Cross-Linking DOI Author URIs Using Research Networking Systems

Nick Benik¹, Timothy Lebo², Griffin M. Weber¹
¹ Harvard Medical School, Boston, MA; ² Rensselaer Polytechnic Institute, Troy, NY

Background

Many academic publications are assigned a universally unique identifier called a Digital Object Identifier (DOI). DOIs can be entered at http://doi.org to find publications regardless of their location on the web.

The doi.org website also caters to semantically-aware web agents by returning metadata in RDF format describing the document identified by the DOI. These RDF metadata documents include unique author URIs generated by doi.org for each author-publication combination. This means that a person will have a different author URI for each of his or her publications, making each author URI a semantically disparate entities.

Many institutions are adopting researcher networking (RN) systems to allow their researchers to manage their personal web page and biosketch online. Several of these RN systems are based on semantic web technologies and use the VIVO ontology when exposing their researchers’ profile information in RDF format. These sites are an excellent source of semantic data containing curated lists of researchers’ publications. VIVO is an NIH-funded ontology built to be a standard for representing the scholarly activities of scientists.

The evolution of the semantic web, along with its various ontologies and growing amounts of published data, would logically allow semantically-aware intelligent agents to generated original datasets containing new assertions based on existing triple data.

Method

A semantically-aware web spider crawled RN websites to collect Linked Data documents describing researchers’ publications. This information was then passed through a custom web service that attempts to populate any missing DOI information using a publication's title, PubMed / PubMed Central ID or other identifiers. Linked Data was then retrieved for publications with DOIs from doi.org. Identification triples (prov:specializationOf) are then generated, stating that the author URIs from doi.org and the RN profile URI represent the same person. Provenance information is then added to the newly inferred data using PROV-O ontology and it is segmented and packaged using the VoID vocabulary.

Results

Linked Data documents for publications were extracted from different VIVO-compliant RN sites. From those publications' documents, DOI references were found for many of them. Each publication containing a DOI had its corresponding Linked Data retrieved from doi.org. Matching author URIs were identified from this data using fuzzy matching and prov:specializationOf records were generated to 1) state that the doi.org author URI and the RN author/person URI represent the same person, 2) to state that the document described on the RN site is the same publication described in the DOI.org site. This new dataset of cross-linking triples is publicly available at the address below.

A visualization tool at http://research.hackerceo.org/ANGEL/Datasets/VIVO2DOI allows researchers to explore the value these new cross-links provide to the larger Linked Data community.

Data Processing Pipeline

- Extract Person RDF from RN sites
- Extract each Person's Publication RDF
- Extract or lookup each Publication's DOI
- Extract Publication Metadata from doi.org
- Crawl PubMed to create publication-to-DOI reference DB
- DOI References
- Segment results using VoID and document provenance using PROV-O
- Publish as Linked Open Data
- Hard-Match RN and DOI Publication RDF Documents
- Fuzzy-Match RN Author with DOI Authors in each Publication's RDF Documents

This project was funded by NSF SciSIP Award #1238469, NIH grant 8UL1TR000170-05, and Harvard University and its affiliated academic health centers.

The original RDF dataset generated by the process described is available at: DOI.org Author URI Disambiguation Dataset. Nick Benik. Figshare. http://dx.doi.org/10.6084/m9.figshare.96533