MEtaData Format for Open Reef Data (MEDFORD)

Polina Shpilker[1][0000−0002−6761−7326], John Freeman[1][0000−0003−1047−9273], Hailey McKelvie[1][0000−0001−6265−5326], Jill Ashey[2][0000−0001−5499−9786], Jay-Miguel Fonticella[1][0000−0001−9910−1933], Hollie Putnam[2][0000−0003−2322−3269], Jane Greenberg[3][0000−0001−7819−5360], Lenore Cowen[1][0000−0001−6698−6413], Alva Couch[1][0000−0002−4169−1077], and Noah M. Daniels[4][0000−0002−9538−825X]

[1] Department of Computer Science, Tufts University, Medford MA, USA
[2] Department of Biological Sciences, University of Rhode Island, Kingston RI, USA
[3] Department of Information Science, Drexel University, Philadelphia, PA, USA
[4] Department of Computer Science and Statistics, University of Rhode Island, Kingston RI, USA noah_daniels@uri.edu

Abstract. Reproducibility of research is critical for science. Computational biology research presents a significant challenge, given the need to track critical details, such as software version or genome draft iteration. Metadata research infrastructures, while greatly improved, often assume a level of programming skills in their user community, or rely on expert curators to ensure that key information is not lost. This paper introduces MEDFORD, a new human-readable, easily-editable and templatable metadata language for scientists to collocate all the details relevant to their experiments. We provide an overview of the underlying design principles, language, and current and planned support infrastructure for parsing and translating MEDFORD into other metadata formats. MEDFORD 0.9 has been specifically designed for the coral research community, with initial metadata generated from RNA-Seq analyses of coral transcriptomes and coral photo collections. Notably, the format is generally applicable and useful for many types of scientific metadata generated by non-computer science experts.

Keywords: Metadata · Research accessibility · Coral Reef Data.

1 Introduction

Corals comprise thousands of different organisms, including the animal host and single celled dinoflagellate algae, bacteria, viruses, and fungi that coexist as a holobiont, or metaorganism [2]. Thus, corals are more like cities than individual animals, as they provide factories, housing, restaurants, nurseries, and more for an entire ecosystem. Research on coral reefs is ever more pressing, given their local and global contributions to marine biodiversity, coastal protection, and economics, plus their sensitivity to climate change [6,16]. Research in this area

requires integration of interdisciplinary data across multiple environments and a range of data types: 'omic data such as gene expression data generated using RNA-Seq (RNA transcript sequencing), image and time-lapse video, and physical and environmental measurements including light and water temperature, to name but a few. The coral research community has long been committed to sharing and open data formats, and both individual researchers and large funding agencies have invested heavily in making data available [4,9,11,17].

Effective data sharing for coral research, as in all data-intensive domains, requires metadata, which is essential for data organization, discovery, access, use, reuse, interoperability, and overall management [8]. The growing amount of digital data over the last several decades has resulted in a proliferation of metadata standards supporting these functions [1,12]. However, the mechanisms to create metadata have been focused primarily on ease of machine parsing and have tolerated schema that are cumbersome and difficult for humans to understand. If creating metadata in the appropriate format is difficult, or requires expert curators, then fewer scientists can comply with metadata standards, leading to loss of scientific data if neither discoverable nor reusable. Meanwhile, much scientific data in multiple countries now falls under mandated data sharing policies that include metadata requirements. Thus, there is a need for a format that streamlines the process of providing what is mandated by law and policy.

We propose to address this need by developing and implementing MEtaData Format for Open Reef Data (MEDFORD). The MEDFORD markup language is simultaneously human and machine writable and readable. MEDFORD is designed to work in conjunction with the BagIt [10] filesystem convention, enabling easily accessible and interoperable bundles of data and metadata.

MEDFORD is initially targeted at coral holobiont transcriptomics data and coral image collections, with the subsequent goal of supporting metadata for additional research fields. The urgent need for international collaboration around saving coral reefs, plus the sheer complexity of the types and modalities of data the coral scientific community generates (from omics data, to image data with geospatial and temporal components, to temperature and color measurements), make corals a good domain choice. This short paper provides the rationale for current work and introduces the MEDFORD (version 0.9) metadata scheme.

MEDFORD will enable interdisciplinary coral reef data to be discoverable, accessible, and interoperable. Further, we are currently building the back-end infrastructure to translate between MEDFORD and make it compatible with other databases and systems such as Resource Description Framework (RDF) [7], ultimately supporting the interopability and reusability in FAIR [15] as well.

2 MEDFORD Design Principles

MEDFORD's design principles are informed by the those underlying highly successful metadata standards, such as the Dublin Core [14], Ecological Metadata Language (EML) [5], and the Data Document Initiative (DDI) [13], while addressing additional requirements enabling ease of metadata creation and other

aspects. The design requirements for creating MEDFORD are, generally, to have a mechanism scientists can use at the point of data collection which is a simple, human-readable format with a simple syntax that does not require programming expertise. In addition, we wish for scientists to be able to create and reuse templates to minimize manual effort, use editing tools of their choice, and be able to produce output in a number of machine-readable formats including Resource Description Format (RDF), Extensible Markup Language (XML), and JavaScript Object notation (JSON), as well as database formats.

These requirements reflect the work of metadata experts [12]. Web-based metadata interfaces can become cumbersome when one is entering metadata for many similar data files or publications. The machine-readable formats XML, RDF, and JSON are difficult to understand and edit for scientists who are not programmers, and error messages for these formats are cryptic. Plain-text specification of metadata, such as the NSF's BCO-DMO resource [3] must then be manually translated into machine-readable form. Thus, the evidence is clear that there is a need for an intermediate format that is both machine-readable and human-readable and understandable by the scientists most qualified to specify the metadata correctly.

MEDFORD is aimed specifically at solving the problems of specifying interdisciplinary research metadata and is first applied to coral reef omics data. In general, however, the principles above apply to any scientific metadata specification problem, and the specific extensions identified here may be supplemented for other scientific disciplines. Thus MEDFORD can be used as a tool for metadata creation in any scientific discipline. These requirements are realized by MEDFORD by adding design elements that satisfy the above principles: first, a contextual grammar, devoid of parentheses or lexical scopes; second, a simple concept of a metadata definition statement: starting with an '`@`', a keyword, and a value; third, a simple concept of hierarchy, in which subparts of a clause start with the same keyword prefix; and fourth, a simple concept of user-extensible formatting, in which metadata details not covered by the main keywords can be added via notes.

All MEDFORD files are defined in reference to a BagIt [10] bag (where there is a special use case where the MEDFORD file refers only to external data by reference). The BagIt bag binds a set of files to the MEDFORD file, where these files can be any type: including source code, scientific papers, or raw data, each represented by a major tag. The provenance of that file is marked using a secondary major tag, where the tag can represent that the bag is the primary and authoritative source for the data or resource. Other secondary major tags describe the file as either a copy of an existing source, or simply a pointer to a URI where the resource can be obtained.

3 The MEDFORD File Format

MEDFORD's file format is plain text, with tags starting with the `@` character. Anything after an `@` character, until the next space in the file, is read as a tag.

There are two other protected symbols that have special meanings: `#` which is treated as a comment character: characters after a `#` on the same line are ignored and do not carry through to any destination format. Finally, the `$$` string (two dollar signs) is used to delimit a string to be parsed by LaTeXmath mode.

The following design principles are important in MEDFORD file syntax: MEDFORD files use the ASCII character set whenever possible. The characters `@`, `#`, and `$$` are reserved and protected (`@` only when it starts a string; `#` only when followed by a space). MEDFORD tags are referred to as `@`tags and always start with the `@` character. Particular `@`tags are given meanings and specific recommended or required rules. The MEDFORD parser passes any unfamiliar `@`tags and their associated text through verbatim. Here, we present an excerpt from a MEDFORD description for a coral paper.

```
# This is just some example fragments of the MEDFORD file we built to
# record the metadata associated with this publication.

@Paper_Primary Coral bleaching response is unaltered following
acclimatization to reefs with distinct environmental conditions
@Paper_Primary-DOI https://doi.org/10.1073/pnas.2025435118
@Paper_Primary-Journal PNAS

@Contributor Katie L. Barott
@Contributor-ORCID 0000-0001-7371-4870
@Contributor-Role First Author, Corresponding author
@Contributor-Contribution Wrote paper, designed & performed research, analyzed data
@Contributor-email kbarott@sas.upenn.edu

@Contributor Hollie M. Putnam
@Contributor-ORCID 0000-0003-2322-3269
@Contributor-Contribution designed research, analyzed data

@Data_Primary Reef Seawater Sample No. 1
@Data_Primary-Location Kane'ohe Bay lagoon
@Data_Primary-Coord 21.4343°N, 157.7991°W
@Data_Primary-ResidenceTime 30+ days
@Data_Primary-ShoreDistance 0.75km
@Data_Primary-SampleDepth 2m

@Data_Reference Paper supplement
@Data_Reference-DOI 10.5281/zenodo.4315627
@Data_Reference-Notes Raw data and R scripts used for statistical analysis and PCA
```

4 Reusability

MEDFORD supports several features aimed at reusability of metadata among projects or datasets. The first of these is tag extensibility; MEDFORD has a set of pre-defined tags, but also accepts user-defined tags with no declaration

needed. The MEDFORD parser will infer structure from the tag and any subtags. MEDFORD also supports templates; a MEDFORD file can be partially filled out, saved, and re-used. For instance, a lab may create a template containing common contributors and funding sources, which can be used for future contributions. MEDFORD recognizes an "invalid data" token, `[..]` that can ensure users complete all fields of a template, such as `@Image-Date [..]`. MEDFORD also supports macros, similar to a variable defined in Bash or a `#define` in C, a macro is a string name that is directly substituted with another, longer string. Finally, MEDFORD can act as a common source format for many possible destination formats; users may wish to submit their data to a database such as BCO-DMO [3]. Other formats can be added to the MEDFORD parser easily; tutorials will be available in the github repository. The ability of MEDFORD to act as an intermediary allows for a lab to write a single MEDFORD file to describe their research and export it to a multitude of different formats.

5 Availability

The MEDFORD parser is open-source and available under the MIT license at https://github.com/TuftsBCB/medford. The authors welcome suggestions for other output formats that may be beneficial for data storage submissions through the Issues tab on GitHub. Some example MEDFORD files are available at https://github.com/TuftsBCB/medford-examples.

6 Discussion and Future Work

This manuscript presents MEDFORD, a lightweight metadata format initially targeted at coral reef research data, intended to be easy for researchers without programming expertise to create and maintain. Initially supporting interoperability and reuse, MEDFORD aims to support all FAIR [15] principles. MEDFORD is currently at version 0.9; we intend to begin testing with users in the coral research community by the time this paper appears, and have already received input from potential users. We are working on finishing initial versions of the basic documentation and the parser.

Currently, MEDFORD relies on editing ASCII or UTF-8 text, but it is capable of extracting text content from Microsoft Word files. One possible critique of MEDFORD is the variety of possible tags. A rich template library can mitigate this, by providing examples that a user can simply fill in. A searchable template library portal (similar to LaTeX's CTAN) would enable users to find applicable templates as the template ecosystem grows.

A visual user interface for writing, editing, and viewing MEDFORD files is also in early stages of development. Another goal will be the development of a front-end tool that does identifier lookup from authoritative sources, such as ORCID, Grant ID, DOI, populating all redundant data fields.

A major future goal will be output of RDF and support for linked open data; EML [5] is another reasonable target specification. We hope to add the ability

to translate a MEDFORD file (and created bag, if applicable) to RDF, as well as the data-1 compliance this involves.

7 Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work is supported in part by funds from the National Science Foundation under grants NSF-OAC #1939263, #1939795 and #1940233.

References

1. A. Ball, J. Greenberg, K. Jeffery, and R. Koskela. RDA metadata standards directory working group. 2016.
2. Thomas CG Bosch and Margaret J McFall-Ngai. Metaorganisms as the new frontier. *Zoology*, 114(4):185–190, 2011.
3. C. L. Chandler et al. BCO-DMO: Stewardship of marine research data from proposal to preservation. *Amer. Geophysical U.*, 2016:OD24B–2457, 2016.
4. Simon D Donner, Gregory JM Rickbeil, and Scott F Heron. A new, high-resolution global mass coral bleaching database. *PLoS One*, 12(4):e0175490, 2017.
5. E. H. Fegraus, S. Andelman, M. B. Jones, and M. Schildhauer. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, 86(3):158–168, 2005.
6. T. P. Hughes et al. Coral reefs in the Anthropocene. *Nature*, 546:82–90, 2017.
7. Ora Lassila, Ralph R Swick, et al. Resource description framework (RDF) model and syntax specification. 1998.
8. J. Leipzig, D. Nüst, et al. The role of metadata in reproducible computational research. *CoRR*, abs/2006.08589, 2020.
9. Y. J. Liew, M. Aranda, and C. R. Voolstra. Reefgenomics.Org - a repository for marine genomics data. *Database*, 2016, 12 2016. baw152.
10. J Littman, L Madden, and B Vargas. The BagIt file packaging format (v0. 97) draft-kunze-bagit-07. txt. 2012.
11. J. S. Madin, M. O. Hoogenboom, S. R. Connolly, E. S. Darling, D. S. Falster, D. Huang, S. A. Keith, T. Mizerek, J. M. Pandolfi, H. M. Putnam, et al. A trait-based approach to advance coral reef science. *Trends in Ecology & Evolution*, 31(6):419–428, 2016.
12. J. Qin, A. Ball, and J. Greenberg. Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. In *International Conference on Dublin Core and Metadata Applications*, pages 62–71, 2012.
13. M. Vardigan. The DDI matures: 1997 to the present. *IASSIST Quarterly*, 37(1-4):45–45, 2014.
14. S. L. Weibel and T. Koch. The Dublin core metadata initiative. *D-lib magazine*, 6(12):1082–9873, 2000.
15. M. D. Wilkinson, M. Dumontier, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.
16. A. Woodhead et al. Coral reef ecosystem services in the Anthropocene. *Functional Ecology*, 33(6):1023–1034, 2019.
17. L. Yu, T. Li, L. Li, et al. SAGER: a database of Symbiodiniaceae and Algal Genomic Resource. *Database*, 2020, 07 2020. baaa051.