Storing Bibliographic Data in Multiple Formats with the NPDS Cyberinfrastructure

S. K. Taswell, K. Uhegbu, S. Mashkoor, S. Dutta, and C. Taswell
Brain Health Alliance, Ladera Ranch, California, USA

***Abstract*—The PORTAL-DOORS Project (PDP) aims to develop the Nexus-PORTAL-DOORS-Scribe (NPDS) Cyberinfrastructure as a distributed network of data repositories that communicate with each other using a common message exchange standard. These data repositories include a collection of servers with a system of registries, directories, and diristries for diverse resources including bibliographic information records. Examples of resource metadata representations can be viewed at PDP participating websites. Until now, PDP has not supported convenient import or export of bibliographic records to or from any of the common bibliographic standards. In this report, we describe our progress on our new PDP utilities for interoperability between the format for NPDS records and various bibliographic formats such as BIBFRAME, MARC, RIS, and BibLaTeX. We will detail the import process when using a converter that transforms bibliographic citations in other formats and stores them in an NPDS diristry. Improved interoperability for conversion between bibliographic records in other traditional formats with the NPDS format will support a variety of use cases that require either lexical and/or semantic parsing of cited references.**

***Index Terms*—BIBFRAME, MARC, RIS, BibLaTeX, NPDS Cyberinfrastructure, PORTAL-DOORS Project, bibliographic data, citation format converter, lexical web, semantic web.**

I. Introduction

The semantic web provides a network of information intended for machine learning and understanding data and metadata shared with semantic markup. However, there remains a challenging problem with inadequate software tools for this semantic markup with applications enabling interoperability between lexical and semantic data. One ongoing project to address this problem has been the PORTAL-DOORS Project, which develops the Nexus-PORTAL-DOORS-Scribe (NPDS) Cyberinfrastructure (Taswell, 2014). With its diristry-registry-directory system, descriptive metadata records are stored in their corresponding problem-oriented registries and directories (Taswell, 2008). An important and desired use case for the NPDS Cyberinfrastructure has been application to bibliographic records with easy export of NPDS records to other citation formats, and conversely, the easy import of bibliographic records into the NPDS repositories. However, if an NPDS system user wanted to access and cite a metadata record in a research paper, the user would not be able to do so easily due to a current lack of import/export utilities for PDP and NPDS. As an example, when writing papers with the LaTeX document preparation system, most authors prefer using BibTeX or BibLaTeX, with a bibliography format that allows for the creation, storage, and management of citations in various bibliographic styles (Patashnik, 1984). By devising conventions with a structured approach and mapping to convert readily between the formats for BibTeX, BibLaTeX, generic citation metadata, and NPDS, it will become easier to import and export bibliographic information to and from NPDS repositories.

II. Bibliographic Formats

Three of the major bibliographic formats commonly in use today are MARC, RIS, and BibTeX. MARC, or Machine Readable Cataloging, is a set of standards made by the Library of Congress for cataloging bibliographic data that is internationally recognized since the 1970s (Rudi and Surla, 2009). There are several different versions of MARC used today including MARC 21 and UNIMARC (Hopkinson, 2008; Das, 2004). Although MARC 21, UNIMARC, and other MARC variants have been fully supported for many years, a more recent effort called BIBFRAME from the Library of Congress has begun the process of updating the older system of MARC to a more modern world of the semantic web (Xu et al., 2018). Another major bibliographic data format is RIS by Research Information Systems (Reuters, 2012). For the BibTeX format, each record consists of the reference type, a citation key, and then a list of fields with their field names and field values. Each of these bibliographic data formats can be abstracted to contain a reference type with a list of different fields each representing some particular kind of metadata. Various organizations support web APIs that allow users to import citation metadata via software tools from their databases. Examples include IEEE Xplore, NLM PubMed, and Unpaywall.

III. Mappings Between Formats

In order to move towards full interoperability between all of the major bibliographic formats, we have continued development on the NPDS Cyberinfrastructure by addressing the import and export of bibliographic records. Since each format contains different metadata specifications, a unique mapping must be created to translate fields with data from each format into their appropriate NPDS fields. There are three main approaches that PDP-Aoraki software uses to convert a bibliographic metadata record into an NPDS record:

- MINIMAL redundancy mapping: store the entire bibliographic metadata record intact in the OtherText field of an NPDS record and generate a PrincipalTag using either the citation key or an acronym extracted from the title.

- MODERATE redundancy mapping: also stores DOI, ISBN, and other identifiers in the CrossReferences fields of an NPDS record, as well as the locations of ecopy instances of the reference if online, or physical addresses of physical copies if offline.
- MAXIMAL redundancy mapping: also parses the bibliography metadata record into all of the lexical PORTAL fields and semantic DOORS fields of the NPDS record.

Because the entire original bibliographic metadata record has been retained in the OtherText field of the NPDS record, it remains possible to re-parse the imported metadata from minimal, to moderate, to maximal, in an idempotent manner. This approach enables re-parsing with new semantic parsing algorithms as they become available. It also enables avoiding the conversion and parsing from minimal to maximal redundancy when data storage space remains costly. To accompany these mappings, originally designed for import of BIBTEX metadata records, we have also developed utilities for importing records from generic bibliographic metadata records by retrieval with identifiers including the digital object identifiers (DOI) from the doi.org service, and what we have generically defined and called service-unique identifiers (SUI) for direct imports (without a DOI) from various citation services such as IEEE Xplore and NLM PubMed databases. Note that these services maintain their own unique identifiers. For example, PubMed maintains NLM identifiers including pmid and pmcid, whereas Xplore maintains an IEEE article number. Therefore, the benefit and utility for our software to support what we have called an SUI. Finally, we have implemented initial versions of import utilities for RIS and MARC metadata records.

IV. Current Status and Future Work

Any registered user with author access who wishes to import and convert bibliographic metadata records into NPDS metadata records may do so at participating PDP web sites (including www.PORTALDOORS.net, www.BrainHealthAlliance.net, www.TeleGenetics.net). In addition to the fields described in Table I, the necessary NPDS EntityLabels are generated automatically to identify and access the NPDS metadata record. Users can also import records from other databases such as IEEE Xplore, NLM PubMed, Unpaywall, and the loc.gov MARC services. Our continuing and future work on these bibliographic import and export utilities for the NPDS Cyberinfrastructure will focus on improving their robustness and stability, while maintaining their interoperability with all major bibliographic formats.

References

Das D. (2004). Marc 21: The standard exchange format for the 21 st century.

Hopkinson A. (2008). *Unimarc manual: Bibliographic format* (Vol. 36). Walter de Gruyter.

Patashnik O. (1984). BIBTEX 101. *TUGboat*.

Reuters T. (2012). Ris file format — researcherid.com. https://web.archive.org/web/20170707033254/http://www.researcherid.com/resources/html/help_upload.htm

Rudi G., & Surla D. (2009). Conversion of bibliographic records to marc 21 format. *The Electronic Library*.

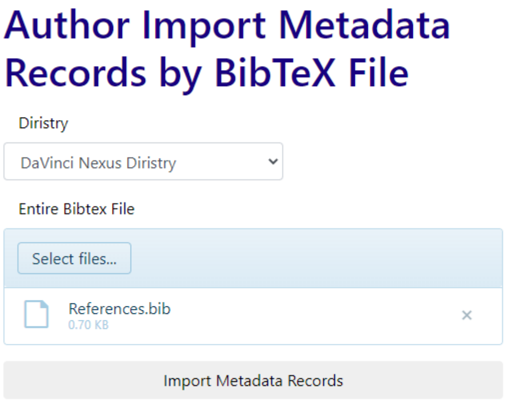
Taswell C. (2008). Doors to the semantic web and grid with a portal for biomedical computing. *IEEE Transactions on Information Technology in Biomedicine*, *12*(2), 191–204.

Taswell C. (2014, November 11). *Management of multilevel metadata in the PORTAL-DOORS system with bootstrapping* (U.S. pat. No. 8,886,628).

Xu A., Hess K., & Akerman L. (2018). From marc to bibframe 2.0: Crosswalks. *Cataloging & Classification Quarterly*, *56*(2-3), https://doi.org/10.1080/01639374.2017.1388326, 224–250. https://doi.org/10.1080/01639374.2017.1388326

Table I: Citation to NPDS field mappings for three different modalities with minimal, moderate, and maximal redundancy

NPDS Field	Citation Field to NPDS (min)	Citation Field to NPDS (mod)	Citation Field to NPDS (max)
PrincipalTag	CitationKey or TitleAcronym	CitationKey or TitleAcronym	CitationKey or TitleAcronym
Name	Title	Title	Title
Nature	Keywords	Keywords	Keywords
OtherTexts	Entire citation record	Entire citation record	Entire citation record
CrossReferences	—	DOI, ISBN, other identifiers	DOI, ISBN, other identifiers
Locations	—	Ecopy URLs, other addresses	Ecopy URLs, other addresses
Descriptions	—	Abstract (lexical)	Abstract (semantic)
Provenances	—	—	Citations, origins reported
Distributions	—	—	Licensing, permissions


Author Import Metadata Records by BibTeX File

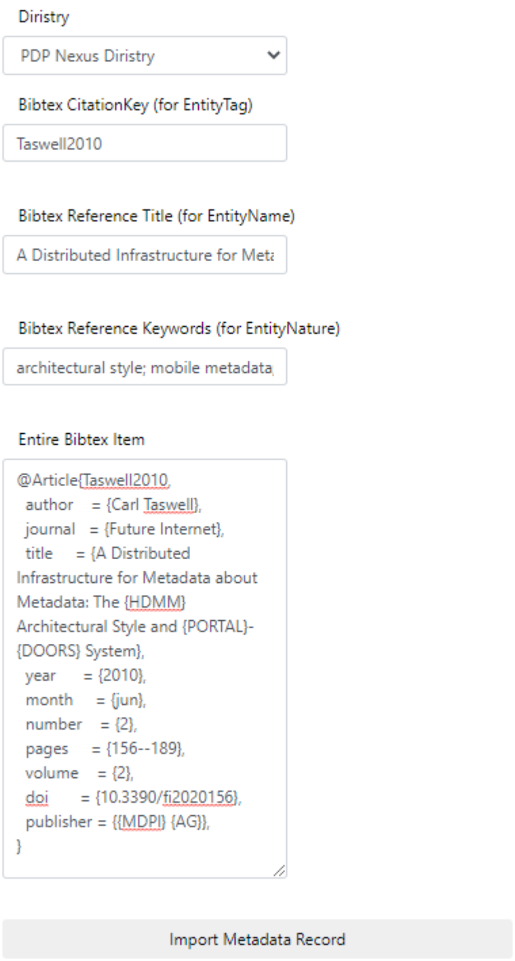
Diristry

DaVinci Nexus Diristry

Entire Bibtex File

References.bib
0.70 KB

Figure 1: BibTeX file import form.



Diristry

PDP Nexus Diristry

Bibtex CitationKey (for EntityTag)

Taswell2010

Bibtex Reference Title (for EntityName)

A Distributed Infrastructure for Met

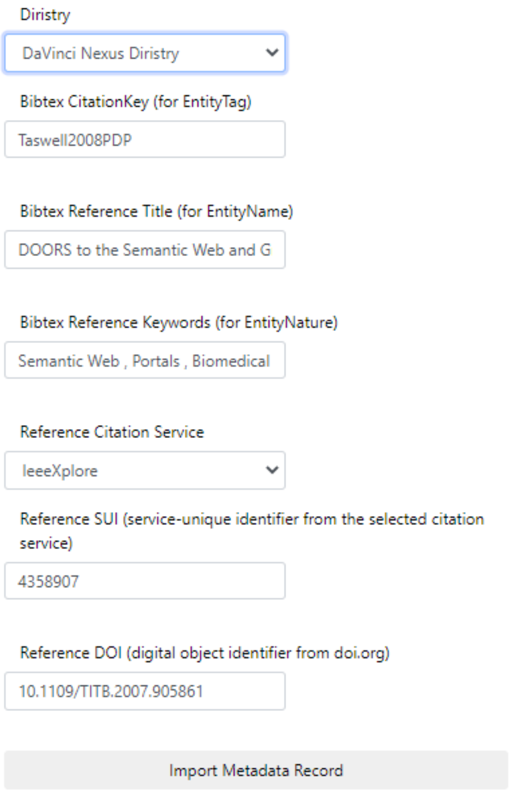
Bibtex Reference Keywords (for EntityNature)

architectural style; mobile metadata

Entire Bibtex Item

```
@Article{Taswell2010,
  author    = {Carl Taswell},
  journal   = {Future Internet},
  title     = {A Distributed Infrastructure for Metadata about Metadata: The {HDMM} Architectural Style and {PORTAL}-{DOORS} System},
  year      = {2010},
  month     = {jun},
  number    = {2},
  pages     = {156--189},
  volume    = {2},
  doi       = {10.3390/fi2020156},
  publisher = {{MDPI} {AG}},
}
```

Figure 2: BibTeX record as form field data before import.



Diristry

DaVinci Nexus Diristry

Bibtex CitationKey (for EntityTag)

Taswell2008PDP

Bibtex Reference Title (for EntityName)

DOORS to the Semantic Web and G

Bibtex Reference Keywords (for EntityNature)

Semantic Web , Portals , Biomedical

Reference Citation Service

IeeeXplore

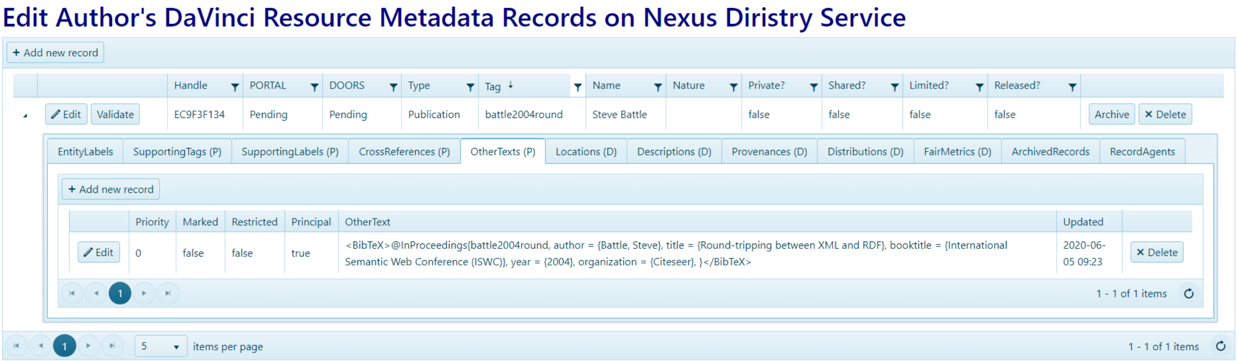
Reference SUI (service-unique identifier from the selected citation service)

4358907

Reference DOI (digital object identifier from doi.org)

10.1109/TITB.2007.905861

Figure 3: Metadata record imported via DOI and/or SUI from a citation service.



Edit Author's DaVinci Resource Metadata Records on Nexus Diristry Service

		Handle	PORTAL	DOORS	Type	Tag	Name	Nature	Private?	Shared?	Limited?	Released?	
	Edit Validate	EC9F3F134	Pending	Pending	Publication	battle2004round	Steve Battle		false	false	false	false	Archive ✕ Delete

	Priority	Marked	Restricted	Principal	OtherText	Updated	
Edit	0	false	false	true	<BibTeX>@InProceedings{battle2004round, author = {Battle, Steve}, title = {Round-tripping between XML and RDF}, booktitle = {International Semantic Web Conference (ISWC)}, year = {2004}, organization = {Citeseer}, }</BibTeX>	2020-06-05 09:23	✕ Delete

1 - 1 of 1 items

5 items per page

1 - 1 of 1 items

Figure 4: BibTeX metadata record embedded as OtherText field after import to an NPDS metadata record.