The Linkable Neil Armstrong: Using BIBFRAME to Increase Visibility of Digital Collections

Carolyn Hansen
University of Cincinnati
United States
carolyn.hansen@uc.edu

Sean Crowe
University of Cincinnati
United States
sean.crowe@uc.edu

Abstract

This report describes the initial phase of an experimental project to increase Web visibility of the Neil Armstrong Commemorative Archive, a digital collection of archival materials concerning astronaut Neil Armstrong's tenure at the University of Cincinnati. The project description includes explanation of the mapping process from Qualified Dublin Core to BIBFRAME as well as data reconciliation and linking to external authorities such as id.loc.gov, VIAF, and Wikipedia. Next steps in the project, such as integrating related MARC datasets from local library catalogs, are also discussed.

Keywords: linked data, BIBFRAME, Dublin Core, metadata, digital collections

1. Introduction

Neil Armstrong, celebrated astronaut and the first person to walk on the moon, was also a professor of aerospace engineering at the University of Cincinnati (UC). In October 2013, the UC Libraries' Digital Collections and Repositories Department published the Neil Armstrong Commemorative Archive, a digital collection of unique archival materials concerning Armstrong's tenure at UC.[1] The collection contained two hundred and eighteen items, including letters, photographs, artifacts, and ephemera. Although the collection was extensively described using established information standards such as the Qualified Dublin Core (DC) metadata standard and Library of Congress Name and Subject Authority Headings (LCNAF and LCSH, respectively), its discoverability outside of library catalogs and repositories was limited by the structured metadata schemas that those systems required. In order to capitalize on the power of linked open data to improve the collection's visibility on the Web, an experimental project was undertaken by UC library faculty to map the original DC metadata to the Bibliographic Framework (BIBFRAME) data model, reconcile and link the data to external authorities using the OpenRefine application, and publish the data as expressed in the Resource Description Framework (RDF).

This report will describe the initial phase of the project, including explanation of the mapping process from Qualified Dublin Core to BIBFRAME as well as data reconciliation and linking to external authorities such as id.loc.gov, the Virtual International Authority File (VIAF), and Wikipedia. In addition, next steps in the project, such as integrating related MARC datasets from local library catalogs, will be discussed.

2. Methodology: Metadata for Discovery

Although data is often considered the unbiased product of research, the environment in which it is created and stored impacts its content, structure, and meaning. In this project, the original dataset consisted of Qualified DC records created for UC's DSpace repository, the Digital Resource Commons (DRC).[2] For purposes of this project, the original records were

[1] https://drc.libraries.uc.edu/handle/2374.UC/713357
[2] https://drc.libraries.uc.edu/

https://doi.org/10.23106/dcmi.952137143

conceptualized as both metadata (abstract representations of digital objects) and data (a set of elements and values generated during the cataloging process). Viewing the metadata in the context of the larger dataset impacted the mapping approach from DC to BIBFRAME; specifically, a lossless migration between standards was not sought. Instead of focusing on comprehensive or archival mapping that preserved the authenticity and content of the original data in a one-to-one mapping, a flexible approach was taken in which a core set of properties (see Table 1) needed for discovery were identified and mapped.

TABLE 1. Discovery Metadata Mapping

Qualified DC (UC Armstrong Collection)	**Simple DC[3]**	**BIBFRAME Core Class**	**BIBFRAME Property**
dc.contributor dc.contributor.author dc.contributor.photographer dc.contributor.other	dc.contributor	bf:Work bf:Authority	bf:contributor
dc.date.available	dc.date	bf:Instance	bf:providerDate
dc.identifier.uri	dc.identifier	bf:Instance or bf:Annotation[4]	bf:uri
dc.publisher.digital	dc.publisher	bf:Instance	bf:providerName
dc.subject dc.subject.lcsh	dc.subject	bf:Work bf:Authority	bf:subject
dc.title	dc.title	bf:Work bf:Authority	bf:title
N/A	N/A	bf:Instance	bf:providerPlace

There are many benefits to a "metadata for discovery" approach.[5] First, being able to omit properties from the mapping provides a clean dataset without idiosyncratic data. For example, UC's DRC repository contained legacy data that conformed to outdated OhioLINK consortial practices (see examples of non-mapped properties from the original dataset in Appendix I); this data did not increase discoverability or add value in a linked data environment. Second, since BIBFRAME is an emerging model that is relatively unstable, eliminating properties that are not crucial for discovery reduces the amount of data cleanup needed as the model changes. Third, creating a lightweight dataset for discovery is time-efficient, allowing for mapping alterations to be made on the fly. Lastly, mapping for discovery simplifies working with multiple instances of physical objects and digital surrogates. Instead of accounting for the various instances of one work, focus can be placed on the digital surrogate. For example, the Armstrong dataset contained a digital surrogate for a photograph that had three instances in its lifecycle: it was created by the

[3] For description of the 15 properties of Simple DC, see: http://dublincore.org/documents/dces/

[4] The BIBFRAME model and vocabulary are still being defined and there is room for interpretation in how to conceptualize and map certain properties. In the Armstrong sample data, bf:uri property is entered under the bf:Instance class, but a case could also be made to enter it under bf:Annotation.

[5] The authors acknowledge that this is only possible if there is an existing system to store and make the comprehensive records accessible.

https://doi.org/10.23106/dcmi.952137143

photographer,[6] published as a reproduction in a magazine, and published as a digital surrogate in the Neil Armstrong Commemorative Archive. Archival materials and museum objects often have multiple stages in their lifecycles, which can be difficult or cumbersome to express in BIBFRAME, since the objects differ from traditional forms of publication. By relying on existing platforms and specialized descriptive standards such as Encoded Archival Description (EAD) or the Visual Resources Association (VRA) standard for comprehensive description, BIBFRAME mappings can be simplified. This is a significant distinction; to paraphrase Nancy Fallgren, Metadata Specialist Librarian at the National Library of Medicine (NLM), "MARC became a descriptive scheme in addition to an encoding standard. We should not do that with BIBFRAME."[7]

3. Mapping Dublin Core to BIBFRAME

BIBFRAME was initially created as a replacement standard for MARC, but it has been advertised as a more inclusive model that can accommodate a broader user community. This may be true in the future; however, working with BIBFRAME outside of text-based materials and MARC record migration is challenging in the current environment. Part of the problem is that this work is very new and there are few example datasets available from BIBFRAME early adopters. The datasets that are available via LC's BIBFRAME website[8] all focus on MARC record migration using LC's Transformation Tool. For those working with non-MARC metadata, digital collections, or archival materials described at the item level, these datasets are of limited assistance. To the authors' knowledge, the mapping process described in this report is the first to work with BIBFRAME and non-MARC metadata in digital collections.

The first step in the mapping process was to eliminate idiosyncratic properties from the dataset. For example, the DC properties referring to events in the lifecycle of the physical object such as dc.date.created were removed (see Appendix I for a full list of unmapped properties). Then, the remaining properties that could not be expressed in BIBFRAME as Uniform Resource Identifiers (URIs) were isolated and examined. If these properties provided information that was structurally important for the BIBFRAME core classes, they were retained.[9] Next, the remaining DC properties were mapped to the four core BIBFRAME classes: Work, Instance, Annotation, and Authority (see Table 1.1 for discovery metadata mapping and Table 1.2 for visual representation of the BIBFRAME model). Initial mapping to the core classes helped to intellectually organize the data; this was important when working with BIBFRAME data serialized as RDF because the heavy use of URIs made the RDF difficult to read. Finally, the DC properties were mapped to corresponding BIBFRAME properties. As a result of the "metadata for discovery" approach in the mapping and the work-centric nature of BIBFRAME, this project used few properties that mapped to BIBFRAME Instance or BIBFRAME Annotation. In this dataset, BIBFRAME Instance was used to describe the publication of the digital surrogate on UC's DRC Repository; it was not used to represent earlier publications or other events in the object's lifecycle (for more information on describing archival materials and digital surrogates, see Section 2).

[6] The act of creation could be considered two separate instances. The physical act of taking the photograph could be one instance, and the development of the film into a print would be the second. For simplicity, the authors define this as one instance.

[7] This is paraphrased from Nancy Fallgren's presentation "Experiments in BIBFRAME: A Modular Approach." given at the American Library Association Midwinter Meeting in January 2015.

[8] http://www.loc.gov/bibframe/implementation/

[9] Currently, www.bibframe.org does not specify input requirements for properties.

https://doi.org/10.23106/dcmi.952137143

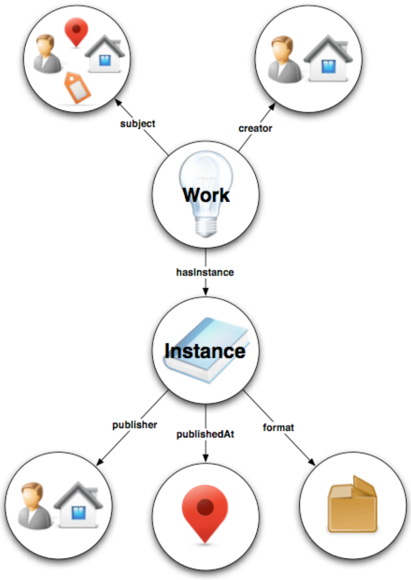


FIG. 1. "BIBFRAME model" by Zepheira, under contract from the U.S. Library of Congress - http://www.loc.gov/bibframe/docs/images/bibframe.png. Licensed under Public Domain via WikimediaCommons http://commons.wikimedia.org/wiki/File:BIBFRAME_model.png#/media/File:BIBFRAME_model.png

4. Defining Authorities

In traditional description, authorities refer to controlled vocabularies that bring together variant forms of a name for people, organizations, subjects, etc. The BIBFRAME concept of authority is more flexible; it is defined only as "Representation of a key concept or thing."[10] In practice, this representation is expressed as both strings (bf:label, bf:titleValue, bf:authorizedAccessPoint) and things (bf:creator, bf:subject, bf:title) in the form of URIs. There are currently no guidelines or best practices regarding what constitutes a reliable authority in terms of site content, although stable URIs are needed from the technical perspective. In this project, LC authorities were contained in the original DC dataset, so id.loc.gov was used as the primary authority (bf:hasAuthority). Reference authorities (bf:referenceAuthority) included VIAF, Wikipedia, and organizational websites for corporate bodies (see Table 2 for authority mapping).

TABLE 2. Authority Mapping

BIBFRAME Property	**Authority Used**
bf:hasAuthority	LCNAF; LCSH
bf:referenceAuthority	VIAF; Wikipedia; organizational websites

5. Linking and Data Reconciliation

The authors investigated several enrichment services for reconciling the core metadata set. Since the DC metadata included LC subject headings and names, the team first explored the possibility of using id.loc.gov for reconciliation; unfortunately, there was no SPARQL endpoint or other batch query interface for id.loc.gov that could be used. As an attempted workaround, the

[10] http://bibframe.org/vocab-list/#Authority

https://doi.org/10.23106/dcmi.952137143

authors downloaded the LC name authority file and using Apache Jena tools,[11] loaded the file into a TDB, to spin up a local SPARQL endpoint for name reconciliation. The LC name authority file was sizable (> 30 GB); even with local access to query the TDB, name reconciliation for one column with ~70 unique entries had a runtime in excess of three hours.

Since the results were not optimal and in the interest of time,_the authors manually created BIBFRAME authority objects and included links from several enrichment services such as VIAF, id.loc.gov, and Wikipedia. The BIBFRAME authorities were then integrated with the core dataset using the OpenRefine reconciliation function to link the separate files. In an ideal process, there would be a reconciliation service for id.loc.gov, since much of the legacy metadata for the DRC dataset included vocabularies and authorities from LC. However, even if a SPARQL endpoint was available, id.loc.gov does not contain a complete dataset of LCNAF and LCSH. Problems also arise when a URI is available for the parent body of an organization, but not the subordinate body as found in the local dataset. For this project, the authors linked to parent bodies, even though the matches were not exact (see Table 3 for examples). This approach also worked for LCSH subject strings when a primary topical heading had an authority but the string did not.

TABLE 3. Example of id.loc.gov partial matches to DC dataset

Entity From DC Dataset	**Partial Match (id.loc.gov)**
American Institute of Aeronautics and Astronautics. Student Chapter	http://id.loc.gov/authorities/names/n79053067 (parent body)
University of Cincinnati. Board of Trustees	http://id.loc.gov/authorities/names/n79034519 (parent body)

6. Transformation Process

One of the goals of this project was to develop scripts for conversion of data from DC to BIBFRAME. Encoding the conversion process into scripts offered the advantage of reuse and easy adaptation. However, faced with challenges in conceptualizing the process and in the interest of experimentation, much of the work was done manually for this first phase. The authors chose to focus on outlining the model and closely curating a small dataset (218 records) as a proof of concept. The input data consisted of a blend of DC metadata in CSV format and manually created RDF/XML. The DC metadata comprised the foundation of the dataset, augmented and linked with additional, hand-curated BIBFRAME elements in RDF/XML.

The output dataset was self-contained, comprised of BIBFRAME Work, Instance, and Authority data, concatenated from separate files. The BIBFRAME Works were mapped directly from the DC dataset; each type was generated with OpenRefine and the Digital Enterprise Research Institute (DERI)'s RDF extension. The DERI RDF extension[12] includes RDF skeleton functionality to map data to namespaced elements for export. For this project, a custom skeleton based on the discovery metadata map was created.

7. Next Steps and Recommendations

The vision for this project is to package, publish, and optimize linked data for all collections at UC Libraries. The authors agree with the philosophy of the LibHub initiative[13] in describing

[11] https://jena.apache.org/
[12] http://refine.deri.ie/
[13] http://www.libhub.org/

https://doi.org/10.23106/dcmi.952137143

efforts to use the Semantic Web and Search Engine Optimization (SEO) techniques to better position cultural heritage institutions for discovery via commercial Web search engines. Achievement of this vision will require several subsequent steps.

1. Investigate other enrichment services for streamlining of the reconciliation process. For example, using OCLC FAST for subject authorities and ISNI for name authorities.
2. Map other ontologies in the dataset; ex. FOAF, SKOS, etc.
3. Server space - For further experimentation, the department has procured a public-facing virtual server for hosting local linked data sets on an ongoing basis and has plans to post linked data sets for public consumption.
4. Process MARC records for UC Libraries' physical collections into BIBFRAME and link with special collections datasets to improve discovery. Ultimately, we will want to take steps to review systems, enterprise-wide, and assess fitness for modeling and exposing linked data. Where possible, integrate linked data publishing at the system level and implement tools for working with linked data natively.

8. Conclusion

This project represents the first linked data initiative for UC Libraries. The authors spent time experimenting with tools and technologies to convert and reconcile legacy metadata for a high-interest special collection. Emerging trends in library linked data and the Semantic Web are central to several of the UC Libraries' strategic initiatives; touching on issues of access, discovery and preservation. Libraries house a wealth of data in many formats, most of which, because of structure or format, are not easily adapted for linking and sharing on the Web. The BIBFRAME initiative offers a core standard for expressing MARC but remains flexible enough to encompass other flavors of metadata. The task of migrating Qualified Dublin Core to BIBFRAME, even with the loosened constraints of our focus on discovery rather than comprehensive representation, is a demonstration of that flexibility. Although letting go of traditional ideas about metadata and description is difficult, thinking in terms of system needs for successful identification and linking of data is an essential step to discovery.

Acknowledgements

The authors would like to thank the following individuals for their help and support for this project: Elna Saxton, Head of Content Services, University of Cincinnati; Leslie Schick, Associate Dean of Library Services & Director of the Health Sciences Library, University of Cincinnati; and Victoria Mueller, Senior Information Architect and Systems Librarian, Zepheira, LLC.

Bibliography

Dublin Core Metadata Initiative (2012). *Dublin Core Metadata Element Set, Version 1.1.* Retrieved from http://dublincore.org/documents/dces/.

Fallgren, Nancy (2015). *Experiments in BIBFRAME*: *A Modular Approach*. Retrieved from http://connect.ala.org/node/68263.

Library of Congress (2015). *Bibliographic Framework Initiative*. Retrieved from http://bibframe.org.

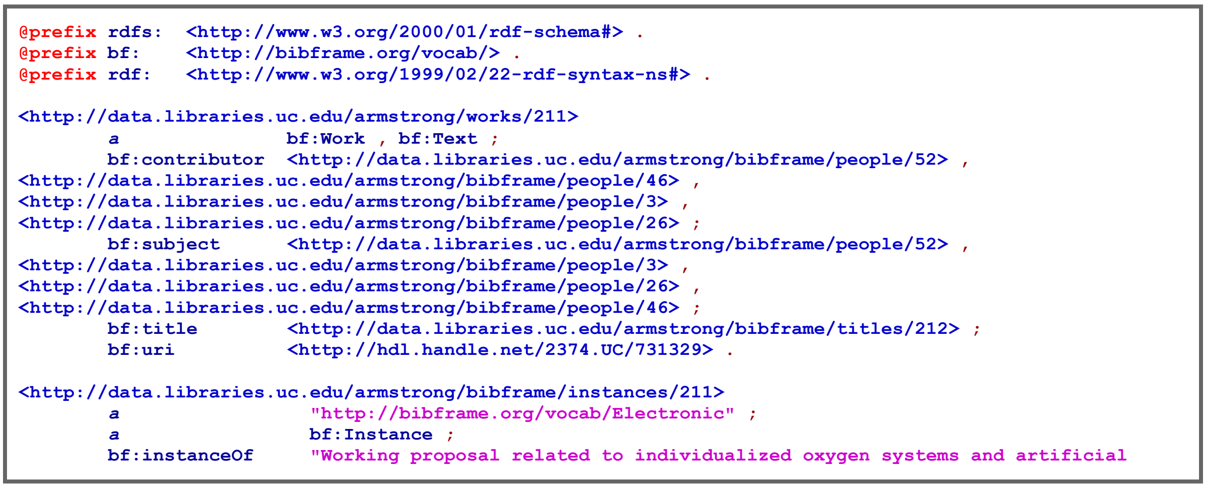
University of Cincinnati Libraries (2013). *Neil Armstrong Commemorative Archive.* Retrieved from https://drc.libraries.uc.edu/handle/2374.UC/713357.

https://doi.org/10.23106/dcmi.952137143

Appendix I: Properties From Original Dataset Not Mapped to BIBFRAME

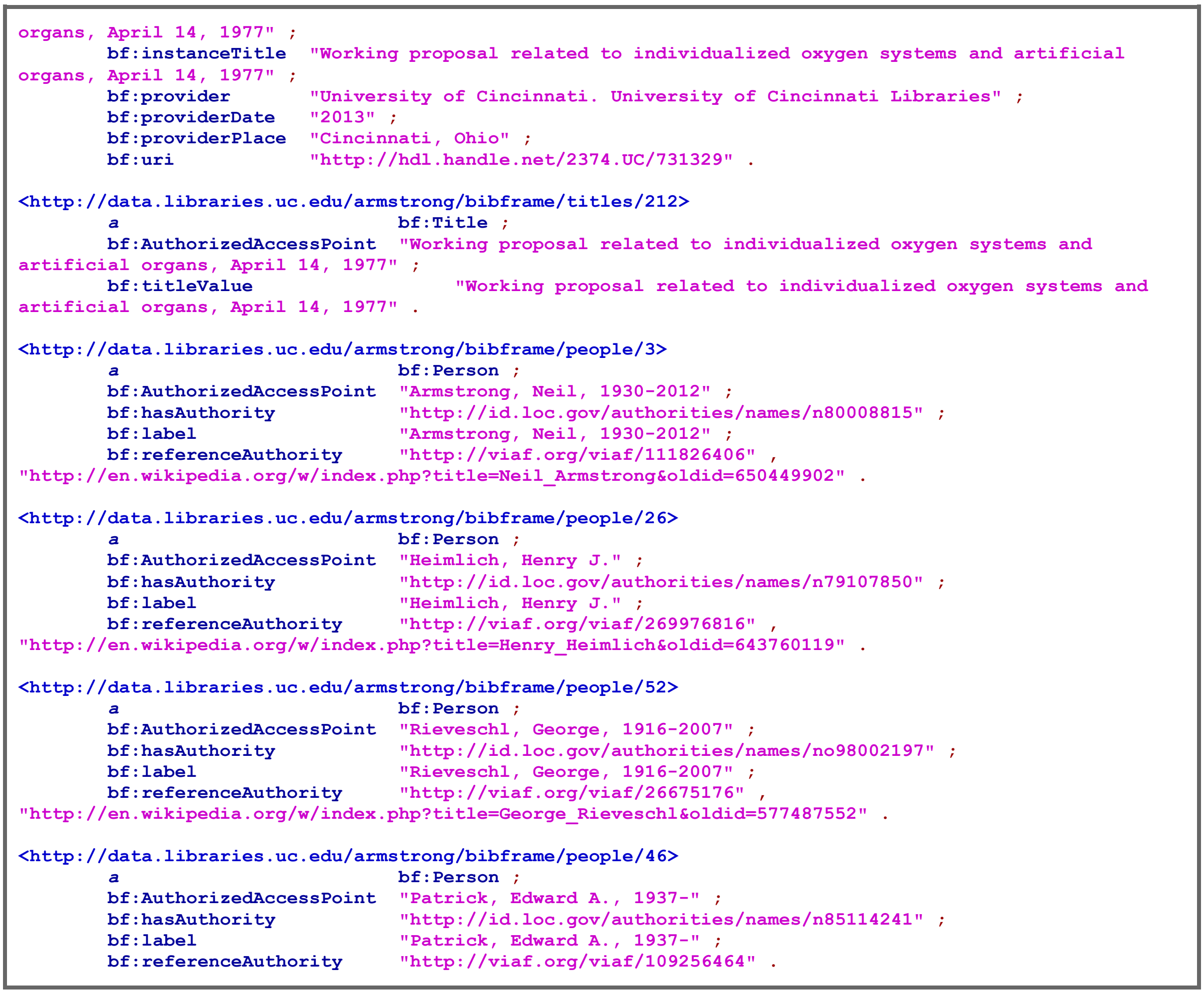
Qualified DC (UC Armstrong Collection)	SIMPLE DC
dc.date.created *Note:* Refers to original object, not digital surrogate	dc.date
dc.date.digitized *Note:* date.available was used as publication date for digital surrogate	dc.date
dc.description	dc.description
dc.format	dc.format
dc.language.iso	dc.language
dc.publisher	dc.publisher
dc.publisher.OLinstitution *Note:* Legacy OhioLINK property	dc.publisher
dc.relationispartof	dc.relation
dc.relationispartofseries	dc.relation
dc.relation.uri	dc.relation
dc.rights	dc.rights
dc.rights.uri	dc.rights
dc.source	dc.source

Appendix II: Sample Data Serialized as Turtle



```
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix bf:    <http://bibframe.org/vocab/> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://data.libraries.uc.edu/armstrong/works/211>
        a                bf:Work , bf:Text ;
        bf:contributor  <http://data.libraries.uc.edu/armstrong/bibframe/people/52> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/46> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/3> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/26> ;
        bf:subject      <http://data.libraries.uc.edu/armstrong/bibframe/people/52> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/3> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/26> ,
<http://data.libraries.uc.edu/armstrong/bibframe/people/46> ;
        bf:title        <http://data.libraries.uc.edu/armstrong/bibframe/titles/212> ;
        bf:uri          <http://hdl.handle.net/2374.UC/731329> .

<http://data.libraries.uc.edu/armstrong/bibframe/instances/211>
        a                 "http://bibframe.org/vocab/Electronic" ;
        a                 bf:Instance ;
        bf:instanceOf     "Working proposal related to individualized oxygen systems and artificial
```

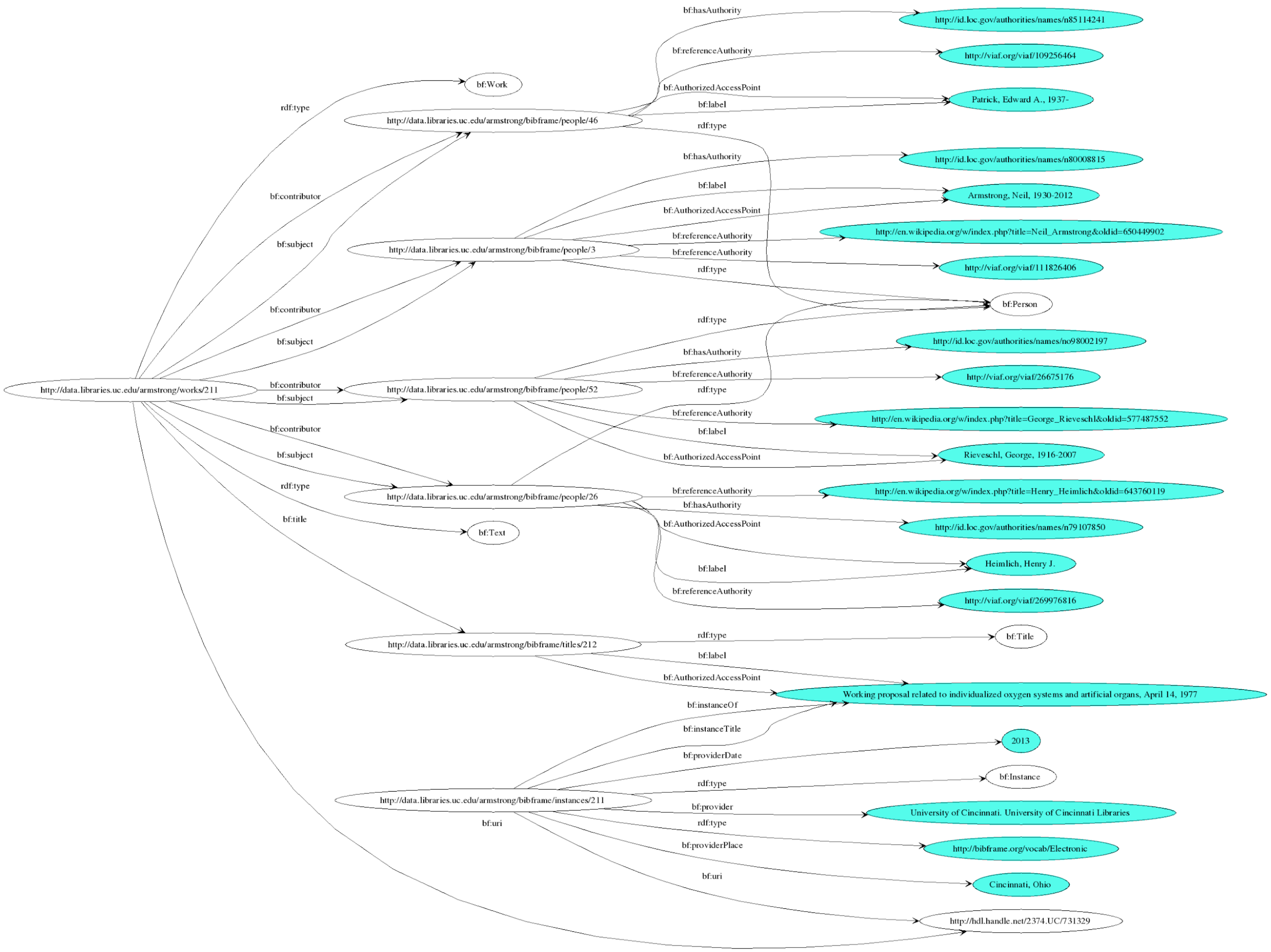
https://doi.org/10.23106/dcmi.952137143



```
organs, April 14, 1977" ;
        bf:instanceTitle  "Working proposal related to individualized oxygen systems and artificial
organs, April 14, 1977" ;
        bf:provider       "University of Cincinnati. University of Cincinnati Libraries" ;
        bf:providerDate   "2013" ;
        bf:providerPlace  "Cincinnati, Ohio" ;
        bf:uri            "http://hdl.handle.net/2374.UC/731329" .

<http://data.libraries.uc.edu/armstrong/bibframe/titles/212>
        a                         bf:Title ;
        bf:AuthorizedAccessPoint  "Working proposal related to individualized oxygen systems and
artificial organs, April 14, 1977" ;
        bf:titleValue                  "Working proposal related to individualized oxygen systems and
artificial organs, April 14, 1977" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/3>
        a                         bf:Person ;
        bf:AuthorizedAccessPoint  "Armstrong, Neil, 1930-2012" ;
        bf:hasAuthority           "http://id.loc.gov/authorities/names/n80008815" ;
        bf:label                  "Armstrong, Neil, 1930-2012" ;
        bf:referenceAuthority     "http://viaf.org/viaf/111826406" ,
"http://en.wikipedia.org/w/index.php?title=Neil_Armstrong&oldid=650449902" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/26>
        a                         bf:Person ;
        bf:AuthorizedAccessPoint  "Heimlich, Henry J." ;
        bf:hasAuthority           "http://id.loc.gov/authorities/names/n79107850" ;
        bf:label                  "Heimlich, Henry J." ;
        bf:referenceAuthority     "http://viaf.org/viaf/269976816" ,
"http://en.wikipedia.org/w/index.php?title=Henry_Heimlich&oldid=643760119" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/52>
        a                         bf:Person ;
        bf:AuthorizedAccessPoint  "Rieveschl, George, 1916-2007" ;
        bf:hasAuthority           "http://id.loc.gov/authorities/names/no98002197" ;
        bf:label                  "Rieveschl, George, 1916-2007" ;
        bf:referenceAuthority     "http://viaf.org/viaf/26675176" ,
"http://en.wikipedia.org/w/index.php?title=George_Rieveschl&oldid=577487552" .

<http://data.libraries.uc.edu/armstrong/bibframe/people/46>
        a                         bf:Person ;
        bf:AuthorizedAccessPoint  "Patrick, Edward A., 1937-" ;
        bf:hasAuthority           "http://id.loc.gov/authorities/names/n85114241" ;
        bf:label                  "Patrick, Edward A., 1937-" ;
        bf:referenceAuthority     "http://viaf.org/viaf/109256464" .
```

https://doi.org/10.23106/dcmi.952137143

Appendix III: Visual Representation of Sample Data


http://data.libraries.uc.edu/armstrong/works/211
rdf:type
bf:Work
bf:contributor
bf:subject
bf:contributor
bf:subject
bf:contributor
bf:subject
bf:contributor
bf:subject
rdf:type
bf:title
bf:Text
http://data.libraries.uc.edu/armstrong/bibframe/people/46
bf:hasAuthority
http://id.loc.gov/authorities/names/n85114241
bf:referenceAuthority
http://viaf.org/viaf/109256464
bf:AuthorizedAccessPoint
bf:label
Patrick, Edward A., 1937-
rdf:type
http://data.libraries.uc.edu/armstrong/bibframe/people/3
bf:hasAuthority
http://id.loc.gov/authorities/names/n80008815
bf:label
Armstrong, Neil, 1930-2012
bf:AuthorizedAccessPoint
bf:referenceAuthority
http://en.wikipedia.org/w/index.php?title=Neil_Armstrong&oldid=650449902
bf:referenceAuthority
http://viaf.org/viaf/111826406
rdf:type
bf:Person
http://data.libraries.uc.edu/armstrong/bibframe/people/52
rdf:type
bf:hasAuthority
http://id.loc.gov/authorities/names/no98002197
bf:referenceAuthority
http://viaf.org/viaf/26675176
rdf:type
bf:referenceAuthority
http://en.wikipedia.org/w/index.php?title=George_Rieveschl&oldid=577487552
bf:label
Rieveschl, George, 1916-2007
bf:AuthorizedAccessPoint
http://data.libraries.uc.edu/armstrong/bibframe/people/26
bf:referenceAuthority
http://en.wikipedia.org/w/index.php?title=Henry_Heimlich&oldid=643760119
bf:hasAuthority
http://id.loc.gov/authorities/names/n79107850
bf:AuthorizedAccessPoint
Heimlich, Henry J.
bf:label
bf:referenceAuthority
http://viaf.org/viaf/269976816
http://data.libraries.uc.edu/armstrong/bibframe/titles/212
rdf:type
bf:Title
bf:label
bf:AuthorizedAccessPoint
Working proposal related to individualized oxygen systems and artificial organs, April 14, 1977
http://data.libraries.uc.edu/armstrong/bibframe/instances/211
bf:instanceOf
bf:instanceTitle
bf:providerDate
2013
rdf:type
bf:Instance
bf:provider
University of Cincinnati. University of Cincinnati Libraries
rdf:type
http://bibframe.org/vocab/Electronic
bf:providerPlace
Cincinnati, Ohio
bf:uri
bf:uri
http://hdl.handle.net/2374.UC/731329

https://doi.org/10.23106/dcmi.952137143