Documentation



Prospective record annotation

JSON-LD, OpenAnnotation compliant tagger API available

in collaboration with Dr. S. Pyssalo and Prof. Lars Juhl Jensen

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

SPECIES v2.0, EXTRACT v2.0

New version of SPECIES (rule based extension for improved bacteria strain name matching)

- Top place in BioNLP – Shared Task 2016 BB3 (April 2016)

  Bacteria Biotope 3 - Event extraction of microorganisms and habitats with ontologies and their linking, in collaboration with Ms. Helen V. Cook and Prof. Lars Juhl Jensen (BB-cat-ner subtask)

  https://sites.google.com/site/bionlpst2016/tasks/bb2

EXTRACT v2.0: More entity types

- Tissues
- Chemicals
- Biological Processes
- Molecular Functions
- Cellular Components
- talk coming up in ICBO/BioCreative 2016 by Dr. Lars Juhl Jensen

DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–10
doi: 10.1093/database/bav126
Original article

Original article

Value, but high costs in post-deposition data curation

Petra ten Hoopen[1,*], Clara Amid[1], Pier Luigi Buttigieg[2], Evangelos Pafilis[3], Panos Bravakos[3], Ana M. Cerdeño-Tárraga[1], Richard Gibson[1], Tim Kahlke[1], Aglaia Legaki[3], Kada Narayana Murthy[5], Gabriella Papastefanou[3], Emiliano Pereira[6], Marc Rossello[1], Ana Luisa Toribio[1] and Guy Cochrane[1]

LifeWatchGreece Research (e-)Infrastructure

https://www.lifewatchgreece.eu/ https://portal.lifewatchgreece.eu/

Slide by Dr. Christos Arvanitidis

LIFE WATCH GREECE

LifeWatchGreece – **Data Management Team**

CLICK HERE

https://www.lifewatchgreece.eu/?q=content/collaborators -

✓Coordination of a team of **20 scientists;**

✓Representatives of **different disciplines, as well as different Institutes & Universities all over Greece;**

✓Acting as **data collectors, data providers and data managers**

Slide by Dr. Dimitra Mavraki

LIFE WATCH GREECE

LifeWatchGreece – **Data Management**

data management tasks can be summarized as:

- ✓**Collecting & Receiving data** from national, regional and international programmes
- ✓Collecting & **Digitizing Historical biogeographic data**;
- ✓**Verifying the quality** of the data (using agreed upon standards);
- ✓**Reporting the results** of quality control directly to data providers as part of the quality assurance process;
- ✓**Making data available**, nationally and internationally through MedOBIS and GBIFgreece ipt and
- ✓Ensuring the **long term preservation** of the data and associated information required for correct interpretation of the data

Slide by Dr. Dimitra Mavraki

LIFE WATCH GREECE

LifeWatchGreece – c. Managing Data

Taxonomic Quality Control

Nemalion helmintoides	Nemalion helminthoides
Didemnids	Didemnidae
Littorina saxatillis	Littorina saxatilis
Gastropods	Gastropoda
Prorocentrum triestnum	Prorocentrum triestinum
Indetermined	Biota or Higher Classification if given
Corallina sp Corallina spp	Corallina
Jania rubens/Haliptilon virgatum	???

- NEMERTINEA ------- NEMERTEA
- Polydontes frons ----- Polyodontes frons
- Amaena tribolata ----- Amaena trilobata
- Asyhis biceps ----- Asychis biceps
- Mangelona papilicornis ----- Magelona papillicornis
- Ochnesoma steenstrupi ----- Onchnesoma steenstrupi
- Aplakophora ----- Aplacophora
- GASTEROPODA ---- GASTROPODA
- Strobiformis subulata ---- Strombiformis subulatus

- ✓ typical taxonomic errors are checked (misspelling, invalid names, inconsistencies, misidentifications)
- ✓ taxon names are linked to one of the following bases: World Register of Marine Species (WoRMS), Catalogue of Life (CoL), IT IS, IRMNG, Paleobiology Data, Internation Plant Name Index (IPNI) & Euro+Med Plantbase, Index Fungorum, Fishbase, Barcode of Life Database (BOLD) and Web of Science.
- ✓ If no link possible relevant papers and data provider(s) are consulted
- ✓ originally delivered name is always safeguarded!

Slide by Dr. Dimitra Mavraki

LIFE WATCH GREECE

LifeWatchGreece – c. Managing Data

Spatial Quality Control

Long/Lat expression	
Decimal Degrees	lat: 35.241078° long: 25.153364°
Degrees, Minutes, Seconds	35°14'27.88" N 25° 9'12.11" E
Degrees, Decimal Minutes	35° 14.465' N 25° 7.202' E
UTM	Zone 35N

CONVERSION

Degrees, Minutes, Seconds → **Decimal Degrees**

Degrees + ((minutes/60)+ (seconds/3600)) = dd

35°14'27.88" N 25° 9'12.11 " E

35+(14/60)+(27.88/3600)= 35.241078

25+(9/60)+(12.11/3600)= 25.153364

Coordinates are always:

- ✓ converted in WGS84 geodetic reference coordinate system (decimal degrees)
- ✓ plotted on map and visual checks are performed

When not available:

- ✓are derived from provided map, http://www.geodata.gov.gr , related paper etc
- ✓Marine Regions webpage is checked to assign coordinates
- ✓Data provider is contacted

Slide by Dr. Dimitra Mavraki

LIFE WATCH GREECE

LifeWatchGreece – c. Managing Data

Darwin Core:

http://rs.tdwg.org/dwc/terms/index.htm#theterms

✓ **a body of standards** intended to facilitate the sharing of information about biological diversity

✓ **a glossary of terms** meant to provide stable semantic definitions with the goal of being maximally reusable

Region	Country	Institution or organisation	Data owner	Sampling area	Name of sampling station	Geographic coordinates of the station (WGS 84 - decimal format)
DwC term: **waterBody**... *The name of the water body in which the location occurs*	DwC term: **country**...*The name of the country or major administrative unit in which the location occurs*	DwC term:**institutionCode**. ..*The name (or acronym) in use by the institution having custody of the object(s) or*	*DwC term: rightsHolder... a person or organization owning or managing rights over the*	DwC term: **locality**... *The specific description of the place*	DwC term: **locationRemarks** ....*Comments or notes about the Location.*	DwC term: **verbatimCoordinates** ...*The verbatim original spatial coordinates of the Location*

Weather type (sunny, cloudy, rain, storm)	Sample track code (country-station-level-plot-replicate)	Species (number of individuals)/Stations	Number per sample-unit	WORMS AphiaID
DwC term: *fieldNotes* ...*the text of notes taken in the field about the Event*	DwC term: **fieldNumber**...*an identifier given to the event in the field*	DwC term: **taxonRemarks** ...*the original name received*	DwC term: **individualCount**... *The number of individuals represented present at the time of the Occurrence*	DwC term: **taxonID** ...*An identifier for the set of taxon information (data associated with the Taxon class)*

Slide by Dr. Dimitra Mavraki

IMBBC
Institute of Marine Biology
Biotechnology and Aquaculture

hcmr
ΕΛΚΕΘΕ

med OBIS MEDOBIS DATA REPOSITORY (IPT)
free and open access to marine biodiversity data

Logged in as dmavraki@hcmr.gr Account Logout ENGLISH

Home | Manage Resources | About

Resource Title CINCS: Pelagic-Benthic Coupling In the oligotrophic Cretan Sea

Basic Metadata

The resource title and description are required. The resource's three main contact's must also be described here: Resource contact, resource creator and metadata provider. For each contact you must supply at least a last name, a position or an organisation before you can make the resource public. The person(s) or organisation(s) responsible for the creation of the resource as it appears in the IPT and for effectively publishing the resource should add themselves as an associated party with role 'publisher'.

Title*

CINCS: Pelagic-Benthic Coupling In the oligotrophic Cretan Sea

Description*

The main goal of CINCS was to obtain a better understanding of the structure and function of the oligotrophic Cretan Sea. The project involved 7 laboratories and approximately 30 scientists. Publication divided on 7 chapters and subsection. The study is focused on macrofaunal community structure of the continental shelf, slope and deep sea basin of the Cretan Sea. Data contain 454

Metadata Language: English

Resource Language: English

Type: Occurrence

Subtype: Observation

Section

- Basic Metadata
- Geographic Coverage
- Taxonomic Coverage
- Temporal Coverage
- Keywords
- Associated Parties
- Project Data
- Sampling Methods
- Citations
- Collection Data
- External links
- Additional Metadata

Slide by Dr. Dimitra Mavraki

DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–10
doi: 10.1093/database/bav126
Original article

Original article

Value, but high costs in post-deposition data curation

Petra ten Hoopen[1,*], Clara Amid[1], Pier Luigi Buttigieg[2], Evangelos Pafilis[3], Panos Bravakos[3], Ana M. Cerdeño-Tárraga[1], Richard Gibson[1], Tim Kahlke[4], Aglaia Legaki[3], Kada Narayana Murthy[5], Gabriella Papastefanou[3], Emiliano Pereira[6], Marc Rossello[1], Ana Luisa Toribio[1] and Guy Cochrane[1]

OSD
Ocean Sampling Day

MicroB3 Summer School

Philia RV, HCMR

Niskin bottle

www.osil.co.uk

CTD

http://www.seabird.com/sbe25plus-sealogger-ctd

Heraklion, May 2014, MicroB3 Summer School

OSD
Ocean Sampling Day

PHILIA
ΕΛΚΕΘΕ

Ocean Sampling Day

OSD Main Event 2014 Pipeline

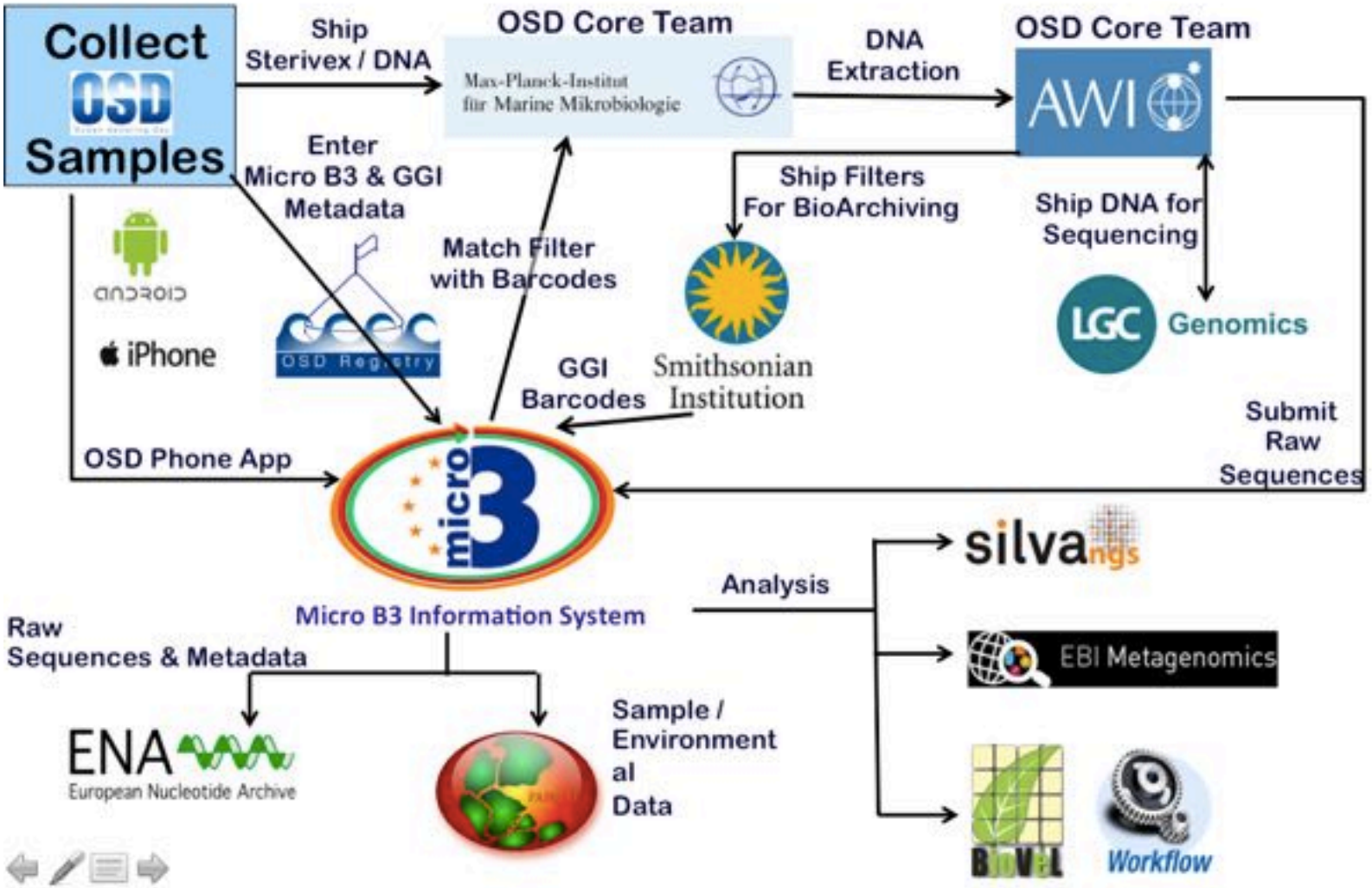


image: http://oceansamplingday.blogspot.com

Annex I
OSD Logsheets

BEFORE

SAMPLING

SAMPLING_Site (ID, Name):	OSD-SUMMER SCHOOL
SAMPLING_Platform:	RV Philia
SAMPLING_Campaign:	OSD-SUMMER-SCHOOL
SAMPLING_Project :	MicroB3-OSD2014 - summer-school

SAMPLING_Investigators : GROUP 2	Last name	First name	email
	CHRISTAKIS	CHRISTOS	christakis@hcmr.gr

SAMPLING_Objective	summer school training

DURING

EVENT

		yyyy	mm	dd	hh	mm	ss	UTC
EVENT_DateTime	Start	2014	05	27	14	15	00	+2
	End	2014	05	27	16	30	00	+2

		+N/-S	dd.0000		+E/-W	ddd.0000
EVENT_Lat/Long	Start		35.012172		+	25.158333
	End		35.011903		+	25.119953

EVENT_Device :	NISKIN BOTTLE
EVENT_Method :	BOTTLES IN-LINE WITH MULTISENSOR
EVENT_Comment :	

AMBIENT

ENVIRONMENT

ENVIRONMENT_ Marine_Region	e.g. Adriatic Sea	HERAKLION BAY - SEA OF CRETE
ENVIRONMENT_ Biome	e.g. ENVO:00000447 for "marine biome"	
ENVIRONMENT_ Feature	e.g. ENVO:00000569 for "marine habitat"	
ENVIRONMENT_ Material	e.g. ENVO:00002042 for "surface water"	

Stephane Pesante
Renzo Kottman
Christos Christakis
MicroB3 Summer School Team

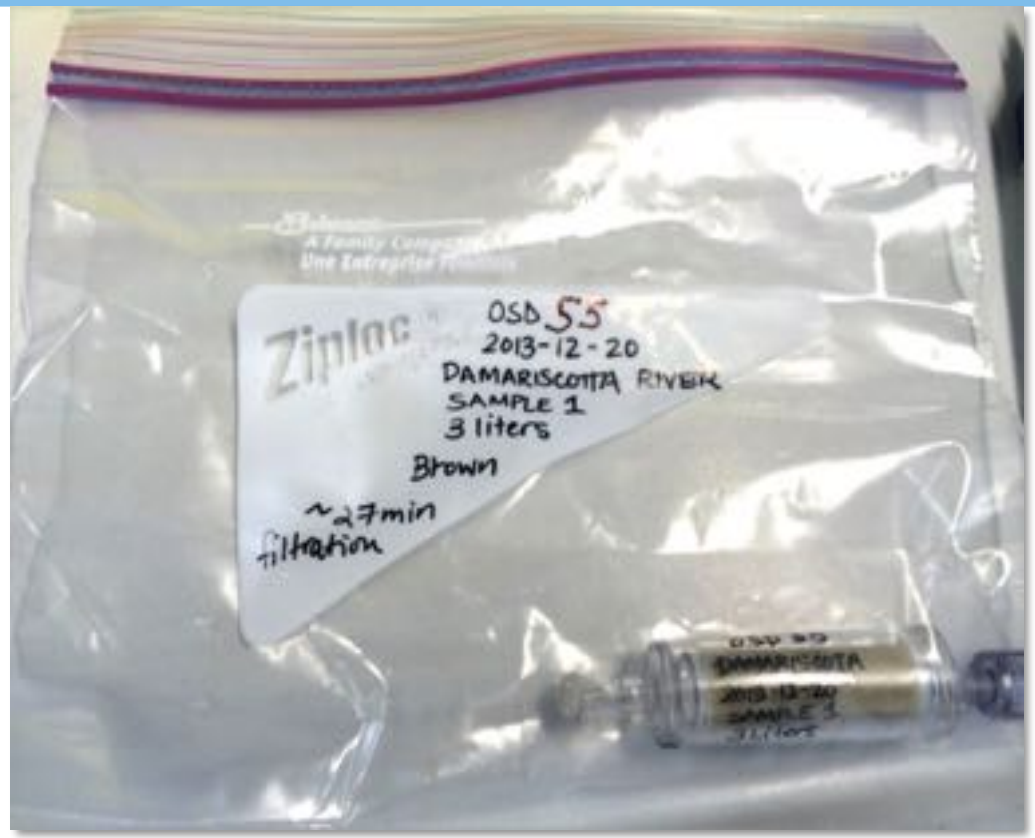
OSD: Sample Packaging and Shipping

- One sampling event **per bag**
- Use **freezable** bags
- Label **filter and bag**
- Seal the label
  - Use **adhesive** tape
  - not parafilm
- International shipping
  - Pack filters on **dry ice**
  - Dispatch samples on a **Monday**
  - Specify**: no commercial value content**, fill-in: **custom forms**

Photo and slide input by Dr. A. Klindworth

OSD: Sample Labels

Label filter and bag



<OSD_SiteID>_<Month>_<Year>_<SiteName>_<SampleNo>_<Depth>

e.g. OSD3_06_14_Helgoland_1_surface

e.g. OSD3_06_14_Helgoland_1_20m

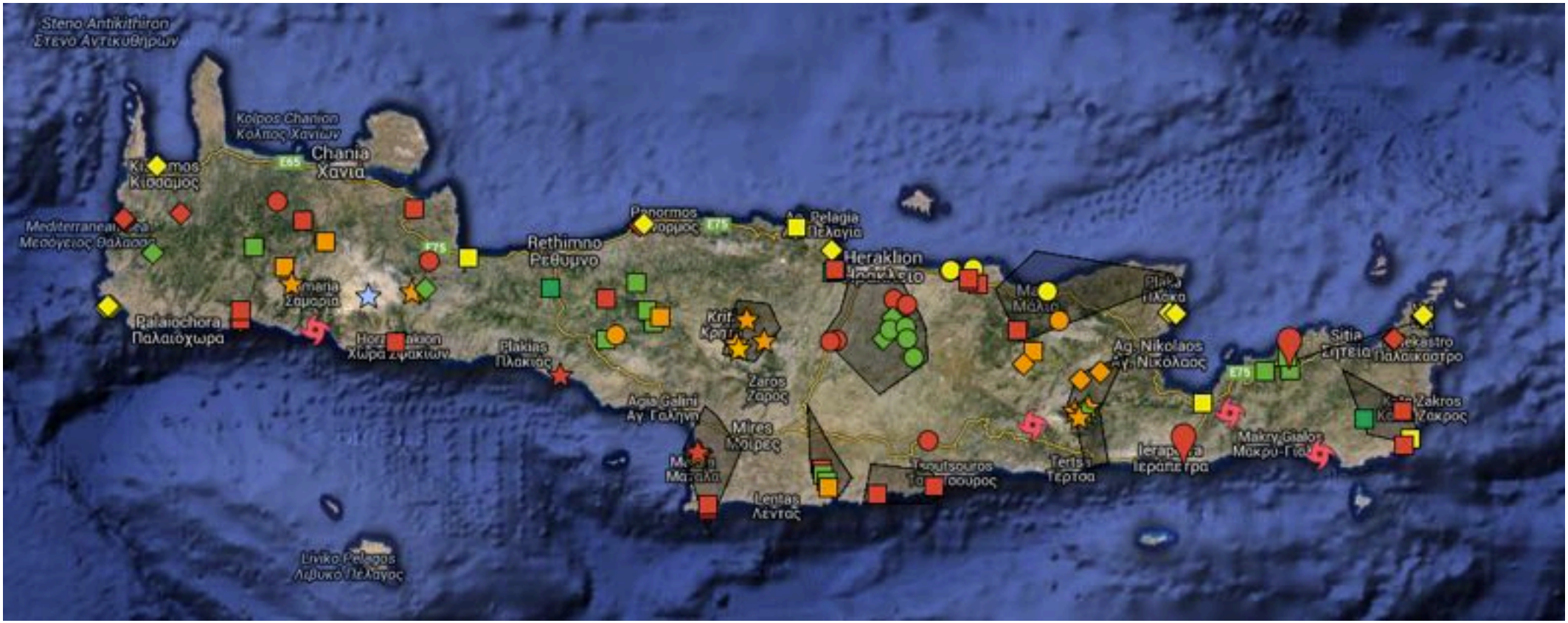
e.g. OSD5_06_14_Crete_1_2m

Photo and slide input by Dr. A. Klindworth

IMBBC
Institute of Marine Biology
Biotechnology and Aquaculture

ISD: Island Sampling Day

hcmr
ΕΛΚΕΘΕ



https://www.google.com/maps/d/edit?mid=1ZZ3mzc7yr0X2Sn3o5FYCUHuApyY

LIFE WATCH

Thank You!

EXTRACT is based on components from: ENVIRONMENTS, SPECIES,Reflect http://reflect.ws Developed by the LifeWatchGreece Research Infrastructure and the group of Dr. Lars Juhl Jensen (Uni Copehagen); with input from the groups of Genomes OnLine Database, Virome / Metagenomes Online, MegX.net and Dr. Pier L. Buttigieg (AWI). **BioCreative**: All evaluators, Dr. Cecilia Arighi and Dr. Lynette Hirschman (DoE Award No DE-SC0010838)

LifeWatchGreece/IMBBC: Christos Arvanitidis, Dimitra Mavraki, Stamatina Nikolopoulou, Sarah Faulwetter, Nikos Minadakis (ICS/FORTH) and many more (https://www.lifewatchgreece.eu/)

NNF CPR: Dr. Sune Frankild, Helen Cook, **MM-MPI:** Frank Oliver Glöckner, Dr. A. Klindworth, Renzo Kottman, E. Pereira, Dr. Julia Schnetzer *et al.* **Uni Delaware:** Dr. Barbra Ferrel, **EBI:** Dr. Guy Cochrane, Dr. Petra ten Hoopen *et al.* **Pangea**: Dr. Stephane Pesant, **HCMR/IMBBC**: Dr. P. Polymenakou, Dr. G. Kotoulas, Dr. Anastasios Oulas, Anastasia Tsiola, Christina Pavloudi, Gabriela Papastefanou, Aglaia Legaki, **UoA:** Panos Bravakos. Many more I might be omitting