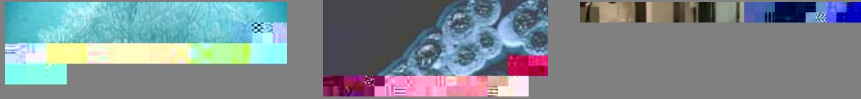
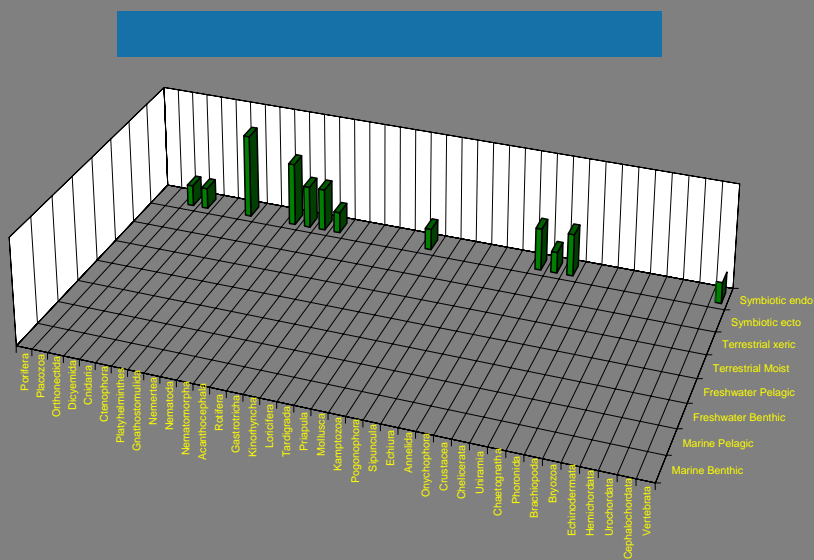


# Towards a practical knowledgebase for marine genetic resources

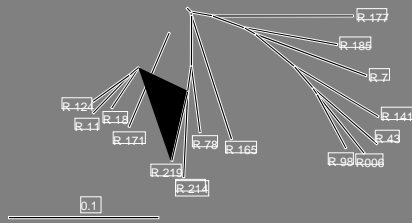


Libby Evans-Illidge

Manager Bioresources Library, Australian Institute of Marine Science,  
 PMB 3 Townsville MC, 4810, Queensland, Australia.  
[e.evansillidge@aims.gov.au](mailto:e.evansillidge@aims.gov.au)



## Superimposed with bacterial symbiont diversity



## Natural Products “Renaissance” (with a twist)

Paterson and Anderson, *Science* 21 October 2005

### A significant cumulative effort in ocean exploration

‘Parents’ of marine science:

- Indigenous observations over millennia
- Aristotle 384-322BC
- Charles Darwin  
HMS *Beagle* 1831-36
- Challenger 1872-75





## Where is the data? How can we access it?

- Specialist data
- Published Literature
- Portals for metadata
- Networked datasets
- Integrated datasets

### The published literature

#### Literature Databases

- Cambridge Scientific Abstracts
- Aquatic Sciences & Fisheries Abstracts
- Zoological Record
- Biosis
- MarineLit
- Patent databases

#### Citation & Indexing services

- Web of Science
- Google Scholar

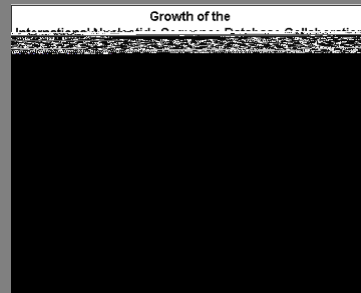
#### Citation data management

- Ref works
- Endnote



## Data Networks/Repositories Nucleotide sequences

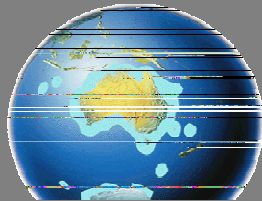
- **INSDC** (international Nucleotide Sequence Database Collaboration)  
[www.insdc.org](http://www.insdc.org)
  - **GenBank<sup>®</sup>, DDBJ, EMBL**
    - Annotated collection of all publicly available DNA sequences
    - Submission of sequences required by many journals prior to publication
    - Online submission, update and review
    - Country of origin identified
    - No restriction on use or distribution
    - 73078143 loci, 77248690945 bases, from 73078143 reported sequences (June 15 2007)
- **Data searching & analysis tools**
  - eg. BLAST



[www.ncbi.nlm.nih.gov/Genbank/](http://www.ncbi.nlm.nih.gov/Genbank/)

## Specialist data – find the people

Location of specimen and data holdings  
reflects historical location of the specialists

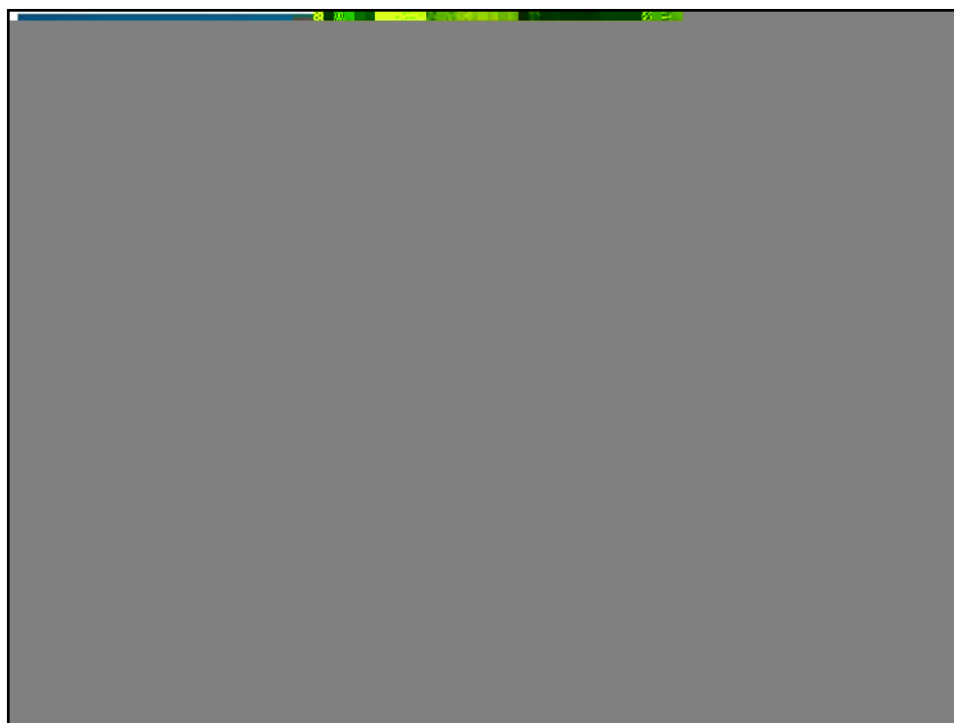


# UN Atlas of the Oceans

- UN-Oceans coordination portal
- 14 Global partners and 8000 individual members.

The slide features a header with the word "Atlas" in a purple box. Below it, five boxes list data sources: "Data from taxonomic names lists eg. APNI, AFD, ITIS, Species2000", "Specimen data from museum & herbarium & culture collections, vouchered specimens used for DNA analysis", "DNA sequence data held in DNA banks and barcodes from CBOL projects", "Character datasets: e.g., morphology, biochemistry, growth and", and "From image banks and repositories". The main content area is a large black rectangle.

- It's expensive (\$40M, over 5 years)
- Consistency in taxonomy is an issue



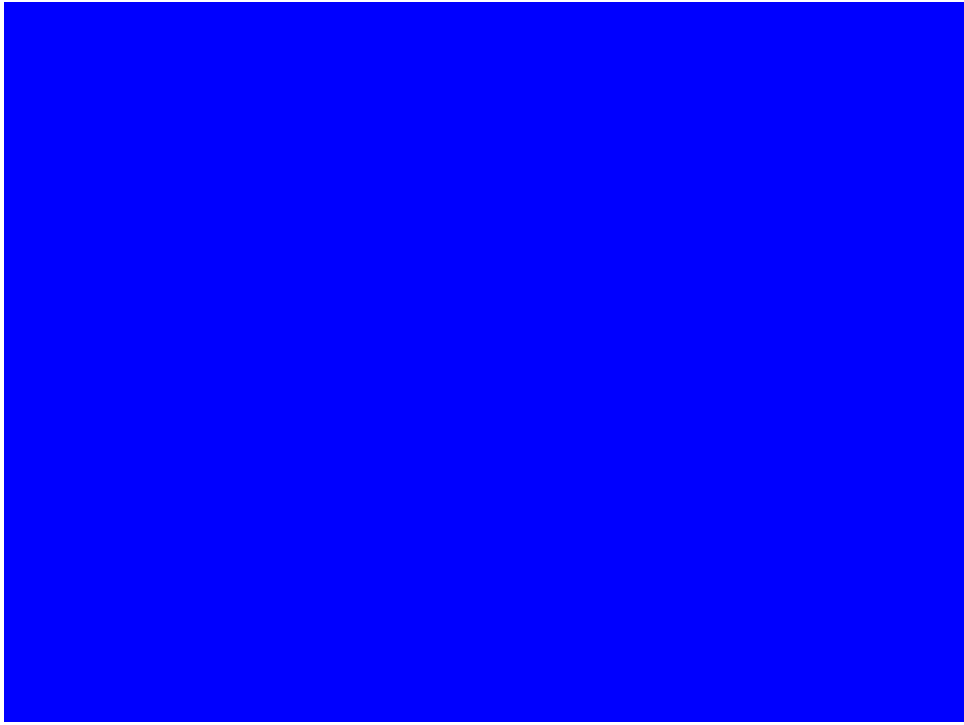


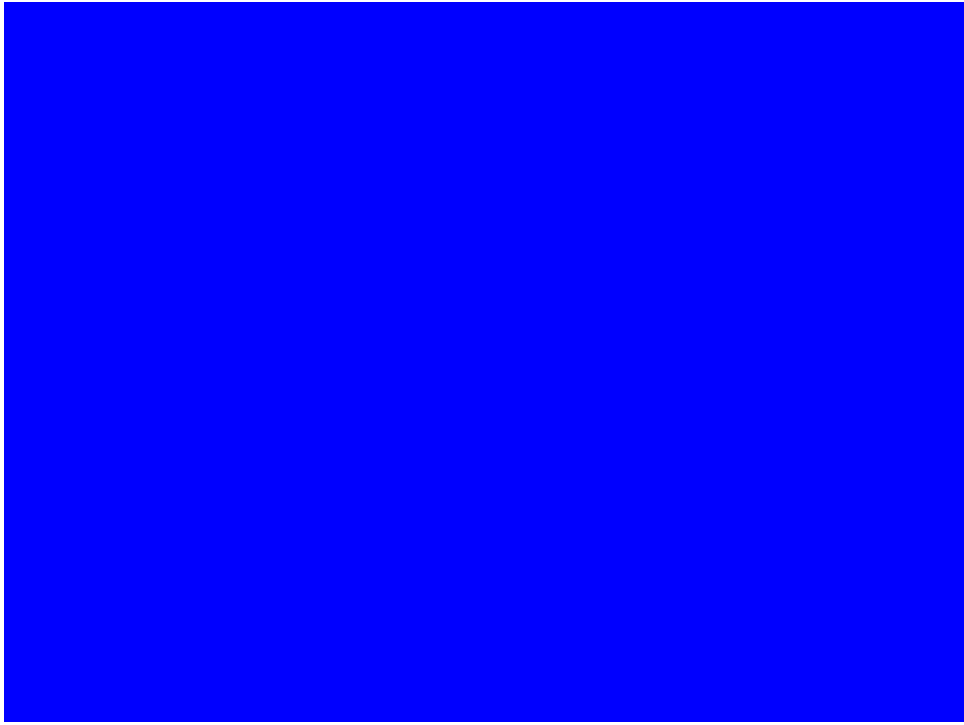
- Growing global network
- Assess and explain the diversity, distribution and abundance of marine life in the world's oceans
- past, present and future
- 2000 to 2010
- 50+ countries, 300+ scientists, 17 major projects
- Link to the Barcode of Life initiative

## Oceans Present: Realm Projects

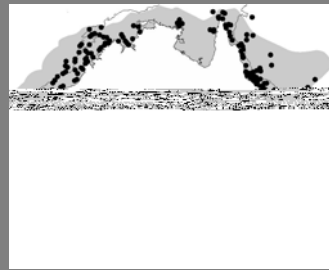
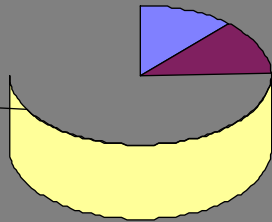
CoML defines its realms & zones in 2003 Baseline Report, *The Unknown Ocean*



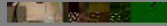




# AIMS Bioresources Library – integrated MGR dataset



## A Bioresources Library for screening and biodiscovery (past, present, future)



Cancer

Environmental  
remediation

Viral

Industrial Enzymes

Antibiotics

Agrichemical

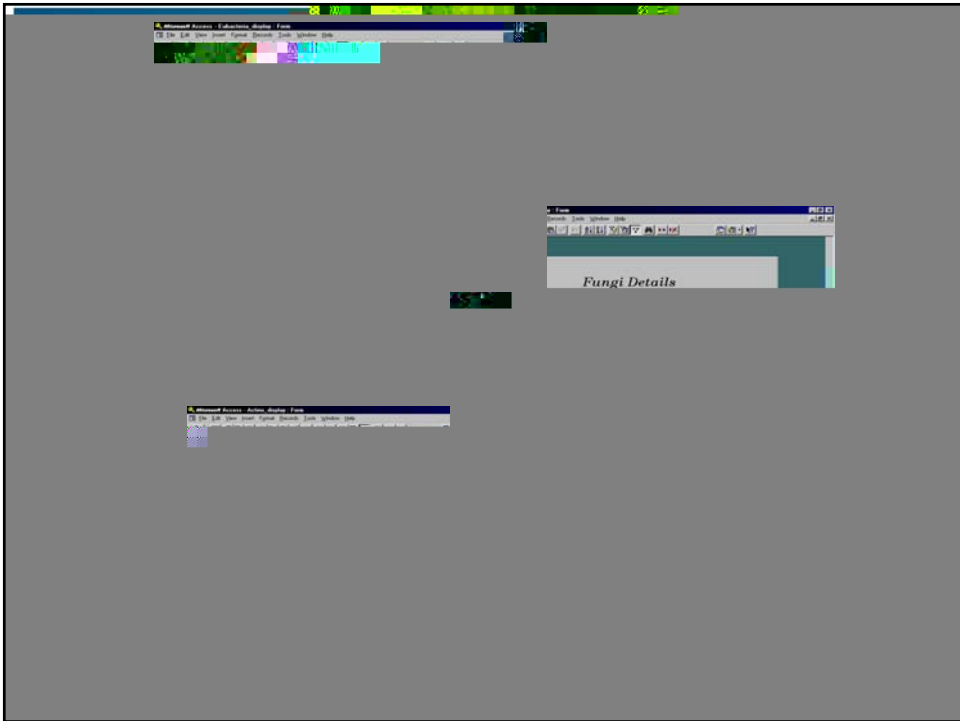
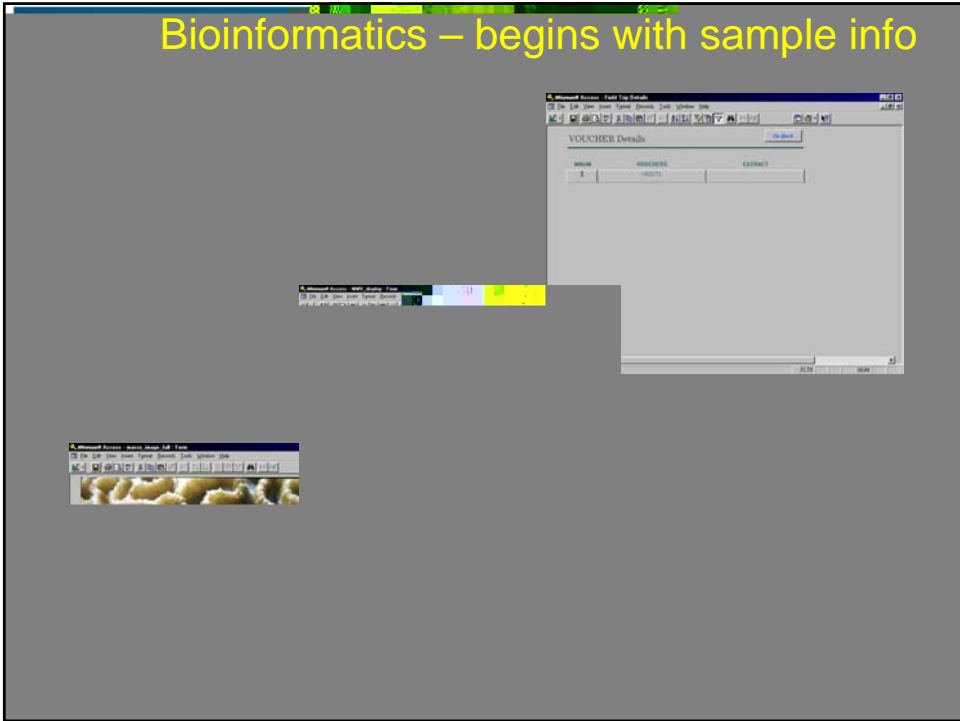
Toxin detection

Paints

Mineral Processing

Central Nervous System

# Bioinformatics – begins with sample info



The image shows a screenshot of a software interface with a dark grey background. At the top, there are two windows: the left one is titled 'Microsoft Access - chun\_mwsp\_101\_Folk' and contains a 'Peyton Spectrum' plot; the right one is titled 'Microsoft Access - chun...'. Below these, there is a window titled 'Microsoft Access - Chemistry' containing a ball-and-stick molecular model of a complex organic molecule. At the bottom right, there is another window titled 'Access - chun\_mwsp\_101\_Folk'. The text '...integrated with other research outputs' is written in yellow at the bottom of the screenshot.

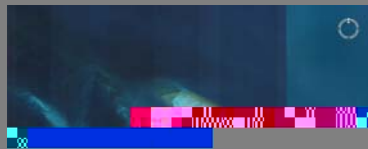
...integrated with other research outputs

## Integrated tool for data mining.

eg. Regional and taxonomic patterns in anti-microbial activity



## Analysis at a range of scales



Google Earth interface facilitates good visualisation of multi-variate data



## Understanding the chemical ecology Apply data-mining to enhance biodiscovery

- Elaborate leads
  - identifying other material with similar taxonomy/ecology/ screening profile/chemistry
  - Naturally occurring analogues
  - Re-supply without re-collecting
- Predict results of future screening based on past profile
  - Compile list of pre-leads
  - Targeted biodiscovery with ex-situ material

## Summary

- Tools exist to access marine biodiversity and genetic resources data that is in the global public domain
- Major networking projects underway to bring together independent geo-referenced datasets (CoML, ALA)
- Consistency in taxonomy is a big issue in networking independent datasets
- Integrated informatics is the ideal
  - Integrate biodiversity data (ecosystems, species (macro and micro), genomes (and meta-genomes)), with natural products research outputs (instrument outputs, structures of compounds, proteins, enzymes etc), and screening results

