Interlinking Datasets with Content: Perspectives on Next Steps

Martin Halbert

Thursday, June 27, 2024

CHORUS Forum: Linking Datasets to Content

Presentation Overview

- About NSF Public Access Initiative
- The Complexities of Linking Content with Data
- Context of Nelson Memo
- Agency policy & researcher burden considerations in linking Datasets and Content
- Considerations in possible applications of AI to Data/Content interlinkages
- This is also an international issue of aligning standards and national PID strategies

About NSF Public Access Initiative

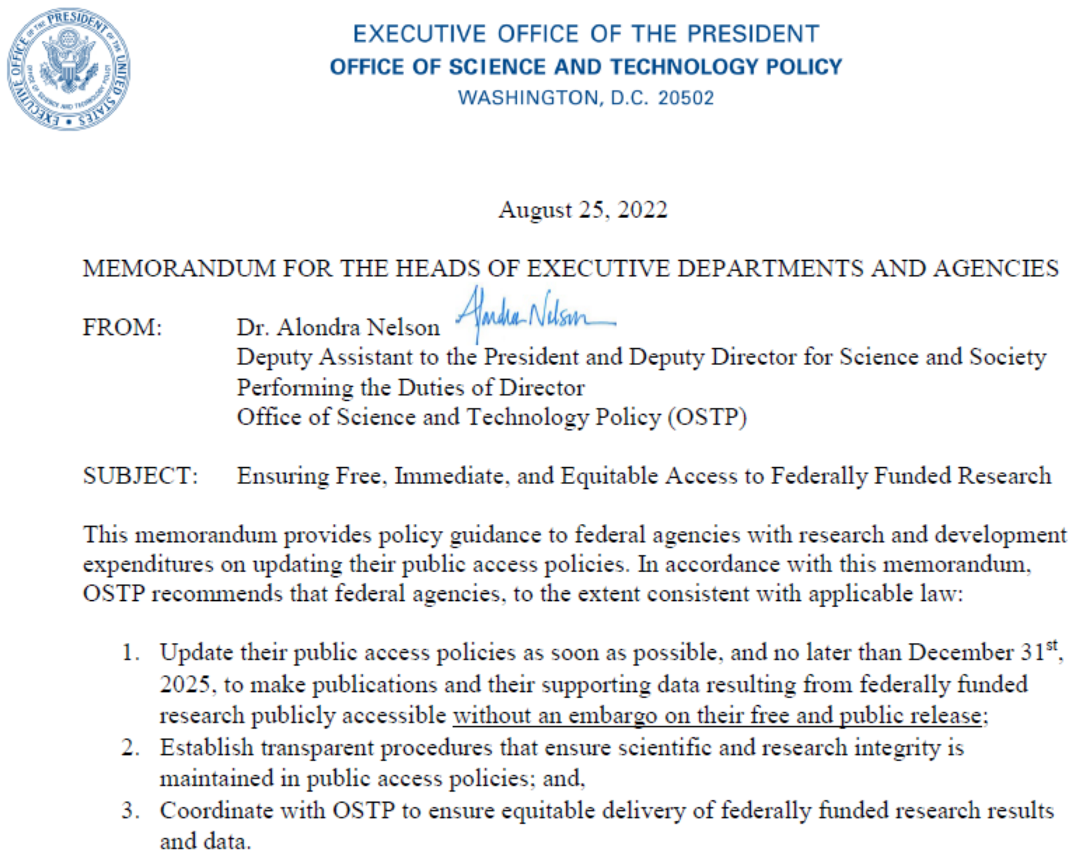
- The NSF PAI encompasses the agency's efforts to ensure that publicly funded research outputs are made publicly accessible
- The Science Advisor for Public Access has broad coordinating responsibilities for this area across the NSF and in the interagency space
- Now working on implementation of the OSTP Nelson Memo requirements and various associated enhancements to the NSF Public Access Repository and our annual reporting processes

Complexities of Linking Content with Data

- There are many potential relationships between various kinds of content and data (36 distinct relationtypes in DataCite metadata schema 4.5!)
- All of these types of inter-relationships are semantically meaningful, useful, and powerful for scientific understanding
- However, such complex interlinkages present a number of practical challenges

Context of the Nelson Memo

- The 2022 White House Office of Science & Technology Policy memorandum issued by Dr. Alondra Nelson directs federal research funding agencies to implement new Public Access requirements for data & content in two phases:
  - Section 3 – 2025
  - Section 4 – 2026
- The Memo also encourages agencies to consider a number of additional measures



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

Nelson Memo Section 4

- The second phase of implementation is significantly more complex in terms of metadata and interlinkage requirements
- 4(a) sets higher quality control requirements for eight metadata elements assigned to the two types of research products the memo is focused on, articles and the data underlying the articles
- 4(b) tells agencies to “Instruct federally funded researchers to obtain a digital persistent identifier”, otherwise known as an ORCID
- 4(c) directs agencies to assign persistent identifiers (DOIs) to awards, enabling awards to be referenced via machine-readable citations
- Considered as a whole, the purpose of Section 4 effectively leads to the concept of an end-to-end machine-readable knowledge graph for all scientific articles and data arising from federal research funding awards

But, Recording all this Metadata takes Work

- And work is not free; it entails labor & cost
- What is the most effective / reasonable amount of metadata effort (aka researcher reporting burden) to expend / require?
- Can we assess the beneficial impacts of implementing FAIR data principles and Open Science practices?
- Who is the most appropriate type of researcher to create & manage metadata? Principle Investigator? First-year Graduate Student? Data Scientist? Repository Managers? What training / preparation is required?

Could we Use New AI Tools to Create Data/Content Interlinkages?

- Possibly, but what are the ramifications of that?
- AI “hallucinations” (incorrect assertions) are real; do we want to inject such errors into knowledge graph citation networks? Can we tolerate that?
- If we don’t use AI tools, who is going to encode all these inter-relationships? Can we afford that?

This is also an International Challenge for Alignment of Standards, National PID Strategies, and Citation Expectations

To be effective, these kinds of scenarios will require cultivating an emerging consensus of practices in the world scientific community (no trivial task!)

Core intermediaries (such as DataCite) can assist in the process of such alignment, but it will require concerted efforts by all stakeholders (scientists, funders, and publishers)