Linked Open Dime Novels; or, 19th Century Fiction and 21st Century Data

Matthew Short, Northern Illinois University; Demian Katz, Villanova University

I'm Matt Short at Northern Illinois University and joining me shortly will be Demian Katz from Villanova University. We're co-principal investigators on the CLIR-funded Johannsen project and the NEH-funded Street & Smith and Tousey projects. We're going to talk about some of the work we've been doing to publish a bibliography of nineteenth-century popular fiction as linked data. But first, I'm going to provide some background about our digitization projects and dime novels themselves in order to provide a bit of context.

Dime novels were the best sellers of the mid- to late-nineteenth century. Costing between 5 and 25 cents each, they were especially popular among the working classes, immigrants, and children, who often couldn't afford conventional hardback fiction. While meant to entertain, they also deal with issues of gender, class, and race, providing unique insight into what a diverse group of Americans were thinking and feeling towards the end of the nineteenth century. Most early genre fiction also started in dime novels, including westerns, detective stories, and even science fiction. So today, they tend to be especially interesting to social and cultural historians, as well as people studying the history of the novel and evolution of popular fiction.

In 2016, NIU and Villanova received a Digitizing Hidden Collections grant from the Council on Library and Information Resources to digitize NIU's Johannsen Collection of dime novels and to publish Albert Johannsen's bibliography as open linked data. The Johannsen project has since expanded to include additional partners at Stanford University, Bowling Green State University, and Oberlin College, and has received two Humanities Collections and Reference Resources awards from the NEH, first in 2020 for the Street & Smith project and again in 2022 for the Frank Tousey project.

Because it was much less expensive to ship periodicals through the mail, dime novel publishers would disguise their novels by issuing them in a series, often with backup features like short stories, serialized stories, or advice columns. There might be several editions of any given dime novel, usually in slightly different formats and published in different series over the span of decades. Sometimes these were published complete in one volume, but they might also be serialized as a backup feature in a different series. Title, author, and even the text sometimes vary between editions. Many dime novels are also attributed to pseudonyms, which might be shared by several real-world persons. Bibliographers may have documented these relationships in print bibliographies, but they are very difficult to express in a MARC or MODS record.

We decided to use linked data in our project to overcome some of the limitations of traditional metadata.

First, we had to build an ontology describing what entities exist, the properties they have, and how everything is related. This provides flexibility when talking about complex bibliographic relationships, but also lets us map the schema each project is currently using to a single shared understanding.

We committed to using shared identifiers (in the form of URIs) in all of our descriptive metadata, including names and subjects, to simplify matching up data between systems.

The ultimate goal was to build a central pool of data which could be harvested and used to enhance individual institutional collections by exposing relationships directly to end users as they're browsing.

Having designed our ontology, we needed a way to publish it online. We decided to use Villanova's dimenovels.org bibliography, which is built on Geeby-Deeby, an open-source project designed to publish bibliographies of books published in series. The Geeby-Deeby database schema was compatible with our ontology, and its existing URL structure was well suited for minting identifiers, so implementation was fairly straightforward. We just enhanced the software with the capability to render displays as RDF using HTTP content negotiation.

With our platform in place, the next challenge was loading data into it. Fortunately, the use of common identifiers simplified this problem. Because NIU's MODS records include commonly used identifiers, it was relatively straightforward to build a data loading process which extracts relevant details from MODS records and matches everything up with existing data in Geeby-Deeby's database. While requiring some human intervention, this tool enabled efficient loading of thousands of records into the bibliography, and its various integrity checks helped to identify and correct problems on both sides of the transaction.

This screen shot shows the data loading process in action; as you can see, many data elements are matched and loaded automatically, but human interaction is occasionally required for disambiguation.

With dimenovels.org fully populated with useful linked data, the next challenge was finding a way to utilize the data in aggregate. A new open-source tool called Murpoint, named after the villain of Charles Garvice's *The Spider and the Fly*, was built to crawl through all of the data and collect it into a single file on disk. This file can then be loaded into any triplestore for easy querying. In NIU's case, they chose Blazegraph as a repository.

NIU then customized their Islandora instance to query Blazegraph using shared work identifiers in order to conveniently display all editions or parts of the same work in a single place.

There's only so much we can say in five minutes, but we hope this has shown some of the potential of linked data. Check out the links here to see more of our content and learn more about our technology stack.

Note: The video corresponding to this transcript is located at **https://youtu.be/BzfEQFjU45U**.