Sustainable DwC-MIxS Interoperability

TG Report

Keywords: microbiome, eDNA, biodiversity information standards, omics, metadata, harmonization, FAIR, MIxS, DarwinCore

Conveners

Raïssa Meyer, Pier Luigi Buttigieg

Members and affiliations

Ward Appeltans	https://orcid.org/0000-0002-3237-4547	Ocean Biodiversity Information System
Pier Luigi Buttigieg	https://orcid.org/0000-0002-4366-3088	Helmholtz Metadata Collaboration \\ GEOMAR Helmholtz Centre for Ocean Research
William D. Duncan	https://orcid.org/0000-0001-9625-1899	Lawrence Berkeley National Laboratory, Berkeley CA
Mariya Dimitrova	https://orcid.org/0000-0002-8083-6048	Pensoft Publishers & Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Yi Ming Gan	https://orcid.org/0000-0001-7087-2646	Royal Belgian Institute of Natural Sciences
Thomas Stjernegaard Jeppesen	https://orcid.org/0000-0003-1691-239X	Global Biodiversity Information Facility
Raïssa Meyer	https://orcid.org/0000-0002-2996-719X	Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research
Chris Mungall	https://orcid.org/0000-0002-6601-2165	Lawrence Berkeley National Laboratory, National Microbiome Data Collaborative
Pieter Provoost	https://orcid.org/0000-0002-4236-0384	Ocean Biodiversity Information System
Tim Robertson	https://orcid.org/0000-0001-6215-3617	Global Biodiversity Information Facility
Saara Suominen	https://orcid.org/0000-0001-9401-8460	Ocean Biodiversity Information System
Maxime Sweetlove	https://orcid.org/0000-0003-3770-3714	Royal Belgian Institute of Natural Sciences
Anton Van de Putte	https://orcid.org/0000-0003-1336-5554	Royal Belgian Institute of Natural Sciences

Ramona Walls	https://orcid.org/0000-0001-8815-0078	University of Arizona, Critical Path Institute
John Wieczorek	https://orcid.org/0000-0003-1144-0290	Darwin Core Maintenance Group, Biodiversity Information Standards

Aim

This Task Group aimed to produce an approach to sustainably align the Minimum Information about any (x) Sequence (MIxS) and the Darwin Core (DwC) (meta)data specifications to enhance more efficient and interoperable exchange across their user communities[1].

Executive summary

This Task Group (TG) was convened to consolidate previous work that aimed to align (meta)data standards in the omics and broader biodiversity communities. This TG brought together experts from the Biodiversity Information Standards (TDWG[2]) and the Genomic Standards Consortium (GSC[3]) - alongside key external stakeholders - to develop an approach to promote sustainable interoperability between the Minimum Information about any (x) Sequence (MIxS[4]; maintained by the GSC) and the Darwin Core (DwC[5]; maintained by TDWG) specifications. In addition to this approach, the TG generated an initial mapping of DwC[6] keys to MIxS[7] keys using the Simple Standard for Sharing Ontology Mappings (SSSOM[8]), including detailed notes on issues faced and opportunities for further work. Further, the TG developed an extension to DwC (MIxS-DwC extension) which includes MIxS core terms distinct from existing DwC terms (i.e. with no mappings). Together, the SSSOM mapping and the MIxS-DwC extension provide a translation layer between MIxS and DwC-compliant metadata records.

---

[1] Our original aim was to ensure that “data produced in either MIxS- or DwC-compliant form can be automatically brokered between user communities”. Upon reflection we have narrowed our aim (as the original one would additionally require the development of tools to do the transformations of any terms that don’t map exactly to make them compliant at the destination). This, however, can be the basis for another TG focusing on using the outputs of this one.

[2] https://www.tdwg.org

[3] https://gensc.org

[4] https://gensc.org/mixs/

[5] https://www.tdwg.org/standards/dwc/

[6] When referring to DwC, we are referring to the most current official version as of the writing of this document: http://rs.tdwg.org/dwc/doc/list/2021-03-29

[7] When referring to MIxS, we are referring the most current official version as of the writing of this document: MIxS version 5, http://press3.mcs.anl.gov/gensc/files/2020/02/mixs_v5.xlsx

[8] https://github.com/mapping-commons/SSSOM

A note on terminology

A note on terminology

MIxS and DwC both use terms (strings associated with a meaning) to identify elements of data structures. That is, terms (such as “elevation”) are used to identify the intended meaning of, for example, 1) the attributes/columns in tabular data or 2) keys in key-value pairs. Both specifications provide metadata about their terms, clarifying their intended meaning and the expected values that should be associated with them once they are cast in a data structure (i.e. values in table cells, or values in key-value pairs).

Typically, in both MIxS and DwC data exchanges between human agents, (meta)data is arranged in spreadsheets or tabular form. The terms are thus used as attribute names / column headers. When archived in the INDSC (MIxS) and/or GBIF/OBIS (DwC), terms are rendered as keys in key-value pairs.

Below, for precision, we default to the usage of “key” (e.g. “temperature”) and its associated “value” (e.g. “18”[9])[10].

Glossary

Term	Definition
Darwin Core (DwC)	A specification released by TDWG which includes a glossary of terms intended to facilitate the sharing of information about biological diversity by providing identifiers, labels, and definitions (Version 2021-03-29) [11]
Darwin Core Archive (DwC-A)	A dataset which 1) contains data about species occurrences, checklists, sampling events and/or material sample data and 2) makes use of Darwin Core terms to qualify fields. DwC-A records comprise a set of text (CSV) files with a simple descriptor record (i.e. meta.xml) to inform others how your files are organized. The format is defined in the Darwin Core Text Guidelines. It is the preferred format for publishing data to the GBIF and OBIS network.
Darwin Core Extension	A list of defined keys to be used in combination with / in addition to DwC keys to create a more complete metadata record for a given situation.[12]
Minimum Information	A collection of checklists released by the GSC to define both

[9] This example assumes that the corresponding unit of the value is defined in the metadata associated with the key. See recommendations for semantic and syntactic alignment.
[10] In the proceedings of this TG, it was noted that the loose usage of such terms referencing the linguistic artifacts (e.g. “terms”) and the more technical data structures (“key-value pairs”) can produce confusion during tasks that require semantic precision, including this mapping. Thus our clarification here.
[11] http://rs.tdwg.org/dwc/doc/list/2021-03-29
[12] https://rs.gbif.org/extension/

about any (x) Sequence (MIxS)	the minimal and extended metadata associated with any sequencing record (Version 5)[13].
MIxS core	A MIxS checklist providing minimal (and extended) sets of metadata keys directly related to the sequences.
MIxS environmental packages	A collection of MIxS checklists providing extended sets of metadata keys about different sampling environments, deemed important by the MIxS user community.
Simple Knowledge Organization System Reference (SKOS)	A common data model for sharing and linking knowledge organization systems via the Web. It provides a lightweight, intuitive language for developing and sharing new knowledge organization systems.
Simple Standard for Sharing Ontology Mappings (SSSOM)	A catalog of minimal and standard metadata elements for the dissemination of mappings between ontology terms.

Outcomes

Mapping

Note: The final form of MIxS IRIs and identifiers has not been established by the GSC. This TG sourced MIxS identifiers from the working document preceding the MIxS v6 release[14].

Following our mapping approach (Appendix I), we mapped 32 DwC keys to 12 MIxS keys. Our resulting SSSOM records are accessible through the GBWG DwC-MIxS GitHub repository[15]. As detailed below (see Recommendations for using the SSSOM mapping matrix and Approach: Mapping), we created three SSSOM records to disaggregate our results:

1. DwC-MIxS_mappingSemantic.tsv[16]: this record contains mappings based on the meanings of the terms associated with the DwC and MIxS keys.
2. DwC-MIxS_mappingSyntactic.tsv[17]: this record contains mappings based on the syntactic similarity of the DwC and MIxS keys.
3. DwC-MIxS_mappingSupport.tsv[18]: this record includes both the semantic in syntactic mappings, as well as the supporting information used to determine both.

[13] http://press3.mcs.anl.gov/gensc/files/2020/02/mixs_v5.xlsx
[14] https://github.com/tdwg/gbwg/issues/11
[15] https://github.com/tdwg/gbwg/tree/v1.0.0/dwc-mixs/mapping
[16] https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/mapping/DwC-MIxS_mappingSemantic.sssom.tsv
[17] https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/mapping/DwC-MIxS_mappingSyntactic.sssom.tsv
[18] https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/mapping/DwC-MIxS_mappingSupport.sssom.tsv

Memorandum of Understanding

To ensure that our mapping and approach are integrated into the procedures and workflows of both TDWG and the GSC, we drafted and circulated a Memorandum of Understanding (MoU; see Appendix 1) to the executive bodies of each organisation.

Once ratified, the MoU will incorporate processes sustaining and furthering interoperability between these specifications and organisations. It is in this way, we hope that the work of our TG can lay the foundation for ever-closer alignment, ultimately allowing precise machine-to-machine translation of metadata using GSC and TDWG specifications.

DwC extensions

MIxS-DwC extension

We created a DwC extension[19] including the MIxS core keys that do not have a counterpart in DwC, and thus were not included in the mapping (see above). Used in combination with the SSSOM record generated by our TG, the MIxS-DwC extension allows a complete encapsulation of MIxS core in a DwC Archive (modulo some semantic and syntactic mismatches, see Recommendations for Semantic and Syntactic Mapping)

Of the 96 keys contained in MIxS core, we included the 82 terms that were not mapped in the extension.

The TG's GitHub repository hosts both, the list of keys[20], as well as a list of excluded (mapped) keys[21]. For the keys included in the extension, we have developed a Darwin Core Archive (DwC-A) extension definition in XML[22], which provides the standard set of terms that are available, onto which one can map one's own CSV[23].

Following the terms of our MoU draft, this extension will be bilaterally endorsed by the GSC and TDWG to assure users that they are implementing an officially recognised recommendation. The manner in which this is declared (e.g. as a header in the DwC-A reference implementation) will be decided upon by the relevant bodies in the GSC and TDWG.

Variations of the MIxS-DwC extension

While the bilaterally endorsed GSC-TDWG extension provides stability, we recognise that the needs of the biodiversity community are more diverse and require more nimble forms of data exchange. In the creation of these more ad hoc extensions, the risk of creating siloed / bespoke data products (and thus reducing global interoperability) is often countered by the practicality of advancing with fewer overheads and at a more rapid pace than standards

---

[19] https://github.com/tdwg/gbwg/tree/v1.0.0/dwc-mixs/dwc
[20] https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/dwc/extension/mixs_darwin_core_extension.xml https://tdwg.github.io/gbwg/dwc-mixs/dwc/extension/mixs_darwin_core_extension.xml
[21] https://github.com/tdwg/gbwg/tree/v1.0.0/dwc-mixs/dwc#mixs-terms-excluded-from-the-extension
[22] https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/dwc/extension/mixs_darwin_core_extension.xml
[23] See the meta.xml file of the Korean Peninsula Flora as an example of how an XML file is used as part of the DwC-A: https://www.gbif.org/dataset/e09e1e1f-2460-4017-a964-e999abd2bf66

bodies can be expected to match. Here, without taking a position on the “better” route, we recognise the reality of this scenario.

To demonstrate how metadata fields relevant to sequence-based biodiversity data can relate to the core outputs of this TG, we include a variation of the MIxS-DwC extension - the DNA-derived data extension - developed by GBIF[24] as an example of the use (and customization) of the MIxS-DwC extension introduced above. Note again, that this DNA-derived data extension is not built on standards-body synchronisation.

The DNA-derived data extension includes all keys of the MIxS-DwC extension, but brings in additional keys necessary to satisfy the exchange needs of the GBIF/OBIS/Atlas of Living Australia (ALA) networks. The additional keys originate primarily from the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) recommendation and Global Genome Biodiversity Network (GGBN).

Additionally, the DNA-derived data extension also takes measures to optimise the formatting and machine-readability of keys from MIxS. This stems from the fact that some MIxS key-value pairs are not atomic, i.e. they include multiple values in the same field (e.g. the MIxS key “*pcr_primers*” requires the user to enter a value which comprises a string that represents both the forward and reverse primer sequence, separated by a semicolon). This value-level formatting creates a bespoke data structure which then requires custom software or code to parse, limiting interoperability with external systems. Thus, in the case of *pcr_primers*, the DNA derived data extension uses alternative keys, based on the MIxS key, which are associated with atomic values: *pcr_primer_forward* and *pcr_primer_reverse*. This allows for more efficient and unambiguous data ingestion into search indices, relational databases, or similar solutions with minimal processing.

We acknowledge that it is a balance for application profiles to both comply with community standard specifications, while also satisfying the needs of the systems using them. To include and represent the evolving needs of the community and applications in existing community standards, we encourage that requests for changes or new keys are directed directly to the GSC[25] or TDWG[26].

Recommendations

In the sub-sections below, we offer several recommendations based on the proceedings and outcomes of this TG. We see our TG’s diverse membership and perspectives as a strong model to follow in future work developing or interlinking community standard specifications used by many stakeholders. Through this, operational realities, technical soundness, and policy-level perspectives can be better integrated and built upon.

Recommendations for using the SSSOM mapping matrix

The Simple Standard for Sharing Ontology Mappings (SSSOM) offers a framework to represent ontology mappings in a precise way, with a structured way to include rich

[24] https://rs.gbif.org/extension/gbif/1.0/dna_derived_data_2021-07-05.xml
[25] https://github.com/GenomicsStandardsConsortium/mixs/issues
[26] https://github.com/tdwg/dwc/issues

provenance. For the work of this TG, we have implemented an SSSOM mapping between the DwC standard and the MIxS checklist.

SSSOM provides a minimal set of standard elements for the dissemination of mappings between terms. This helps to ensure a reliable interpretation of mappings and enables sharing and data integration between human and machine agents.

As described in the Recommendations on semantic and syntactic alignment, even closely related MIxS and DwC terms, may have semantic variance, and expect values with different syntax. To manage that variance, we propose extending the list of SSSOM metadata elements to include elements to capture the syntactic mapping (*syntax_predicate_id*, *syntax_comment*; see Approach: Mapping) in addition to the existing semantic mapping metadata elements.

During the process of mapping, it is very useful to include additional attributes / columns in the SSSOM matrix in which information, upon which the mapping is based, can be stored. We thus propose adding such columns during the process (e.g. definitions [*subject_definition*, *object_definition*], syntax requirements [*subject_valueSyntax*, *object_valueSyntax*]; see Approach: Mapping). Once the process is over, a leaner SSSOM product can be released omitting these supporting attributes.

For mapping keys from metadata standards to one another, this task group recommends:

1. Follow the SSSOM guidance[27].
2. Until official guidance is offered from the SSSOM team, apply the extension proposed below (see Approach: Mapping) to additionally capture the mapping of syntax requirements:
    a. using the SSSOM *predicate_id* and corresponding *comment* to capture the semantics, and the *syntax_predicate_id* and corresponding *syntax_comment* to capture the syntactic mapping of terms.
3. Communicate any needed extensions to the SSSOM team via their issue tracker[28].

Recommendations for many-to-one, many-to-many, one-to-many mappings

Due, in part, to the different approaches to atomization described above and below, many of the proposed mappings between MIxS and DwC keys required one-to-many or many-to-one. This usually occurred when one specification offered multiple similar alternative keys for a phenomenon (e.g. DwC offers five keys relevant to “depth” measurements, while MIxS only offers one).

Recognizing that many keys in DwC or MIxS have community- and development-specific legacies, we recommend:

1. A mapping between metadata standards should be all-encompassing, and may thus include many-to-one, many-to-many, or one-to-many mappings.
2. Implementers, which represent a community of practice, can add notes on what keys they think are the most sensible.

---

[27] https://github.com/mapping-commons/SSSOM/blob/master/SSSOM.md
[28] https://github.com/mapping-commons/SSSOM/issues

3. In the long term, the standards agencies should aim to reduce the complexity of keys, moving towards atomization, to support more one-to-one relationships, eventually supporting full convergence.

Recommendations for semantic and syntactic alignment

DwC and MIxS specifications both offer guidance on the syntax expected for each value in a given key-value pair, alongside general notes on the expected semantics. In DwC, a value's expected semantics[29] are captured in the *Definition* and *Notes* attributes the List of Darwin Core Terms[30], while the *Examples* attribute shows expected value syntax. MIxS offers similar semantic guidance in the *Definition* attribute, with syntax and similar conventions specified in the *Expected value*, *Value syntax*, *Preferred unit*, and *Examples* attributes of the MIxS checklist[31]. Since DwC and MIxS have been developed independently from one another, and complex/bespoke syntax are common to both specifications, there is considerable divergence in their conventions. These include:

- For measured values, MIxS expects the unit to be included as part of the value, while DwC does not (optional for verbatim fields[32]).
- For measured values, MIxS offers a *"preferred" unit* option, which - as the label implies - is not mandatory, while DwC clearly defines the expected unit of each value (except for verbatim fields).
- Some MIxS keys, such as lat_lon, expect values which capture two or more measured/derived values. DwC typically separates these measured/derived values across two or more keys (e.g. decimalLatitude and decimalLongitude).
- Also, several MIxS fields allow for a numeric value or a range, followed by a measurement unit (*size_frac*, *samp_size*, *temp*, *depth* etc.). Darwin Core generally opts for atomic values associated with its keys.

Incompatibilities, such as those above, create (meta)data silos between communities using one or the other specification. Mappings built upon these can (in general) only be semantically and syntactically loose, and implementers must create and maintain converters or automated translators between the two, severely limiting and likely causing errors propagation in machine-to-machine exchanges.

To secure improved semantic and syntactic alignment, this TG recommends the following:

1. The use of more explicit labels (terms), associated with less ambiguous definitions (many of which are more descriptive than definitional).
    a. Additionally, further cross-organization effort to align the semantics of their fields in successive releases, using their obsolescence/change processes as appropriate.
2. Examples or descriptions of what is within and outside of the semantic scope / range of each field.
3. For any non-verbatim fields, clear guidance on what syntax is expected in each field (e.g. how many terms, separated how, with or without which unit?).

---

[29] In both MIxS and DwC, multiple definitions suffer from ambiguity, circularity, or other semantic aberrations. An effort to improve these would also improve future mapping and (meta)data (re)use efforts.
[30] https://dwc.tdwg.org/list/
[31] https://gensc.org/mixs/
[32] verbatim fields are essential to collect specimen data from museums, etc.

4. Re-use of existing and established terms from more general standards organizations within each specification (e.g. using dc:license to capture licensing information within MIxS and DwC).
5. Alignment to official external standards (e.g. using ISO 8601 to capture the time and date of an event)[33].
6. Synchronisation between standards-bodies ahead of new releases for closer syntactic alignment.
7. Semantic stability and standard syntax so stable converters can be written.
8. Atomic key-value structures, such that no complex or bespoke data structure exists in each value. For example, splitting ranges into dedicated start and stop fields.
    a. With advancement towards RDF- or JSON-based representations, allowing lists to be rendered as repeated key-value pairs.
9. Removing units from values by, e.g. requiring a standard unit in the definition of each key.

Recommendations for the mapping of MIxS environmental package terms

In addition to MIxS core, MIxS contains numerous “environmental packages” which bundle keys which improve the contextualisation of sequences in a given sampling environment. These are especially relevant for associating specific chemical and physical environmental measurements with specimens collected from these environments. Examples include marine, soil, food, and host-associated packages. These packages were created as a means to keep the core set relatively small, while rapidly accounting for the needs of sub-domains. These keys, and specifications of expected values, however, have not been harmonised or otherwise made interoperable with information standards published and used in Earth and environmental sciences.

Thus, this task group created SSSOM mappings and harmonisation notes only for MIxS keys which directly pertained to sequences (MIxS core), rather than the specific environment they were obtained from.

Recognising that the standardisation domain/mandate of the GSC does not extend to standards of environmental parameters, this task group recommends that:

1. Any sustained reference implementation of a MIxS extension of DwC - endorsed by the GSC and TDWG - is limited to those MIxS keys which closely pertain to sequences (MIxS core), rather than the environments they originate from (MIxS environmental packages).
2. The GSC, as it begins to transition MIxS into RDF, should make efforts to map and eventually replace their environmental keys with equivalent, well-described keys from an information standards body working in the Earth and environment domain. We strongly advise that this is done as a joint activity with TDWG, to prevent decoupling and the need for downstream re-alignment.
3. Users wishing to use the MIxS environmental package keys in DwC Archives should use the MeasurementOrFact (MOF)[34] collection of keys (cast as an MoF class and associated properties, see Appendix 3 for technical clarification) in the DwC

[33] the rare occasion where DwC and MIxS semantically and syntactically matched exactly was due to external standards (ISO 8601)
[34] https://dwc.tdwg.org/terms/#measurementorfact

specification. In our analysis, we found it valid to include a qualified mapping to a MIxS key URI as a value associated with the DwC “measurementRemark” key. This - alongside the other MOF key-value pairs - would allow any key in the MIxS environmental packages (either directly measured [measurement] or asserted to be true [fact]) to be represented in DwC.[35]

   a. While we demonstrate how to link MixS environmental package keys to DwC’s MoF, we draw attention to the fact that the GSC’s mandate is not within the standardisation of Earth and environment metadata. Thus, where possible, users should attempt to use values from more Earth and environmental vocabularies, thesauri, ontologies, etc.
   b. Please see Appendix 3 for an example of the above, and note the measurementType

4. TDWG and the GSC, in partnership with one or more standards bodies in the Earth and environmental sciences (e.g., the Earth Science Information Partners), convene a task group (or extend and expand this TG with a new mandate) to provide recommendations on how to sustainably and FAIRly incorporate well-adopted and more formally standardised environmental parameters into both MIxS and DwC.

To our knowledge, there is no sustained attempt to secure interoperability between the competing standards (most of which are informal, ad hoc, or de facto, as are MIxS and DwC) in this space. Some organisations and efforts of interest are listed below.

- Parameter vocabularies
    - The British Oceanographic Data Centre (BODC)[36] Natural Environment Research Council (NERC) Vocabularies[37], e.g. BODC Parameter Usage Vocabulary[38]
- The Open Geospatial Consortium (OGC)[39]
- Climate and Forecasting Variables[40]

We note that, while this vacuum exists, implementers will create their own internal standards for expediency[41]. This does provide some basis for later alignment, but also creates overhead as more unaligned information standards are released, compete for users, and decouple information systems and communities. We therefore re-emphasise the need for both TDWG and the GSC to engage with information standards communities in the Earth and environment domain to integrate their specifications.

---

[35] Please see Appendix 3 for an example of this.
[36] https://www.bodc.ac.uk
[37] http://vocab.nerc.ac.uk
[38] http://vocab.nerc.ac.uk/collection/P01/current/
[39] https://www.ogc.org/
[40] https://www.w3.org/2005/Incubator/ssn/ssnx/cf/cf-property
[41] e.g. GBIF is building basic vocabularies in SKOS, based on the values they see in their system. The objective here is more to clean data than to build rigorous vocabularies. Such internal efforts would greatly benefit from having a consolidated, appropriately endorsed, and standardised specification of environmental terms to align to.

Recommendations for licensing information

Information on licensing is critical for data reusability (as declared in the FAIR Principles[42]). Such information is captured in DwC through the import of the Dublin Core key dcterms:license[43]; however, there is no equivalent key provided in the MIxS specification.

Recognising that the GSC does not currently intend to extend their core checklist to include a key for licensing information[44], this task group recommends that implementers extend MIxS records with the Dublin Core key http://purl.org/dc/terms/license to capture data reuse restrictions.

Conclusion and outlook

In concluding this document, we emphasise the importance of convening a diverse and multi-stakeholder TG. With representatives from established biodiversity data infrastructures, domain experts, data generators, and publishers, we - ab initio - bridged the conceptual to the application space. We leveraged this to 1) generate, and internally review, a fine-grained mapping in a standard format, 2) implement new extensions to DwC, and 3) develop recommendations on how to expand on and sustain these. We have also identified areas of concern, which are in need of further attention and follow-up TGs.

Despite the achievements above, the work of this TG falls short of making an automated conversion possible. For this to be achievable, both community standards require further semantic and syntactic alignment, both between one another and with external data-on-the-web standards and best practices. In general, avoiding bespoke value syntax and complex semantics associated with keys (e.g. by unpacking complex keys into a number of simpler ones) will help this effort.
As stated in our draft MoU, the sustainability of this TG's output must be ensured through aligned processes within the community standards bodies involved. As noted below, we recommend that the sustainability of this TG's outputs are further secured, and protected from ad-hoc changes, by creating a follow-up TG to develop a MIxS-driven vocabulary enhancement[45] based on the MIxS-DwC extension. All of this is working towards a state where, as soon as an updated specification is released, the possibility of automatic data translation between standards exists and is validated.
In the long term, as sequence based (meta)data becomes more central to biodiversity observing, we anticipate a full convergence of these standards. Simultaneously, tools to converge records built from these specifications into more machine-readable forms (e.g. RDF triples), would increase their value, scalability and portability.

We trust that the activities of this TG will inspire similar activities between other metadata standards in this space, to break down silos and open a path to a more collaborative and interoperable future.

---

[42] https://doi.org/10.1038/sdata.2016.18
[43] License information is additionally captured on the dataset level in a DwC-A in EML, however, this declaration may not carry through automatically to the record in the dataset.
[44] https://github.com/GenomicsStandardsConsortium/mixs/issues/111#issuecomment-790759090
[45] Similar to the Chronometric Age vocabulary enhancement https://tdwg.github.io/chrono/terms/

Appendix 1

Approach

Mapping

Simple Standard for Sharing Ontology Mappings (SSSOM) provides a list of minimal and standard metadata elements[46]. These are used in combination with standard predicate terms, such as the Simple Knowledge Organization System (SKOS) terms to provide mappings between terms in differing terminologies (or ontologies).

We performed a comprehensive mapping from DwC to MIxS, capturing differences in both semantics and syntax between corresponding keys using the format of the SSSOM.

The semantic mapping[47] was based on the minimal and standard set of metadata elements provided by SSSOM, in combination with the relevant SKOS predicates.

As the SSSOM standard set of metadata elements does not yet[48] include means to capture information about the syntactic alignment of terms[49], we expanded the list of metadata elements to additionally capture information on the syntactic alignment of mapped terms (see table 1). The additional metadata elements were added to our syntactic mapping document[50] in replacement of the semantic mapping metadata attributes.

Table 1: Metadata elements additionally added to the DwC-MIxS mapping document to capture the syntactic mapping between keys.

Element ID	Description	TSV Example	RDF example
syntax_predicate_id	The ID of the predicate or relation that relates the syntax of the subject and object of this match.	skos:relatedMatch	skos:relatedMatch
syntax_comment	Free text field containing either curator notes or text generated by tool providing additional informative information on the syntactic mapping.	The subject expects a verbatim input (so anything really), while the object expects a {float} {unit} entry.	The subject expects a verbatim input (so anything really), while the object expects a {float} {unit} entry.

[46] https://github.com/mapping-commons/SSSOM/blob/master/SSSOM.md#sssom-metadata-elements

[47] https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/mapping/DwC-MIxS_mappingSemantic.sssom.tsv

[48] Currently the SSSOM community is working to provide best practice for these situations; see https://github.com/tdwg/gbwg/issues/54, https://github.com/mapping-commons/SSSOM/issues/52, https://github.com/mapping-commons/SSSOM/issues/56.

[49] For example, one of the challenges with mapping different term lists is that frequently we see that one system bakes in a unit to the meaning of the term, and the other system has a corresponding term whose value is a compound of value plus unit.

[50] https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/mapping/DwC-MIxS_mappingSyntactic.sssom.tsv

To facilitate the mapping process during our working period, we additionally added further metadata elements to capture definitions and value syntax (see Table 2). This working document is also available through our GitHub repository[51]. This is a secondary output which might be of relevance for future TGs performing mappings between metadata standards.

Table 2: Metadata elements additionally added to the working document for the SSSOM mapping between DwC and MIxS keys. These metadata elements were additionally added to facilitate the mapping process by having all the information needed as part of one spreadsheet.

Element ID	Description	TSV Example	RDF example
subject_definition	The definition of the subject of this mapping.	The original description of the depth below the local surface.	The original description of the depth below the local surface.
subject_valueSyntax	The value syntax expected for the subject of this mapping.	verbatim	verbatim
syntax_predicate_id	The ID of the predicate or relation that relates the syntax of the subject and object of this match.	skos:relatedMatch	skos:relatedMatch
syntax_predicate_label	The label of the predicate/relation of the syntactic mapping.	related match to	related match to
object_definition	The definition of the object of this mapping.	Depth is defined as the vertical distance below local surface, e.g. For sediment or soil samples depth is measured from sediment or soil surface, respectively. Depth can be reported as an interval for subsurface samples	Depth is defined as the vertical distance below local surface, e.g. For sediment or soil samples depth is measured from sediment or soil surface, respectively. Depth can be reported as an interval for subsurface samples
object_valueSyntax	The value syntax expected for the object of this mapping.	{float} {unit}	{float} {unit}
syntax_comment	Free text field containing either curator notes or text generated by tool providing additional informative information on the syntactic mapping.	The subject expects a verbatim input (so anything really), while the object expects a {float} {unit} entry.	The subject expects a verbatim input (so anything really), while the object expects a {float} {unit} entry.

For each mapping, group consensus was reached through a combination of structured discussions in the GitHub issue tracker and online video-chat meetings. Mappings can be found in the TDWG/GBWG GitHub repository[52], with related discussions captured on the issue tracker[53].

---

51 https://github.com/tdwg/gbwg/blob/v1.0.0/dwc-mixs/mapping/DwC-MIxS_mappingSupport.sssom.tsv

52 https://github.com/tdwg/gbwg/tree/v1.0.0/dwc-mixs/mapping

53 https://github.com/tdwg/gbwg/issues?q=is%3Aissue+label%3A%22DwC-MIxS+TG%22+is%3Aclosed

The SSSOM compliance of the mapping products was validated by Chris Mungall and Harshad Hegde.

The results of the mapping process were to provide:

- qualifications, which explain if discrepancies in semantics or syntax are to be expected and suggest how these can be resolved.
- DwC and MIxS keys identified by IRIs as opposed to labels.
- semantic mappings between DwC and MIxS keys following the SSSOM specification, using SKOS predicates (e.g., SKOS:exactMatch).
- semantic predicates and comments on the semantic mapping in the SSSOM matrix.
- augmentation of the SSSOM matrix to also include information on the level of syntactic compatibility. For example, the DWC key *decimalLatitude* expects values in the key-value pair to be decimals, whereas the MIxS key does not.

Extension

Darwin Core Archives are generally built on a combination of a core CSV file and zero or more extension CSV files. The schemas of the core and extensions are defined by XML documents maintained in the GBIF GitHub repository for machine-readable resources (https://github.com/gbif/rs.gbif.org). Core files act as the primary focus of a data set (e.g., Occurrences of organisms in nature), while the extensions add information relevant for specific uses (e.g., the proposed MIxS extension). The MIxS extension contains the list of keys that are orthogonal (have no equivalent mappings) to keys in the Darwin Core standard. Being orthogonal and defined by GSC, the keys in the extension are identified by IRIs from a namespace (fully qualified namespace pending, will be available with the release of MIxS V6) distinct from that of Darwin Core (http://rs.tdwg.org/dwc/terms/).

This was achieved by 1) documenting the relevant MIxS terms in the XML format specified by GBIF[54] and 2) creating vocabulary definitions in the XML format specified by GBIF[55] that contain the thesauri for the terms that should be controlled.

Testing

To test technical interoperability and simulate the ingestion of MIxS compliant metadata into a Darwin Core based database environment (e.g. OBIS or GBIF), a marine 'omics dataset (Franco et al. 2017) was selected from the www.biodiversity.aq/POLA3R portal. This dataset was previously published to GBIF as metadata-only, and represents a typical use-case where the community composition of microbes was profiled by high-throughput amplicon sequencing of the 16S rRNA gene. This generates microbial occurrences of both known and unknown species that are exclusively based on environmental DNA sequences. These sequences are available under the Bioproject PRJNA335729 on the databases of the International Nucleotide Sequence Database Consortium. The sequence metadata was provided compliant to MIxS v5, and sequences along with corresponding taxonomic

[54] http://rs.gbif.org/schema/extension.xsd
[55] http://rs.gbif.org/schema/thesaurus.xsd

annotation were downloaded from MGnify[56] in BIOM and FASTA formats and converted to DwC occurrences using a script.
Similar tests were performed using data representing Pico- to Mesoplankton along the 2000 km Salinity Gradient of the Baltic Sea[57] and data from a study demonstrating how nets mounted on rooftops of cars (car nets) and DNA metabarcoding can be applied to sample flying insect richness and diversity across large spatial scales within a limited time period[58].

We were able to successfully ingest the data into GBIF's user agreement test environment (www.gbif-uat.org). These test cases show it is possible for 'omics data to be incorporated along human observation based occurrence datasets using data processing by MGnify. This advancement is especially relevant for microbial groups, some of which are only known from environmental DNA sequences. It opens up new opportunities to include the vast biodiversity of micro eukaryotes, Bacteria and Archaea in repositories that up to now have been dominated by plants and animals.

Additionally OBIS will be working on a first test case of the DNA-derived data extension utilising Autonomous Reef Monitoring Structures (ARMS) datasets (https://doi.org/10.3389/fmars.2020.572680), which will link occurrences derived from genetic samples, morphological identifications and photographic evidence to each sampling device. To facilitate addition of sequencing datasets to the database, OBIS is also developing a bioinformatics pipeline, which will output a dataset formatted to the DwC-A including the MIxS extension.

Community feedback

This task team will solicit feedback from the GSC steering group and TDWG executive committee upon finalisation of this report. The feedback will be collected and appended to this report for the consideration of future TGs.

Memorandum of Understanding (MoU) between TDWG and the GSC

Preamble

The Biodiversity Information Standards (TDWG) group and the Genomic Standards Consortium (GSC) have emerged as de facto (meta)data standards authorities in the biodiversity domain. The former’s scope spans biodiversity data at large, while the latter focuses on genomic, and then multi-omic, data and metadata such as lab protocols or chemical/physical measurements. Their activities, technologies, and management structures have been largely parallel, with some notable exceptions catalyzed through joint interest groups such as the Genomic Biodiversity Working Group (GBWG).

The overlap of TDWG and the GSC in multi-omic biodiversity data is an opportunity to begin sustainable convergence of the (meta)data standards these organizations maintain. Most notably among these, are the Darwin Core (DwC) and the Minimal Information about any (x) Sequence (MIxS) specifications. This memorandum builds on the output of a GBWG task

[56] https://www.ebi.ac.uk/metagenomics/
[57] https://doi.org/10.3389/fmicb.2016.00679
[58] https://doi.org/10.1098/rsbl.2020.0833

group to propose a solution for sustained mapping and scalable interoperation of both DwC and MIxS. Its goal is to ensure that TDWG and the GSC create a lasting and continuous model to synchronize their standards, eventually promoting full bi-lateral integration.

Memorandum

Recognizing that both the Biodiversity Information Standards (TDWG) group and the Genomic Standards Consortium (GSC) have established well-adopted and community-driven (meta)data specifications for sequence-based biodiversity data;

Further recognizing that users of one standard specification should not have to invest additional effort in independently translating their (meta)data into another;

It is resolved that:

- The GSC and TDWG will maintain and endorse an authoritative and machine-readable mapping of the fields in their MIxS and DwC (meta)data standard specifications;
- These authoritative mappings (in SSSOM-compliant tab-separated value files) and other digital references will be maintained in the GBWG GitHub repository within the TDWG organization and with TDWG-issued IRIs;
- Further, both organizations will provide bilaterally endorsed reference implementations of how to use their counterpart's specification in their data structures (e.g. a DwC Archive incorporating fields mapped to MIxS in a DwC extension);
- Any necessary modification of identifiers (URNs, URLs, URIs, IRIs, etc) or other component of a standard issued by one organisation for the purposes of the other should be declared and the particulars agreed upon in documented appendices to this MoU;
- When one specification is updated, the TDWG DwC Maintenance Group and the GSC Compliance and Interoperability Group (CIG) will hold joint sessions to update and validate any mappings and reference implementations to ensure clarity in the multi-omic biodiversity data community.

Additionally recognizing that unilateral innovation and research actions will propose and implement alternative mappings and extensions to sequence-based metadata specifications.

It is further resolved that:

- Only those modifications which have been reviewed and endorsed by mechanisms bi-laterally convened by TDWG and the GSC will be considered standardized;
- Innovation is still welcome, and both organisations will welcome input and inspiration from application-driven modifications of the base standard.

Signatories:

Representative of GBWG
Representative of TDWG Executive
Representative of the GSC Board

Ensuring sustainability

GitHub releases of new versions of either DwC or GSC shall trigger a notification to the maintainers of the mapping created by this task group, who will review the new release and update the mapping if needed. As both standards have a release approximately annually, we estimate that long-term maintenance should require approximately 10-30 combined person hours for mapping review per year, plus review by the TDWG DwC Maintenance Group and the GSC Compliance and Interoperability Group (CIG), each of which can be accomplished as part of one of their regular monthly meetings.

As part of the MOU, both GSC and TDWG have agreed to provide personnel to maintain this mapping in perpetuity and to provide ongoing development to automate the mapping process as possible.

- **DwC release process**: TDWG has an official process for the maintenance of standards embodied in the Vocabulary Maintenance Standard (http://www.tdwg.org/standards/642) and documented in the Vocabulary Maintenance Specification (https://github.com/tdwg/vocab/blob/master/vms/maintenance-specification.md). The Darwin Core Maintenance Group (https://www.tdwg.org/community/dwc/) is responsible for the maintenance and evolution of the standard, including extensions to it, of which MIxS would be one. Updates to the standard result in releases on GitHub (https://github.com/tdwg/dwc/releases), which are backed up on Zenodo. GBIF maintains a repository (https://github.com/gbif/rs.gbif.org/tree/master/extension) of the production versions of the Darwin Core Archive extension XML files which are available to be used to create Darwin Core Archives using the GBIF Integrated Publishing Toolkit (IPT, https://www.gbif.org/ipt).
- **GSC release process**: By July 2021, and premiering with the release of MIxS V6, the GSC will have a workflow in place on GitHub which automatically builds new versions of the standards from code, releases stable versions, and backs them up on Zenodo. The normal release cycle for MIxS is about 1 time per year, but with the new release technology, there may be minor releases during the year. The minor releases will always be backward compatible with their major releases and will only include the addition of new terms. Furthermore, new keys can be created for MIxS between releases and be approved as individual keys with a stable URI, but not be considered part of an official MIxS release. This allows the rapid minting of keys while still providing time for thorough review before changing official releases.

Appendix 2

Relation of interoperable standards to the future of data-driven publishing

Standards alignment can facilitate data and metadata exchange between infrastructures and during the academic publishing process. In fact, the DwC-A format has already been used to exchange biodiversity information across different aggregators using Scratchpads user communities. In addition, DwC Archives have been actively used as supplementary data files associated with research papers in journals such as Zookeys to enrich traditional publications with structured data and for liberating structured content from journals by Plazi. Reuse of standard-compliant biodiversity metadata within the publishing process has also been realised for the Ecological Metadata Language (EML) which can be imported directly into a data paper manuscript from various sources (GBIF, LTER, DataONE). More recently, a workflow for import of genomic metadata from European Nucleotide Archive (ENA), BioSamples and ArrayExpress, part of which is MIxS-compliant, in the narrative of an omics data paper manuscript has also been developed.

All of these advancements in semantic publishing and data exchange signify that the proposed DwC-MIxS mapping would further improve the interoperability between infrastructures and facilitate the reusability of omics and biodiversity metadata. Improved interoperability between standards could also drive scholarly publishers and database managers to implement workflows for standard-compliant data reuse.

Appendix 3

Using MIxS environmental package keys in DwC Archives

To illustrate the Recommendation for the mapping of MIxS environmental package terms, we have added an example for moving from a MIxS-compliant temperature measurement to a DwC-compliant temperature record (see Table 3 and the RDF serialization below).

The simple temperature key-value pair from MIxS offers information on the following DwC properties from the MoF class: dwc:MeasurementOrFact, dwc:measurementType, dwc:measurementValue, dwc:measurementUnit (see Table 3). However, the MIxS specification currently falls short from providing fields to capture the, not less relevant, information on the measurement method [dwc:measurementMethod], the person taking the measurement [dwc:measurementDeterminedBy], and the accuracy in the measurement [measurementAccuracy], or any other remarks [dwc:measurementRemarks]. While there are some MIxS fields to capture e.g. the method used for a certain procedure (e.g. MIXS:0000002 samp_collect_device, which expects the method or device employed for collecting the sample type), there are no such options for any of the environmental package fields.

Especially anticipating the broader use of Essential Ocean Variables (EOVs) from the Global Ocean Observing System (GOOS), we will need to find a coherent and consistent way of reporting not only on the value of a certain measurement, but also the method.[59] To accommodate this need identified by the community, sustainable solutions will need to be developed and implemented to capture the exact method used to obtain a measurement value also in MIxS-compliant records (e.g. through an extension of the MIxS specification, or by using a combination of MIxS and the DwC MoF extension, ...).

Table 3: Translation of MIxS environmental package key-value pairs into DwC MeasurementOrFact key-value pairs. Prefix expansions: 1) “mixs:” = “https://w3id.org/gensc/terms/” 2) “dwc:” = “http://rs.tdwg.org/dwc/terms/”

MIxS key-value pair		**DwC key-value pairs**	
mixs:MIXS:0000742	17 °C	dwc:MeasurementOrFact	Temperature of the sample at the time of sampling
		dwc:measurementType	temperature
		dwc:measurementValue	17
		dwc:measurementUnit	°C
		dwc:measurementRemark	The values of dwc:measurementValue and dwc:measurementUnit captured here are a near

[59] Within the GOOS EOV framework, methods identified by the GOOS Expert Panel to deliver superior results for any given EOVs, will be endorsed in the UNESCO Ocean Best Practices System (OBPS). To know whether a given measurement value can be used to report on an EOV, information on the method used will be essential, and will thus need to be included as part of the metadata.

			match to the expected value of MIXS:0000742, which combines the two in a string

Exemplar RDF to express MIxS environmental keys as DwC Measurement or Fact triples

Below, we have translated the information captured in Table 3 into RDF. We are using very expanded RDF to make sure that we are as unambiguous as possible. **Please note that this is a draft, and that the MIxS IRIs may still be subject to change, as they have not yet been officially released.**

```
PREFIX : <http://example.dwc-mixs.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX mixs: <https://w3id.org/gensc/terms/>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
PREFIX dwcA: <http://rs.tdwg.org/dwc/terms/attributes/>

# OUR_IRI is an IRI that dereferences to a value in a data table.
# MIxS is currently castings its keys as owl:ObjectProperties,
# see https://github.com/cmungall/mixs-source for up-to-date
# information.

mixs:MIXS:0000742 rdf:type     rdf:Property .
mixs:MIXS:0000742 rdfs:range   xsd:string   .

# the value of the MIxS property must be string
# because of unit splicing, see comments on atomicity above

OUR_IRI  mixs:MIXS:0000742  "17 °C" .

# We now define some DwC MoF classes and properties

dwc:MeasurementOrFact  rdf:type rdfs:Class   .
dwc:measurementValue   rdf:type rdf:Property .
dwc:measurementUnit    rdf:type rdf:Property .
dwc:measurementType    rdf:type rdf:Property .
dwc:measurementRemarks rdf:type rdf:Property .

# We use a property from DwC attributes to link these
```

```
# properties to the MoF class. See the following for
documentation:
#
https://github.com/tdwg/rdf/blob/master/Beginners4Vocabularies.md#
461-informational-properties
# In terms of RDF, this is not strictly necessary, but reflects
# the internal organisation of the DwC specification which may
# ease discovery of which properties to use when using a DwC
# class

dwc:measurementValue   dwcA:organizedInClass dwc:MeasurementOrFact
.
dwc:measurementUnit    dwcA:organizedInClass dwc:MeasurementOrFact
.
dwc:measurementType    dwcA:organizedInClass dwc:MeasurementOrFact
.
dwc:measurementRemarks dwcA:organizedInClass dwc:MeasurementOrFact
.


# We now map the MIxS key to the DwC key with a skos
# predicate, loosely.
# Because the syntax and content of the MIxS value is
# non-atomic and the MoF splits units and the quantities
# they qualify, we can't state stronger equivalence.
# Here, we state that the MIXS:0000742 key is narrower
# in conceptual scope than dwc:measurementType

dwc:measurementType skos:narrowMatch mixs:MIXS:0000742 .



# State that OUR_IRI is an instance of dwc:MeasurementOrFact
OUR_IRI rdf:type dwc:MeasurementOrFact .


# As the MoF class expects a description as a value, state what
the data is about using the MIxS definition

OUR_IRI    rdf:value  "Temperature of the sample at the time of
sampling"^^xsd:string .

# We add some remarks to explain the measures taken to
# cast the MIxS field to DwC

OUR_IRI dwc:measurementRemarks "The values of measurementValue and
measurementUnit captured here are a near match to the expected
```

```
value of https://w3id.org/gensc/terms/MIXS:0000742, which combines
the two in a string" .


# Now add values using our DwC MoF properties as predicates
# Note that we define the datatypes in line as these may
# vary across MIxS keys

# We use an IRI for temperature from PATO for machine-readability
OUR_IRI    dwc:measurementType
http://purl.obolibrary.org/obo/PATO_0000146  .

# If we had a more specific type of temperature
# measurement , e.g. sea surface temperature, we could
# (and should) use an IRI with more specific semantics
#
# OUR_IRI    dwc:measurementType
http://purl.obolibrary.org/obo/ENVO_04000002 .

# We now set the remaining properties using literals

OUR_IRI    dwc:measurementValue  17^^xsd:decimal  .
OUR_IRI    dwc:measurementUnit   "°C"^^xsd:string .
```

Appendix 4

Issues noted for future TGs

The following issues were noted during the proceedings of this TG, which require the convening of subsequent TGs with appropriate scope.

MIxS-driven vocabulary enhancement

The development of a MIxS-driven vocabulary enhancement to TDWG/DwC would provide additional protection from ad-hoc changes and improve the sustainability of this TG's outputs by bringing them into the official TDWG standards space. A TG aiming to create a such a vocabulary enhancement could follow the example of the Chronometric Age vocabulary enhancement[60].

The development of a MIxS-driven vocabulary enhancement to TDWG/DwC would entail the following steps:

1. Based on this TG's MIxS-DwC extension, create a csv file that contains the complete descriptions and definitions of the MIxS core keys that cannot be mapped to DwC and that are currently part of the MIxS-DwC extension; re-using the MIxS IRIs to identify the MIxS keys.
2. Use existing scripts to produce a quick reference guide from 1) and a term list document containing normative content.
3. Follow the Vocabulary Maintenance Standard specification for a public review of the vocabulary enhancement.
    a. As this is part of aligning two standards bodies - with the proposed terms being pre-defined and governed by a different standards body - the review process would be limited to the overall concept of the extension and usage comments and examples noted as part of DwC.
    b. Requested changes regarding the term definitions and descriptions could be suggested to the GSC directly via the MIxS issue tracker and could be considered for future MIxS versions.
    c. Precedent: Adoption of existing Dublin Core terms in DwC, for which the definitions are out of the TDWG jurisdiction, but for which the usage comments about how to use it with Darwin Core and the examples are open for change.
4. Incorporate review comments.
5. Move the vocabulary enhancement into operation.

Controlled vocabularies in DwC to promote improved consistency and DwC-MIxS alignment / Improved semantic control through term lists from a curated list of ontologies

The general benefits of using open, sustained, community-driven, and quality controlled vocabularies, thesauri, or ontologies aligned to the FAIR Principles are many. In the context

60

of this TG, doing so would greatly enhance the stability of semantic and syntactic mappings between keys and values, as well as conversions between them.

Previous efforts to incorporate FAIR terminological resources have been pursued by both communities, e.g. recommendations to use the Environment Ontology (ENVO) or ontologies interoperating with it in the mandatory elements of a MIxS-compliant record. In this case, interoperation potential is increasing as DwC considers similar content for the dwc:biome key (see below). There are other keys where agreeing upon a relatively small controlled vocabulary (either extant or yet-to-be-developed) for the values of a given key-value pair would be quite straightforward.

The following observations and considerations are offered to TG conveners who wish to take this issue forward:

1. As a result of the Darwin Core Public Review concluded on 2021-05-31, the recommendation was to commit the issue proposing a new key 'dwc:biome' and its dwciri: analog to a task group. Though there was general agreement on the utility of the new key for Darwin Core, there were several concerns raised about using a community ontology like ENVO directly. These included that the syntax used to add terms from ENVO (or other ontologies) to MIxS (e.g. *tropical moist broadleaf forest biome* [ENVO:01000228]) was not an approach which the DwC community currently uses, and concerns were raised about the chance of input errors arising. An alternative proposed by Steve Baskauf involves 1) The creation of a local, DwC controlled vocabulary with terms following patterns that have been adopted for other Darwin Core terms (e.g., dwc:establishmentMeans), 2) Linking these terms to OBO ontologies for the purpose of interoperability and definition.
2. The new key proposal for dwc:environmentalMaterial was not included in the recent Darwin Core Public Review as there was insufficient demand demonstrated via the DwC proposal process. With support from the MIxS stakeholders, this proposal could be promoted for inclusion in the TG described above for dwc:biome, as it would be of service in aligning the two specifications.
3. Though not included in the outputs of this task group, controlled vocabularies for a selection of mapped MIxS key-value pairs should also be created and socialised by both TDWG and the GSC.
   The recommendation from this group is to create these controlled vocabularies as a vocabulary enhancement set under a new task group. Part of the TG's mission will be to include a way for the vocabularies associated with each key to be accessible to both MIxS users (e.g. submitting to the INSDC) and DwC users (submitting to OBIS or GBIF)
   Potential technical and operational issues were identified in how these vocabularies should be encoded and distributed. For example:
   - In order to include the terms from controlled vocabularies in software such as the Integrated Publishing Toolkit (IPT), "dummy" extension to a key, IRIs are created, such as dc:URI='https://rs.gbif.org/vocab/dna/decontam_software/anvi_o'. To minimise confusion, it was decided to use the GBIF namespace for this.
   - In the above, "/contigs" is one of the values expected in a controlled vocabulary for https://w3id.org/gensc/terms/MIXS:0000005

    - However, the IRI with the dummy suffix is not maintained or endorsed by the GSC, risking decoupling, misleading references, and technical confusion
    - A TG working on this issue must also ensure that technical solutions do not prioritise technical convenience over coherent and unambiguous standard specifications.

  While these recommendations and standards are being developed the Task Group recommends that these “dummy” IRIs are kept internal and inline documentation clearly states expectations around their maintenance and resolvability.

Representing replicates and derived samples/specimens and the relationships between them

As noted in Issue 24 of this TG’s proceedings, both DwC and MIxS require advancement to reliably and clearly relate replicates (be they technical [i.e. generated to verify stable signals during downstream processing of a sample], or elements of a sampling/experimental design) to one other using metadata fields.

Participants noted that community portals or other users of both standard specifications, rather than the standards bodies themselves, are defining links between replicates and derived samples/specimens. The interoperability of these solutions is questionable. A subsequent TG should investigate how to reconcile and render these interoperable by further coordination between DwC and MIxS, and other specifications in the TDWG and GSC scope.
Further, participants noted that several working groups have explored potential models for this task, and we recommend the subsequent TG should engage them.

Resources:

- http://www.obofoundry.org/ontology/ro.html
- https://www.tdwg.org/community/interaction/
- https://www.tdwg.org/community/cd/
- Droege et al. (2016): The Global Genome Biodiversity Network (GGBN) Data Standard. Database baw125 doi: 10.1093/database/baw125
- DwC Term relationshipOfResource: https://github.com/tdwg/dwc/issues/194

Recommendations for richer data exchange formats beyond DwC-A

The widely-used Darwin Core Archive (DwC-A) format arranges data into a simple "star schema" containing core records, such as species occurrences which can be extended in a many-to-one manner, such as multiple images for the occurrence. With no ability to relate records across extensions, the only feasible arrangement is to use an Occurrence core, with an extension that holds the sequence metadata supporting the claim of species occurrence. This is the same conclusion that the GBIF *DNA-related data* task group documented[61].

[61] https://doi.org/10.35035/doc-vf1a-nr22

The use of Occurrence core, limits the ability to easily track sampling event data such as arranging a hierarchy of nested samples, and forces unnecessary repetition of data in the DwC-A. In practice, this poses limitations, such as:

1. Metabarcoding of sediment samples, sediment parameter MeasurementOrFact records (porosity, grain size, solutes, organic carbon) will need to be repeated for every occurrence as there are no events to link to
2. Sediment or plankton samples being processed with microscopy, but subsamples are used for metabarcoding. In this case there's no way to indicate parent child relationship between samples and subsamples.

The task group recommends that work is done to explore more expressive data exchange formats ensuring lossless exchange is easily possible.