OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2008.0A10

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

Abstract

The Genomic Contextual Data Markup Language (GCDML) is a core project of the Genomic Standards Consortium (GSC) that implements the "Minimum Information about a Genome Sequence" (MIGS) specification and its extension, the "Minimum Information about a Metagenome Sequence" (MIMS). GCDML is an XML Schema for generating MIGS/MIMS compliant reports for data entry, exchange, and storage. When mature, this sample-centric, strongly-typed schema will provide a diverse set of descriptors for describing the exact origin and processing of a biological sample, from sampling to sequencing, and subsequent analysis. Here we describe the need for such a project, outline design principles required to support the project, and make an open call for participation in defining the future content of GCDML. GCDML is freely available, and can be downloaded, along with documentation, from the GSC Web site (http://gensc.org).

Introduction

It is well known that the entire collection of genomic and metagenomic DNA sequences is a complex and valuable resource (Field and Hughes, 2005). Hundreds of archaeal, bacterial, and eukaryotic genomes have been sequenced since the first microbial genome, *H. influenzae* (Fleischmann et al., 1995), was published a decade ago. Additionally, DNA sequences of thousands of organelles, plasmids, and viruses are available. In recent years, metagenomic sequencing has also become more prominent, with the largest single input of new sequences provided by the Global Ocean Survey (GOS) (Rusch et al., 2007). Taking into account the immense genomic diversity found in the natural world (Binnewies et al., 2006), it is anticipated that large-scale metagenomic sequencing projects, such as GOS, flag just the beginning of a new era in molecular microbiology.

In particular, in light of the rapidly growing number of environmental and microbiome sequencing projects, it is increasingly clear that the biological interpretation of such data, especially in a comparative context, is dependent on the quantity and quality of associated information (Field et al., 2008a; Raes et al., 2007). It is the aim of the Genomic Standards Consortium (GSC) to support the capture of a richer set of data. The first step of this international community has been to define the "Minimum Information about a Genome Sequence" (MIGS) and "Minimum Information about a Metagenome Sequence" (MIMS) specifications. Use of MIGS/MIMS will provide a mechanism for capturing a consensus-driven minimum set of metadata describing aspects of genomes and metagenomes, such as geographic location and habitat type from which the sample was taken, as well as the details of the sequencing method used. The support of maximum reporting of such projects, though, will require a much richer set of descriptors (Raes et al., 2007). Such descriptors must cover both the origin and processing of a sample, from the time of sampling up to sequencing, and the subsequent analysis. This suite of metadata is collectively referred to here as contextual data.

It is the aim of the GSC to provide support for the capture of richer contextual data describing genomes and metagenomes by developing the Genomic Contextual Data Markup Language (GCDML). This project is the natural extension of original efforts to implement the MIGS/MIMS check-list as an XML Schema (MIGS.xsd) (Field et al., 2008a). The scope of this restricted schema evolved into the scope of GCDML to specifically support "maximal" reporting of contextual data and the desire of groups in the GSC to include local descriptors in the original MIGS/MIMS schema.

There are two key aspects to the development of GCDML. First are the technical aspects of "how to build it." Second is the issue of "scope," or "what to put in" (the exact descriptors to be included). Here, the focus is primarily on the former and outline the design principles of GCDML as a technical shell for future content development in the coming years. The current core scope of GCDML has been defined

[1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, 28359 Bremen, Germany.
[2]NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford, Oxfordshire, United Kingdom.
[3]J. Craig Venter Institute, Rockville Maryland.

by the MIGS/MIMS checklist, and covers the minimum description of "nucleic acid sequence source," "environment," and "sequencing methodology" (Field et al., 2008a). Beyond that, it will take future GSC workshops, telecoms, and significant participant contribution to develop the full content of GCDML based on the needs and interests of the community.

The Scope of GCDML and Design Considerations

GCDML aims to take full advantage of the benefits of an XML representation of genomic contextual data. XML provides a machine-readable representation of metadata that facilitates the capture, exchange, and comparison of large amounts of data. XML is widely used to describe data capture and exchange format. Numerous XML definitions exist in bioinformatics (for an overview, see Seibel et al., 2006). XML definitions such as BSML (Cerami, 2005) and INSDSeq (http://www.insdc.org/page.php?page=documents last verified on 15.02.2008) have applications in sequence annotation; PSAML (Su-Hyun et al., 2002), ProML (Hanisch et al., 2002) and PSI MI (Hermjakob et al., 2004) in protein modeling and interaction; MAGE-ML (Spellman et al., 2004) in gene expression. The Functional Genomics (FuGE) project aims to develop a single generic data model that will underpin a variety of XML-based formats by providing a single common framework (Jones et al., 2008). XML is also widely used in many other scientific fields, and for example, the Ecological Modeling Language standard is described in this issue (San Gil et al., 2008).

In overview, MIGS/MIMS will be central to GCDML, and GCDML will provide the GSC's official implementation of the checklist. This requires GCDML to specifically support the validation of different subsets of descriptors because MIGS/MIMS is applied differently across taxa (e.g., eukaryotes vs. bacteria vs. viruses or metagenomes) (Field et al., 2008a). Beyond the minimum descriptors of MIGS/MIMS, GCDML will be open and extensible to evolve with the needs of the community. In mature versions of GCDML, it is envisioned that it will be possible to describe the exact origin and processing of a biological sample from the time of sampling up to sequencing, and subsequent analysis. This should support a range of desired applications, such as tracking the geographic origin and habitat of a sample or a set of organisms, for example, comparative ecological genomic studies, to capture the pathogenicity of a sequenced organism, or to describe the host–microbiome relationship in human microbiome studies. GCDML should also be viewed as a single, community developed specification for exchanging data between databases and providing integrated information from the resources to the wider community.

At the technical level, GCDML must be designed to enable support for the above requirements, facilitate a broad adoption, and support the future inclusion of a far wider range of descriptors. GCDML is therefore being built to have a strongly typed and clear structure that embodies the following design principles. GCDML must be:

- ”application agnostic” (to facilitate the mapping of data from diverse resources and applications, and the use by applications with their own needs, e.g., Web services)
- compliant with (MIGS/MIMS), while allowing for richness of expression
- support the integration of terms lists, including those from ontologies
- allow the recording of legacy data, even when fields are missing
- open and extensible to allow evolution of the MIGS/MIMS specification and all associated optional descriptors of genomes and metagenomes, as well as the evolution of new databases and sources of related metadata.
- open to link, map, or incorporate other standards as required, and in particular, provide integration with of Geography Markup Language (GML), which became ISO standard 19136
- support versioning of GCDML and any reports generated from it.

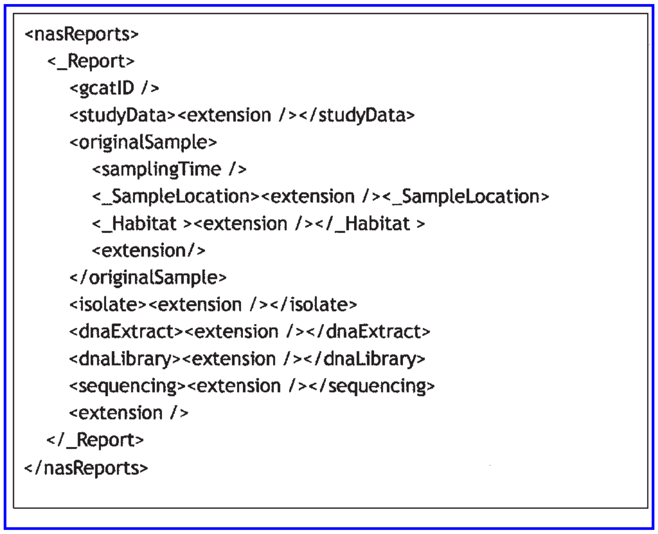
These complex requirements first necessitated a shift in the design of the prototype MIGS.xsd schema from a Russian doll model to an open flat design (van der Vlist, 2002), which makes the schema highly modular (no nesting of elements). The hierarchical structure of the different report types is subsequently created by appropriately nesting references to the existing elements (see below). This provides significant practical benefits. First, it facilitates implementation of one kind of genome report for all types of MIGS/MIMS (eukaryote, bacteria and archaea, plasmid, virus, organelle, or metagenome) project types (Field et al., 2008a). Now, any report can, using the relevant subsets of MIGS/MIMS descriptors, be validated by any XML Schema-capable XML parser. Second, it allows definition of additional descriptors, which do not interfere with MIGS/MIMS genome reports. Third, the flat design is the basis for an open, extensible design (see below), which allows both the necessary evolution of the MIGS/MIMS specification and the additional descriptors (Field et al., 2008a).

Supporting MIGS/MIMS compliance: report structure

Although reports can drastically differ in their details, as they do for MIGS/MIMS compliant reports, in GCDML they all follow the same semantic structure. The main XML elements are simplified models of samples that are produced during genomic studies. The order of the XML elements follows a simplified and generalized protocol for sample collection (<originalSample>), in the case of a single organism, the isolation (<isolate>), DNA extraction (<dnaExtract>), DNA library construction (<DNALibrary>), and sequencing details, including assembly (<sequencing>). The descriptors of the MIGS/MIMS checklist are associated with the main XML elements via nesting. An example, as depicted in Figure 1, is the nesting of sampling time (<samplingTime />), sample location (<_SampleLocation />), and habitat (<_Habitat />) within <_originalSample />.

MIGS/MIMS-compliant reports share the same root element

The nasReports (Nucleic Acid Sequence reports) element is the root element of all five MIGS/MIMS compliant report types for Eukarya, Bacteria, Archaea, Plasmids, Viruses, Organelles, and Metagenomes (Fig. 1). The root element acts as



```
<nasReports>
  <_Report>
    <gcatID />
    <studyData><extension /></studyData>
    <originalSample>
      <samplingTime />
      <_SampleLocation><extension /><_SampleLocation>
      <_Habitat ><extension /></_Habitat >
      <extension/>
    </originalSample>
    <isolate><extension /></isolate>
    <dnaExtract><extension /></dnaExtract>
    <dnaLibrary><extension /></dnaLibrary>
    <sequencing><extension /></sequencing>
    <extension />
  </_Report>
</nasReports>
```

FIG. 1. General structure of nucleic acid sequence (nas) reports including extension points (<extension/>) within various elements.

a container for any number of NAS Reports per XML document. GCDML does not put any constraint on this container. Thus, NAS reports can be grouped in any way, for example, from groups of different reports of the same genome (but from different sources), to groups of genome reports of the same taxonomic group.

Keeping "free text" to a minimum and the use of categorical terms lists for report values

Validation of categorical terms with standard XML parsers is achieved through the enumeration of valid terms. XML schema allows the construction of simple types with a restricted sets of terms, which can come from ontologies, taxonomies, gazetteers, and other sources of controlled vocabularies. This general approach allows mixing of terms from different sources. GCDML makes wide use of categorical terms for many MIGS/MIMS descriptors. An XML schema example is given in Figure 2a, which restricts an XML type to the terms "soil" and "water," as depicted in Figure 2d. This technique allows syntactical restriction of the set of useful terms and validation with standard XML parsers. Therefore, the use of free-text fields is decreased to an absolute minimum.

Supporting the use of ontologies for categorical terms

The GSC is strongly committed to the use and development of ontologies. For example, the GSC is a member community in the Ontology for Biomedical Investigations (Whetzel et al., 2006), a founding member of the is Environment Ontology project (http://environmentontology.org), and is driving the development of a minimum controlled vocabulary of habitat terms (Hirschman et al., 2008). The GSC has agreed that, just as for MIGS/MIMS (Field et al., 2008a), there is a need to provide semantic transparency for GCDML through the integration of ontologies. An outstanding issue is how to do it best. The following consensus emerged from discussions within the GSC: finding a solution that does not add complexity to NAS reports, does not depend on the availability of ontologies, and does not force users of GCDML, and/or NAS reports, to become acquainted with ontologies.

GCDML currently uses SAWSDL, which stands for "Semantic Annotation for WSDL and XML Schema" (Kopecký et al., 2007). The World Wide Web Consortium recommendation (W3C; August 2007) allows annotation of XML Schema with references to ontological concepts (independent of the type and format of the ontology). SAWSDL allows separation of the syntactic modeling of data and semantic modeling of a knowledge domain by first focusing on the use of XML Schema to model data. Next, SAWSDL is used to state within the GCDML schema which XML schema construct has a meaningful relationship to which ontological entity. Thus, enumeration of categorical terms can be used to ensure the syntactic consistency of NAS reports, while allowing semantic applications to utilize NAS reports with ontological concepts. Figure 2c shows the XML Schema from Figure 2a with SAWSDL annotation, which still validates the same XML documents as the non-annotated XML schema. SAWSDL can be introduced at any time, because it would not affect NAS reports in any way.

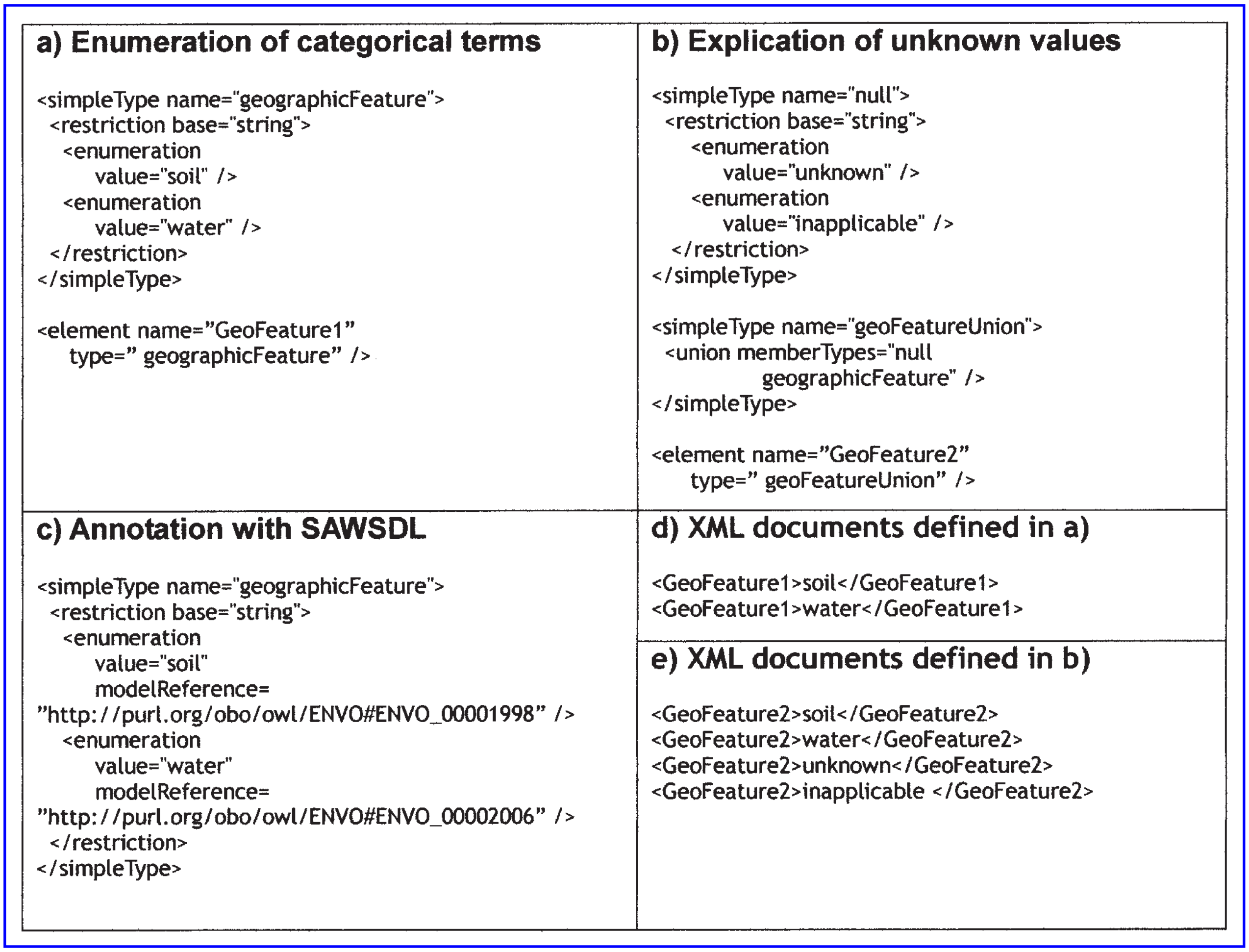
Integration of legacy data with "missing fields"

The question of how to integrate noncompliant legacy data, without relaxing rules and constraints on future data incorporation, is an important issue in the application of any new standard. A case study is the issue of reporting sample location and sampling time, which are mandatory according to the MIGS/MIMS check-list (Field et al., 2008a). Although the community agreed it is necessary to require sample location and time in the future, it was realized that these data are often not available for legacy data.

To overcome this, XML Schema allow for the creation of unions of simple types. GCDML introduces a special simple type for the enumeration of categorical terms that indicate where and why data is missing. A simplified example is given in Figure 2b: first, a simple type, "null," is created; second, a new simple type, "geoFeatureUnion," is created as the union of the "null" simple type and the "geographicFeature" simple type, as given in Figure 2a. GCDML uses these union types to explicitly state reasons for missing data throughout the schema.

Openness and extensibility of GCDML

The extensibility of GCDML is defined by the ability of other applications to use elements from GCDML as a basis for its own element definitions (van der Vlist, 2002). This is facilitated by the flat design of GCDML, use of substitution groups, and almost exclusive derivation by extension (van der Vlist, 2002). Openness of an XML schema is achieved if the schema allows addition of any content at well-defined extension points in XML documents. GCDML adds extension points to <nasReports>, <_Report>, <studyData>, <sequencing>, <habitat>, and each element that models a sample (see Fig. 1). This allows applications generating MIGS/MIMS compliant <nasReports> to add additional self-defined XML within its extension points, which adhere to GCDML.



a) Enumeration of categorical terms

```
<simpleType name="geographicFeature">
  <restriction base="string">
    <enumeration
        value="soil" />
    <enumeration
        value="water" />
  </restriction>
</simpleType>

<element name="GeoFeature1"
    type=" geographicFeature" />
```

b) Explication of unknown values

```
<simpleType name="null">
  <restriction base="string">
      <enumeration
          value="unknown" />
      <enumeration
          value="inapplicable" />
   </restriction>
</simpleType>

<simpleType name="geoFeatureUnion">
  <union memberTypes="null
              geographicFeature" />
</simpleType>

<element name="GeoFeature2"
    type=" geoFeatureUnion" />
```

c) Annotation with SAWSDL

```
<simpleType name="geographicFeature">
  <restriction base="string">
    <enumeration
        value="soil"
        modelReference=
"http://purl.org/obo/owl/ENVO#ENVO_00001998" />
    <enumeration
        value="water"
        modelReference=
"http://purl.org/obo/owl/ENVO#ENVO_00002006" />
  </restriction>
</simpleType>
```

d) XML documents defined in a)

```
<GeoFeature1>soil</GeoFeature1>
<GeoFeature1>water</GeoFeature1>
```

e) XML documents defined in b)

```
<GeoFeature2>soil</GeoFeature2>
<GeoFeature2>water</GeoFeature2>
<GeoFeature2>unknown</GeoFeature2>
<GeoFeature2>inapplicable </GeoFeature2>
```

FIG. 2. XML examples demonstrating different implementation approaches of GCDML: (**a**) a simplified XML schema using enumerations to restrict values to the categorical terms "hot spring" and "hydrothermal vent" only; (**b**) a simplified XML Schema for handling missing data; (**c**) the same simpletype as in (**a**) annotated following the SAWSDL standard; (**d**) simplified XML documents showing all possible uses of <GeoFeature1> as defined in (**a**); and (**e**) simplified XML documents showing all possible uses of <GeoFeature2> as defined in (**b**).

Versioning GCDML and NAS reports

Versioning is the process of assigning unique version numbers to either unique states of GCDML or unique states of NAS reports, and other applications of GCDML. However, versioning GCDML is separate from versioning NAS reports in the sense that a version of GCDML indicates a certain state of the grammatical structure and constraints that apply to NAS reports. Whereas versioning of NAS reports indicate the state of the information content of the report.

GCDML will use the common *major.minor.release* versioning scheme, with version "1.5.0" serving as the first "public" release version. The criterion for an increase of the *major* number is a feature extension to GCDML. A *minor* version update will be triggered by a set of changes to GCDML that require NAS reports to be transformed in order to concur with the new version. Finally, the *release* number is set to increase in the case of changes to GCDML that have no effect on any existing NAS report.

Early Adopters of GCDML

Several groups presented on their intentions to adopt GCDML at the fifth GSC workshop (Field et al., 2008c). The first true example of adoption of GCDML within a database is that of the GSC's Genome Catalog (GCat). GCat is an online database designed to collect MIGS/MIMS compliant reports. It is run with the generic GenCat software, which creates an online database system with data input, edit, browse, and search functions from XML Schemas to allow capture of XML schema-compliant reports. Adoption has involved replacement of the original MIGS.xsd with the gcdml.xsd and the addition of support for new features found in GCDML. GCat will help the GSC to demonstrate and clarify the details of GCDML to the wider community, as it renders the GCDML as an input form. The GCDML-driven input form identifies optional elements, repeatable elements, substitution groups, and choices, as well as lists the values of enumerated elements. In addition, element names and associated documentation contained in the XML schema are displayed in the input form. GCat further supports the ongoing development of GCDML by, for example, its support of term capture. This will allow newly submitted terms in the input form to be included as enumerated elements in GCDML. The terms can later be reviewed by a third party, for potential inclusion in the next revision of GCDML.

As described in this special issue, the use of GCDML is now also being explored by the Long-Term Ecological Research (LTER) Network (San Gil et al., 2008). The LTER is a collaborative effort funded by the National Science Foundation to investigate ecological processes over long temporal and broad spatial scales. Increasingly, the microbial observatories within the LTER are generating genomic and metagenomic data. The LTER has produced the Ecological Metadata Language (EML) standard, an XML Schema that provides a metadata specification for describing data relevant to the ecological discipline (Fegraus et al., 2005).

A Roadmap for GCDML and an Open Call for Participation

GCDML was initially proposed at the fourth GSC Workshop in June 2007 (Field et al., 2008b). Subsequently, a GCDML prototype was developed and further improvements made based on community feedback, including that of experts on the biology of particular types of genomes. At the fifth GSC Workshop, GCDML was officially accepted as the GSC implementation of MIGS/MIMS and as a candidate standard mechanism for data exchange across a range of databases (Field et al., 2008c). As the next top-priority core GSC project, following the publication of the MIGS specification (Field et al., 2008c), our top aim is to develop a stable release of GCDML and additional working examples of the use of GCDML within the community.

The construction of GCDML will largely progress in three overlapping phases:

1. Setup of GCDML with a primary focus on supporting the creation of MIGS/MIMS-compliant XML reports
2. Integration of ontologies
3. Enlargement of the set of descriptors based on the needs of a variety of databases and additional specifications (e.g., MINIMESS)

Phase 1 is largely complete, and MIGS/MIMS version 2.0 is implemented in the current schema. Phase 2 is in the early stages, and will depend heavily on the maturation of a variety of key controlled vocabularies, like Habitat-Lite (Hirschman et al., 2008), and ontologies (e.g., the Environment Ontology (http://environmentontology.org) upon which Habitat-Lite is based). Phase 3 will start in the summer of 2008.

A key purpose of the description of GCDML presented here is to serve as an open call for participation in the GCDML project. Just as for the GSC as a whole (Field et al., 2008a) our activities are open to new contributors and end users at any time; only through wide participation of the community will GCDML evolve to be a valuable consensus-driven tool of use beyond the GSC's Genome Catalog. In particular, this work should proceed through the development of a series of case studies. As GCDML matures, we aim to shift to a more formal mechanism of vetting the inclusions of terms and changes to the core schema. We plan to produce a fully MIGS/MIMS compliant version of GCDML with support for ontologies by the end of 2008 for presentation at the combined GSC/ ”Metagenomics 2008” meeting. All versions of the GCDML schema and additional documentation are available at http://sourceforge.net/projects/gensc.

Discussion

GCDML meets the two key goals of the GSC, providing both richer descriptions of genomes and metagenomes and facilitating open and transparent data exchange (Field et al., 2008a). The purpose of GCDML is to provide a mechanism for capturing MIGS/MIMS-compliant reports and a larger XML vocabulary for transparent contextual data exchange between specialized local databases, such as the GSC's Genome Catalog, Genome Reviews (Sterk et al., 2006), GOLD (Liolios et al., 2008), IMG (Markowitz et al., 2008), Genomes Mapserver (Lombardot et al., 2006), and CAMERA (Seshadri et al., 2007). The technical design principles outlined here will enable this in the future.

The sample-centric design of the NAS Reports is in accordance with the harmonization efforts of the different standardization initiatives of the “omics” community (Morrison et al., 2006), especially the “Investigation, Study, Assay” concept of the Reporting Structures for Biological Investigations (RSBI) working group (Sansone et al., 2008). Moreover, the sample-centric structure of NAS reports (Fig. 1) has proven to be quite stable through a round of several iterations of the schema. The numerous improvements of GCDML based on community feedback from experts on the biology of particular types of genomes did not change the general structure.

The sample-centric, open, and extensible design of GCDML serves well as the basis for applications beyond MIGS/MIMS-compliant reports as well. For example, at the last GSC meeting, members of the Alpine Microbial Observatory proposed the extension of GCDML to report contextual data for single rRNA sequences (Field et al., 2008c). This would undoubtedly benefit the “molecular diversity” community, which is faced with more than 500,000 rRNA sequences in the public nucleotide databases (Cole et al., 2007; DeSantis, 2006; Pruesse et al., 2007).

Further community acceptance of GCDML can be expected through the integration of existing industrial strength standards, such as Geography Markup Language (Lake et al., 2004) and SAWSDL (Kopecký et al., 2007). Advanced programming interfaces exist for SAWSDL to facilitate ontological use of GCDML and many geographical tools already support GML and derived standards (http://www.opengeospatial.org/resource/products, last verified on 15.02.2008). Besides public visibility, one key to broad acceptance of GCDML is comprehensive user and developer documentation, including cookbook style best practice guides.

Summary

GCDML serves as the official XML schema implementation of the MIGS/MIMS specification, as defined by the Genomic Standards Consortium (GSC) (Field et al., 2008a). The design of GCDML is guided by several principles. First, it is strongly typed, avoids free-text fields through the use of enumerations, and supports the use of terms from ontologies. This is expected to have a significant impact on data integrity. Second, GCDML complies with the minimum requirements of MIGS/MIMS, but additionally allows a much richer contextual description of genomic studies. Third, it allows the incorporation of nonstandard legacy data without weakening the strong typing. Fourth, the flat design allows other applications to use selected parts of GCDML for their own needs, and is therefore well suited, for example, for

Web-service applications. Finally, the open and extensible design of GCDML allows feature extensions with minimal effort to update existing reports. Thus, making GCDML prepared for extensions, such as the inclusion of descriptors proposed by MINIMESS (Raes et al., 2007) and additional descriptors for more comprehensive description of the annotation processes of genomic and metagenomic sequences.

Acknowledgments

We would like to thank the experts on the biology of particular types of genomes and metagenomes for their valuable comments during the development of GCDML, namely, Jack Gilbert, Keith James, Gareth Wilson, Sarah Turner, Christiane Hertz-Fowler, Peter Dawyndt, Phillip Goldstein, and Melissa Duhaime. We thank the EU Sixth Framework Programme (FP6-NEST) for providing financial support (MetaFunctions project, contract no. 511784). Tanya Gray is support by NERC Grants NE/E007325/1 and NE/3521773/1 to D.F.

Author Disclosure Statement

The authors declare that no competing finacial interests exists.

References

Binnewies, T.T., Motro, Y., Hallin, P.F., Lund, O., Dunn, D., La, T., et al. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries.. Funct Integr Genomics 6, 165–185.

Cerami, E. (2005). *XML for Bioinformatics* (Springer Science+ Buisness Media, Inc., New York).

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., et al. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res 35, D169–D172.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72, 5069–5072.

Fegraus, E.H., Andelman, S., Jones, M.B., and Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bull Ecolo Soc Am 86, 158–168.

Field, D., and Hughes, J. (2005). Cataloguing our current genome collection. Microbiology 151, 1016–1019.

Field, D., Garrity, G.M., Gray, T., Morrison, N., Selengut, J.D., Sterk, P., et al. (2008a). The "Minimum Information about a Genome Sequence"(MIGS) specification. Nat Biotechnol 26, 541–547.

Field, D., Glöckner, F.O., Garrity, G.M., Gray, T., Sterk, P., Cochrane, G., et al. (2008b). Meeting report: the 4th Genomic Standards Consortium (GSC) workshop. OMICS (this issue).

Field, D., Garrity, G.M., Sansone, S.-A., Sterk, P., Gray, T., and Glöckner, F.O. (2008c) The fifth Genomic Standards Consortium Workshop meeting report. OMICS (this issue).

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Hanisch, D., Zimmer, R., and Lengauer, T. (2002). ProML—the protein markup language for specification of protein sequences, structures and families. In Silico Biol 2, 313–324.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., et al. (2004). The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat Biotechnol 22, 177–183.

Hirschman, L., Clark, C., Bretonnel, C., Mardis, S., Luciano, J., Cole, J., et al. (2008). Habitat-Lite: a GSC case study based on free text terms for environmental metadata. OMICS (this issue).

Jones, A.R., Miller M., Aebersold, R., Apweiler, R., Ball, C.A., Brazma, A., et al. (2008). The functional genomics experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol 25, 1127–1133.

Kopecký, J., Vitvar, T., Bournez, C., and Farrel, J. (2007). SAWSDL: semantic annoations for WSDL and XML schema. IEEE Internet Comput 11, 60–67.

Lake, R., Burggraf, D., and Trninic, M. (2004). *Geography Mark-Up Language (GML): foundation for the Geo-Web* (John Wiley & Sons Ltd., West Sussex, London).

Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N.C. (2008). The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36, D475–D479.

Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C. et al. (2006). Megx.net—database resources for marine ecological genomics. Nucleic Acids Res 34, D390–D393.

Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.A., et al. (2008). The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res 36, D528–D533.

Morrison, N., Cochrane, G., Faruque, N., Tatusova, T., Tateno, Y., Hancock, D., et al. (2006). Concept of sample in OMICS technology. OMICS 10, 127–137

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188–7196.

Raes, J., Foerstner, K.U., and Bork, P. (2007). Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10, 490–498.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5, e77.

San Gil, I., Sheldon, W., Schmidt, T., Servilla, M., Aguilar, R., Greis, C., et al. (2008). Defining linkages between the GSC and NSF's LTER program: how the ecological metadata language relates to GCDML and other outcomes. OMICS (this issue).

Sansone, S.-A., Rocca-Serra, P., Brandizi, M., Brazma, A., Field, D., and Fostel, J. (2008). The first RSBI (ISA-TAB) workshop: "Can a simple format work for complex studies?" OMICS (this issue).

Seibel, P.N., Krüger, J., Hartmeier, S., Schwarzer, K., Löwenthal, K., Mersch, H., et al. (2006). XML schemas for common bioinformatic data types and their application in workflow systems. BMC Bioinformatics 7, 490.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: a community resource for metagenomics.. PLoS Biol 5, e75.

Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. 3, 9.

Sterk, P., Kersey, P.J., and Apweiler, R. (2006). Genome reviews: standardizing content and representation of information about complete genomes. OMICS 10, 114–118.

Su-Hyun, L., Jin-Hong, K., Geon-Tae, A., and Myung-Joon, L. (2002). An XML representation of protein data for efficient structure comparison. *Proc. of Second International Conference on Computer and Information Science*, No. 1, p. 313.

van der Vlist, E. (2002). *XML Schema*. (O'Reilly & Associates, Inc., Sebastopol, CA).

Whetzel, P.L. Brinkman, R.R., Causton, H.C., Fan, L., Field, D., Fostel, J., et al. (2006). Development of FuGO: an ontology for functional genomics investigations. OMICS 10, 199–204.

Address reprint requests to:
Renzo Kottmann
Microbial Genomics Group
Max Planck Institute for Marine Microbiology and
Jacobs University Bremen
Celsiusstr. 1
28359 Bremen, Germany

E-mail: rkottman@mpi-bremen.de

This article has been cited by:

1. Peter C Marks, Marc Bigler, Eric B Alsop, Adrien Vigneron, Bart P Lomans, Renato De Paula, Brett Geissler, Nicolas Tsesmetzis. 2018. MetaHCR: a web-enabled metagenome data management system for hydrocarbon resources. *Database* **2018**. . [Crossref]
2. André Antunes, Marta F. Simões, Stefan W. Grötzinger, Jörg Eppinger, Judith Bragança, Vladimir B. Bajic. Bioprospecting Archaea: Focus on Extreme Halophiles 81-112. [Crossref]
3. Nicolas Tsesmetzis, Pelin Yilmaz, Peter C. Marks, Nikos C. Kyrpides, Ian M. Head, Bart P. Lomans. 2016. MIxS-HCR: a MIxS extension defining a minimal information standard for sequence data from environments pertaining to hydrocarbon resources. *Standards in Genomic Sciences* **11**:1. . [Crossref]
4. Xiaoli Li, Lai Song, Guoliang Wang, Lufeng Ren, Dan Yu, Guanjun Chen, Xumin Wang, Jun Yu, Guiming Liu, Zongjun Du. 2016. Complete genome sequence of a deeply branched marine Bacteroidia bacterium Draconibacterium orientale type strain FH5T. *Marine Genomics* **26**, 13-16. [Crossref]
5. G. Droege, K. Barker, O. Seberg, J. Coddington, E. Benson, W. G. Berendsohn, B. Bunk, C. Butler, E. M. Cawsey, J. Deck, M. Döring, P. Flemons, B. Gemeinholzer, A. Güntsch, T. Hollowell, P. Kelbert, I. Kostadinov, R. Kottmann, R. T. Lawlor, C. Lyal, J. Mackenzie-Dodds, C. Meyer, D. Mulcahy, S. Y. Nussbeck, É. O'Tuama, T. Orrell, G. Petersen, T. Robertson, C. Söhngen, J. Whitacre, J. Wieczorek, P. Yilmaz, H. Zetzsche, Y. Zhang, X. Zhou. 2016. The Global Genome Biodiversity Network (GGBN) Data Standard specification. *Database* **2016**, baw125. [Crossref]
6. Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R. Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, Suresh Poudel, David W. Ussery. 2015. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics* **15**:2, 141-161. [Crossref]
7. Ramona L. Walls, John Deck, Robert Guralnick, Steve Baskauf, Reed Beaman, Stanley Blum, Shawn Bowers, Pier Luigi Buttigieg, Neil Davies, Dag Endresen, Maria Alejandra Gandolfo, Robert Hanner, Alyssa Janning, Leonard Krishtalka, Andréa Matsunaga, Peter Midford, Norman Morrison, Éamonn Ó. Tuama, Mark Schildhauer, Barry Smith, Brian J. Stucky, Andrea Thomer, John Wieczorek, Jamie Whitacre, John Wooley. 2014. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE* **9**:3, e89606. [Crossref]
8. Jared Bischof, Travis Harrison, Tobias Paczian, Elizabeth Glass, Andreas Wilke, Folker Meyer. 2014. Metazen – metadata capture for metagenomes. *Standards in Genomic Sciences* **9**:1, 18. [Crossref]
9. Sebastian Jan Janowski, Barbara Kaltschmidt, Christian Kaltschmidt. Biological Network Modeling and Analysis 203-244. [Crossref]
10. J. Ison, M. Kalas, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, P. Rice. 2013. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* **29**:10, 1325-1332. [Crossref]
11. Renzo Kottmann. Integrated Database Resource for Marine Ecological Genomics 1-6. [Crossref]
12. Marcin Radom, Agnieszka Rybarczyk, Renzo Kottmann, Piotr Formanowicz, Marta Szachniuk, Frank Oliver Glöckner, Dietrich Rebholz-Schuhmann, Jacek Błażewicz. 2012. Poseidon: An information retrieval and extraction system for metagenomic marine science. *Ecological Informatics* **12**, 10-15. [Crossref]
13. H. Teeling, F. O. Glockner. 2012. Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. *Briefings in Bioinformatics* **13**:6, 728-742. [Crossref]
14. Ramiro Logares, Thomas H.A. Haverkamp, Surendra Kumar, Anders Lanzén, Alexander J. Nederbragt, Christopher Quince, Håvard Kauserud. 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. *Journal of Microbiological Methods* **91**:1, 106-113. [Crossref]
15. Neela Enke, Anne Thessen, Kerstin Bach, Jörg Bendix, Bernhard Seeger, Birgit Gemeinholzer. 2012. The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —. *Ecological Informatics* **11**, 25-33. [Crossref]
16. Jörg Bendix, Jens Nieschulze, William K. Michener. 2012. Data platforms in integrative biodiversity research. *Ecological Informatics* **11**, 1-4. [Crossref]
17. Robert J. Robbins, James Beach, Stan Blum, Peter Dawyndt, John Deck, Renzo Kottmann, Norman Morrison, Éamonn Ó Tuama, Inigo San Gil, David Vieglas, John Wieczorek, John Wooley. 2012. RCN4GSC Meeting Report: Initiating a Testbed for Managing Data at the Interface of Biodiversity and Genomics/Metagenomics, May 2011. *Standards in Genomic Sciences* **6**:3, 171-174. [Crossref]
18. Jack A. Gilbert, Yiming Bao, Hui Wang, Susanna-Assunta Sansone, Scott C. Edmunds, Norman Morrison, Folker Meyer, Lynn M. Schriml, Neil Davies, Peter Sterk, Jared Wilkening, George M. Garrity, Dawn Field, Robert Robbins, Daniel P. Smith, Ilene

Mizrachi, Corrie Moreau. 2012. Report of the 13th Genomic Standards Consortium Meeting, Shenzhen, China, March 4–7, 2012. *Standards in Genomic Sciences* **6**:2, 276-286. [Crossref]

19. Jesse R.R. Zaneveld, Laura Wegener Parfrey, Will Van Treuren, Catherine Lozupone, Jose C. Clemente, Dan Knights, Jesse Stombaugh, Justin Kuczynski, Rob Knight. 2011. Combined phylogenetic and genomic approaches for the high-throughput study of microbial habitat adaptation. *Trends in Microbiology* **19**:10, 472-482. [Crossref]
20. Pelin Yilmaz, Jack A Gilbert, Rob Knight, Linda Amaral-Zettler, Ilene Karsch-Mizrachi, Guy Cochrane, Yasukazu Nakamura, Susanna-Assunta Sansone, Frank Oliver Glöckner, Dawn Field. 2011. The genomic standards consortium: bringing standards to life for microbial ecology. *The ISME Journal* **5**:10, 1565-1567. [Crossref]
21. Wolfgang Hankeln, Norma Johanna Wendel, Jan Gerken, Jost Waldmann, Pier Luigi Buttigieg, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner. 2011. CDinFusion – Submission-Ready, On-Line Integration of Sequence and Contextual Data. *PLoS ONE* **6**:9, e24797. [Crossref]
22. Dawn Field, Linda Amaral-Zettler, Guy Cochrane, James R. Cole, Peter Dawyndt, George M. Garrity, Jack Gilbert, Frank Oliver Glöckner, Lynette Hirschman, Ilene Karsch-Mizrachi, Hans-Peter Klenk, Rob Knight, Renzo Kottmann, Nikos Kyrpides, Folker Meyer, Inigo San Gil, Susanna-Assunta Sansone, Lynn M. Schriml, Peter Sterk, Tatiana Tatusova, David W. Ussery, Owen White, John Wooley. 2011. The Genomic Standards Consortium. *PLoS Biology* **9**:6, e1001088. [Crossref]
23. Elizabeth M. Glass, Folker Meyer. The Metagenomics RAST Server: A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes 325-331. [Crossref]
24. Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison, Philippe Rocca-Serra, Peter Sterk, Manimozhiyan Arumugam, Mark Bailey, Laura Baumgartner, Bruce W Birren, Martin J Blaser, Vivien Bonazzi, Tim Booth, Peer Bork, Frederic D Bushman, Pier Luigi Buttigieg, Patrick S G Chain, Emily Charlson, Elizabeth K Costello, Heather Huot-Creasy, Peter Dawyndt, Todd DeSantis, Noah Fierer, Jed A Fuhrman, Rachel E Gallery, Dirk Gevers, Richard A Gibbs, Inigo San Gil, Antonio Gonzalez, Jeffrey I Gordon, Robert Guralnick, Wolfgang Hankeln, Sarah Highlander, Philip Hugenholtz, Janet Jansson, Andrew L Kau, Scott T Kelley, Jerry Kennedy, Dan Knights, Omry Koren, Justin Kuczynski, Nikos Kyrpides, Robert Larsen, Christian L Lauber, Teresa Legg, Ruth E Ley, Catherine A Lozupone, Wolfgang Ludwig, Donna Lyons, Eamonn Maguire, Barbara A Methé, Folker Meyer, Brian Muegge, Sara Nakielny, Karen E Nelson, Diana Nemergut, Josh D Neufeld, Lindsay K Newbold, Anna E Oliver, Norman R Pace, Giriprakash Palanisamy, Jörg Peplies, Joseph Petrosino, Lita Proctor, Elmar Pruesse, Christian Quast, Jeroen Raes, Sujeevan Ratnasingham, Jacques Ravel, David A Relman, Susanna Assunta-Sansone, Patrick D Schloss, Lynn Schriml, Rohini Sinha, Michelle I Smith, Erica Sodergren, Aymé Spor, Jesse Stombaugh, James M Tiedje, Doyle V Ward, George M Weinstock, Doug Wendel, Owen White, Andrew Whiteley, Andreas Wilke, Jennifer R Wortman, Tanya Yatsunenko, Frank Oliver Glöckner. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* **29**:5, 415-420. [Crossref]
25. Melissa Beth Duhaime, Renzo Kottmann, Dawn Field, Frank Oliver Glöckner. 2011. Enriching public descriptions of marine phages using the Genomic Standards Consortium MIGS standard. *Standards in Genomic Sciences* **4**:2, 271-285. [Crossref]
26. M. Kalas, P. Puntervoll, A. Joseph, E. Bartaseviciute, A. Topfer, P. Venkataraman, S. Pettifer, J. C. Bryne, J. Ison, C. Blanchet, K. Rapacki, I. Jonassen. 2010. BioXSD: the common data-exchange format for everyday bioinformatics web services. *Bioinformatics* **26**:18, i540-i546. [Crossref]
27. Frank Oliver Glöckner, Ian Joint. 2010. Marine microbial genomics in Europe: current status and perspectives. *Microbial Biotechnology* **3**:5, 523-530. [Crossref]
28. Bert Verslyppe, Renzo Kottmann, Wim De Smet, Bernard De Baets, Paul De Vos, Peter Dawyndt. 2010. Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. *Research in Microbiology* **161**:6, 439-445. [Crossref]
29. Willy A. Valdivia-Granda. 2010. Bioinformatics for Biodefense: Challenges and Opportunities. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* **8**:1, 69-77. [Abstract] [Full Text] [PDF] [PDF Plus]
30. John C. Wooley, Adam Godzik, Iddo Friedberg. 2010. A Primer on Metagenomics. *PLoS Computational Biology* **6**:2, e1000667. [Crossref]
31. R. Kottmann, I. Kostadinov, M. B. Duhaime, P. L. Buttigieg, P. Yilmaz, W. Hankeln, J. Waldmann, F. O. Glockner. 2010. Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Research* **38**:Database, D391-D395. [Crossref]
32. Jack A. Gilbert, Folker Meyer, Rob Knight, Dawn Field, Nikos Kyrpides, Pelin Yilmaz, John Wooley. 2010. Meeting report: GSC M5 roundtable at the 13th International Society for Microbial Ecology meeting in Seattle, WA, USA August 22-27, 2010. *Standards in Genomic Sciences* **3**:3, 235-239. [Crossref]

33. Tanja Davidsen, Ramana Madupu, Peter Sterk, Dawn Field, George Garrity, Jack Gilbert, Frank Oliver Glöckner, Lynette Hirschman, Eugene Kolker, Renzo Kottmann, Nikos Kyrpides, Folker Meyer, Norman Morrison, Lynn Schriml, Tatiana Tatusova, John Wooley. 2010. Meeting Report from the Genomic Standards Consortium (GSC) Workshop 9. *Standards in Genomic Sciences* **3**:3, 216-224. [Crossref]

34. Elizabeth Glass, Folker Meyer, Jack A Gilbert, Dawn Field, Sarah Hunter, Renzo Kottmann, Nikos Kyrpides, Susanna Sansone, Lynn Schriml, Peter Sterk, Owen White, John Wooley. 2010. Meeting Report from the Genomic Standards Consortium (GSC) Workshop 10. *Standards in Genomic Sciences* **3**:3, 225-231. [Crossref]

35. Dawn Field, Iddo Friedberg, Peter Sterk, Renzo Kottmann, Frank Oliver Glöckner, Lynette Hirschman, George M. Garrity, Guy Cochrane, John Wooley, Jack Gilbert. 2009. Meeting Report: Metagenomics, Metadata and Meta-analysis; (M3) Special Interest Group at ISMB 2009. *Standards in Genomic Sciences* **1**:3, 278-282. [Crossref]

36. Johannes Wagener, Ola Spjuth, Egon L Willighagen, Jarl ES Wikberg. 2009. XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services. *BMC Bioinformatics* **10**:1. . [Crossref]

37. Oranmiyan W. Nelson, Scott H. Harrison, George M. Garrity. 2009. Meeting report for SIGS1: First Conference of the Standards in Genomic Sciences eJournal. *Standards in Genomic Sciences* **1**:1, 72-76. [Crossref]

38. Dawn Field, Peter Sterk, Nikos Kyrpides, Renzo Kottmann, Frank Oliver Glöckner, Lynette Hirschman, George M. Garrity, John Wooley, Paul Gilna. 2009. Meeting Report from the Genomic Standards Consortium (GSC) Workshops 6 and 7. *Standards in Genomic Sciences* **1**:1, 68-71. [Crossref]

39. John C. Wooley, Dawn Field, Frank-Oliver Glöckner. 2009. Extending Standards for Genomics and Metagenomics Data: A Research Coordination Network for the Genomic Standards Consortium (RCN4GSC). *Standards in Genomic Sciences* **1**:1, 85-90. [Crossref]

40. Stephen A. Chervitz, Helen Parkinson, Jennifer M. Fostel, Helen C. Causton, Susanna-Assunta Sanson, Eric W. Deutsch, Dawn Field, Chris F. Taylor, Philippe Rocca-Serra, Joe White, Christian J. Stoeckert. Standards for Functional Genomics 293-329. [Crossref]

41. David E. Whitworth. 2008. Genomes and knowledge – a questionable relationship?. *Trends in Microbiology* **16**:11, 512-519. [Crossref]

42. David W. Ussery, Trudy M. Wassenaar, Stefano Borini. The Challenges of Programming: A Brief Introduction 69-91. [Crossref]