Open access • Journal Article • DOI:10.1080/19386389.2018.1443414

Consortial Geospatial Data Collection: Toward Standards and Processes for Shared GeoBlacklight Metadata — Source link

Andrew Battista, Karen Majewicz, Stephen Balogh, Darren Hardy

Institutions: New York University, University of Minnesota, Stanford University

Published on: 02 Oct 2017 - Journal of Library Metadata (Routledge)

Topics: Metadata, Geospatial metadata and Geospatial analysis

Journal of Library Metadata

ISSN: 1938-6389 (Print) 1937-5034 (Online) Journal homepage: http://www.tandfonline.com/loi/wjlm20

Consortial Geospatial Data Collection: Toward Standards and Processes for Shared GeoBlacklight Metadata

Andrew Battista, Karen Majewicz, Stephen Balogh & Darren Hardy

To cite this article: Andrew Battista, Karen Majewicz, Stephen Balogh & Darren Hardy (2018): Consortial Geospatial Data Collection: Toward Standards and Processes for Shared GeoBlacklight Metadata, Journal of Library Metadata, DOI: 10.1080/19386389.2018.1443414

To link to this article: https://doi.org/10.1080/19386389.2018.1443414

Published online: 13 Mar 2018.

JOURNAL OF LIBRARY METADATA
2018, VOL. 0, NO. 0, 1–18
https://doi.org/10.1080/19386389.2018.1443414

Consortial Geospatial Data Collection: Toward Standards and Processes for Shared GeoBlacklight Metadata

Andrew Battista[a], Karen Majewicz[b], Stephen Balogh[a], and Darren Hardy[c]

[a]New York University, New York, New York, USA; [b]University of Minnesota, Minneapolis, Minnesota, USA; [c]Stanford University, Stanford, California, USA

ABSTRACT

Consortial geospatial data communities, such as the OpenGeoPortal federation and the GeoBlacklight initiative, facilitate contextualized discovery and promote metadata sharing to disperse hosting and preservation responsibilities across institutions. However, the challenges of communal metadata are manifold; they include proliferating standards, varying levels of completeness, mutable technology infrastructures, and uneven availability of human labor. Drawing from literature on metadata quality control, we outline a procedure for "scoring" GeoBlacklight records to establish a Domain Specific Language for metadata best practices. We propose strategies for authorship and management conducive to functionally interoperable geospatial metadata, that is versioned and enhanceable by the collective.

KEYWORDS

geospatial metadata; discovery; GeoBlacklight; interoperability; best practices; quality control

Introduction

GeoBlacklight and OpenGeoPortal are world-class, open-source web applications that provide solutions for contextual geospatial data discovery, detailed layer previewing, and data downloads across multiple formats. While these platforms can be adopted for use in a variety of settings, within and beyond academic libraries, they are more than free-standing software solutions. Rather, they emerge from communities that provide robust technical and social frameworks, and which coordinate cross-institutional software development sprints, host annual conferences, and facilitate governance over spatial data storage and delivery practices (Florance, McGee, Barnett, & McDonald, 2015). These projects began with the realization that full-scale geospatial metadata, such as the International Standards Organization (2007) 191xx series (ISO) and the Federal Geographic Data Committee (1998) Content Standard for Digital Geospatial Metadata (FGDC), do not crosswalk well with the underlying platforms and search indices typically required for discovery catalogs (Hardy & Durante, 2014; Poore & Wolf, 2013). Furthermore, the communities behind both projects have acknowledged that the proliferating array of geospatial

CONTACT Andrew Battista ✉ ab6137@nyu.edu Elmer Holmes Bobst Library, 70 Washington Square South, New York, NY 10012, USA.

data formats and standards makes it difficult to collect, curate, and present data as durable digital library assets (Hardy & Durante, 2015; Hardy, Reed, & Sadler, 2016). Authoring and maintaining rich geospatial metadata is a formidable task; the ISO and FGDC standards contain over 300 element sets each. In contrast, GeoBlacklight is built around a schema of its own—comprised of seventeen Dublin Core elements and augmented by six domain-specific geospatial terms—that emphasizes discovery and interoperability (GeoBlacklight Schema, 2017).

To date, academic libraries, municipalities, and research centers have deployed OpenGeoPortal or GeoBlacklight instances to expose geospatial holdings at scale. Many of these institutions have also contributed metadata records to (OpenGeoMetadata, 2017), an organization of GitHub repositories that is used to share metadata in an open, standards-agnostic way, and which has become an essential piece of infrastructure for building cross-institutional catalogs. The goal of OpenGeoMetadata is to foster collaboration that allows individual libraries to increase the breadth of geospatial data discoverable within a single search interface. By publishing metadata to these repositories, records from a variety of peer-institutions can easily be brought into a single catalog. OpenGeoMetadata began in 2014 with eight metadata collections and has grown to encompass the holdings of at least 18 academic institutions. As of August 2017, OpenGeoMetadata contained 43,000 GeoBlacklight records, representing the work of Stanford University, Harvard University, Princeton University, New York University, and members of the Big Ten Academic Alliance. The majority of OpenGeoMetadata represents geospatial data objects held within the respective institution's repository, but the organization also includes records extracted from government open data portals, such as the Federal Government's dataset catalog, Data.gov.

Although the consortial geospatial metadata approach has gained traction, several factors indicate that the community should do more to shape standard practices concerning metadata quality and operability. Many records currently found on OpenGeoMetadata present numerous obstacles that prevent them from being directly indexed into a catalog. These issues include disparate metadata standards, varying file formats, improperly formatted or missing required elements, broken data access links, inconsistent use of controlled vocabularies, and sparse or absent content in the descriptive fields. Unfortunately, neither the GeoBlacklight nor the OpenGeoPortal schema has been released with a complete metadata application profile, with agreed-upon standards of completeness. Furthermore, the tools and workflows for metadata remediation offered by these collectives remain somewhat unresolved, as not all participants add new metadata regularly or provide updates when functional elements of GeoBlacklight metadata change. In early iterations, the OpenGeoPortal metadata working group established principles and best practices to generate metadata for discovery, but the group has yet to agree upon a formalized standard or to establish a protocol for vetting contributions (OpenGeoPortal Metadata, 2015).

In this article, we offer a prescriptive approach to assessing the completeness and functionality of GeoBlacklight metadata to promote parity of user experience across institutions. As the GeoBlacklight consortium grows, a consistent standard of metadata quality needs to be established so that institutions can benefit more fully from the records comprising the OpenGeoMetadata collection. We begin with an environmental scan of collaborative metadata projects and approaches to quality control. We have deployed the terms and concepts of existing projects to propose a rubric, or a set of metrics, used as a framework for automated evaluation that measures how GeoBlacklight records conform to emerging quality norms. After performing a manual assessment of nine existing records, we propose several steps that the geospatial data community might take that would establish a baseline quality standard for metadata records to be indexed into discovery portals. Finally, we propose expectations for contributed metadata to be enhanced by other members of the OpenGeoMetadata community. These suggestions intend to normalize GeoBlacklight metadata for stronger contextual preview and functional data access.

Metadata quality control for consortial metadata collections: A situated overview

Systematic quality assessment requires a framework for metadata to be evaluated, a process that has traditionally involved establishing criteria that represent characteristics of a record's content. Such frameworks are often a composite of quantitative and qualitative methods. In the domain of consortial metadata, there are a number of existing models to assess the quality and interoperability of records.

An early study that applied a metadata quality framework to digital information from multiple sources was Moen, Stewart, and McClure (1997). This study reviewed 23 assessment criteria that had been developed for bibliographic resources and identified those that would be applicable to metadata for networked resources found in the Government Information Locator Service (GILS). The authors distilled the criteria into the categories of Completeness, Accuracy, and Serviceability. The analysis process involved pulling 87 records from GILS and manually examining them. For Completeness and Accuracy, they were able to apply a simple count of how many elements and controlled vocabulary terms were utilized and tally the number of spelling, typographical, or formatting errors. Assessing a record's Serviceability required the researchers to use subjective analysis to determine if a record included a sufficiently descriptive Title or Abstract.

The development of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) necessitated further investigations into metadata interoperability and authoring practices, as records were increasingly being shared across collections. Bruce and Hillman (2004) responded to this growth by cultivating a metadata quality model for digital libraries that identifies seven categories of assessment:

- Completeness (how many elements are included?)
- Provenance (who created the metadata and how?)
- Accuracy (is the metadata structure valid and are the element values appropriate?)

- Conformance to expectations (is the metadata aligned with audience expectations?)
- Logical consistency and coherence (are the element values consistent with one another and the wider collection?)
- Timeliness (is the metadata out of date with the actual resource?)
- Accessibility (is the metadata readable and understandable?).

Stvilia, Gasser, Twidale, and Smith (2007) created a framework of assessing general information quality using example Dublin Core and encyclopedic records. This framework is split into three large sections: Intrinsic, Relational, and Reputational, with 22 total dimensions. The metrics for each of these dimensions are generated primarily by computing "counts," such as number of empty elements, number of broken links, or number of elements used.

Evaluative frameworks have several commonalities, which inform our approach to GeoBlacklight metadata remediation. First, as Park (2009) notes, most of the frameworks contain categories for Accuracy, Completeness, and Consistency, which are based on quantitative approaches. Second, these frameworks often reduce evaluative questions to a binary "yes" or "no" answer, or to a computed count whenever possible. For example, a title of a record may be declared either descriptive or not descriptive based on a raw character count. However, the reliance upon binary designations alone may not provide enough nuance to produce metrics that are meaningful or usable outside of a given collection. Thus, most frameworks integrate a significant degree of qualitative assessment. Bruce and Hillman (2004) acknowledge the difficulty of establishing concrete definitions of "quality" for metadata records, especially in domains where controlled vocabularies and traditional practices abound. Thus, they proposed a "tiered" model that considers both the "administrative wrapper" (i.e., will the schema validate in an index?) and a more complex process of human assessment, which includes interpretations of logical consistency and comprehensiveness. Subjectivity is inevitable, but frameworks, which consider the general intellectual worth of descriptive metadata, help to codify assessment into actionable categories.

Despite the need to reveal a full picture of metadata worth, the qualitative element of assessment has often been deemphasized, which is partially due to the advent of harvesting protocols, such as OAI-PMH. Such protocols enable researchers to aggregate large quantities of records programmatically, without curatorial controls. Bui and Park (2006) harvested over one million records from the National Science Digital Library and used spreadsheets to tally the elements that were most commonly utilized by contributors. Kapidakis (2012) performed a similar analysis on millions of metadata records in Europeana by collection but also counted controlled vocabulary usage and the length of descriptive fields.

Some metadata quality assessments have restricted analysis to certain domains. For example, Renteria-Agualimpia, Lopez-Pellicer, Lacasta, Zarazaga-Soria, and Muro-Medrano (2016) focused solely on assessing geospatial consistency. They analyzed metadata records for over ten thousand maps and atlases in the Library of Congress to test for inconsistencies in the collection and identified three patterns of error: syntactic (incorrect syntax for bounding box coordinates), geosemantic (a mismatch in a record between the place name keywords and location as indicated by the bounding box coordinates), and contextual (a discrepancy across multiple records between which place name keywords are used for matching bounding box coordinates.) Another example of a topic specific metadata quality assessment was the DPLA evaluation of its Rights field. Although a rights statement was required for submission to the DPLA, the content of the field was undefined. As a result, the International Rights Statements Working Group (2016) discovered that over 87,000 different rights statements were in use by the mid-2010s.

The outcomes of these metadata evaluations has largely led to refinements with existing metadata authoring practices, rather than to structural changes in authoring processes. Bruce and Hillman (2004) concluded that metadata quality assessments should result in improved metadata creation guidelines and recommended the practice of developing application profiles to document a community consensus for specific elements and values. Park (2009) suggested that detailed metadata guidelines were the best quality assurance and that automated metadata creation tools, such as Omeka, should be used as much as possible. Metadata quality assessments can also indicate when a new initiative is required. The discovery of the proliferation of rights statements in the DPLA led to the creation of the RightsStatements.org project and the establishment of 12 statements that can be used for records in both the DPLA and Europeana. This approach is also evident in the refinements over time for the Europeana Data Model (EDM). Europeana organized a Metadata Quality Task Force in 2013 (Dangerfield, 2015) to go beyond the baseline requirements of the EDM and identify what entails higher quality metadata and strategies for achieving it. By April 2015, the Task Force had manually checked 1,809 records for errors, first by simply reading through them in XML form, and then by loading them into customized validation tools. The Task Force concluded that data providers should be more selective about what they submit, that the EDM should be more thoroughly documented, and that metadata should be regularly spot checked by both the data providers and the aggregators. Metadata analysis was also discussed by Harper (2016), who illustrated how commercial visualization platforms and quantitative analysis can be used to optimize metadata for large-scale consortial collections. Harper's approach establishes that there are efficient ways to develop "metadata fingerprints," or observations gleaned from natural language processing and other quantitative methods that can improve the discoverability and searchability of records.

Bulk remediation strategies for fixing existing collections of metadata are mentioned less frequently in the literature. Hillman (2008) observed that most library-based metadata professionals focus solely on authoring metadata, and may not be prepared to investigate strategies for improving metadata that they did not create. Hillmann further notes that, despite this mindset, metadata aggregation projects will inevitably face quality problems that will require some degree of subsequent normalization. The DPLA, a massive aggregation project, has developed strategies for remediating metadata at the time of ingest (Matienzo & Rudersdorf, 2014). The remediation piece of their metadata transformation and enrichment pipeline involves several automated actions, including cleaning up semantic variations, such as capitalizations, reconciling terms in certain fields with controlled vocabularies, and geocoding by place name keywords. Stein, Applegate, and Robbins (2017) noted the "surprising dearth of literature on retroactive metadata analysis and remediation" (p. 647). Their study described workflows for cleaning messy metadata values for the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS) with OpenRefine.

Existing challenges for GeoBlacklight metadata

The challenges associated with geospatial metadata are numerous. For GeoBlacklight to behave consistently as a discovery interface, metadata elements must be formatted consistently, as there is a symbiotic relationship between software function and metadata construction. As Schaffner (2009) phrased it, the metadata *is* the interface. However, as individual institutions establish GeoBlacklight instances, they will make decisions about the look, feel, and functionality for their users. Furthermore, they will determine the way items are presented and how they fit within the larger landscape of library collections. Aspects of metadata construction that affect user experience were outlined in a multi-institution usability study of the Big Ten Academic Alliance Geospatial Data Project's GeoBlacklight (Blake, Majewicz, Tickner, & Lam, 2017). The authors noted that users can become confused and frustrated when confronted with inconsistent data accessibility within search results. They also noted that semantic inconsistencies, such as subjects with irregular capitalization, spellings, or terms, make it difficult for users to discover content via text searches or keyword browsing. The study further provided directions for what to emphasize in the metadata to improve discovery. It found that Description, Spatial Coverage, Subject, and Temporal Coverage are the most valuable fields for successful geospatial data discovery.

With these challenges in mind, we undertook a cross-institutional critical examination of the records currently found on OpenGeoMetadata. We identified several issues that cause problems with function, data discovery, interoperability, and user experience. First, many of the records were found to be nonfunctioning (in part or entirely) in the GeoBlacklight application. In extreme cases, improperly formatted records are unusable because they do not conform to the schema of a search index, either due to syntactical errors, or because of broken or missing technical values. Functional metadata that has missing extent coordinates, non-resolving web services URLs, or absent direct download links will still greatly inhibit user interaction. These deficient records may even prohibit access altogether, which creates "noise" in a catalog. For example, all links to access the data and metadata supplements are provided in the References field. This field must contain a nested object of multiple key-value pairs, which activates many features in the GeoBlacklight interface. Thus far, this has been the community standard for modeling many pieces of information (additional metadata, codebooks, web services) related to a record. Because the References value is a complex object, none of the information contained within it is indexed by Solr. Nor is it directly verifiable by a schema validator. If an institution does not populate these values within a GeoBlacklight metadata record, it will still “work” within the search index of the application; however, it will result in the unsatisfying discovery of impoverished metadata records that do not provide data access or supplemental documentation links. The fact that information modeled within References is not a part of the search index means it is difficult to exclude automatically (or to dampen the search relevance) records that are deficient in terms of its references.

Metadata scoring process

The geospatial data community would benefit from more specific guidelines about metadata quality. In an effort to elucidate the process of quality assessment for consortial use, we have developed a rubric to measure the completeness, accuracy, and functionality of GeoBlacklight metadata. This rubric provides an outline of trait-based metadata characteristics, with attention toward elements that affect the functionality and discoverability within the GeoBlacklight interface. The overarching goal of a rubric-based approach is to provide transparency about how the disparate parts of a record can affect composite function and quality. Such transparency allows the curator of another repository to determine whether or not to ingest records, individually or at the batch level, depending on local platform requirements. The approach also highlights areas in which institutions could prioritize efforts to collaborate and improve the completeness and consistency of a given set of records. By providing several examples of the rubrics we developed, we hope that others in the community can execute evaluative and remedial work on their own metadata records (see Appendix A).

We began by establishing six categories for analysis based upon the characteristics of GeoBlacklight’s current functionality. The first category, *Structural/Functional*, refers to the presence and accuracy of required fields that directly affect the function and operability of the record within GeoBlacklight. This category checks the existence of the essential technical fields: Title, Bounding Box, Identifier, Slug, Provenance, Rights, and Schema. If one of these elements is not present or has invalid content, the record will exhibit unpredictable behavior or simply fail to load in GeoBlacklight. The second category, *Data Access*, addresses the availability of web references and the presence of active download links for a “preservation copy” of the data source. Although this category seemingly indicates the available data formats and provenance alone, it can also be thought of a measure of whether or not the metadata represents the resource with fidelity. To cite a prominent example, the public spatial data infrastructure suffers from ephemeral access to government data, and records may become orphaned if the dataset disappears or its links change. This is an important consideration for institutions looking to provide a reliable catalog that is not hampered by broken data access links.

The third category, *Bibliographic*, refers to the general completeness of the elements that are informational rather than functional, and includes a subjective assessment of the quality of these fields' content. An evaluation of the most important discovery elements, Description, Spatial Coverage, Subject, and Temporal Coverage are emphasized, while elements that are useful for interpretation of the data, but not as frequently accessed for discovery, such as Modified Date and Format, would receive less scrutiny. Although not every element will be used in each record, the Bibliographic category is designed to encourage the use of as many elements as possible in order to make sure that the metadata is not only robust, but also interoperable. For example, a record may include dates and formats within the description field, but this information is not indexable for search facets unless it is included in the assigned elements. The fourth category, *Adherence to Authorities*, addresses the degree to which certain fields conform to the respective ontologies recommended by the GeoBlacklight schema. These elements include but are not limited to: Subject (Library of Congress Subject Headings), Publisher (Library of Congress Authorities), Spatial Coverage (GeoNames), and Language (RFC 5646). Agreeing upon controlled vocabularies greatly improves interoperability, since these values are grouped together into facets within the default GeoBlacklight application.

The fifth category, *Ancillary Information*, addresses the extent to which supplemental materials, such as codebooks, data dictionaries, and contextual information are paired with the data object and are formatted appropriately within the references metadata element. This is a significant consideration, because GeoBlacklight is a pared down, discovery schema. However, the baseline documentation for geospatial data includes additional information that does not fit directly in the schema, such as attribute table definitions, accuracy reports, data-collection methods, and spatial reference systems. These can be provided as external links to downloadable codebooks or a geospatial metadata standard file. The sixth category, *Rights/Usage*, addresses whether or not a clear, valid license and usage statement is made available in a discrete metadata element, or anywhere within the record, such as the Description. This category also checks if the record uses a standardized rights statement, such as those provided by Creative Commons.

Once the categories were established, we examined a small selection of existing GeoBlacklight records to help us calibrate an evaluative framework. We randomly selected one record from the nine institutions that had contributed a large collection of records to OpenGeoMetadata as of August 2017. This allowed us to examine the local practices of each institution synoptically and determine ideal metadata standards for the community as a whole. To view these records and all documents associated with our process, see Appendix B and the Open Science Framework Project cited in the references. We then manually examined each record and graded each category on the scale of 1–4 according to the criteria established. As per the recommendations of Blake, Majewicz, Tickner, and Lam (2017), we concluded that the categories of Structure/Function and Data Access were more vital to assess than others, since they have the most direct effect on whether or not the metadata record would be serviceable. Therefore, these critical categories were adjusted to be weighted double. For example, the "Chicago, IL and vicinity" record is only available as an XML file in the FGDC format in OpenGeoMetadata. Since it requires transformation to a GeoBlacklight JSON format, a process that is lossy when performed automatically via a script, this record scores poorly in the Structural/Functional category. Another low scoring record is the "National Wetlands Inventory, 2009." Although it is valid and reasonably well-documented bibliographically, it does not include a web service preview or a direct download link, which are data access points that have been deemed important by usability studies within the GeoBlacklight community.

Some elements of the scoring process can be evaluated objectively. For example, records that have no value in the References field received a "1," for scant or absent development in the Structural/Functional category. Other assessments were more subjective and required that we make a qualitative appraisal of the content of each layer. For example, "Arlington, VA 2003 Bridges" received a score of "3" in the Adherence to Authorities category because it does have a subject value from the Library of Congress Subject Headings. However, it scored an average of "1.7" for Bibliographic because its description merely reiterates the title and does not add any meaningful contextualizing information about the origins and use of the data. Further, the record lacks any values for place names or publisher. In terms of functionality, this sample record is strong, but it needs significant qualitative enhancement to make it fully discoverable within a collective index of GeoBlacklight records. The results showed that higher scoring records are generally those representing data objects that have been archived in an institutional repository, are made available for preview, and download through web services, such as Web Mapping Service (WMS) or Web Feature Service (WFS). High scoring records also include additional documentation in the form of codebooks or attribute tables, and they use the full metadata schema populated by authority values. Lower scoring records have lapses or vacant values in fields associated with data access. They often have an unstable hosting source and may be hampered by discovery limitations due to the omission of keywords, publishers, or temporal information.

Toward a Domain Specific Language

The rubric serves as a conceptual basis for a potential Domain Specific Language (DSL), or an expression of rules that allow for verification, optimization, and transformation of metadata records at scale (Mernik et al., 2005). Although a concept that emerges from computer science literature, DSLs are capable of expressing scoring functions and could have many benefits for the open source geospatial metadata community, particularly if the set of functions is implemented within the context of a software solution. Many aspects of the assessment that were performed manually could be automated by using a DSL, for example. If provided with a system that takes GeoBlacklight records and reports scores about various dimensions of metadata quality, curators, and metadata professionals could prioritize forms of remediation and determine which records may be better candidates than others for highlighting (or inclusion in a catalog at all). Curators could offer an overall better experience of discovering and using geospatial data, even if not every record is perfect. Furthermore, a DSL-based metric provides a useful measure which curators can use to ask more nuanced questions about collection metadata (i.e., are there correlations between "completeness" or "adherence to authorities" in particular fields and the presence of cartographic metadata, and access statistics?). Our hope is that by providing the basis for some quantification of error in metadata documents, we can help establish norms that were previously nonexistent or ill-defined. This process may prove crucial for any attempt to build a distributed and massively cross-institutional discovery environment.

To demonstrate this potential, we drew from the rubric to devise an example implementation that retains the overall categories but checks for specific conditions within individual elements and assigns each a weighted score (see Appendix C). To accommodate the increased number of specific conditions, the total possible score is expressed as a percentage. This approach allows for more quantitative assessments and can indicate a more precise score for elements that are deemed important. However, this approach still requires some qualitative assessment, especially of the descriptive fields—namely, Title, Description, and Subjects—which would need to be manually scanned to determine if they were accurate to the dataset and if they were sufficiently formed.

When mechanisms for scoring and improving metadata records are clearly defined, and when geospatial data collectors have the ability to quantify the impact of standardizing terms within a metadata element, cross-institutional collaboration on metadata creation becomes a much more tenable endeavor. For instance, in 2017, Harvard University migrated over 5000 of its records to a new web services solution, thus "breaking" existing versions of their geospatial metadata. The Harvard case presents a clear opportunity to capitalize on some of the techniques we are proposing: community-developed DSLs for assessing completeness or validity of fields could be used to detect the problematic technical metadata that is within OpenGeoMetadata. Because the records exist in a public Git repository, which is a version control interface that provides a mechanism for collaborative work and allows for history tracking of file updates, any institution with an interest in doing so could make necessary technical changes and contribute back a "corrected" version of the metadata corpus, which would then be available for review, use, and propagation into shared catalogs. This could save significant duplication of labor on the part of the many institutions that make use of Harvard's geospatial data holdings, all of whom otherwise would have to manually implement identical corrections in absence of an authoritative modification from Harvard.

Possibilities for further action

After scoring the sample metadata records and postulating strategies for selecting higher quality records from OpenGeoMetadata, we propose the following suggestions for community adoption. First, we recommend that OpenGeoMetadata establish a culture in which anyone who participates in the collective is encouraged to enhance contributed records. This has implications that are technical and sociocultural. Following on the model of software development that occurs openly, and popularized by services like GitHub, metadata maintenance might rightly be seen as a process capable of being performed transparently, and where contributions from institutional collaborators are welcomed as routine occurrence. Open data in this sense would not mean merely publishing the results of a metadata unit or curator's work to a repository where it is publically visible to all; rather, it would mean that community members should be capable of actually making suggestions, proposing modifications, and contributing enhancements to the record data itself, because a mechanism to do that exists. The use of Git, specifically, as a method for tracking metadata change within the OpenGeoMetadata community exemplifies a tangible step in this direction. Many successful cooperative metadata efforts operate with the expectation of communal contribution already. For instance, the DPLA views metadata enrichment as a precondition for participation within their program, and this approach also undergirds the assumptions of the Europeana Data Model ("Europeana Data Model Primer," 2013; "Metadata Application Profile," 2017). Performing steps like harmonizing subjects and place names makes the list of subjects clearer and more consistent, which will become increasingly important as geospatial data collections grow and encompass more institutions. Furthermore, such enrichments follow the recommendations of Blake, Majewicz, Tickner, and Lam (2017) and can contribute to cleaner discovery catalogs that minimize cognitive load. By remediating the metadata of other institutions, the community can achieve a truly interoperable metadata record that is current and functional.

Second, and related, we suggest that the OpenGeoMetadata community adopt a policy that would establish the expectations for contributing and maintaining records. This policy could stipulate that submitting to the repository is a *de facto* invitation for other members of the collective to propose changes or enhancements that would align records with the community standards. Right now, the implication is that metadata enhancement is an altruistic process, taken on by those who want to serve the "greater good." The ideal is that a more enriched toolkit for harvesting metadata and omitting records that do not meet a baseline quality standard would provide a larger incentive for systematic contributions by members of the community. The set of proposed policies for OpenGeoMetadata could evolve to encompass the territory of the DPLA's collaborative model, in which the workflow for remediating or enhancing metadata is a dialogic process, a conversation between contributing partners, called Hubs, in which data is constantly refined until it meets an operable standard that can function within the DPLA's architecture (Matienzo & Rudersdorf, 2014).

Third, we suggest that the OpenGeoPortal Solr Schema and the GeoBlacklight schema become formally interoperable standards, without disparities that would hinder widespread metadata sharing between the two platforms. Both schemas exist as mechanisms for making GIS layers discoverable in Solr, and accordingly there is significant overlap in the type of information being represented across the schemata. Unfortunately, there has been little effort spent in trying to make these formats interoperable, despite the potential for building cross-institutional catalogs with significantly larger collections. Clearly documented procedures, and tooling which implements these procedures in user-friendly ways, need to exist in order to support the lossless transformation of metadata between GeoBlacklight and OpenGeoPortal instances.

Finally, we encourage the continued development of tools and further infrastructure for the management and conversion of discovery metadata. Our proposed metadata rubric can serve as a fundamental reference for the development of these processes. The creation of better metadata tooling could significantly streamline the OpenGeoMetadata-to-discovery catalog pipeline by providing a mechanism for performing normalizations, substitutions, filtering records, and ingesting. Tooling should also address the need for automated validation procedures; for instance, it should be possible to determine not only schema validity of a record, but also the degree to which a record adheres to community norms concerning ontologies and controlled vocabularies. We take inspiration from UC Santa Barbara's largely automated Git workflow for metadata creation and modification, where alterations are validated in the context of a continuous integration (CI) suite on GitHub (Dunn, Critchlow, & Rissmeyer, 2017).

Conclusion

The volume of available geospatial data, the range of approaches taken to create metadata for discovery, and the complex constellation of technologies needed to facilitate data access collectively contribute to the challenge of meaningful geospatial data discovery. The metadata scoring process we propose ideally moves the GeoBlacklight and OpenGeoPortal communities toward formalizing a unified metadata application profile that establishes clear standards and remediation procedures. The goal is to create and maintain an infrastructure housing a body of metadata that is interoperable with a broad variety of discovery applications and environments, and to provide tools to lower barriers to performing the relevant work. We hope that our work will further concretize the efforts of the OpenGeoMetadata collective and enable more institutions to ingest geospatial metadata that fully capitalizes on the features of the GeoBlacklight discovery platform. As the GeoBlacklight community has grown to include regular web meetings and annual development sprints, we hope that the emergence of governance structure associated with the group will begin to play a larger role in the stewardship and management of metadata that members of the community submit to the collective.

References

Blake, M., Majewicz, K., Tickner, A., & Lam, J. (2017). Usability analysis of the Big Ten Academic Alliance Geoportal: Findings and recommendations for improvement of the user experience. *Code4Lib Journal*, (38). Retrieved from http://journal.code4lib.org/articles/12932/.

Bui, Y., & Park, J. R. (2006). An assessment of metadata quality: A case study of the National Science Digital Library metadata repository. Paper presented at the Annual

Conference of CAIS/Actes du congrès annuel de l'ACSI. Retrieved from http://www.cais-acsi.ca/ojs/index.php/cais/article/viewFile/524/168/.

Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.), *Metadata in practice*. Chicago: ALA Editions. (pp. 238–256).

Dangerfield, M. C. (2015). Report and Recommendations from the Task Force on Metadata Quality. Retrieved from https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/metadata-quality-report.pdf.

Dunn, A., Critchlow, M., & Rissmeyer, C. (2017). "Using Git for Metadata – Part 1." Samvera Connect 2017 workshop. Retrieved from https://docs.google.com/presentation/d/1-4MBBZkSmYGseGsvh1fY2N9I9zNTCvpbBL2oE9HeeM4/edit?usp=sharing

Europeana Data Model Primer. (2013). Retrieved from http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf/.

Federal Geographic Data Committee. (1998). Content Standard for Digital Geospatial Metadata. (FGDC-STD-001-1998). Retrieved from https://www.fgdc.gov/standards/projects/metadata/base-metadata/v2_0698.pdf

Florance, P., McGee, M., Barnett, C., & McDonald, S. (2015). The OpenGeoportal Federation. *Journal of Map and Geography Libraries*, *11*(3), 376–394. doi:10.1080/15420353.2015.1054543

GeoBlacklight Schema. (2017). Retrieved from https://github.com/geoblacklight/geoblacklight/blob/master/schema/geoblacklight-schema.md/.

Hardy, D., & Durante, K. (2014). A metadata schema for geospatial resource discovery use cases. *Code4Lib Journal*, (25). Retrieved from http://journal.code4lib.org/articles/9710/.

Hardy, D., & Durante, K. (2015). Discovery, management, and preservation of geospatial data using Hydra. *Journal of Map and Geography Libraries*, *11*(2), 123–154. doi:10.1080/15420353.2015.1041630

Hardy, D., Reed, J., & Sadler, B. (2016). Geospatial resource discovery. In K. Varnum (Ed.), *Exploring Discovery: The Front Door to your Library's Licensed and Digitized Content* (pp. 47–62). Chicago, IL: ALA Editions.

Harper, C. A. (2016). Metadata analytics, visualization, and optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA). *Code4Lib Journal*, (33). Retrieved from http://journal.code4lib.org/articles/11752/.

Hillmann, D. I. (2008). Metadata quality: From evaluation to augmentation. *Cataloging and Classification Quarterly*, *46*(1), 65–80. doi:10.1080/01639370802183008

International Rights Statements Working Group. (2016). Recommendations for Standardized International Rights Statements [White Paper]. Retrieved from http://rightsstatements.org/files/180108recommendations_for_standardized_international_rights_statements_v1.2.1.pdf.

International Standards Organization (2007). ISO/TS 19139:2007 Geographic information – Metadata – XML schema implementation. Retrieved from https://www.iso.org/standard/32557.html

Kapidakis, S. (2012). Comparing metadata quality in the Europeana context. In *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments*. (p. 25). New York, NY: ACM. doi:10.1145/2413097.2413129.

Matienzo, M. A., & Rudersdorf, A. (2014). The Digital Public Library of America ingestion ecosystem: Lessons learned after one year of large-scale collaborative metadata aggregation. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, *16*. arXiv:1408.1713.

Mernik, M., Heering, J., & Sloane, A. M. (2005). When and how to develop domain-specific languages. *ACM Computing Surveys*, *37*(4), 316–344. Retrieved from. doi:10.1145/1118890.1118892

Metadata Application Profile. (2017). Digital Public Library of America. Retrieved from https://dp.la/info/developers/map/.

Moen, W. E., Stewart, E. L., & McClure, C. R. (1997). The Role of Content Analysis in Evaluating Metadata for the US Government Information Locator Service (GILS): Results from an exploratory study. Retrieved from http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm/.

OpenGeoMetadata. (2017). Retrieved from https://github.com/OpenGeoMetadata

OpenGeoPortal Metadata: Best Practices Guide. (2015). OpenGeoPortal. Retrieved from https://docs.google.com/document/d/1IeFjKMqXYhNwG6q8DiG7f3EthPpAo_KTOQM52obLH04/.

Park, J. R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, *47*(3-4), 213–228. doi:10.1080/01639370902737240

Poore, B. S., & Wolf, E. B. (2013). Metadata squared: Enhancing its usability for volunteered geographic information and the GeoWeb. In: D. Sui, S. Elwood, & M. Goodchild (Eds.) *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. (pp. 43–64). New York: Springer.

Renteria-Agualimpia, W., Lopez-Pellicer, F. J., Lacasta, J., Zarazaga-Soria, F. J., & Muro-Medrano, P. R. (2016). Improving the geospatial consistency of digital libraries metadata. *Journal of Information Science*, *42*(4), 507–523. doi:10.1177/0165551515597364

Schaffner, J. (2009). The metadata is the interface: Better description for better discovery of archives and special collections, synthesized from user studies. OCLC. Retrieved from https://www.oclc.org/content/dam/research/publications/library/2009/2009-06.pdf/.

Stein, A., Applegate, K. J., & Robbins, S. (2017). Achieving and Maintaining Metadata Quality: Toward a Sustainable Workflow for the IDEALS Institutional Repository. *Cataloging & Classification Quarterly*, *55*(7-8), 644–666. doi:10.1080/01639374.2017.1358786

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the Association for Information Science and Technology*, *58*(12), 1720–1733.

Appendix A. Metadata scoring rubric

All metadata records and scoring tables available in Open Science Foundation at https://osf.io/7auts/.

Category	Fields to Check	Substantial (4)	Moderate (3)	Minimal (2)	Scant-to-Absent (1)
Structural/ Functional (2)	Identifier Slug Title Bounding box Provenance Rights Schema version	All necessary components relating to function of GeoBlacklight exist in record, including valid bounding box coordinates, date value(s), geom type, and a stable unique identifier	Most necessary functional components of a record exist, even if there are some predictable lapses that are easily remediated	Record has missing or erroneously formatted values for one or two required elements, which could be programmatically remediated, such as Provenance	Record missing multiple required elements that can't be easily remediated or determined by a secondary processor (e.g., identifier, title, bounding box, or year);Record isn't in a valid .JSON format
Data Access (2)	References	Record includes valid preview web serv service hosted by provenance institution and downloadable files, including a "preservation copy"	Item is hosted by provenance institution and is downloadable	Item is hosted by an external agency with a history of stability, and includes a web service and/or downloadable data	Item is hosted by external agency without a history of stability or the metadata record does not include access to the data
Ancillary Information (1)	References	Layer is documented with fulsome resources (codebook, data dictionary, and information on context and methodology of data) and formatted in metadata (e.g., the described by field has a value); files of ancillary information are housed in a stable location (e.g., with preservation copy of data)	Record contains additional information of minimal completeness or lacking in context for the lineage, methodology, or other important ancillary information; some lapses in providing a reference to a web-hosted codebook	The record contains external links from which a user may be able to search for additional information, but such links are interpolated in description only	The record is only documented with information in the GeoBlacklight schema; no context is available for meaning of attribute table columns or methods behind the development of the data

(*Continued on next page*)

(Continued).

Category	Fields to Check	Substantial (4)	Moderate (3)	Minimal (2)	Scant-to-Absent (1)
Adherence to Authorities (1)	Spatial coverage Subject Type Format Creator Publisher	All GeoBlacklight fields that adhere to a specified ontology do so completely; no evident duplication or noise from typos, leading or trailing spaces, etc.	There are some lapses in adherence to an authority; multiple kinds of authorities within the same field coexist (e.g., LC subjects and ISO topics)	Fields contain words or strings from a folksonomy or other tagging system that are inconsistent but do have some informational value	Fields are completely blank or are populated with acronyms or stray terms that are very inconsistent, so as to clutter a search index
Rights/Usage (1)	Description Custom field	Record includes standardized rights statement in discrete metadata field that indicates fully the license and usage permissions of the data	Record includes a rights statement in a discrete metadata field, but it may be customized or otherwise nonstandard	There is some mention of rights or usage in the item description, but it is not parsed out into a discrete metadata field	Record does not include a rights statement, either in the description or any other metadata element
Bibliographic (1)	Title Description Subject Creator Publisher Spatial coverage Temporal coverage	Item is well documented within GeoBlacklight schema and includes very robust coverage of elements like Title, Description, Subjects, and Publisher as prescribed by best practices	Most fields are filled out; relatively rich inclusion of subjects and place names, even if there are some lapses in following best practices	Record has few fields filled out, and there are significant gaps in fields that would aid in the discovery and interpretation of the data	Record only includes bare minimum requirements and is sparsely populated

Appendix B. List of layers selected for calibration.

Record Title	Data Access	Structural/Functional	Bibliographic	Authorities	Ancillary	Rights/Usage	Total Weighted Score
Arlington, VA 2003 Bridges	3.3	3.7	1.7	3.0	2.3	1.0	22.0
Brooklyn bus routes	4.0	4.0	3.7	4.0	3.3	1.3	28.3
Canada VMap1	4.0	4.0	3.7	4.0	4.0	1.0	28.7
Chicago, IL and vicinity	2.0	1.0	2.0	2.0	2.0	1.7	13.7
Habitat, Offshore of San Francisco, CA, 2013	4.0	4.0	4.0	3.7	3.3	4.0	31.0
Income and Employment United States 1982	3.7	3.7	3.0	1.0	3.3	1.0	23.0
MUSA Composite, 2015	2.3	3.7	3.7	3.7	3.3	1.0	23.7
National Wetlands Inventory, 2009	1.0	2.7	2.7	3.7	2.0	1.0	16.7
Netherlands COROP Region Boundaries, 2005	4.0	4.0	3.3	3.0	3.0	1.0	26.3

All documents associated with the scoring process, including original metadata records, are available at https://osf.io/7auts/.

Appendix C. Example programmatic implementation of the rubric.

Category	Element Name	URI	Condition	If Condition Is Met, Assign This Score
Structural/Functional			Record is formatted as JSON	10
	Title	dc_title_s	Present	5
	Bounding Box	solr_geom	Present	5
	Identifier	dc_identifier_s	Present	5
	Slug	layer_slug_s	Present	5
	Provenance	dct_provenance_s	Present	5
	Rights	dc_rights_s	Present	5
				Category Total: 40
Data Access	References	dct_references_s	Hosted/archived by provenance	5
	References	dct_references_s	Web service	5
	References	dct_references_s	Direct download	5
	References	dct_references_s	Canonical landing page	5
				Category Total: 20
Bibliographic: Completeness	Description	dc_description_s	Present	2.5
	Spatial Coverage	dct_spatial_sm	Present	2
	Subject	dc_subject_sm	Present	1
	Temporal Coverage	dct_temporal_sm	Present	1
	Format	dc_format_s	Present	0.5
	Geometry Type	layer_geom_type_s	Present	0.5
	Date Issued	dct_issued_dt	Present	0.5
	Creator	dc_creator_sm	Present	0.5
	Publisher	dc_publisher_s	Present	0.5
	Modified Date	layer_modified_dt	Present	0.25
	Language	dc_language_s	Present	0.25
	Type	dc_type_s	Present	0.25
	Is Part Of	dct_isPartOf_sm	Present	0.25
				Category Total: 10
Bibliographic: Qualitative Assessment	Title	dc_title_s	Includes place name	2
	Title	dc_title_s	Includes temporal extent	2
	Description	dc_description_s	Sufficient/matching	2
	Place keywords	dct_spatial_sm	Sufficient/matching	2
	Subject keywords	dc_subject_sm	Sufficient/matching	2
				Category Total: 10
Adherence to Authorities	Spatial Coverage	dct_spatial_sm	GeoNames	2
	Subject	dc_subject_sm	LCSH and/or ISO	2
	Creator	dc_creator_sm	LOC Name Authority	2
	Publisher	dc_publisher_sm	LOC Name Authority	2
				Category Total: 8
Ancillary Information	References	dct_references_s	Codebook/attribute table	3
	References	dct_references_s	Standards metadata	3
	Description	dc_description_s	Link to other external documentation	2
				Category Total: 8
Rights/Usage	Any		Rights statement present	2
	Any		Rights statement standardized	1
	Custom Rights Field		Present	1
				Category Total: 4
				All Categories Total: 100