Toward a Flexible Metadata Pipeline for Fish Specimen Images*

Dom Jebbia[1,2,3][0000−0002−9587−8718], Xiaojun Wang[2][0000−0002−2995−9050], Yasin Bakis[2][0000−0001−6144−9440], Henry L. Bart Jr.[2][0000−0002−5662−9444], and Jane Greenberg[1][0000−0001−7819−5360]

[1] Drexel University Metadata Research Center, Philadelphia, PA 19104 USA
[2] Tulane University Biodiversity Research Institute, Belle Chasse, LA 70037 USA
[3] Carnegie Mellon University, Pittsburgh PA 15213, USA
djebbia@andrew.cmu.edu; xwang48@tulane.edu; ybakis@tulane.edu; hbartjr@tulane.edu; jg3243@drexel.edu

Abstract. Flexible metadata pipelines are crucial for supporting the FAIR data principles. Despite this need, researchers seldom report their approaches for identifying metadata standards and protocols that support optimal flexibility. This paper reports on an initiative targeting the development of a flexible metadata pipeline for a collection containing over 300,000 digital fish specimen images, harvested from multiple data repositories and fish collections. The images and their associated metadata are being used for AI-related scientific research involving automated species identification, segmentation and trait extraction. The paper provides contextual background, followed by the presentation of a four-phased approach involving: 1. Assessment of the Problem, 2. Investigation of Solutions, 3. Implementation, and 4. Refinement. The work is part of the NSF Harnessing the Data Revolution, Biology Guided Neural Networks (NSF/HDR-BGNN) project and the HDR Imageomics Institute. An RDF graph prototype pipeline is presented, followed by a discussion of research implications and conclusion summarizing the results.ite this need, researchers seldom report their approaches for identifying metadata standards and protocols that support optimal flexibility. This paper reports on an initiative targeting the development of a flexible metadata pipeline for a collection containing over 300,000 digital fish specimen images, harvested from multiple data repositories and fish collections. The images and their associated metadata are being used for AI-related scientific research involving automated species identification, segmentation and trait extraction. The paper provides contextual background, followed by the presentation of a four-phased approach involving: 1. Assessment of the Problem, 2. Investigation of Solutions, 3. Implementation, and 4. Refinement. The work is part of the NSF Harnessing the Data Revolution, Biology Guided Neural Networks (NSF/HDR-BGNN)

* Supported by NSF-HDR-OAC: Biology-guided Neural Networks for Discovering Phenotypic Traits: 1940233 and 1940322m, NSF HDR-OAC:Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning: 2118240, and the Institute of Museum and Library Services (IMLS) RE-246450-OLS-20.

project and the HDR Imageomics Institute. An RDF graph prototype pipeline is presented, followed by a discussion of research implications and conclusion summarizing the results.

Keywords: Metadata pipelines · Open data · Metadata workflows · FAIR data. · Digital images · Biodiversity Collections

1 Introduction

Digital technology, cyberinfrastructure, and the full open research movement have enabled new pathways for scientific research. This is particularly true with digital images of scientific specimens. Scientists are able to examine and compare samples on a scale that was not possible in the analog world. Moreover, computational methods enable new modes of inquiry. Although the research opportunities seem endless, researchers face obstacles as they try to sample the correct type of scientific specimen, or develop efficient pipelines to support their work. Many of these challenges stem from metadata quality issues, or simply the absence of metadata, associated with the life-cycle of the digital specimen [24, 29, 41, 53].

A range of metadata challenges in this area became quite apparent as a group of researchers associated with the NSF supported Harnessing the Data Revolution, Biology Guided Neural Networks (HDR-BGNN) project began their work. A key goal of this research is to examine images of fish specimens and their morphological traits via segmentation followed by feature extraction to determine differences among images representing fish from different taxonomic groups. Combining state-of-the-art image segmentation techniques with Phenoscape ontologies for algorithmic analysis [21, 39, 23, 38, 5], researchers could potentially identify undescribed species grouped within currently described species. The collections of images for training neural networks and developing an image-processing workflow revealed many metadata challenges, which led BGNN collaborators at Drexel University's Metadata Research Center (MRC) and Tulane University's Biodiversity Research Institute (TUBRI) to develop a flexible, extensible metadata pipeline.

This paper reports on the efforts of the MRC-TUBRI collaboration. The next section of the paper provides background context, followed by the underlying goals and objectives. The four-phased approach that framed the work is explained, along with the current RDF-graph prototype model. Finally, the discussion addresses the extensibility of the current model, and the conclusion summarizes the key results.

2 Background context

Digital technology and data sharing have motivated the development of national and global repositories that provide global access to digital images of biological specimens. Even so, connecting to these repositories and taking advantage of

this new infrastructure can be obstructed by a range of challenges associated with metadata and pipeline models [33, 13].

2.1 Open Science Repositories

Over the past two decades, researchers have supported the proliferation of digital repositories. The growth of these collections has been motivated by a number of key factors, including the open science movement and, most recently, the international embrace of the Findable, Accessible, Interoperable, and Reusable (FAIR) [60] data principles. For the purposes of this paper, it is important to note the role of government policy, which first encouraged and now requires publicly funded data to be made available. These evolving mandates can take several forms.

Europe The European Union was the first major government body to develop policy regarding the availability of publicly funded research. It first did so through the Public Sector Information Directive in 2003 [6], and later by its 2019 amendment as the Open Data Directive [8]. The European Commission (EC) has supported these directives by developing infrastructure such as OpenAIRE and Europeana [27, 28, 33, 32, 51, 40, 7].

United States of America Similarly, in 2013 the U.S. Office of Science and Technology Policy (OSTP) mandated that federal agencies with more than $100 million in research and development should make their data available within one year of publication [12]. Most recently, in August 2022 the same agency issued a White House supported memo stating that all federally funded research should be available without delay [45].These policies and similar developments worldwide have created an imperative for academic organizations to make researchers' data available. They have also encouraged the development of metadata standards that support open data and data interoperability on a global scale.

2.2 Metadata for Open Science and Digital Scientific Specimens

Open science and open data sharing have motivated the development of many metadata standards, and the adaptation of existing standards. At the general domain level, researchers can apply the Dublin Core (DC) metadata following the extensive list of metadata properties registered at the DCMI Terms namespace [9]. Researchers may also develop a Dublin Core Metadata application profile by integrating metadata properties from other standards with Dublin Core properties. Two well-known examples include the Virtual Open Access Agriculture and Aquaculture Repository VOA3R metadata application profile [22] developed to support the description and reuse of research results in the fields of agriculture and aquaculture as part of a larger federation of open access repositories; and the Dryad metadata application [31], which underlies a global repository that publishes research data underlying scientific publications. On a more specific

level, there are hundreds of metadata schemes developed for different research domains and types of scientific data. Examples include the Ecological Metadata Language (EML) [42] for ecology data, the Darwin Core [59] (DWC) for scientific museum specimens, and the Data Document Initiative (DDI) [61] for social science research. There are also a wide array of metadata standards associated with the type (e.g., static image, X-ray, moving image), preservation status, and rights specifying data access and usage.

The overabundance of metadata standards that can be used to describe scientific data can be both exciting and overwhelming for scientists trying to determine which standards support their data needs. In response to this challenge, various communities have developed directories and registries to help inform decision making and pipeline design. Key examples include the Digital Curation Center's Disciplinary Metadata Directory [14, 15], the Research Data Alliance's Metadata Standard Directory [15, 50], the National Consortium of Biological Ontologies Bioportal [4], and the FAIR Sharing Standards Registry [1]. These are significant efforts; however, these extensive resources require human examination, which can be daunting. This challenge is quite evident when looking specifically at the metadata for individual specimens. The 'Life Sciences' class in the RDA directory includes 32 sub-topic areas. Most of the sub-topics identify five or more metadata standards, and a number of subtopics refer to ten or more applicable metadata standards for any given area. This is also simply within the 'Life Sciences' class, and does not include the applicable metadata standards listed in the 'Physical Sciences & Mathematics' and 'Social & Behavioral Science' classes–both of which may include metadata standards that are applicable to physical or other types of scientific specimens. The challenges associated with identifying an appropriate metadata standard further impact metadata pipeline development, data sharing, and the FAIR principles.

The FAIR principles motivated this work. FAIR establishes that data should be findable, accessible, interoperable, and reusable. Scientific images, particularly images of specimens housed in digital repositories may be findable and accessible, but the data associated with them is not always interoperable or reusable. These limitations are grounded in metadata [18, 58, 37, 43, 20, 16]. Moreover, they interfere with being able to leverage rich resources for scientific research. One key solution is to develop better metadata pipelines to support FAIR, which is key to the work presented here.

2.3 Metadata Pipelines

The concept of pipelines denotes a workflow or systems approach to how materials, information, or other types of resources flow from one place to the next, and the stops along the way. Computing and informatics frequently refer to data pipelines to describe the flow of data throughout an information system. A metadata pipeline is, essentially, a type of data pipeline. Metadata pipelines are key to supporting reproducible computational research [11], and the overall execution of the FAIR principles. A metadata pipeline frequently begins with the harvesting of existing metadata or creation of new metadata in the absence of

metadata, followed by the transport of the metadata, often with the associated object, through a series of operations. While a metadata pipeline is intended to support a workflow, the operation is frequently inhibited by inconsistent application of metadata, the absence of key metadata, and conflicting metadata —all of which impact metadata quality [56, 47, 48]. Finally, the identification and implementation of a metadata workflow model presents challenges. Researchers can work with the common workflow language and look at developments, such as the metadata underlying the Open Archival Information System (OAIS) reference model, Digital Asset Management System (DAMS) workflows, or potentially more sophisticated developments, such as the Unified Modeling Language (UML) information model. Another way that may be more comprehensible to researchers is the Resource Description Framework (RDF) model, which underlies the Semantic Web and linked data. All of this has informed the work reported in this paper.

3 Goals and Objectives

Metadata challenges along with associated metadata model complexities impact the development of successful metadata pipelines. The current circumstance has helped shape the overall goals and objectives that inform our work, the overall goal of which is to develop a flexible and extensible metadata pipeline to support the HDR-BGNN effort. The flexibility allows TUBRI to align the final output of the pipeline with FAIR principles, increasing the impact of the data. Furthermore, the work is also necessary for BGNN to interconnect with the recently established HDR Imageomics Institute. Key objectives shaping our work include:

1. Understanding the scope of TUBRI's data flow and metadata needs to accommodate AI research across the BGNN project and the connected Imageomics Institute.
2. Designing a plan to improve the current metadata pipeline.
3. Implementing, assessing, and modifying the metadata pipelines as needed.
4. Demonstrating a proof-of-concept using RDF to align data pipelines and their outputs with FAIR principles.

Our work is presented in the next section.

4 Designing a Flexible, Extensible Metadata Pipeline

Our approach to addressing the above objectives and our overall goal was carried out in four phases,identified and discussed here.

4.1 Phase 1: Assessment of the problem

The process started by evaluating the BGNN metadata lifecycle. First, we determined potential sources of future image collections and what associated meta-

data elements could potentially be included. Then we considered the future internal needs at TUBRI. Throughout the process, we weighed how these workflows could be restructured to make the dataset useful to the largest audience.

This was explored by evaluating the previous data pipelines, workflows, and internal practices at TUBRI and BGNN. Figure 1 demonstrates the fish specimen image pipeline developed by BGNN, as well as the challenges to creating fully automated computational workflows. We contextualized these observations through interviews and collaboration with researchers in the groups. This information was then compared with the practices of other organizations that contribute to the HDR Imageomics Institute [10], oceanographic data organizations [3, 2], and open science repositories. This assessment identified several deficiencies within the data pipeline. The two most significant were the number of organizations providing collection event metadata and the sparse, irregular conditions of the raw datasets. This created difficulties adapting the ingestion process to normalize metadata with recognized standards (DC, DWC, Exchangeable Image File (EXIF), etc) communicating those choices to users of the dataset. Many organizations have developed various approaches for making their data FAIR, unfortunately those solutions are generally project specific and not often shared in the literature. It was clear that workflows for making datasets FAIR needed to be made more FAIR.

4.2 Phase 2: Investigation of Solutions

TUBRI researchers determined that there were two approaches to improving the flexibility of the new database structure for BGNN. One is to modify the database schema based on the relational (table-based) database. The other is to switch from the relational database to a document-oriented NoSQL database. Table 1 details the solutions identified during the investigation.

The second option is a document-oriented NoSQL database, which is a non-tabular database structure to store the data like a relational database. It offers a fast and flexible schema that enables data models to evolve with frequent changes. The database can use JSON, XML, BSON, and YAML formats to define and manage data.

The first approach was chosen because it built upon the relational database structure in use, rather than conceptually redesigning the database structure. Furthermore, this builds upon the semantic interoperability work pursued by earlier collaboration between the MRC and TUBRI on the BGNN project [36, 35, 25, 49]. Of the identified techniques, the EAV model was the most adaptable as a database design pattern to restructure the relational databases containing BGNN data. There were also concerns that the JSON and XML data types solutions may cause problems such as poor performance or making the database structure difficult to manage. EAV model was the most abstract of the solutions, but offered methods to redesign the databases to make it more adaptable to new workflows and extensible when ingesting new metadata elements. There are also numerous ways to implement EAV using JSON or XML. EAV implementation can may use an XML column in a table to capture the incomplete information

or variable information, while similar principles apply to databases that support JSON-valued columns.

RDF was chosen to implement EVA because:

- It offers an extensible solution to the ongoing ingestion of new data from disparate sources.
- Major repositories have already adopted some form of RDF, for instance many cultural heritage organizations have adopted it built on the Europeana Data Model, or science data through OpenAIRE.
- It makes FAIR principles foundational to the design of data pipelines.

4.3 Phase 3: Implementation

Two methods were examined to create an RDF graph to represent the metadata. One is an implementation using Python libraries; the other uses the desktop version of Protégé. Python libraries offer numerous applications and workflows, but Protégé was chosen to create the prototype because the graphical user interface (GUI) provides an interface to directly interact with the graph. Moreover, the Protégé-OWL batch import plug-in offered an efficient, if limited, way to transform spreadsheet data into RDF schema.

Team members evaluated the standards in use and chose new control schemes to more accurately describe the metadata. Selected standards are described in Table 3. This included the removal of duplicated, redundant, or deprecated elements. The remaining elements were checked for accurate usage and adjusted accordingly, such as changing `<dwc:AccessConstraints>` to `<dc:accessRights>`. Finally, new schemes were chosen to align the image data with standards used in other photographic applications, for instance, the adoption of standard maintained by Adobe, the International Press Telecommunications Council, and the PLUS Registry, amongst others. The RDF model theoretically allows for the adoption of any standard to normalize data. A rights statement and IRI was also included as an `rdfs:comment` in the graph.

Figure 2 represents the RDF graph prototype that was generated. Table 2 lists the previous data containers and the updated database structure and the sources from which the metadata were derived. In the new structure, metadata is grouped into classes based on the kind of metadata in use. For example, `Multimedia` represents the administrative metadata related to the raw image and its capture event. `IQ metadata` refer to the elements generated by BGNN through computational workflows; the training dataset was created by humans and then metadata was generated through segmentation and trait extraction. The `ExtendedImageMetadata` class encompasses image quality metadata for processed images. `Collection event` metadata refers to the specimen data gathered by researchers in the field. `Bach` contains administrative metadata for the final dataset. Each of the top-level updated nodes is assigned an Archival Resource Key (ARK) that serves as a persistent identifier. The ARK associated with `Multimedia` is the parent identifier for the set of images and metadata generated from the workflow.

4.4 Phase 4. Refinement

Phase 1 assessed the metadata ecosystem at TUBRI and identified pipeline features that create data bottlenecks and barriers for image processing and segmentation masking. Phase 2 investigated the potential solutions to these identified problems. Phase 3 implemented a prototype RDF model to restructure the BGNN databases. As of this writing, the project is in Phase 4, which synthesizes the results of the previous stages to design more robust and sustainable workflows.

Some of the barriers identified in the previous phases include:

- Determining which technology or approach is effective and scalable to different collections.
- Creating an RDF structure that will enhance the metadata pipeline and make the final datasets more FAIR.
- Designing processes and techniques that are applicable to many different fields, rather than domain specific solutions.

Phase 4 seeks to further investigate and resolve these challenges by:

- Creating programmatic workflows that make it easier to create and maintain RDF graphs.
- Employing the prototype RDF schema to implement a system that accesses the relational databases as virtual RDF graphs. This allows the query a non-RDF database using SPARQL, access the content of the database as Linked Data over the Web, create custom dumps of the database in RDF formats for loading into an RDF store, and access information in a non-RDF database using the Apache Jena API.
- Using a Python wrapper to make the data more accessible to researchers through an application programming interface (API).

Table 1: Relational database solutions.

Solution	Description
Add columns to tables	Extend existing database.
Entity-Attribute-Value (EAV)	Restructure database using EAV model.
JSON data type	Convert database structure to JSON/XML.
XML data type	Convert database structure to XML.

5 Discussion

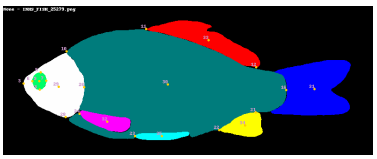
5.1 RDF

This paper demonstrates how RDF's flexibility and extensibility can be used to streamline the (meta)data creation process, in addition to providing a database

(a) A raw image with a ruler and specimen label.

(b) A bounding box image is created from the raw image.

(c) A segmentation mask is generated from the bounded image.



(d) Trait features are labeled on the segmentation mask.

Fig. 1: The BGNN image processing pipeline featuring a *Carassius auratus* specimen image. Optical character recognition (OCR) is used to extract metadata from the specimen label and validate against the collection event metadata associated with the raw image.

Table 2: Changes in the database structure.

Original data containers	Updated RDF nodes	Metadata source
Media	Multimedia	Raw image
Collection event	Collection event	Specimen
ImageQualityMetadata	IQ metadata	Bounding box image
	ExtendedImageMetadata	Labeled segmentation mask
	Batch	Administrative

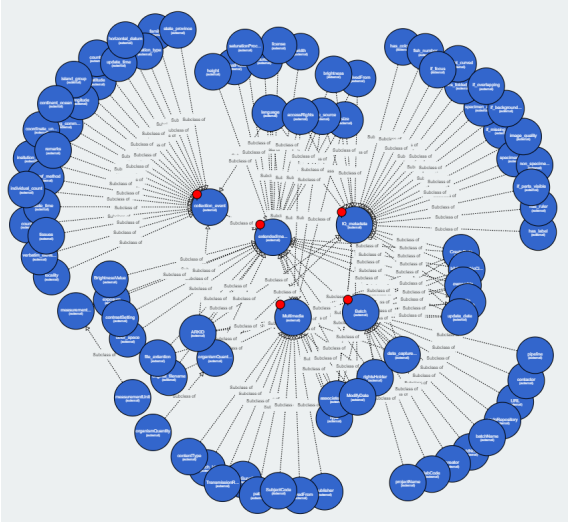


Fig. 2: A visualization of the RDF prototype created with Protege.

Table 3: Standards added to the RDF prototype.

Standard	Namespace prefix	IRI
Audobon Core	ac	http://rs.tdwg.org/ac/terms/
Camera Raw	crs	http://ns.adobe.com/camera-raw-settings/1.0/
Darwin Core	dwc	http://rs.tdwg.org/dwc/terms/
Darwin Core	dwciri	http://rs.tdwg.org/dwc/iri/
Exchangeable Image File	exif	http://ns.adobe.com/exif/1.0/
IPTC* Core	Iptc4xmpCore	http://iptc.org/std/Iptc4xmpCore/1.0/xmlns/
Photoshop	photoshop	http://ns.adobe.com/photoshop/1.0/
Picture Licensing Universal System	plus	http://ns.useplus.org/ldf/xmp/1.0/
Extensible Metadata Platform	xmp	http://ns.adobe.com/xap/1.0/
Basic Job Ticket	xmpBJ	http://ns.adobe.com/xap/1.0/bj/
XMP Media Management	xmpMM	http://ns.adobe.com/xap/1.0/mm/

*International Press Telecommunications Council

design pattern that can adapt to the changing needs of research investigations. The frameowrk makes research output more FAIR by providing the foundational infrastructure for computational analysis. Specifically, RDF provides a means to express the metadata elements in relation to the resources they describe and to each other, rather than the arbitrary location of the information in a database structure. Investigators spend most of their research time cleaning data [57], so pipeline design is an important part of making the final output of a project reusable and ultimately affects the results of later machine learning and neural networks. RDF provides a means of rapidly responding to the changing technical and structural parameters of a project. One major reason for this RDF implementation was to make the database structure able to respond to changing technical requirements, for example, case sensitivity in programming languages. It also provides a means to communicate complex licensing, attribution, and usage rights that accumulate during data reuse.

However, the flexibility and extensibility of RDF can present a number of challenges. Although schema can be useful in database design [34], if RDF is implemented without consideration of FAIR principles the resulting database structure may inhibit its ability to link data through the Semantic Web [44]. One way that this could occur is by applying an RDF structure to a database without curating the standards defining the data. Because data collection is laborious and time-consuming, as research approaches evolve it can be difficult to maintain structured data. RDF can make this data findable and accessible to some degree, however without a contextual data model scientists may need to further analyze and clean the data, contact the original investigators for further information, or guess as to the details of the original investigation. This can severely impact quality and reproducibility. Furthermore, the last few decades of digitization efforts by galleries, libraries, archives, and museums have produced corpora of semi-structured historical data relating to every domain of science documented before the digital age. These datasets are important to climate and biological scientists as they attempt to understand climate change and biodiversity [54, 46, 26, 52]. However, the people who created these early datasets had no idea how the data they were collecting could be used by others in the future. The cost of data management and geopolitical remnants of colonialism are also significant barriers to making data FAIR [17, 46, 55]. RDF can help make these datasets findable and accessible if the circumstances allow, however significant curation is necessary to make them interoperable and reusable by machines.

5.2 FAIR Pipelines for Open Science Repositories

One of the desired outputs of the BGNN project is a dataset of processed images, segmented masks, and rich metadata for others to reuse in future studies. The RDF database structure makes it easy to manage and update data structures, resulting in the ability to accept new metadata elements and adjust them as the requirements of the project evolve. The workflows that RDF makes possible improve the quality and quantity of metadata associated with the images in

the dataset. This helps align the research output with FAIR, while also designing pipelines that can be adopted by other studies interested in building FAIR aligned workflows. Although researchers conceptually understand metadata, it is difficult to stay up-to-date with the technical and practical nuances of metadata creation, even among metadata professionals [19]. As technology becomes more sophisticated and metadata standards proliferate, there is a growing need for researchers to use adaptable schemes in their pipelines to make their data interoperable with machines. Open science mandates from governments and funders will further encourage scholars to house research datasets in open repositories. Data repositories have a role encouraging the adoption of RDF schemes that will make curated data more FAIR.

5.3 Future Research

As discussed in section 4.4, the Phase 4 Refinement will continue to refine the RDF protype. The prototype has already been used to construct a demo REST API to interact with the BGNN dataset [30]. The API provides both a GUI to search the dataset by genus or ARK identifier, as well as command line access using `cURL` and /tt Wget. An API call will download a zip file that contains:

- CSV files containing the metadata associated with each image.
- XML files containing he metadata associated with each image.
- A text document with the preferred citations.
- An OWL file containing the RDF graph.

The future focus of the the MRC-TUBRI collaboration is to continue refining the RDF model and testing the the API. Further investigation into different modes of RDF adoption for data management and metadata creation is needed to understand other database implementations using RDF automatic workflows for creating and managing knowledge graphs.

6 Conclusion

This paper reports on an initiative targeting the development of a flexible metadata pipeline through a collaborative effort involving the MRC-TUBRI. A key contribution is a four-phased approach covering the 1. Assessment of the Problem, 2. Investigation of Solutions, 3. Implementation, and 4. Refinement. The other key contribution is the presentation of the RDF graph prototype. The work presented has been applied to over 300,000 digital images of scientific specimens, specifically fish images, drawn from multiple collections. While we are in the early stage of the RDF graph prototype, the biologist and computer scientists are finding that the workflow and the model expedites their work to service the larger BGNN team in seeking image samples for training the bio-generated neural network. Our next steps include extending our model to other images in the Imageomics institute, given the broad applicability of this work.

As already stated, open data sharing has motivated development of many metadata standards, and a range of metadata models. Indeed these standards aim to ensure smooth operations, whether the goal is resource discovery, support for other aspects of FAIR, or integrating into an AI operation. Metadata is a form of data intelligence, and significant time and money are involved in developing, reviewing, endorsing and implementing standards. With respect to the work reported on in this paper, the initial metadata spreadsheet reviewed was loosely structured around the database containers where the various elements were stored as a result of the pipeline structure. This metadata was roughly organized by metadata creation or modification date. Our four-phased approach and adoption of RDF presents a proof of concept for expressing the metadata elements and their relationship to each other rather than the specific location of the data. This work has helped the team achieve a flexible and extensible metadata pipeline. Our overall conclusion is that the RDF graph prototype and our 4-phased approach is flexible and extensible to the wider variety of analysis of a full range of images being examined in the Imageomics institute. In addition, the proof-of-concept is applicable to other metadata pipelines, and supports computational analysis.

Acknowledgments We thank the Integrated Digitized Biocollections (iDigBio), Global Biodiversity Information Facility (GBIF) and MorphBank data repositories, and the curators of the fish collections in the Great Lakes Invasives Network – Field Museum of Natural History, Illinois Natural History Survey, J. F. Bell Museum of Natural History, Ohio State University Museum of Biological Diversity, University of Michigan Museum of Zoology, and University of Wisconsin-Madison Zoological Museum – for sharing images of their fish specimens with us. We also thank Anuj Karpatne and team at Virginia Tech University who developed and trained the fish feature segmentation ANN component of the workflow, Joel Pepper for automated image quality feature extraction workflow and Bahadir Altintas for developing automated landmark extraction workflow.

Bibliography

[1] FAIR Sharing Standards Registry, https://fairsharing.org/search?fairsharingRegistry=Standard

[2] Introduction to BCO-DMO | BCO-DMO, https://www.bco-dmo.org/

[3] Marine Environmental Research Infrastructure for Data Integration and Application Network, https://meridian.cs.dal.ca/

[4] National Center for Biomedical Ontology BioPortal, https://bioportal.bioontology.org/

[5] Phenoscape, https://phenoscape.org

[6] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information (Nov 2003), http://data.europa.eu/eli/dir/2003/98/oj

[7] EU-funded projects go public www.openaire.eu. MRS Bulletin **37**(8), 714–714 (Aug 2012). https://doi.org/10.1557/mrs.2012.193

[8] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast) (Jun 2019), http://data.europa.eu/eli/dir/2019/1024/oj/eng

[9] DCMI Metadata Terms (Jan 2020), https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[10] Imageomics Institute (Aug 2021), https://imageomics.osu.edu/

[11] Arencibia, E., Martinez, R., Marti-Lahera, Y., Goovaerts, M.: On Metadata Quality in Sceiba, a Platform for Quality Control and Monitoring of Cuban Scientific Publications. In: Garoufallou, E., Ovalle-Perandones, M.A., Vlachidis, A. (eds.) Metadata and Semantic Research, vol. 1537, pp. 106–113. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-98876-0_9

[12] Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M., Messerschmitt, D.G., Messina, P., Ostriker, J.P., Wright, M.H.: Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Tech. rep., National Science Foundation (Jan 2003), https://www.nsf.gov/cise/sci/reports/atkins.pdf

[13] Bailey, C.B., Balakirev, F.F., Balakireva, L.L.: Closing the Gap between FAIR Data Repositories and Hierarchical Data Formats. The Code4Lib Journal **empty**(52) (Sep 2021), https://journal.code4lib.org/articles/16223

[14] Ball, A.: Metadata Standards Directory (Aug 2016), https://www.youtube.com/watch?v=Lh8w2_TpFP8

[15] Ball, A., Chen, S., Greenberg, J., Perez, C., Jeffery, K., Koskela, R.: Building a Disciplinary Metadata Standards Directory. International Journal of Digital Curation **9**(1), 142–151 (Jun 2014). https://doi.org/10.2218/ijdc.v9i1.308

[16] Batista, D., Gonzalez-Beltran, A., Sansone, S.A., Rocca-Serra, P.: Machine actionable metadata models. Scientific Data **9**(1) (2022). https://doi.org/10.1038/s41597-022-01707-6

[17] Brunet, M., Gilabert, A., Jones, P., Efthymiadis, D.: A historical surface climate dataset from station observations in Mediterranean North Africa and Middle East areas. Geoscience Data Journal **1**(2), 121–128 (Nov 2014). https://doi.org/10.1002/gdj3.12

[18] Child, A.W., Hinds, J., Sheneman, L., Buerki, S.: Centralized project-specific metadata platforms: toolkit provides new perspectives on open data management within multi-institution and multidisciplinary research projects. BMC Research Notes **15**(1), 106 (Dec 2022). https://doi.org/10.1186/s13104-022-05996-3

[19] Chuttur, M.Y.: Perceived Helpfulness of Dublin Core Semantics: An Empirical Study. In: Garoufallou, E., Greenberg, J. (eds.) Metadata and Semantics Research. pp. 135–145. Communications in Computer and Information Science, Springer International Publishing, Cham (2013). https://doi.org/10.1007/978-3-319-03437-9_14

[20] Courtot, M., Gupta, D., Liyanage, I., Xu, F., Burdett, T.: BioSamples database: FAIRer samples metadata to accelerate research data management. Nucleic Acids Research **50**(D1), D1500–D1507 (Jan 2022). https://doi.org/10.1093/nar/gkab1046

[21] Dececchi, T.A., Balhoff, J.P., Lapp, H., Mabee, P.M.: Toward Synthesizing Our Knowledge of Morphology: Using Ontologies and Machine Reasoning to Extract Presence/Absence Evolutionary Phenotypes across Studies. Systematic Biology **64**(6), 936–952 (2015). https://doi.org/10.1093/sysbio/syv031

[22] Diamantopoulos, N., Sgouropoulou, C., Kastrantas, K., Manouselis, N.: Developing a Metadata Application Profile for Sharing Agricultural Scientific and Scholarly Research Resources. In: García-Barriocanal, E., Cebeci, Z., Okur, M.C., Öztürk, A. (eds.) Metadata and Semantic Research. pp. 453–466. Communications in Computer and Information Science, Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24731-6_45

[23] Edmunds, R.C., Su, B., Balhoff, J.P., Eames, B.F., Dahdul, W.M., Lapp, H., Lundberg, J.G., Vision, T.J., Dunham, R.A., Mabee, P.M., Westerfield, M.: Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes. Molecular Biology and Evolution **33**(1), 13–24 (Jan 2016). https://doi.org/10.1093/molbev/msv223

[24] Elberskirch, L., Binder, K., Riefler, N., Sofranko, A., Liebing, J., Minella, C.B., Mädler, L., Razum, M., van Thriel, C., Unfried, K., Schins, R.P.F., Kraegeloh, A.: Digital research data: from analysis of existing standards to a scientific foundation for a modular metadata schema in nanosafety. Particle and Fibre Toxicology **19**(1) (Dec 2022). https://doi.org/10.1186/s12989-021-00442-x

[25] Elhamod, M., Diamond, K.M., Maga, A.M., Bakis, Y., Bart, H.L., Mabee, P., Dahdul, W., Leipzig, J., Greenberg, J., Avants, B., Karpatne, A.:

Hierarchy-guided Neural Networks for Species Classification. preprint, Evolutionary Biology (Jan 2021). https://doi.org/10.1101/2021.01.17.427006

[26] Fordham, D.A., Jackson, S.T., Brown, S.C., Huntley, B., Brook, B.W., Dahl-Jensen, D., Gilbert, M.T.P., Otto-Bliesner, B.L., Svensson, A., Theodoridis, S., Wilmshurst, J.M., Buettel, J.C., Canteri, E., McDowell, M., Orlando, L., Pilowsky, J., Rahbek, C., Nogues-Bravo, D.: Using paleo-archives to safeguard biodiversity under climate change. Science **369**(6507), eabc5654 (Aug 2020). https://doi.org/10.1126/science.abc5654

[27] Freire, N., Meijers, E., de Valk, S., Raemy, J.A., Isaac, A.: Metadata Aggregation via Linked Data: Results of the Europeana Common Culture Project. In: Garoufallou, E., Ovalle-Perandones, M.A. (eds.) Metadata and Semantic Research. pp. 383–394. Communications in Computer and Information Science, Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-71903-6_35

[28] Freire, N., Voorburg, R., Cornelissen, R., de Valk, S., Meijers, E., Isaac, A.: Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network. Information **10**(8), 252 (Aug 2019). https://doi.org/10.3390/info10080252

[29] Gallas, E.J., Malon, D., Hawkings, R.J., Albrand, S., Torrence, E.: An integrated overview of metadata in ATLAS. Journal of Physics: Conference Series **219**(4), 042009 (Apr 2010). https://doi.org/10.1088/1742-6596/219/4/042009

[30] tubri github: tubri-github/bgnn_api (Nov 2022), https://github.com/tubri-github/bgnn_API, original-date: 2022-10-12T14:03:39Z

[31] Greenberg, J., White, H.C., Carrier, S., Scherle, R.: A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata **9**(3-4), 194–212 (Nov 2009). https://doi.org/10.1080/19386380903405090

[32] Houssos, N., Stamatis, K., Banos, V., Kapidakis, S., Garoufallou, E., Koulouris, A.: Implementing Enhanced OAI-PMH Requirements for Europeana. In: Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (eds.) Research and Advanced Technology for Digital Libraries. pp. 396–407. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24469-8_40

[33] Houssos, N., Stamatis, K., Koutsourakis, P., Kapidakis, S., Garoufallou, E., Koulouris, A.: Enhanced OAI-PMH services for metadata sharing in heterogeneous environments. Library Review **63**(6/7), 465–489 (Jan 2014). https://doi.org/10.1108/LR-05-2014-0051

[34] Kalogeros, E., Gergatsoulis, M., Damigos, M.: Document-based RDF storage method for parallel evaluation of basic graph pattern queries. International Journal of Metadata, Semantics and Ontologies **14**(1), 63 (2020). https://doi.org/10.1504/IJMSO.2020.107798

[35] Karnani, K., Pepper, J., Bakis, Y., Wang, X., Bart, H., Breen, D., Greenberg, J.: Computational Metadata Generation Methods for Biological Specimen Image Collections (Apr 2022). https://doi.org/10.21203/rs.3.rs-1506561/v1

[36] Leipzig, J., Bakis, Y., Wang, X., Elhamod, M., Diamond, K., Dahdul, W., Karpatne, A., Maga, M., Mabee, P., Bart, H.L., Greenberg, J.: Biodiversity Image Quality Metadata Augments Convolutional Neural Network Classification of Fish Species (Jan 2021). https://doi.org/10.1101/2021.01.28.428644

[37] Leipzig, J., Nüst, D., Hoyt, C.T., Ram, K., Greenberg, J.: The role of metadata in reproducible computational research. Patterns **2**(9), 100322 (Sep 2021). https://doi.org/10.1016/j.patter.2021.100322

[38] Mabee, P.M., Balhoff, J.P., Dahdul, W.M., Lapp, H., Mungall, C.J.: Reasoning over anatomical homology in the Phenoscape KB. In: Proceedings of the 9th International Conference on Biological Ontology (ICBO 2018). p. 2. Corvallis, Oregon, USA (2018)

[39] Manda, P., Balhoff, J.P., Lapp, H., Mabee, P., Vision, T.J.: Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. genesis **53**(8), 561–571 (2015). https://doi.org/10.1002/dvg.22878

[40] Manghi, P., Houssos, N., Mikulicic, M., Jörg, B.: The Data Model of the OpenAIRE Scientific Communication e-Infrastructure. In: Dodero, J.M., Palomo-Duarte, M., Karampiperis, P. (eds.) Metadata and Semantics Research, Communications in Computer and Information Science, vol. 343, pp. 168–180. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35233-1_18

[41] Margaritopoulos, M., Margaritopoulos, T., Mavridis, I., Manitsaris, A.: Quantifying and measuring metadata completeness. Journal of the American Society for Information Science and Technology **63**(4), 724–737 (2012). https://doi.org/10.1002/asi.21706

[42] Michener, W.K.: Creating and Managing Metadata. In: Recknagel, F., Michener, W.K. (eds.) Ecological Informatics: Data Management and Knowledge Discovery, pp. 71–88. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-59928-1_5

[43] Mons, B.: Data Stewardship for Open Science: Implementing FAIR Principles. Chapman and Hall/CRC, New York, 1 edn. (Mar 2018). https://doi.org/10.1201/9781315380711

[44] Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B., Wilkinson, M.D.: Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use **37**(1), 49–56 (Mar 2017). https://doi.org/10.3233/ISU-170824

[45] Nelson, A.: Desirable Characteristics of Data Repositories for Federally Funded Research. Tech. rep., Executive Office of the President of the United States (May 2022). https://doi.org/10.5479/10088/113528

[46] Nordling, L.: Scientists struggle to access Africa's historical climate data. Nature **574**(7780), 605–606 (Oct 2019). https://doi.org/10.1038/d41586-019-03202-2

[47] Park, J.R.: Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Cataloging & Classification Quarterly **47**(3-4) (Apr 2009). https://doi.org/10.1080/01639370902737240

[48] Park, J.R., Tosaka, Y.: Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. Cataloging & Classification Quarterly **48**(8) (Sep 2010). https://doi.org/10.1080/01639374.2010.508711

[49] Pepper, J., Greenberg, J., Bakiş, Y., Wang, X., Bart, H., Breen, D.: Automatic Metadata Generation for Fish Specimen Image Collections (Oct 2021). https://doi.org/10.1101/2021.10.04.463070

[50] Perez, C.I.: The RDA's Metadata Standards Directory: Information Gathering. Master's thesis, University of North Carolina at Chapel Hill (Nov 2013), https://www.rd-alliance.org/sites/default/files/CPerez-RDA-Metadata.pdf

[51] Rettberg, N., Schmidt, B.: OpenAIRE: Supporting a European open access mandate. College & Research Libraries News **76**(6), 306–310 (Jun 2015). https://doi.org/10.5860/crln.76.6.9326

[52] Rockembach, M., Serrano, A.: Climate change and web archives: an Ibero-American study based on the Portuguese and Brazilian contexts. Records Management Journal **31**(3) (Jan 2021). https://doi.org/10.1108/RMJ-11-2020-0039

[53] Schöpfel, J.: Adding Value to Electronic Theses and Dissertations in Institutional Repositories. D-Lib Magazine **19**(3/4) (2013). https://doi.org/10.1045/march2013-schopfe

[54] Soltis, P.S.: Digitization of herbaria enables novel research. American Journal of Botany **104**(9), 1281–1284 (Sep 2017). https://doi.org/10.3732/ajb.1700281

[55] Sterner, B., Elliott, S.: The FAIR and CARE Data Principles Influence Who Counts As a Participant in Biodiversity Science by Governing the Fitness-for-Use of Data (Apr 2022), http://philsci-archive.pitt.edu/21039/

[56] Tsiflidou, E., Manouselis, N.: Tools and Techniques for Assessing Metadata Quality. In: Garoufallou, E., Greenberg, J. (eds.) Metadata and Semantics Research. pp. 99–110. Communications in Computer and Information Science, Springer International Publishing, Cham (2013). https://doi.org/10.1007/978-3-319-03437-9_11

[57] Virkus, S., Garoufallou, E.: Data Science from a Perspective of Computer Science. In: Garoufallou, E., Fallucchi, F., William De Luca, E. (eds.) Metadata and Semantic Research. pp. 209–219. Communications in Computer and Information Science, Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-36599-8_19

[58] Vlachidis, A., Antoniou, A., Bikakis, A., Terras, M.: Semantic metadata enrichment and data augmentation of small museum collections following the FAIR principles. In: Information and Knowledge Organisation in Digital Humanities, pp. 106–129. Routledge (2021), https://doi.org/10.4324/9781003131816-6

[59] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D.: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE **7**(1), e29715 (Jan 2012). https://doi.org/10.1371/journal.pone.0029715

[60] 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.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., 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.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**(1), 160018 (Mar 2016). https://doi.org/10.1038/sdata.2016.18

[61] Wong, E.Y.: Data Documentation Initiative. Technical Services Quarterly **33**(1) (Jan 2016). https://doi.org/10.1080/07317131.2015.1093852