

Request for Information: NASA Public Access Plan

John Beverley, PhD
Vice President of Next Generation Web Division, ScienceCast, Inc.
Vice President, National Center for Ontological Research
Assistant Professor, University at Buffalo

Andrew Jiranek
President of Operations, ScienceCast, Inc.

0. Introduction

NASA's Public Access Plan aims to promote equitable opportunities for researchers regardless of career stage of affiliation, democratize research access by encouraging machine-readability and dropping 12-month access embargos, improve metadata use in the interest of transparency, track fairness of fees and costs associated with research, as well as strategies for software sharing, reuse, and archiving. Pursuit of these commendable goals is, we believe, best facilitated by leveraging existing, vetted, knowledge representation artifacts – ontologies, knowledge graphs - and semantic web technologies. We outline this vision within the context of the five issues helpfully identified in *Request for Information: NASA Public Access Plan for Increasing Access to the Results of NASA-Supported Research*.¹

1. Ensuring Equity in Publication Opportunities for NASA-supported Investigators

Background

NASA's Public Access Plan promotes equity by allowing flexibility in choosing where to publish, allowing submission on platforms like CHORUS, STRIVES, ADS, or PubSpace, regardless of journal publishing model. This flexibility addresses concerns stemming from publishing models disadvantaging groups of researchers. NASA also permits researchers to allocate reasonable publishing costs to their awards.

Within the Public Access Plan, inequity may arise through publication costs and journal credibility. While NASA allows researchers to charge “reasonable” publishing costs to their awards, what counts as “reasonable” may be quite high for certain high-value venues, resulting in researchers exhausting funds on fees. Additionally, the flexibility in choosing where to publish may lead to bias towards established journals, which may disadvantage new, potentially more innovative, journals.

Knowledge Representation Perspective

Knowledge representation has been used effectively, in domains such as bioinformatics, healthcare, industrial manufacturing, and intelligence analysis, to promote data interoperability, standardization, explainability, provenance tracking, and logical rigor. Knowledge representation artifacts, methodologies, and tools can be leveraged to mitigate potential inequities stemming from publication costs. For example, ontologies – logically defined controlled vocabularies of terms and relations representing a domain of interest - can be used to represent general financial information associated with research publishing across publisher, as well as open access charges and ancillary costs. Similarly, knowledge graphs – ontologies combined with concrete data about a domain - can

¹<https://www.federalregister.gov/documents/2023/05/18/2023-10643/request-for-information-nasa-public-access-plan-for-increasing-access-to-the-results-of>

support visualization and analysis of cost distributions across demographics, in the interest of identifying disproportionate burdens on historically disadvantaged groups.

To balance flexibility in choosing where to publish against the potential for inadvertently favoring established journals, knowledge graphs can be employed to visualize publication trends across different platforms. By integrating data from CHORUS, STRIVES, ADS, and PubSpace, we can derive insights into publication patterns and biases. This will allow stakeholders to make informed decisions, ensuring that newer platforms and journals receive equitable attention and support. By comparing their growth and acceptance rates to those of established journals, knowledge representation strategies can be used to identify disparities, explicit and implicit.

2. Improving Equity in Access and Accessibility of Publications

Background

NASA has eliminated the 12-month embargo period and has emphasized the need to make content available in formats suitable for both human readers and automated text processing.

Automated information processing is challenged by the potential for misinterpretation of processed data that may result in inequities, as well as the complexity of scientific content, e.g., equations, figures, etc. that may not translate easily into machine-readable formats.

Knowledge Representation Perspective

Ensuring that content is not only machine-readable but also machine-*interpretable* is paramount. While machine-readability ensures that content can be accessed and processed by automated systems, machine-interpretability ensures that meaning and context are preserved and understood by these systems. This distinction is crucial, especially when dealing with complex scientific content that NASA funds. An effective strategy to achieve machine-interpretability is via a semantic layer over data sources, in the form of ontologies or knowledge graphs, which represent data and the relationships and context surrounding that data. Roughly speaking, while a machine-readable format might allow a system to recognize an equation, a machine-interpretable format would enable the system to understand the significance and application of that equation within a broader scientific context.

Adopting established ontology architectures, such as the Basic Formal Ontology (BFO) - the ISO 21838-2 top-level standard for ontology development - can provide a solid foundation for this semantic layer. That said, just adopting an architecture isn't enough. Robust and reliable ontologies require careful curation of both textual and logical definitions of terms and relations used to represent a given domain, involving consensus-building exercises with subject-matter experts to ensure that the ontology accurately represents the domain of knowledge it covers.

Consensus-building serves a dual purpose. First, it ensures ontologies are comprehensive and accurate, capturing all nuances of the domain. Second, they act as a validation mechanism, where interpretations are rigorously tested for coherence, accuracy, and practicality. By grounding the ontology in the expertise of domain specialists, we can ensure that automated systems using the ontology for interpretation are drawing from a well of trusted and validated knowledge. In this manner, challenges posed by the complexity of scientific content and the potential for misinterpretation can be effectively addressed, ensuring that NASA funded research is not only accessible to machines but also meaningfully interpreted by them.

3. Monitoring Evolving Costs and Impacts

Background

NASA intends to proactively track the progression of publication fees and policies and is seeking insights on efficient strategies to observe these trends, particularly concerning publishing equity.

Monitoring the evolving costs and impacts on affected communities presents several challenges. Distinct publishers have a myriad of fee structures, making direct comparisons problematic. The lack of transparency in some publishing fee structures can obscure actual costs, while the absence of a standardized definition for "equity" can lead to inconsistent assessments. Changes to publication policies exacerbate these issues.

Knowledge Representation Perspective

Here too knowledge representation can effectively supplement existing technologies to aid tracking and the identification of trends. In particular, ontologies may be developed that encapsulate various publisher metrics – such as base fees, discounted fee schedules, etc. – compare fee schedule differences across publishers and researcher demographics. Because knowledge representation artifacts are often used to represent provenance, spatial, and temporal data, trends can be identified over time and geographic region. Leveraging code libraries – such as RDFLib - designed to interact with knowledge representation artifacts, moreover, can facilitate monitoring for when publishers change fee structures or policies as they are updated.

With a rich dataset in place, SPARQL queries can be engineered to extract meaningful trends. These queries can unravel patterns, highlight anomalies, and discern disparities in publication fees and opportunities. To make this information accessible for NASA representatives, building on knowledge representation successes in other fields - a dedicated dashboard can be developed, to visualize policy and fee differences, similarities, and updates.

4. Increasing Findability and Transparency of Research

Background

NASA is exploring ways to enhance the findability and transparency of research, with emphasis on exploring use of PIDs and metadata.

The lack of standardized metadata structures can result in incomplete or misleading representation of data. Diverse naming conventions and identifier formats may cause duplication or misidentification of research outputs. Integrating new PIDs with existing systems sometimes results in compatibility issues, and the varied acceptance and recognition of different identifiers can hamper their universal adoption and trustworthiness.

Knowledge Representation Perspective

Inconsistent terminologies across different platforms and databases are a significant concern, but can be mitigated by appealing to knowledge representation solutions. For example, ontologies can be leveraged to serve as standardized vocabularies, ensuring that even if diverse terminologies are used, they can be mapped back to a consistent set of terms. These ontologies can then be extended to knowledge graphs constructed in the interest of linking researchers, their publications, datasets, etc. to PIDs. In effect, this strategy would provide a centralized semantic layer through which such information can be connected, which would in turn facilitate the identification and remedying of incomplete or misleading data representations.

Moreover, publicly-available semantic annotation tools and standards can be leveraged to address the diversity of naming conventions, and indeed to identify format duplication and misidentification. When such annotation tools are used in concert with ontologies and knowledge graphs, they improve ease of access and findability, as research content can be easily queried. Additionally, semantic web technologies can streamline the integration of new PIDs with existing

systems, by mapping them to existing ontologies or data formats and evaluating such updates for compatibility issues in real-time. Lastly, by fostering consensus among researchers as to the content of these knowledge representation artifacts, and through promoting adoption of these semantic technologies, the trustworthiness of systems using them will be bolstering.

5. Suggestions on Sharing and Archiving of Software

Background

NASA is exploring ways to enhance software archiving, sharing, and maintenance for reuse, considering platforms like GitHub and Zenodo.

Knowledge Representation Perspective

Establishing knowledge representation artifacts that represent software metadata promises to aid archiving, sharing, and maintenance for reuse strategies. This includes modeling phenomena such as version history, code dependencies, usage guidelines, and licensing requirements. Ontologies and knowledge graphs used in such a manner would provide insights into the lifecycle of a given piece of software, as well as to its suitability for reuse. Consensus-building exercises – so integral to good ontology and knowledge graph design practices – can ensure that the relevant knowledge representation artifacts are comprehensive and align with real-world development and maintenance nuances.

Improving discoverability may also be facilitated by knowledge representation, for example, by the semantic annotation of software repositories. Integrating RDF generators into prominent platforms like GitHub and Zenodo ensures that every software action, be it an upload or an update, can be complemented by associated semantic data, tracking provenance, maintenance requirements, etc. The ultimate vision here would be a centralized semantic layer which pools software metadata from various sources, accompanied by pre-built, sophisticated SPARQL querying capabilities providing researchers ways to access explicit and implicit information about the software lifecycle.

6. Conclusion

By leveraging ontologies and semantic web technologies, NASA can effectively monitor publication trends and enhance software sharing and archiving practices. These solutions not only provide a technological edge but also ensure that the processes are transparent, equitable, and efficient. Regular engagement with the research and developer communities will further refine these strategies, ensuring they remain relevant and effective.