MAINSTREAMING METADATA INTO RESEARCH WORKFLOWS TO ADVANCE REPRODUCIBILITY AND OPEN GEOGRAPHIC INFORMATION SCIENCE

J. Holler[1,*] and P. Kedron[2]

[1]Department of Geography, Middlebury College - josephh@middlebury.edu
[2]School of Geographical Sciences and Urban Planning, Arizona State University - peter.kedron@asu.edu

Commission IV, WG IV/4

KEY WORDS: Metadata, Open Science, Reproducibility, Open Source GIS

ABSTRACT:

Reproducible open science with FAIR data sharing principles requires research to be disseminated with open data and standardised metadata. Researchers in the geographic sciences may benefit from authoring and maintaining metadata from the earliest phases of the research life cycle, rather than waiting until the data dissemination phase. Fully open and reproducible research should be conducted within a version-controlled executable research compendium with registered pre-analysis plans, and may also involve research proposals, data management plans, and protocols for research with human subjects. We review metadata standards and research documentation needs through each phase of the research process to distil a list of features for software to support a metadata-rich open research life cycle. The review is based on open science and reproducibility literature and on our own work developing a template research compendium for conducting reproduction and replication studies. We then review available open source geographic metadata software against these requirements, finding each software program to offer a partial solution. We conclude with a vision for software-supported metadata-rich open research practices intended to reduce redundancies in open research work while expanding transparency and reproducibility in geographic research.

1. INTRODUCTION

Can the scientific community expand knowledge production and the scope of inquiry, accelerate and improve scientific communication, and improve scientific rigour? These are the motivations of open science principles—to expand the availability and usability of scientific research publications, data, and methods (NASEM, 2018), thereby making scientific studies more reproducible and enabling new forms of inquiry based on synthesis and meta-analysis (NASEM, 2018, 2019). However, one key barrier to reproducible open science is a lack of standardised metadata documentation for research projects and associated data, code, and processing environments (Ibid.). The National Academies of Sciences, Engineering and Medicine (NASEM, 2018) *Open Science by Design* report therefore calls for adherence to the FAIR principles (Wilkinson et al., 2016) for sharing research data in a findable, accessible, interoperable, and reusable manner. According to the report, achieving FAIR principles for open science will require infrastructure—data repositories and metadata annotation software—and additional researcher labour to document metadata in the preservation phase of the research life cycle (NASEM, 2018). Leipzig et al. (2021) suggest that metadata may even help resolve the reproducibility crisis in computational research. Therefore, we argue that the crucial task of documenting and sharing data about data must be a continuous part of the research process.

We propose that researchers formally utilise and create metadata from the inception of a research project and maintain metadata throughout the full research life cycle. Open source geographic information systems software should support metadata-rich research life cycles from inception through dissemination and preservation. These changes in research practices and infrastructure should increase the reproducibility of geographic research and indirectly increase the pace and credibility of knowledge production.

Our proposal is based on a growing reproducibility literature in the social and geographical sciences and on our experience conducting reproductions and replications of geographic research and training geography students in reproducible research practices (Kedron et al., 2022). In particular, we have developed templates for pre-analysis plan registrations, reproduction and replication study reports, and reproducible research compendiums (Kedron and Holler, 2022b), and we have applied the templates to seven reproduction and replication studies conducted with teams of students and research assistants (Kedron et al., 2022). We suggest that rather than redundantly writing about their data through all phases of a research work cycle, researchers could formalise metadata at the project inception, use the metadata for much of the required documentation, and use software to maintain and track changes in metadata throughout the research process. Furthermore, we anticipate that a metadata-rich research life cycle would enhance the quality and transparency of metadata in open science by more thoroughly and consistently recording information about data provenance, license, access, and distribution. High quality metadata is also the next best solution for reproducibility in cases of restricted proprietary or confidential research data.

In the following section, we review the most important metadata standards for documenting geographic research and discuss the role of metadata in each phase of an open science research life cycle. We then describe our methods for selecting and reviewing open source software tools for annotating and maintaining metadata in support of metadata-rich research life cycles. We discuss the existing capabilities of open source software and conclude with suggested directions for future development in support of reproducible open science.

* Corresponding author

2. GEOGRAPHIC METADATA FOR REPRODUCIBLE OPEN SCIENCE

Open science aims to enhance the transparency, accessibility, and reproducibility of scientific research (NASEM, 2018). In geographic information science, this can be achieved with open public domain data, open source GIS software, public research workflows, and peer review inclusive of data and workflows (Singleton et al., 2016). Following the National Academies of Science, Engineering and Medicine (NASEM, 2019), a reproduction study aims to find the same results using the same data and methodology as a published study. Once a study is reproducible, it becomes possible to reanalyse the original study design by purposefully altering parameters or procedures using the same data (Christensen et al., 2019). Reanalysis studies provide insight into the sensitivity of original results by testing the internal validity of the finding and demonstrating how the finding compares to a set of findings produced by possible alternative analyses. A replication study aims to test the findings of a published study by collecting new data and following a similar methodology (NASEM, 2019). Whereas a study may be reproducible if original data is provided, replication will require complete metadata in order to create new data following the same procedures (Ostermann and Granell, 2017). Preliminary assessments of replicability and reproducibility in the geographic sciences have excluded many studies because of missing research components (Ostermann and Granell, 2017; Konkol et al., 2019). From the remaining sample of publications, the majority of volunteered geographic information publications were not reproducible (Ostermann and Granell, 2017) and the majority of spatial-temporal figures were not identically reproduced by provided code (Konkol et al., 2019).

Together, reproduction, reanalysis, and replication studies can offer a deep understanding of the original research, test its credibility, and enhance the self-corrective mechanisms of the scientific community (Christensen et al., 2019; NASEM, 2019). Over a series of replication studies, alternative hypotheses can be tested across geographic contexts to develop generalizable theories through the accumulation of evidence (Kedron and Holler, 2022a). However, the classic geographic research challenges of spatial heterogeneity, spatial dependence, and scale dependence imply that geographers will require distinctly geographic approaches and standards to achieve reproducibility (Kedron et al., 2021; Brunsdon and Comber, 2020) and evaluate evidence from replications (Kedron and Holler, 2022a). The process of reproducing existing work starts with using metadata about the research process to understand what was done, so that the research can be repeated. High quality metadata is also required to move on to the more complex processes of reanalysing published work, replicating it in a new context, or critically assessing a set of published studies, because metadata contains information crucial to understanding the logic used to shape the original research claim(s).

Open and reproducible science requires data to be findable, accessible, interoperable, and reusable (FAIR), and each of the four FAIR guiding principles require metadata (Wilkinson et al., 2016; NASEM, 2018). Metadata is information that describes data, providing essential context about the data so that other users can find, access, and use the data appropriately. Kim (1999) summarises seven essential categories of geographic metadata based on common standards: identification (title, keywords, authors), data quality, spatial data organisation (e.g. raster or vector model), spatial reference (coordinate system and datum), entity and attributes (data dictionary), distribution (contact, licenses, and fees), and metadata authorship.

Schuurman and Leszczynski (2006) and Comber et al. (2008) argue that metadata lack sufficient social and ontological context to evaluate data usability and facilitate semantic interoperability. Researchers need more information on data quality, sampling method, attribute name definitions, measurement specifications, classification systems, data models, collection rationales, and policy and legal context (Schuurman and Leszczynski, 2006). This may require ontological and perhaps even ethnographic study of the social context in which data was created (Schuurman, 2008). In the context of metadata for open science and reproducibility, the most recent international standards for data (see section 2.1) have answered many of these critiques with metadata classes for measurement, lineage, data quality, and usage. For additional social, semantic, and ontological information, bundling or linking data with publications and executable research compendiums should add substantial additional context for data reuse.

The important lesson here is that the data, metadata, and publication should be bundled and linked together in the form of a compendium with persistent identifiers. However, according to NASEM (2018, 137-8), "Making data 'interoperable' and 'reusable' can only be achieved if the data are annotated with comprehensive, standardised, high-quality metadata... [T]he absence of necessary metadata standards, appropriate ontologies, and easy-to-use annotation tools is a significant barrier." Although geography does have metadata standards (section 2.1), the open source geospatial software ecosystem may need improved tools for mainstreaming geographic metadata into the full research life cycle (section 2.2).

2.1 Metadata Standards for Geographic Research

In the previous section, we reviewed the critical importance of geographic metadata for open science and reproducibility. Despite this importance, scientific disciplines tend to lack sufficient standards for data sharing, interoperability, and documentation (NASEM, 2019). Fortunately, the geospatial industry has standard formats, protocols, and algorithms for storing, distributing, and analysing geographic data—all coordinated by the Open Geospatial Consortium (OGC: www.ogc.org). However, the OGC has left metadata standards to the spatial data infrastructures (SDIs) of individual states and regions, including the Federal Geographic Data Committee's (FGDC) Content Standard for Digital geographic metadata (CSDGM) in the United States, and the Infrastructure for Spatial Information in Europe (INSPIRE) (Kim, 1999; Bartha and Kocsis, 2011). Individual SDIs are increasingly following and harmonising with the International Standards Organization (ISO) series of geographic information metadata standards, especially the 19115 standard (ISO, 2014) for geographic information metadata and the 19139 standard (ISO, 2019) for encoding metadata in extensible markup language (XML). While the ISO standards are copyrighted and costly to purchase, researchers may access them through the many SDIs and open source geospatial software projects that have implemented them.

At the project level, research publications and compendiums can be documented with the Dublin Core™ standard metadata elements (DCMI and Hillmann, 2005). Elements of the ISO 19115 and Dublin Core standards that are relevant for open science and reproducibility are summarised in Table 1, with

similar concepts arranged on the same row. We have omitted information about language, character sets, and maintenance/accrual common to both standards.

ISO 19115	Dublin Core
Dataset name	Title
Abstract	Description
Purpose	Audience
Keywords	Subject Keywords
Topic Category	–
Unique Identifier	Identifier
Date	Date
Contact / Responsible parties	Author
Credit	Contributors
Citation	Creator, Publisher
Spatial resolution	–
Extent (spatial & temporal)	Coverage
Spatial representation	Type
Temporal resolution	–
Content information	–
Constraints	Rights
Data quality	–
Lineage	Provenance
Usage	–
Distribution and format	Type
Metadata about the data	–

Table 1. Summary of geographic metadata standards.

The ISO 19115 standards are specifically designed for geographic data types, whereas the Dublin Core is a simpler general standard suitable for archived objects and collections. In the ISO standard, lineage information is capable of including multiple source datasets and sequences of processing steps referencing specific software algorithms, whereas the Dublin Core lineage is more like a chain of custody of owners or stewards. The ISO spatial representation types are highly specialised, including raster, vector, topological, and three dimensional formats, whereas Dublin Core offers a single field for resource type. ISO content information may include metadata and descriptive statistics specific to raster data, remote sensing imagery, or vector features, attributes and attribute statistics. Citation information can include bibliographic information and persistent identifiers like the digital object identifier (DOI). Constraints may include many types, including copyright, patent, license, privacy, statutory, confidentiality, and more. Distribution metadata provides space for specific instructions on how to access the original data, including the format in which the data is provided.

In sum, the Dublin Core standards for the overall research project and the ISO 19115 standards for geographic data layers provide a structured foundation to support reproducibility and open science throughout the research life cycle.

2.2 Open Science Research Life Cycle

The 2018 NASEM report *Open Science by Design* envisions open science practices in all six phases of the research life cycle: provocation, ideation, knowledge generation, validation, dissemination, and preservation. However, the report singles out the preservation phase for metadata documentation. In the subsections below, we outline each of the life cycle phases and argue that metadata creation in each plays a key role in achieving the aims of open science.

2.2.1 Provocation In the provocation phase, researchers review literature and data to identify opportunities for novel contributions. As they gather and review prior studies and arguments, researchers would clearly benefit from a metadata-rich open science environment. Easily accessible and rich metadata reduce the time and effort needed to critically assess the logic and argumentation of existing studies. More formally, with open geographic data and metadata for published literature, researchers could design geographically-explicit bibliometric analyses and synthesis studies. For example, quantitative meta-analyses seeking to estimate effect sizes from studies conducted in different geographic contexts could use geographic metadata to introduce simple controls for effect variation across regions or the sensitivity of effect estimates to the scale of original research. This type of regional or scalar differentiation or adjustment is not currently possible in human-environment geography because the geographic metadata does not exist. Margulies et al. (2016) review 437 global change science case studies and find persistently ambiguous descriptions of geographic extents. Researchers would also gain more detailed insight into the data and methodology of the studies they review with human-readable forms of metadata for each component of the study, helping to clarify the ambiguity of communicating complex computational methods with narrative publications.

2.2.2 Ideation In the ideation phase, researchers investigate data; and then plan, prototype, and preregister research designs. Three different types of plans are required of ethical and open research in this phase: 1) protocols for research with human subjects for ethical review (DHEW, 1978), 2) research proposals and their associated data management plans (DMPs)(NSF, 2021), and 3) pre-registered analysis plans (Nosek et al., 2018). The primary purpose of reporting research with human subjects is to protect the privacy and rights of research subjects (DHEW, 1978). The purpose of DMPs is to explain data management protocols (NSF, 2021) but also increasingly to explain how data will be made available to the public according to open science principles (Gil et al., 2016). The purpose of pre-registering analysis plans is to enhance the transparency and replicability of scientific studies, encouraging researchers to objectively carry through with deductive research plans and report results (Nosek et al., 2018). The idea is to avoid unobserved selective inference (e.g., p-hacking), remove false positives from the literature, and increase (or perhaps recover) reliability and inferential power.

These plans are essentially narrative metadata—documentation about the research process and could be supported by project-level and data-level metadata. Each of the three types of plans require project-level identifying metadata: title, authors, personnel and contributors, abstract or summary, location or spatial extent, and temporal extent.

The plans also require metadata about each data layer. If secondary data is to be used, researchers are required to investigate their metadata and report on any use restrictions. If primary human subjects data is to be collected, researchers must specify the sampling methodology. Ethics review additionally requires specification of the recruitment protocol and survey instrument. Data collection methods and data variables should be described for all human and environmental data. Each of the plans also requires researchers to specify protocols for storing, archiving, and disseminating research data.

Plans diverge in some data-specific reporting requirements. The pre-analysis plan requires additional detail on planned data transformation and analysis methods—essentially looking forward to documenting provenance and lineage. Pre-analsyis registration is ideally accomplished without viewing data directly, and therefore relies heavily on fully specified metadata for any

secondary sources (Nosek et al., 2018). The human subjects review requires additional focus on treatment of personally identifiable information, confidentiality, and data security (DHEW, 1978).

In sum, the ideation phase requires researchers to study metadata for any secondary data sources they plan to use, and to specify metadata for any data they plan to create. In the current state of practice, this metadata documentation is required in narrative form in a variety of documents. We propose that formalising metadata documentation in this research phase could mitigate redundancies by first populating metadata with required information, and then reusing that information for each of the planning documents required in this phase. This change in research practice would also increase transparency if the metadata is stored in a research compendium using Git version tracking, so that any changes to intended data creation or dissemination protocols can be visualised across versions of the project.

2.2.3 Knowledge Generation In the knowledge generation phase, researchers use open source software tools to collect and analyse data in interoperable formats with sufficient documentation, metadata, and computational notebooks to enable future reuse and replication.

Ideally, researchers will organise their materials and methods for computational research in a structured executable research compendium (Singleton et al., 2016; Nüst and Pebesma, 2021) containing all of the narrative, data, code, software, and computer scripts required to compile the final publication starting with raw data. Computational notebooks like Jupyter notebooks or R Markdown are commonly used to interweave narrative with code in executable compendiums (ibid). It is recommended to store compendiums in version tracking systems like Git in order to preserve a full history of changes to the research project (Stodden et al., 2014), and to integrate compendiums into the full research workflow from knowledge generation to publishing (Kray et al., 2019). Compendiums should implement a routine structure for research components, including directories for procedures or code, documents, raw data, processed data, and results (Kedron and Holler, 2022b; Christensen et al., 2019; Marwick et al., 2018). In addition to this structure, the compendium components should be well-documented with metadata (Kedron and Holler, 2022b; Marwick et al., 2018).

During the knowledge generation phase, metadata records should be maintained and updated with complete provenance information on the origins of the data and a history of all data transformations (NASEM, 2019; Tullis and Kar, 2021). Provenance is essential for reproducibility (Kedron et al., 2021) and understanding the quality of data within the research and the quality and context of the research data if its to be reused elsewhere (Tullis and Kar, 2021; Schuurman and Leszczynski, 2006). The complexity of computational research in geography implies that provenance metadata is a necessary precursor for communication and reproduction of research methodologies, and therefore software to automate provenance records may improve reproducibility (Kedron et al., 2021). Adhering to this logic, Anselin et al. (2014) created a metadata system for spatial weights matrices in a form that both records human-readable provenance and machine-readable instructions for reproducing the analysis. As such complex computational research methods diverge from the original pre-registered plan, Git can track and visualise changes to metadata updated during the knowledge generation phase, lending transparency to intended changes and unintended deviations.

2.2.4 Validation In the validation phase, researchers analyse, visualise, interpret, and validate results while sharing preliminary findings in working papers and conferences. Surveys of publications presented in the AGILE (Nüst et al., 2018) and GIScience (Ostermann et al., 2021) conferences found the majority of papers irreproducible due to missing metadata, data, and procedures. At this phase, the overall project and any public project component can be registered and assigned a persistent link like the DOI through digital repositories like Open Science Framework (OSF) or figshare. Registration requires project-level metadata to enable archiving and searching. Although researchers may be reluctant or constrained to release complete data at this phase, metadata can be shared for project components that must remain private or embargoed.

2.2.5 Dissemination and Preservation In the dissemination phase, the research is peer reviewed, revised, and published, ideally with associated data and code. A version of the research compendium should be made available for the peer review process (Singleton et al., 2016), complete with research data, procedures, and metadata. Some scholars are calling for reproducibility to become a standard criteria for author guidelines and peer review, with the ideal paper supported by metadata records of data and provenance (Gil et al., 2016; Nüst et al., 2018). In response to reviews, any changes to the research procedures can ideally be documented and tracked through changes in metadata and release of a final version of the compendium.

Finally, in the preservation phase, the manuscripts, data, and code are placed in FAIR digital archives with final revisions of metadata. In other words, an open access research compendium should be published and archived with metadata specifying access, licenses, data quality, and limitations. Wilson et al. (2021) propose a five-star system for rating the reproducibility of such compendiums. Publishing data and code with some metadata is only two-star level reproducibility: complete implementation of international metadata and encoding standards earns four stars. For example, researchers could earn four stars by storing data with Open Geospatial Consortium (OGC) standard formats, document metadata with ISO standards, and encode the metadata in XML. Documenting and containerising the processing environment would earn the fifth star.

Due to the proprietary, private, or voluminous nature of some data, it may not be possible to release a fully reproducible research compendium. Researchers may need to alter or fabricate alternative data for the purposes of reproducibility, e.g. through simulation, jittering, or sampling (Tullis and Kar, 2021; Singleton et al., 2016). In this case, metadata is essential for documentation of original data, means for accessing original data, and methods for creating alternative demonstration data.

In order to maximise the findability and legibility of the research compendium for both humans and machines, the overall repository and each of its components must be meticulously documented with metadata according to international standards (Wilkinson et al., 2016; Wilson et al., 2021). For instance, GitHub repositories have readme and citation files to facilitate this, while OSF projects have project-level metadata and the ability to register DOI persistent identifiers.

In our own template compendium (Kedron and Holler, 2022b), we have used a project-level readme document with links to comma-separated values tables to orient researchers and readers to the compendium contents, including data layers, procedural code, and results. We designate a metadata directory for

storage of more complete information about each data layer, where XML files using international standards can be placed. We have found this compendium design to be sufficient, but maintaining complete and accurate metadata has been tedious. We are therefore looking for open source software options to improve our metadata management in the reproducible research compendium.

3. MATERIALS AND METHODS

In this section, we describe our approach for reviewing metadata capabilities of open source geographic information software.

3.1 Open source geographic information systems

Following Singleton et al. (2016), we focused on open source software for the purposes of open science and reproducibility. We identified software to evaluate by searching for candidates on the FGDC's ISO geographic metadata editors registry (https://www.fgdc.gov/metadata/iso-metadata-editor-review-v2), the OSGEO Projects (https://www.osgeo.org/projects/), packages compatible with spatial data science in R or Python, and literature on reproducibility in geography. We have excluded proprietary software and software that has not been recently updated or maintained. For example, CatMDEdit was last updated in 2014 (version 5.0) and is no longer maintained.

Our metadata software search ultimately discovered several different types of applications. Desktop GIS like QGIS (QGIS Development Team, 2022), GRASS (GRASS Development Team, 2020), and SAGA (Conrad et al., 2015) are designed for interactive data visualization, editing, and processing. The R and Python programming languages are increasingly used for geospatial analysis, prompting development of specialised packages for managing geographic metadata in those languages. Examples include the geometa package (Blondel, 2022) for R and the pygeometa package (pygeometa team, 2022) for Python. Catalogue services like GeoNetwork (GeoNetwork opensource, 2022) are designed for maintaining and sharing databases of searchable geographic metadata. Content Management Systems (CMS) like GeoNode (GeoNode contributors, 2022) are designed to store and share geographic data in searchable web-accessible archives. Specialised metadata authoring software like MetadataWizard (USGS Fort Collins Science Center, 2022) and mdEditor (ADIwg, 2022) allow users to author and maintain geographic metadata in a stand-alone application. Finally, o2r-meta (Nüst, 2021) is python software designed to support metadata in the o2r executable research compendium.

3.2 What do we need from metadata software?

Based upon our review in section 2 above, we have enumerated useful characteristics for open-source software in support of open and reproducible research (see table 2 columns). The specific characteristics fall into three main categories: (1) ease of use and start-up, (2) implementation of metadata standards, and (3) automated features to facilitate metadata management.

First, metadata software should be easy to set up and to use in order to ease the burden of metadata documentation and management on precious research resources. Software support for metadata management varies tremendously with regard to set-up and installation. Stand-alone desktop metadata editors tend to be very easy to install and start using straight away, while internet-based metadata editing services require only a web browser and login. Desktop GIS software is similarly straightforward to install and run, but some systems have additional difficulties in setting up databases or installing required plugins or add-ons. Packages for computer languages require advanced knowledge of metadata and programming in order to install and learn their functions. Finally, content management systems (CMS) require installation, server administration, and user login prior to working with any metadata.

Graphical user interfaces (GUIs) enhance ease of use and learning how to document metadata, especially for novice users. Interactive features can aid users with features like help documentation about each metadata field, auto-populated lists of keywords from standard dictionaries, selection of spatial and temporal extents, highlighting incomplete required fields, and organising complex information into separate sections or tabs.

Second, geographic metadata in an open science framework should be documented with common international standards. Dublin Core is the present standard for documenting the overall research project. ISO 19115 is the present international standard for documenting individual data layers. The United States FGDC CSDGM standard for data layers is similar to ISO 19115, and the federal government is in the process of adopting ISO 19115. The European INSPIRE standard for data layers is an extension to ISO 19115. If metadata software does not support these common standards, then the metadata may prove useful internally to the research team, but it will not easily be integrated with archives or CMSs or included in automated synthesis or meta-analysis research.

Once metadata conforms to international standards, it should also be encoded and stored with open machine- and human-readable standard formats. The use of an open standard and open machine-readable format enables interoperability with CMSs and automated synthesis research algorithms. If metadata is readable by computers with common parsers, then the metadata can be integrated with more general research management tools to perform functions like creating and updating pre-analysis plans, data management plans, human subjects research protocols, or research compendium documentation. In addition, Git can track versions of text-based formats like Extensible Markup Language (XML), JavaScript Object Notation (JSON), and YAML Ain't Markup Language (YAML). This implies that as metadata changes over time, Git can be used to visualise differences in metadata over different phases of a research project, from pre-registration of the analysis plan to reporting results and finalising the peer review process. Git's difference visualisation could highlight changes in spatial extent, the data dictionary of variables to be collected or computed, protocols for access, and more—even for restricted datasets.

Third, metadata software can automate the discovery, creation, and verification of metadata. In terms of discovery, software can catalogue data layers and partially automate documentation of geographic and attribute metadata. Many desktop GIS and geographic data catalogues can automatically catalogue the geographic data in a research compendium by parsing computer directories in search of recognised geographic data types, resulting in a list of any potential geographic data sources and their relative locations on a computer drive. This feature is particularly useful for routinely cataloguing the data sources contained in a research compendium and verifying the completeness of metadata records for the compendium. Software can also be programmed to automatically extract or calculate geographic

Software	Start-up	GUI Editor	Standards	Encoding	Cataloguing	Automated Geographic	Automated Attribute	Validate	Provenance
mdEditor	very easy	yes	ISO, FGDC	JSON	no	no	no	yes	no
Metadata Wizard 2.0.7	very easy	yes	FGDC	XML	no	yes	yes	yes	no
QGIS 3.24.3	easy	yes	none	XML	browser	yes	fields view	yes	no
SAGA 7.8.2	easy	no	none	none	rasters only	yes	yes	no	yes
GRASS 7.8.5	easy-hard	addon	addons for ISO	XML	no	yes	no	no	no
Geometa 0.6-6	hard	no	ISO	XML	no	no	no	yes	no
pygeometa 0.11.0	hard	no	ISO	XML, YAML	no	no	no	—	no
o2r-meta	hard	no	none	XML, JSON	yes	yes	no	yes	no
GeoNetwork 4.2.0	hard	yes	Dublin, ISO	XML	no	no	no	yes	no
GeoNode 3.3.2	very hard	yes	Dublin, ISO, FGDC	XML	no	yes	yes	no	no

Table 2. Spatial metadata software capabilities.

metadata, including coordinate reference systems, data types, and spatial extents. Attribute data can be extracted to facilitate creation and maintenance of data dictionaries, including variable names, attribute data types, feature or observation counts, descriptive statistics for quantitative data, and unique values for categorical data.

If software contains all the features for cataloguing geographic data and much of its geographic and attribute metadata, then it can also be extended to validate individual metadata records or the records for an entire research compendium. This feature would crosscheck metadata documentation with all automatically derived metadata to report any irregularities or missing information. In addition, validation should check for compliance with regard to completeness of other required metadata fields which cannot be automatically derived, e.g. authorship, license, and distribution information.

None of this yet ensures compliance with one of the most important functions of metadata: to record provenance. Researchers working exclusively with Python, R or other computing languages will hopefully have recorded a complete history of data transformations and manipulations in legible code, and this method of provenance documentation requires the metadata to link to permanently accessible code. Another approach to provenance is to use analytical software that records each step of data transformation as metadata attached to the data itself, which can then be included in the formalised metadata record. Depending on the software environment, this metadata may even be used as a script of instructions for the software to reproduce the data transformations.

If any geographic metadata software could implement all of the features described above, it would be easier for researchers to 1) document their projects and data according to interoperable and international standards, 2) control the completeness and accuracy of geographic metadata records, and 3) mobilise metadata to support the full research workflow, thereby reducing redundancy and increasing transparency.

3.3 Results and Discussion

The results of our software review are summarised in Table 2. We found that no single software program provides all of our desired features for metadata authoring and maintenance, although each of the features exists in at least one of the software programs we reviewed.

The stand-alone metadata editors (mdEditor and Metadata Wizard) are very easy to use, but neither has both the support for international standards and semi-automated authoring features that we were looking for. Open source desktop GIS software (QGIS, GRASS, and SAGA) have been poor at implementing international metadata standards, negating their usefulness for FAIR data. SAGA commendably records provenance for each layer, which can be exported as an executable tool chain in XML format. SAGA also automatically generates geographic and attribute metadata for viewing in the GUI, but does not export the metadata. QGIS has good support for documenting metadata for projects and layers, but uses its own non-standard format. Plugins for older versions of QGIS once supported metadata, and our results suggest a need for renewed interest in either updating the QGIS core to conform more precisely to international standards, or adding a metadata project to the public QGIS plugins repository. The packages for spatial data science in R and Python (geometa and pygeometa) support international standards, but are very difficult to learn and would require additional software code to automate any metadata documentation. The package for an executable research compendium (o2r-meta) is similarly difficult to learn and does not support international metadata standards. However, o2r-meta notably has a valuable function to catalogue geographic data layers within the compendium. Finally, the GeoNetwork and GeoNode content management systems have good support for authoring metadata

with international standards, but they have the significant barrier of requiring installation and administration of servers.

In a context in which researchers' time and software tools are already perceived as limitations on reproducible open science, it appears that there is an urgent need for a new open source software tool to facilitate geographic metadata authoring and management in a research compendium. There are also several existing open source projects from which design ideas and code should be useful.

4. CONCLUSIONS

Geographic metadata is an essential component of open reproducible science. Metadata can also contribute to improved transparency and efficiency throughout the full research life cycle and there are currently a number of open source software tools for authoring and maintaining geographic metadata. However, none of the software currently supports the full range of features desired for supporting metadata-rich research life cycles at every phase.

We conclude that development of a new lightweight and extensible software application for cataloguing and authoring geographic metadata would significantly lower the transaction costs for researchers interested in adopting open science practices throughout the research life cycle, resulting in more FAIR data and reproducible research across the discipline. This software tool should include support for the ISO and Dublin Core metadata standards, an intuitive and instructional graphic user interface, functions to automatically catalogue geographic data in a research compendium and automatically populate geographic information and a data dictionary, and validation. In the future, more advanced development should focus on recording provenance and on using metadata to enrich the full research life cycle by facilitating the creation of research documentation for compendiums, pre-analysis registrations, proposals, data management plans, and human subjects research protocols.

ACKNOWLEDGEMENTS

This research is supported by National Science Foundation project BCS-2049837 and made possible with the hard work of research assistants (Derrick Burt, Drew An-Pham, and Junyi Zhou) and the students in our methods courses. Responsibility for errors lies with the authors. Corrections and comments on a living version of this paper (Holler and Kedron, 2022) are welcome.

REFERENCES

ADIwg, 2022. mdEditor. ADIwg www.mdeditor.org.

Anselin, L., Rey, S. J., Li, W., 2014. Metadata and provenance for spatial analysis: the case of spatial weights. *International Journal of Geographical Information Science*, 28(11), 2261–2280. dx.doi.org/10.1080/13658816.2014.917313.

Bartha, G., Kocsis, S., 2011. Standardization of geographic data: The European inspire directive. *European Journal of Geography*, 2(2), 79–89.

Blondel, E., 2022. geometa: Tools for reading and writing iso/ogc geographic metadata in R. Zenodo doi.org/10.5281/zenodo.5907920. Sponsors/Funders: R Consortium (Linux Foundation project).

Brunsdon, C., Comber, A., 2020. Opening practice: supporting reproducibility and critical spatial data science. *Journal of Geographical Systems*. doi.org/10.1007/s10109-020-00334-2.

Christensen, G., Freese, J., Miguel, E., 2019. *Transparent and reproducible social science research : how to do open science*. University of California Press, Oakland.

Comber, A. J., Fisher, P. F., Wadsworth, R. A., 2008. Semantics, metadata, geographical information and users. *Transactions in GIS*, 12(3), 287–291.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, 8(7), 1991–2007. gmd.copernicus.org/articles/8/1991/2015/.

DCMI, Hillmann, D., 2005. Using Dublin Core - The Elements.

DHEW, 1978. The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research. Technical report, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education, and Welfare, Washington D.C.

GeoNetwork opensource, 2022. GeoNetwork opensource Version 4.2.0. GeoNetwork opensource geonetwork-opensource.org.

GeoNode contributors, 2022. GeoNode opensource Version 3.3.2. GeoNode geonode.org.

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., Karlstrom, L., Lee, H., Mills, H. J., Oh, J.-h., Pierce, S. A., Pope, A., Tzeng, M. W., Villamizar, S. R., Yu, X., 2016. Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, 3(10), 388–415. onlinelibrary.wiley.com/doi/10.1002/2015EA000136.

GRASS Development Team, 2020. Geographic Resources Analysis Support System (GRASS) Software Version 7.8.5. Open Source Geospatial Foundation grass.osgeo.org.

Holler, J., Kedron, P., 2022. Mainstreaming metadata into research workflows to advance reproducibility and open geographic information science. osf.io/52j8s.

ISO, 2014. ISO 19115-1 Geographic information - Metadata. Technical report, ISO, Geneva, Switzerland.

ISO, 2019. ISO/TS 19139-1 Geographic information - XML schema implementation. Technical report, ISO, Geneva, Switzerland.

Kedron, P., Holler, J., 2022a. Replication and the search for the laws in the geographic sciences. *Annals of GIS*, 28(1), 45–56. doi.org/10.1080/19475683.2022.2027011.

Kedron, P., Holler, J., 2022b. Template for Reproducible and Replicable Research in Human-Environment and Geographical Sciences. osf.io/w29mq.

Kedron, P., Holler, J., Bardin, S., Hilgendorf, Z., 2022. Reproducibility, Replicability, and Open Science Practices in the Geographical Sciences. osf.io/c5a2r.

Kedron, P., Li, W., Fotheringham, S., Goodchild, M., 2021. Reproducibility and replicability: opportunities and challenges for geospatial research. *International Journal of Geographical Information Science*, 35(3), 427–445. doi.org/10.1080/13658816.2020.1802032.

Kim, T. J., 1999. Metadata for geo-spatial data sharing: A comparative analysis. *Annals of Regional Science*, 33(2), 171–181.

Konkol, M., Kray, C., Pfeiffer, M., 2019. Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science*, 33(2), 408–429.

Kray, C., Pebesma, E., Konkol, M., Nüst, D., 2019. Reproducible research in geoinformatics: Concepts, challenges and benefits. Sabine Timpf, C. Schlieder, M. Kattenbeck, B. Ludwig, K. Stewart (eds), *14th International Conference on Spatial Information Theory (COSIT 2019)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 1–13.

Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., Greenberg, J., 2021. The role of metadata in reproducible computational research. *Patterns*, 2(9).

Margulies, J. D., Magliocca, N. R., Schmill, M. D., Ellis, E. C., 2016. Ambiguous geographies: Connecting case study knowledge with global change science. *Annals of the American Association of Geographers*, 106(3), 572–596.

Marwick, B., Boettiger, C., Mullen, L., 2018. Packaging Data Analytical Work Reproducibly Using R (and Friends). *The American Statistician*, 72(1), 80–88. doi.org/10.1080/00031305.2017.1375986.

NASEM, 2018. *Open Science by Design*. National Academies Press, Washington, D.C.

NASEM, 2019. *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., Mellor, D. T., 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2600–2606.

NSF, 2021. *Proposal and Award Policies and Procedures Guide*. National Science Foundation, Washington D.C.

Nüst, D., Granell, C., Hofer, B., Konkol, M., Ostermann, F. O., Sileryte, R., Cerutti, V., 2018. Reproducible research and GIScience: an evaluation using AGILE conference papers. *PeerJ*, 6, e5072.

Nüst, D., Pebesma, E., 2021. Practical Reproducibility in Geography and Geosciences. *Annals of the American Association of Geographers*, 111(5), 1300–1310. doi.org/10.1080/24694452.2020.1806028.

Nüst, D., 2021. Reproducibility Service for Executable Research Compendia: Technical Specifications and Reference Implementation. Zenodo doi.org/10.5281/zenodo.5106499.

Ostermann, F. O., Granell, C., 2017. Advancing Science with VGI: Reproducibility and Replicability of Recent Studies using VGI. *Transactions in GIS*, 21(2), 224–237. onlinelibrary.wiley.com/doi/10.1111/tgis.12195.

Ostermann, F. O., Nüst, D., Granell, C., Hofer, B., Konkol, M., 2021. Reproducible research and GIScience: An evaluation using GIScience conference papers. *Leibniz International Proceedings in Informatics, LIPIcs*, 208(33), VII.

pygeometa team, 2022. pygeometa. GitHub github.com/geopython/pygeometa.

QGIS Development Team, 2022. QGIS geographic information system version 3.24.3. www.qgis.org.

Schuurman, N., 2008. Database Ethnographies Using Social Science Methodologies to Enhance Data Analysis and Interpretation. *Geography Compass*, 2(5), 1529–1548. dx.doi.org/10.1111/j.1749-8198.2008.00150.x.

Schuurman, N., Leszczynski, A., 2006. Ontology-Based Metadata. *Transactions in GIS*, 10(5), 709–726.

Singleton, A. D., Spielman, S., Brunsdon, C., 2016. Establishing a framework for Open Geographic Information science. *International Journal of Geographical Information Science*, 30(8), 1507–1521. dx.doi.org/10.1080/13658816.2015.1137579.

Stodden, V., Leisch, F., Peng, R. D. (eds), 2014. *Implementing Reproducible Research*. Taylor & Francis, Boca Raton.

Tullis, J. A., Kar, B., 2021. Where Is the Provenance? Ethical Replicability and Reproducibility in GIScience and Its Critical Applications. *Annals of the American Association of Geographers*, 111(5), 1318–1328. doi.org/10.1080/24694452.2020.1806029.

USGS Fort Collins Science Center, 2022. MetadataWizard. GitHub github.com/usgs/fort-pymdwizard.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 1–9.

Wilson, J. P., Butler, K., Gao, S., Hu, Y., Li, W., Wright, D. J., 2021. A Five-Star Guide for Achieving Replicability and Reproducibility When Working with GIS Software and Algorithms. *Annals of the American Association of Geographers*, 111(5), 1311–1317. doi.org/10.1080/24694452.2020.1806026.