The Genomic Standards Consortium (GSC)

Dawn Field\textsuperscript{1}/Peter Sterk\textsuperscript{1,2}
\textsuperscript{1}Oxford Centre for Ecology and Hydrology
\textsuperscript{2}Wellcome Trust Sanger Institute

\textit{On behalf of the GSC}

\textit{Thanks to so many in the GSC for the slides used in this presentation}
Genomic Standards Consortium (GSC)
Opportunities and Challenges

• Querying databases
• Populating databases
• Enhancing computational pipelines
• Hypothesis Testing and Generation
Recording and accelerating scientific advance through standards and databases

Scientific advance...

Databases

Journals

'Knowledge'

Databases

Journals

Distilled data

Software

Ontologies

File formats

(Meta)data

http://gensc.org
From data generator to public databases: the new field of “standards bioinformatics”

Chris Taylor, EBI

Diverse community-specific extensions

Generic Features (origin of biomaterial)

Generic Features (experimental design)
Often challenging to negotiate: technology rich

Chris Taylor, EBI
Reporting standards: definition and scope

• What is a standard?
  • An agreement between several players in a certain area

• Reporting standards
  • Set to enable unambiguous representation, description and communication (among repositories and tools) of both:
    - data: the final results
    - metadata: the laboratory (wet) and computational (dry) workflow, i.e. sample characteristics, experimental design, protocols’ parameters and execution, instrument settings …

Enable interpretation of data and interoperability of resources = advance science
Proliferation of standardization efforts - Good

CONTENT
Minimal/core information to be reported

QUALITY METRICS
Data reliability and confidence

SYNTAX
Format used for communication

SEMANTICS
Terminology used for description

Databases
Journals
http://gensc.org
Proliferation of standardization efforts - Bad

- Fragmentation is an important issue we need to deal with
  - Focused on particular communities’ interests, leads to
    - duplication of effort and development of different standards
    - hindering data integration and interoperability of tools/databases
Consequences of the fragmentation

- These resources, as well as their standards, are not interoperable
  - Different metadata representations and terminologies;
  - Different submission and exchange formats
Synergistic activities: exemplar projects

MIBBI (mibbi.sf.net)

XML

Tabular

Syntactic activities: exemplar projects

Many formats … e.g. GCDML
but several synergise under:

- FuGE XML-based (fuge.sf.net)
- ISA-TAB tabular (isatab.sf.net)
MIBBI project (http://mibbi.sf.net)

- Minimal Information for Biological and Biomedical Investigations
  - International collaboration since 2007
  - Brings together communities developing a ‘minimal information’ checklist

- Portal as a ‘one-stop shop’
  - Serves researchers, journal editors and reviewers, and funders
    - To discover (whether there are) checklists for a particular domain
    - To monitor progress of extant efforts and facilitate collaborations

- Link with health research EQUATOR (http://equator-network.org)

- Foundry for integration
  - Refactor the checklists as integrable, non-overlapping modules

MIBBI paper: Published 2008-08-07 in Nature Biotechnology
<table>
<thead>
<tr>
<th>Project</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIMR</td>
<td>Core Information for Metabolomics Reporting</td>
</tr>
<tr>
<td>MIABE</td>
<td>Minimal Information About a Bioactive Entity</td>
</tr>
<tr>
<td>MIACA</td>
<td>Minimal Information About a Cellular Assay</td>
</tr>
<tr>
<td>MIAME</td>
<td>Minimum Information About a Microarray Experiment</td>
</tr>
<tr>
<td>MIAME/Env</td>
<td>MIAME / Environmental transcriptomic experiment</td>
</tr>
<tr>
<td>MIAME/Nutr</td>
<td>MIAME / Nutrigenomics</td>
</tr>
<tr>
<td>MIAME/Plant</td>
<td>MIAME / Plant transcriptomics</td>
</tr>
<tr>
<td>MIAME/Tox</td>
<td>MIAME / Toxicogenomics</td>
</tr>
<tr>
<td>MIAPA</td>
<td>Minimum Information About a Phylogenetic Analysis</td>
</tr>
<tr>
<td>MIAPAR</td>
<td>Minimum Information About a Protein Affinity Reagent</td>
</tr>
<tr>
<td>MIAPE</td>
<td>Minimum Information About a Proteomics Experiment</td>
</tr>
<tr>
<td>MIARE</td>
<td>Minimum Information About a RNAi Experiment</td>
</tr>
<tr>
<td>MIASE</td>
<td>Minimum Information About a Simulation Experiment</td>
</tr>
<tr>
<td>MIFlowCyt</td>
<td>Minimum Information for a Flow Cytometry Experiment</td>
</tr>
<tr>
<td>MIGen</td>
<td>Minimum Information about a Genotyping Experiment</td>
</tr>
<tr>
<td>MIGS</td>
<td>Minimum Information about a Genome Sequence</td>
</tr>
<tr>
<td>MIMix</td>
<td>Minimum Information about a Molecular Interaction Experiment</td>
</tr>
<tr>
<td>MIMPP</td>
<td>Minimal Information for Mouse Phenotyping Procedures</td>
</tr>
<tr>
<td>MINI</td>
<td>Minimum Information about a Neuroscience Investigation</td>
</tr>
<tr>
<td>MINIMESS</td>
<td>Minimal Metagenome Sequence Analysis Standard</td>
</tr>
<tr>
<td>MINSEQE</td>
<td>Minimum Information about a high-throughput Genomic Sequencing Experiment</td>
</tr>
</tbody>
</table>
ISA-TAB project: simple format for complex studies

- International collaboration (since 2007)
  - Communities developing a tab-delimited format, or where formats do not yet exist or are incomplete and there is an immediate need to collect/exchange datasets

- Investigation /Study /Assay TABular
  - Format (initially) created to address EBI internal ‘issues’
  - It builds on this existing paradigm of the MAGE-TAB (by MGED Society)
    - Has additional features making it a more general framework

- Two use cases for this simple format
  - As submit/export = exchange format
    - For researchers with little/no bioinformatics support to create XML-based files
    - Easily created programmatically or by spreadsheet software
  - User-friendly
### ISA-TAB contributors and documentation

<table>
<thead>
<tr>
<th>System</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArrayExpress</td>
<td>EBI</td>
</tr>
<tr>
<td>ArrayTrack</td>
<td>FDA's NCTR</td>
</tr>
<tr>
<td>BioInvestigation Index</td>
<td>EBI</td>
</tr>
<tr>
<td>CEBS</td>
<td>NIEHS</td>
</tr>
<tr>
<td>Coral</td>
<td>Unilever</td>
</tr>
<tr>
<td>Genome Catalogue</td>
<td>GSC</td>
</tr>
<tr>
<td>Omixed</td>
<td>NERC</td>
</tr>
<tr>
<td>PRIDE</td>
<td>EBI</td>
</tr>
<tr>
<td>Rosetta Resolver®</td>
<td>Rosetta Biosoftware</td>
</tr>
<tr>
<td>Simbioms</td>
<td>EBI, EU Molpage Consortium</td>
</tr>
<tr>
<td>SyMBA</td>
<td>CISBAN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>‘Standards’ Initiative*</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSC</td>
<td>Genomics</td>
</tr>
<tr>
<td>MGED</td>
<td>Transcriptomics</td>
</tr>
<tr>
<td>MIACA</td>
<td>Cellular assay</td>
</tr>
<tr>
<td>MSI</td>
<td>Metabolomics</td>
</tr>
<tr>
<td>PSI</td>
<td>Proteomics</td>
</tr>
<tr>
<td>FuGE</td>
<td>Generic data model</td>
</tr>
<tr>
<td>MIBBI</td>
<td>Modular checklists</td>
</tr>
<tr>
<td>OBI</td>
<td>Common terminology</td>
</tr>
</tbody>
</table>

- **ISA-TAB meetings**
  - 1st meeting: EBI, Dec 2007
  - 2nd meeting: EBI, June 2008
  - 3\textsuperscript{rd} a users meeting: June 2009

- **Specification and stand-alone tool**
  - The ‘release candidate 1, ISA-TAB v1’, with examples
  - ISAcreator and editor (open source) tool

[READNA, Berlin, Germany, Jan 16 2009](http://gensc.org)
Study and Assay metadata repository @ EBI

The BioInvestigation Index – Standards and Infrastructure for Omics Data

Philippe Rocca-Serra, Marco Brandizi, Nataliya Sklyar, Eamonn Maguire, Chris Taylor, Gabriella Rustici and Susanna-Assunta Sansone

http://www.ebi.ac.uk/bioinvindex/
The Genomic Standards Consortium (GSC)

The goal of this international community is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data.

- Established in Sept 2005 (includes participants from DDBJ, EMBL, GenBank, Sanger, JCVI, JGI, EBI and a range of US, UK and EU research institutions)
- GSC 6 held at the EBI, Cambridge, Oct 13-17th 2008
- GSC 7 held at Metagenomics 2008 in San Diego
- Metagenomics, Metadata and Metaanalysis (M3) SIG, ISMB Stockholm, June 27th 2009. Web page and call for papers and proposals out soon.
The goal of this international community is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data.
GSC aim

• The aim of the Genomic Standards Consortium (GSC) is to support the community-based development of a genomic standard that captures a richer set of information about complete genomes and metagenomic datasets.
• "Minimum Information about a Genome Sequence" specification (MIGS).
• “Minimum Information about a Metagenome Sequence” specification (MIMS)

Having this genome metadata available is essential for interpretation of our ‘omics’ data
Achievements

• MIGS Checklist
• MIGS/MIMS Standard
• Establishment of working groups dedicated to descriptions of different taxa (metagenomes, eukaryotes, prokaryotes, viruses, plasmids, and organelles) and concepts (Organism, Environment, Phenotype, Sample Processing, and Data Processing).
• Community support
• "Genomic Contextual Data Markup Language" (GCDML), a much richer XML schema that also implements MIGS/MIMS in addition to containing a wider range of additional elements for describing genomes and metagenomes.
• GSC website, wiki (sourceforge) and mailing lists (see gensc.sf.net)
Monday, November 3

7:30  Shuttles Pick Up at La Jolla Shores Hotel

Enabling Metagenomics and Genomics: Information required for Metagenomics and Genomic Standards
[Session organized by GSC, the Genomic Standards Consortium]

Organizers:
Dawn Field, Oxford Centre for Ecology and Hydrology
Frank Oliver Gloeckner, Max Planck Institute for Marine Microbiology
John Wooley, UCSD

SESSION I: Genomic and Metagenomic Metadata

08:30  John Wooley, Setting the stage: mobilizing the metagenomics
Metagenomics and Genomic Standards
[Session organized by GSC, the Genomic Standards Consortium]

Organizers:
Dawn Field, Oxford Centre for Ecology and Hydrology
Frank Oliver Gloeckner, Max Planck Institute for Marine Microbiology
John Wooley, UCSD

SESSION I: Genomic and Metagenomic Metadata

08:30 John Wooley, Setting the stage: mobilizing the metagenomics community

08:45 Iddo Friedberg, CAMERA interest in community standards for the capture and exchange of metadata: stage setting

09:00 Dawn Field, Overview of the GSC and the Minimum Information about a (Meta) Genome Sequence" (MIGS/MIMS) specification

09:15 Susanna Sansone, The rapidly growing standards landscape in 'omics

09:45 Guy Cochrane, The views of the INSDC on metadata capture
SESSION II: Core GSC Projects

10:30 Renzo Kottmann, Implementing MIGS/MIMS: The Genomic Contextual Data Markup Language (GCDML):

10:45 Peter Dawyndt, StrainInfo and the linkage of organisms, gene and genomes: the Genomic Rosetta Stone

11:00 Lynnette Hirschman, Towards computer assisted markup of data: Habitat-Lite

11:15 George Garrity, Towards a Standards Compliant Literature: the GSC eJournal

11:30 Owen White, Towards transparency of computational analyses: a central SOP repository

11:45 QUESTIONS & COMMUNITY COMMENTS

12:00-1:00 LUNCH
SESSION III: Defining the scope of GCDML

01:00 Jeroen Raes, The importance of context for the design and interpretation of comparative metagenomics studies: the MINIMESS proposal

01:15 Inigo San Gil, Metadata capture for ecological data: The Ecological Markup Language and a harmonization with GCDML

01:30 Frank Oliver, Gloeckner Extending MIGS/MIMS to the description of ribosomal RNA sequences

01:45 QUESTIONS & COMMUNITY COMMENTS

SESSION IV: Ontologies and the description of Habitat and geolocation

02:00 Norman Morrison, The Environment Ontology
02:15 Lynn Schrimal, Towards and open access Gazetteer
02:30 James Cole, RDP Survey of habitats descriptors

02:45 QUESTIONS & COMMUNITY COMMENTS
SESSION V: Metadata Capture - A Key Step for Advancing Understanding

03:30 Nikos Kyrpides, The Genomes Online Database (GOLD): the value of a comprehensive metadata collection

03:45 Victor Markowitz, Metadata capture in the IMG/IMGm: getting scientists to contribute

04:15 QUESTIONS & COMMUNITY COMMENTS

4:30-5:30 Panel Discussion: A community vision
How do we describe the context of the rapidly growing collection of genomes and metagenomes

(1) 1 metagenome
Sargasso Sea
Sanger sequencing
(Venter et al, 2005)

(120) 4 metagenomes & 4 metatranscriptomes
Bergen mesocosm experiment
Pyrosequencing
(Gilbert et al, 2008)

(~80) 41 metagenomes
“Global Ocean Survey”
Sanger sequencing
(Rusch et al, 2007)
The “Minimum Information about a Genome Sequence” (MIGS) specification

MIGS extends what DDBJ/EMBL/GenBank (the INSDC) request upon submission of a genome sequence

- Geographic location of sample and environment - Extended by the “Minimum information about a metagenomic sequence” (MIMS) specification describing measurements that defines an environment
- Unencumbered access to genomic reagents – where?
- Description of organism and phenotype – what?
- Details of how the sequence was generated – how?
“Minimum Information about A Genome Sequence” (MIGS)

<table>
<thead>
<tr>
<th>Investigation</th>
<th>Report type</th>
<th>EU</th>
<th>BA</th>
<th>PL</th>
<th>VI</th>
<th>OR</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Submit to trace archives and INSDC</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Investigation type (i.e., report type)</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Project name</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Study</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Environment</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Geographic location (latitude and longitude), source, taxonomic and region</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Time of sample collection</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Habitat</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>MIMS extension: select to report a set of uniform measurements for a given habitat:</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Water body: Temperature, pH, salinity, pressure, chlorophyll, conductivity, light intensity, dissolved organic carbon (DOC), current, atmospheric data, density, alkalinity, dissolved oxygen, particulate organic carbon (POC), phosphate, nitrate, sulfates, sulfides, primary production</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Nucleic acid sequence source</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Subspecific genetic lineage (lowest rank of NDBI taxonomy, which is subspecies)</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>(e.g. serovar, biotype, ecotype)[NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Ploidy (e.g., allotetraploid, polyploid)</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Number of replicons (EU, BA: chromosomes [haploid count]; VI: segments)</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Extrachromosomal elements</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Estimated size before sequencing: to apply to all draft genomes</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Reference for biomaterial (primary publication if isolated before genome publication; otherwise, primary genome report)</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Source material identifiers: cultures of microorganisms: identifiers</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Culture collections[NCBI]: specimens (e.g., organelles and Eukarya): voucher condition and location[NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Known pathogenicity</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>• Biotic relationship (e.g.: free-living, parasite, commensal, symbiont) [NCBI]</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>• Specific host (e.g., host taxid, unknown, environmental) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Host specificity or range [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Health or disease status of specific host at time of collection (e.g., alive, asymptomatic) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Trophic level (e.g., autotroph, heterotroph) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Propagation (phage, lytic or lysogenic; plasmid, incompatibility group) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Encoded traits (e.g., plasmid, antibiotic resistance, phage: converting genes) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Relationship to oxygen (e.g., aerobic, anaerobic) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Isolation and growth condition [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Biomaterial treatment (e.g., filtering of sea water) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Volume of sample [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Sampling strategy (enriched, screened, normalized) [NCBI]</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Assay</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Sequencing</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Nucleic acid preparation (extraction method[NCBI]; amplification[NCBI] )</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Library construction (library size[NCBI], number of reads[NCBI], vector[NCBI])</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Sequencing method (e.g., diplosequencing, pysequencing, pyrosequencing)</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Assembly (assembly method[NCBI], estimated error rate[NCBI], method of calculation[NCBI])</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Finishing strategy (status—e.g., complete or draft[NCBI], coverage[NCBI], contigs[NCBI])</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Relevant Standard Operating Procedures (SOPs)</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>• Relevant electronic resources</td>
<td></td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>
Reserved keyword: the deal between databases and community

- Databases charge community with responsibility to decide what information should be captured
  - use of source qualifiers and extensible fields

- Community charges databases with data capture and presentation
  - clearly defined compliance rules that can be applied systematically

- Databases report compliance with community’s standard to users
  - use of the reserved keyword system
The MIGS example

A Standard MIGS/MIMS Compliant XML Schema:
Toward the Development of the Genomic Contextual Data Markup Language (GCDML)

Renzo Kottmann,1 Tanya Gray,2 Sean Murphy,3 Leonid Kagan,3 Saul Kravitz,3 Thierry Lombardot,1 Dawn Field,2 Frank Oliver Glöckner,1 and the Genomic Standards Consortium

0.1 microns
0.8 microns
www.environmentontology.org
Ontology Ecology

**Annotator:**
Label data with categories

Goal: Ease of annotation

External measure: Interannotator consistency

**Ontology Developer:**
Build, maintain ontology

Goal: Capture structure & relations

External measures:
- Information content?
- Consistency?
- Adoption?

**Data Generator:**
Enter data into repository

Goal: Ease of data entry

External measure: Completeness of data

**EnvO-Lite:**
Terminology, Guidelines, Mapping tool, Interface

**User:**
Search data set for entries matching search criteria

Goal: Get back just the relevant docs

External measure: Balanced precision/recall

---

READNA, Berlin, Germany, Jan 16 2009

http://gensc.org

Lynnette Hirschman, MITRE
OMICS Version (0.2) of Habitat-Lite (terms + guidelines)

Choose one or more:

Aquatic: freshwater
- soil
Aquatic: marine
- sediment
- oxygen-depleted intertidal marine sediment
- sludge
- thermophilic methanogenic sludge
Aquatic: marine, Extreme
- biofilm
- microbial mat
Aquatic
- sludge
- hot spring
- hydrothermal vent
- the shallow hot vent in IowoJima
- microbial mat
Aquatic, Extreme
- biofilm of drinking water distribution system
Aquatic, hot spring?
- chlorophyll
- cyanobacteria
- sulfur-oxidizing bacteria
- sulphur-oxidizing bacteria
Food
- surface of smear ripened cheese
Organism Associated
- gut of nitidulid beetle
Organism-Associated
- anatomy
- gut
- skin
- gut, skin
Extreme Environment
- extremely alkaline (pH 12 to 13) groundwater
Other
- chalky-salty rendzine in sebkha

Other
Essential URLs

**Wiki:**
Envo wiki: http://gensc.org/gc_wiki/index.php/EnvO_Project

**Gaz and Envo Sourceforge sites:**
http://obo.cvs.sourceforge.net/obo/obo/ontology/environmental/

**Tracker Sites at Sourceforge:**
Tracker for EnvO: https://sourceforge.net/tracker/?group_id=76834&atid=934923

**Gazetteer Tracker:**
http://sourceforge.net/tracker/?group_id=76834&atid=1014856

**OBO-ENVO Discussion List:** obo-envo@lists.sourceforge.net
motivation

**Complete Genomes (1995-2007)**
(source: GOLD)
SIGS is an open-access, standards-supportive publication that rapidly disseminates concise genome and metagenome reports in compliance with MIGS/MIMS standards, and presents coverage of: detailed standard operating procedures, meeting reports, reviews and commentaries, data policies, white papers and other gray literature, that are relevant to genome sciences, but absent from the scholarly literature.
### Editorial office

<table>
<thead>
<tr>
<th><strong>Location</strong></th>
<th>Department of Microbiology &amp; Molecular Genetics, Michigan State University</th>
</tr>
</thead>
</table>
| **Editor-in-Chief** | George Garrity, Sc.D.  
Professor, MSU |
| **Managing editor** | Scott Harrison, Ph.D.  
Research Associate, MSU |
| **Genome editor, EU** | Peter Sterk, Ph.D.  
EBI, Sanger |
| **Production editor** | Position open |
Founding Members of the Editorial Board

Sam Angiuoli  
University of Maryland School of Medicine

Eugene Kolker  
Seattle Children's Hospital

Susanna-Assunta Sansone  
European Bioinformatics Institute

Nikos C Kyrpides  
DOE Joint Genome Institute

Patrick Chain  
DOE Joint Genome Institute

Lynn Schriml  
University of Maryland School of Medicine

Dawn Field  
NERC Centre for Ecology and Hydrology

Peter Sterk  
European Bioinformatics Institute

Frank-Oliver Glöckner  
Max Planck Institute for Marine Microbiology

David Ussery  
Technical University of Denmark

Lynette Hirschman  
Mitre Corp.

Owen White  
University of Maryland School of Medicine
Confirmed Editorial Board

Scott Baker, PNL
Mark Blaxter, U. Edinburgh
Rob Britton, MSU
Guy Cochrane, EMBL
Jonathan Eisen, UC Davis
Pilar Francino, JGI
Iddo Friedberg, UCSD
Phil Hugenholtz, JGI
Janet Jansson, LBNL
Hans Peter Klenk, DSMZ
Ilene Karsch-Mizrachi, GenBank
Jereon Raes, EMBL

Martin Keller, ORNL
Todd Lowe, UCSC
Natalia Maltsev, U Chicago
Victor Markowitz, LBNL
Norm Morrison, NERC
Christos Ouzounis, Kings College
Mircea Podar, ORNL
Martin Poltz, MIT
Sudipto Saha, Case Western Reserve
Gretta Serres, MBL
Nick Thompson, Sanger

Your name could appear here
MIGS/MIMS Adoption

Optional Fields
- MIENS
- MIMS
- MIGS

GCDML for data capture
- <GCDML>

GCDML for data exchange
- <GCDML>

Researchers in the GSC
- StrainInfo
- RDP
- Sliva
- JGI Researchers

Single Global List of Genomes and Metagenomes

READNA, Berlin, Germany, Jan 16 2009

http://gensc.org
Summary

• The GSC is an open membership consortium working towards richer descriptions of the complete collection of genomes and metagenomes

• The mission of the GSC is to work towards richer descriptions of genomes and metagenomes (and environmental sequences)
Summary

• The GSC is working to fulfil this stated mission by building a tightly networked, expert community that will foster the development and implementation of standards for data description and exchange.

• The goal of the GSC, is and will always be, to facilitate and accelerate scientific discovery – the paradigm shift we need is an improved appreciation of contextual data, new mechanisms for dealing with it at both the local and global levels.
Summary

• As a first step, the GSC has published the “Minimum Information about a (Meta) Genome” (MIGS) specification

• We can soon recommend compliance in the form of a structured comment submitted to the INSDC
Summary

• The GSC is now shifting to next phase: “Implementation” of MIGS/MIMS
• To implement MIGS/MIMS the GSC needs core technology projects (GCDML, GRS, Habitat-Lite)
• The GSC is launching a standards-compliant and standards-supportive journal called “Standards in Genomic Sciences”
• Other projects on the horizon – including MIENS
• Ideas, Signatories, Linkages, Feedback and Volunteers welcome
The minimum information for genome (MIGS) specification

Dawn Field¹, George Garrity¹, Nicholas Thomson⁸, Michael AddEnlarge, Sandra Baldauf¹², Stuart Ball¹, Claude dePamphilis¹⁸, Robert Field¹, Frank Oliver Glöckner²³, Philip Henning Hermjakob⁶, Chris Hope⁴, Kathleen Kennedy²⁷, George Kogut²⁸, Jim Leebens-Mack³³, Suzanne Markowitz³⁷, Jennifer Markowitz³⁷, Julian Parkhill⁸, Lita Proctor⁴⁵, Paul Swift¹, Chris Taylor⁶, Yo-Wei Yeh⁴⁵, Naomi Ward⁴⁵, Trish Whetzel⁴⁵, Yvonne Wilkins⁴⁵, Peter Yarwood⁴⁵

The Genomic Standards Consortium has an open membership policy. If you would like to join the GSC, please contact one of the members or email curator@csb.info.