|  |  |
| --- | --- |
|  | Thank you!  The title of my talk today is Mainstreaming Metadata into Research Workflows to advance Reproducibility and Open Geographic Information Science.  I am Joseph Holler from Middlebury College and this work was completed with the collaboration of Peter Kedron at Arizona State University with the support of a National Science Foundation grant for improving reproducibility and replicability in the geographic sciences through a project-based graduate and undergraduate methods curriculum.  These slides and the associated paper are available in the ISPRS archives, on GitHub, and on OSF. |
|  | We need some motivation for a 20 minute talk about metadata!  I am motivated because I have reluctantly concluded that metadata is required to enhance the reproducibility of geographic research.  In turn, enhanced reproducibility is expected to increase the pace and credibility of knowledge production in the geographic sciences.  Furthermore, I am optimistic that integrating metadata into everyday research practices will facilitate more efficient and open research life cycles.  This research is based on experience applying the broad consensus reports by the National Academies on Open Science and Reproducibility and Replicability to the specific challenges in the geographic sciences.  So, what is happening in geography? |
|  | For some context from our research project,  we have surveyed geographers about their research practices, finding that most folks would say, “I am familiar with reproducibility and my research is reproducible”  However, when asked about metadata, most would say, “Metadata, No, I have never used that...”  Unfortunately, providing data or even code alongside a research publication without metadata is like publishing a map with no title or legend… leaving serious questions about the data and its proper use unanswered. (COVID rates as of August 2020)  We have reluctantly and painfully come to this conclusion– and the need for this paper– after attempting seven reproduction studies and publishing each as a reproducible research compendium in our GitHub HEGSRR Organization. |
|  | I am hoping to convince you of these three points:   1. Open Science and Reproducibility require standardized metadata 2. Researchers use, create, and modify information about their research projects and research data throughout the research life cycle 3. We need better open source geospatial software to support metadata-rich research   But first, what is reproducibility? |
|  | Reproducibility is a core motivation of open science, but it’s more than simply repeating another researcher’s computations as illustrated by this matrix.  In the top left, if we use the **same** methods and same data to achieve the **same** results as a prior study, this **REPRODUCTION** provides a check of the internal validity of the study.  Thus far, we have almost always discovered uncertainties or errors in the original procedures, leading us to the top-right, where we REANALYZE the **same** data with **varied** methods and test the sensitivity of the study to researcher decisions and errors.  If we want to externally validate a study, we drop to the bottom left, where we apply the **same** methods to **new** data in a REPLICATION study, which may contribute evidence for the generalizability of a theory across different contexts.  Finally, the bottom right is where science makes further progress by EXTENDING previous studies with **varied** methods and **new** data.  What is the role of metadata in all of this? |
|  | Geospatial metadata is information about spatial data.  Drawing on the iceberg cliché, consider that geospatial data is just the attractive ice floating above the ocean surface,  Meanwhile, a mass of contextual information lies beneath, keeping the ice above water afloat and causing danger for anyone unaware of its presence. |
|  | We can think of metadata as one of the irreplaceable cogs allowing the gears of open science to work.  Metadata provides essential social and ontological context for data’s meaning, interoperability, and appropriate use.  Good metadata is also an ethical issue, particularly with regards to problems of privacy and quality of big data for humanitarian response.  In open science, metadata is the key to FAIR open data.  Data becomes findable with project-level metadata including the study’s geographic and temporal extent.  Data is accessible when metadata specifies an open license or access protocols, or when metadata provides enough detail about inaccessible data for someone to recreate or simulate the data.  Data is interoperable when metadata adheres to machine-readable international standards, and  Data is reusable when metadata provides enough context and detail for re-creation and appropriate re-use  The 5-star guide to reproducible geographic research distinguishes THREE out of five stars with metadata!  Providing code and data with an open license only achieves one star.  Two stars are achieved with SOME metadata and a third star is awarded for COMPLETE metadata.  A fourth star is reserved for data and metadata encoded with international standards.  So, what are metadata standards, and which standards should we use? |
|  | Metadata standards for geographic information may come either spatial data infrastructures, e.g. the Federal Geographic Data Committee for the United States or the Infrastructure for Spatial Information in Europe.  These data infrastructures were developed to enhance data standards and interoperability between governments and government agencies  Metadata standards have also been developed by other organizations.  The 191 series of standards by the International Organization for Standardization (ISO) are increasingly adopted and extended by both the American FGDC and European INSPIRE, making the ISO standard an ideal choice for documenting geographic data layers.  The Dublin Core standard developed by librarians for managing digital archives is ideally suited for the overall research project.  The Open Geospatial Consortium offers plenty of guidance on open data storage formats, but leaves the metadata standard to the ISO.  Let’s look more closely at the content of the ISO and Dublin Core standards for use in geographic research. |
|  | Again, the ISO 19115 and 191 suite of standards are best for geographic data layers and the Dublin Core standard is best for whole research projects.  I compare the two standards in the table on the right.  The first five rows answer the basic what, why, and when questions of the data, and make your projects findable by using a DOI for the unique identifier and controlled vocabularies for topic/subject keywords.  The next two rows answer the who question about the data: who is responsible for authoring, creating, publishing, maintaining, etc. ?  The next row provides legal issues of constrains and rights. Ideally data will be published with open licenses, without which copyright protection is implied and re-use is forbidden.  From this point on, the ISO 19115 standard provides much more detail for geographic data layers than Dublin Core would, including  the spatial data model and spatial and temporal extents and resolutions.  Content information for vector data is essentially a data dictionary of all the attributes, attribute types, measurements, and even descriptive statistics. Content information for gridded imagery or satellite data contains details of the sensors and measurements.  The ISO standard also contains essential information on data quality and proper usage.  Finally, the ISO’s lineage feature is robust enough to record step-by-step data transformations, while the Dublin Core standards are more about recording a chain of custody. |
|  | Hopefully my first point is now clear: Open science and reproducibility require standardized metadata.  How does all of this information fit within a research life cycle? Let’s take a look. |

|  |  |
| --- | --- |
|  | The National Academy’s Open Science by Design report envisions research life cycle with expanded research opportunities and improved research quality enabled by open science practices in each of six phases. |
|  | In the provocation phase, we search and review existing literature to generate new research ideas. |
|  | In the ideation phase, we plan the research, including human subjects protocols for ethics review, research funding proposals with data management plans, and pre-analysis registrations of study protocols. |
|  | In the knowledge generation phase, we collect, create, and analyze data and start documenting metadata. |
|  | In the validation phase, we analyze and share preliminary results, which can be aided with OSF / figshare registration |
|  | In the dissemination phase, we undergo our beloved peer review process, revision, and formal publication. |
|  | In the preservation phase, we archive research data and code in an open access repository and finalize project metadata for searching and data-layer metadata for (re)usability |
|  | If we reimagine this life cycle using a research compendium, then we actually work on the preservation phase with metadata through the whole research life cycle. |
|  | We have developed a template research compendium for reproduction and replication studies.  It provides structure for project-level metadata and for organizing all of the data, metadata, procedures and code, documents and manuscripts, and resulting figures, tables and model outputs related to the research project. |
|  | The compendium is managed as a Git repository for version tracking, comparing differences or changes between versions (as shown in red and green at bottom right), and branching and merging alternative research designs.  At the provocation phase, an open science literature review is enabled by project-level metadata, spatially-explicit synthesis / meta-analysis / bibliometric analysis.  At Ideation phase, we create a new compendium with project-level metadata.  We research and imagine the data that we intend to use and create, and we generate standardized metadata for this in our metadata folder.  We then use our metadata to help us organize and write research proposals, data management plans, ethical human subjects research protocols, and pre-analysis plan documents. |
|  | We register the plan(s) on OSF or similar using project-level metadata and link our OSF project to the Git repository before moving on to knowledge generation.  At the Knowledge Generation phase, we create the data, update our metadata documents, and enable visualization of any changes to the data portion of our research protocols. Ideally, metadata tools support cataloging data, updating metadata, and building a directory for the compendium.  At the validation phase, we document unplanned deviations in the research protocol, write and register reports, and develop open access preprints and conference presentations associated with our compendium.  At the dissemination phase, our compendium provides unprecedented access to the details of our research to reviewers.  We also track changes required by review and ensure our research is more findable through project-level metadata and more usable through data-level metadata.  We make sufficient metadata details of embargoed, restricted, or proprietary data are available to simulate, access, or recreate similar data.  At the preservation phase, we realize that we have been preserving the research in a public compendium all along, and ensure we have added a persistent identifier (DOI) to all of our interlinked research products.  Let’s look at an example compendium… |
|  | Here you see one of our reproduction studies on GitHub, with the data, docs, procedure, and results directory structure, a citation file, an open access license, and project-level read me document.  The readme ideally contains project-level metadata and a directory of data files, and ideally this would be maintained automatically with software. |
|  | Each significant data layer is documented with metadata in the metadata folder, and here is the top of an XML file using the FGDC standard to describe American Community Survey data. |
|  | Currently, we are maintaining a list of data paths, file names, formats, and metadata files in a comma-separated text file that renders as a table on GitHub.  Ideally, software would help maintain such a directory automatically. |
|  | Hopefully my second point is now clear: **Researchers use, create, and modify information about their research projects and research data throughout the research life cycle**  So, what does the FOSS4G software ecosystem offer for metadata-rich research life cycles? |
|  | There are several types of software that may be of use…  Only those software with metadata editing functionality and open licenses with no fees are considered.  In the desktop GIS category, we reviewed QGIS, GRASS and SAGA.  From the spatial data science worlds, we reviewed geometa in R and pygeometa in Python.  OSGEO projects include a catalogue server, GeoNetwork, and a content management server, GeoNode.  In our software search, we discovered two specialized metadata authoring programs: Metadata Wizard by the USGS, and mdEditor, a web-based system created by the Alaska Data Integration Working Group.  Finally, the o2r Opening Reproducible Research program is developing a containerized executable research compendium with a metadata tool, o2r-meta.  Given this suite of possible software tools, what software features do we need? |
|  | Based on our practical experience and literature review, we suggest the following metadata software needs.  First, metadata software must be easy to use for students, research assistants, and faculty with limited time.  Installation, start-up, and learning the software should be easy, and a graphical user interface should support editing and provide guidance through the editing process.  The software must support open standards, including support for all of the fields and controlled vocabularies of the ISO and Dublin Core, as well as encoding in standardized formats, namely XML.  Ideally, working with metadata should be facilitated by as much automation as possible, including features to  parse a directory for spatial data and catalogue it,  extract geographic metadata like the coordinate reference system and extent,  Extract attribute metadata like field names, types, and descriptive statistics,  Validate metadata documents for completeness and conformity to standards,  And track provenance– a detailed history of data revisions. |
|  | How did the FOSS4G software stack up against these needs? Here are our results… |
| Finally, only one software– SAGA– tracks provenance and records it as metadata attached to a layer, which can be copied and executed as a tool chain!  The key message here is my third point: We need better open source geospatial software to support metadata-rich research | In the easy start criteria, metadata editors receive double checks for very fast installation and easy use. Desktop GIS receive a single check for straightforward installation and use. The servers and code packages have numerous barriers for a novice user to install and start using.  In terms of a graphical user interface, the desktop GIS software, servers, and metadata editors all had GUIs, but SAGA and GRASS could not easily use the GUI for metadata editing.  In terms of standards, only the GeoNetwork and GeoNode servers supported both ISO and Dublin Core. A single black check indicates support for ISO only, and Metadata Wizard supports only FGDC.  Most of the software could save metadata encoded in machine-readable XML or JSON formats. These are essential for extending metadata to other purposes, like auto-filling research documents and registrations.  Only o2r-meta had a true cataloging feature meant to discover all spatial data layers in a research compendium directory. QGIS and SAGA are both capable of viewing at least some spatial data formats in a directory, but not in a format usable for metadata documentation.  Six of the software options contained at least semi-automated features for extracting geographic metadata,  and four of those contained features for automatically extracting, or at least viewing, some attribute metadata.  Six of the software options also contained features to validate metadata records, but thus far only Metadata Wizard has full implementation of validation and automated geographic and attribute metadata. |
|  | My third point should now be clear: we need better open source geospatial software to support metadata-rich research, fulfilling the software needs enumerated on the right. |
|  |  |