DCMI 2017
26 October 2017
Washington, D.C.

Extending Legacy Metadata with Linked Open Data

Jacob Jett (jjett2@Illinois.edu)
Timothy W. Cole (t-cole3@Illinois.edu)
Deren Kudeki (dkudeki@illinois.edu)
Myung-Ja Han (mhan3@Illinois.edu)

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Project Context

- *Exploring the Benefits for Users of LOD for Digitized Special Collections*
  - 18-month exploratory study
  - Funded by the Andrew W. Mellon Foundation
- Digitized Library Special Collections
  - Many relegated to information silos largely disconnected from the broader Web
  - How can we better connect these special resources to the Web?
    - Can we use Linked Open Data to help us? If so, how hard is it to do?
- Objectives
  - Map legacy metadata schemas to LOD-compliant schemas
  - Actively link to and from DBpedia, VIAF, Wikidata, and related Web resources

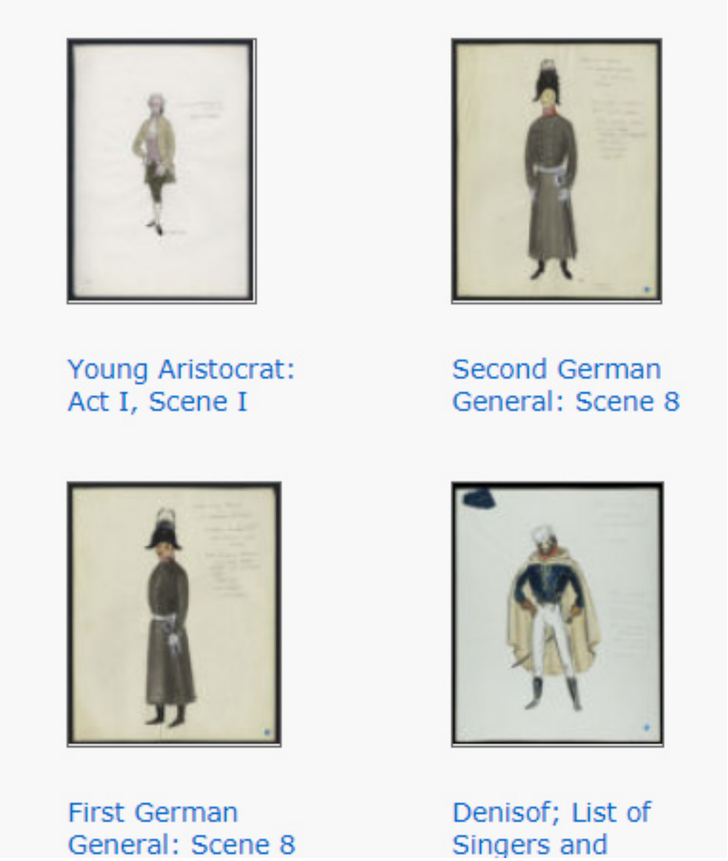
Collections Involved

- The Motley Collection of Theatre & Costume Design
  - About 5,000 images of costume and set designs, sketches, production notes, and similar objects
  - Represents a variety of objects from the Motely Group's career (1932-1976)
- Portraits of Actors, 1720-1920
  - Nearly 3,500 pictures of actors, including Sarah Siddons, Edmund Kean, and others
- Kolb-Proust Archive for Research
  - About 8,700 of Professor Philip Kolb's research notecards on Marcel Proust
    - A chronology of events concerning Marcel's Proust's life
    - A bibliography of works mentioned in Marcel's Proust's correspondences

Motley & Portraits of Actors Metadata Challenges

- 1-to-1 entity description violations
  - Each metadata record contained information about the image, the production or performance it related to, the associated people, and the associated play
  - Had to disambiguate these different entities from one another in order to produce quality Linked Data
- Mixed levels of authority control
  - Name authority control idiosyncratic and localized
  - Subject authority control used standard vocabularies and thesauri
- Conflated data (e.g., associated people: actors, composers, authors, etc., all in the same field)

Motley & Portraits of Actors Collections



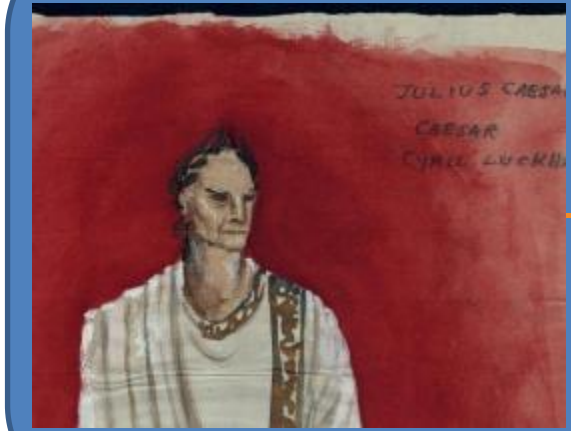
Young Aristocrat: Act I, Scene I

Second German General: Scene 8

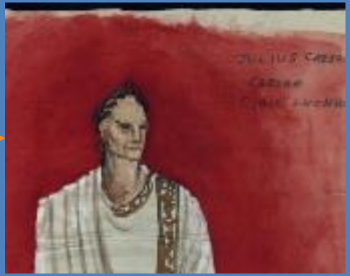
First German General: Scene 8

Denisof; List of Singers and

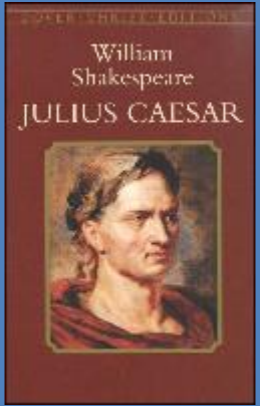
What the CONTENTdm metadata actually describes



Costume design by Motley



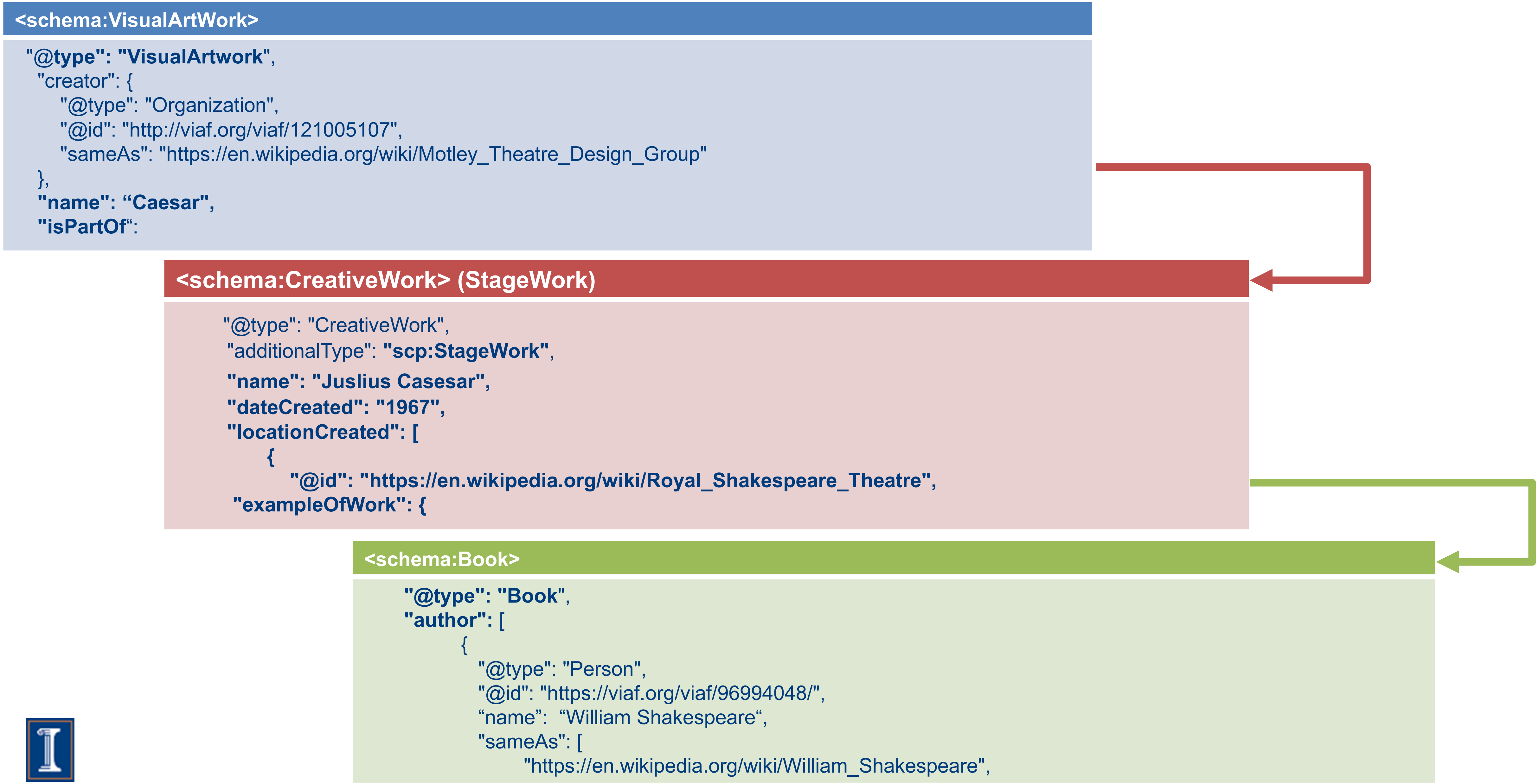
For a stage production directed by Glen Byam Shaw



Of a play by William Shakespeare

Performed at the Royal Shakespeare Theater

Linked Data Adds Context



<schema:VisualArtWork>

```
"@type": "VisualArtwork",
 "creator": {
    "@type": "Organization",
    "@id": "http://viaf.org/viaf/121005107",
    "sameAs": "https://en.wikipedia.org/wiki/Motley_Theatre_Design_Group"
 },
 "name": “Caesar",
 "isPartOf“:
```

<schema:CreativeWork> (StageWork)

```
"@type": "CreativeWork",
 "additionalType": "scp:StageWork",
 "name": "Juslius Casesar",
 "dateCreated": "1967",
 "locationCreated": [
       {
          "@id": "https://en.wikipedia.org/wiki/Royal_Shakespeare_Theatre",
  "exampleOfWork": {
```

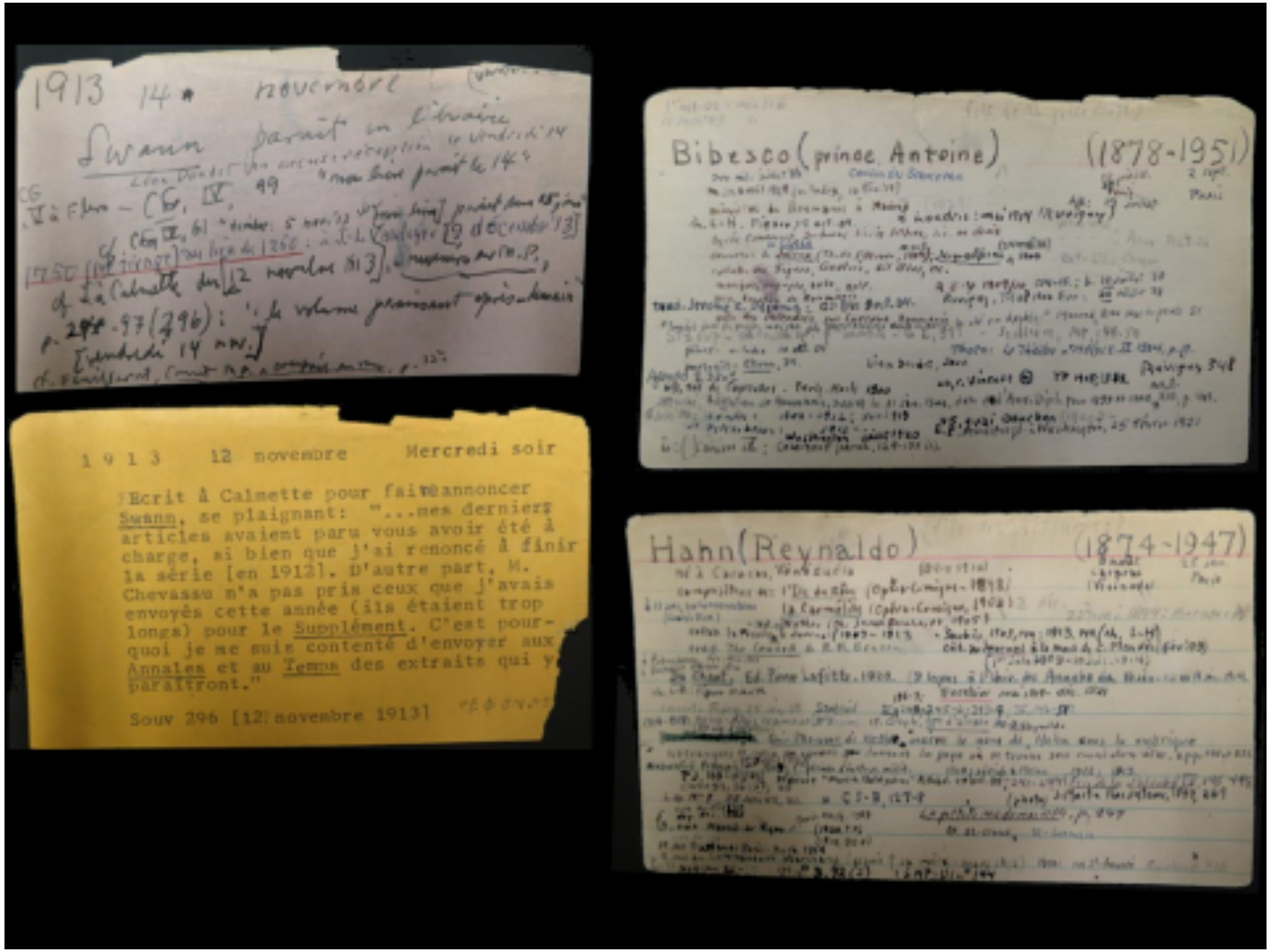
<schema:Book>

```
"@type": "Book",
"author": [
        {
          "@type": "Person",
          "@id": "https://viaf.org/viaf/96994048/",
          “name”:  “William Shakespeare“,
          "sameAs": [
                "https://en.wikipedia.org/wiki/William_Shakespeare",
```

Kolb-Proust Archive Metadata Challenges

- Digital objects are TEI-encoded notecards, not metadata records
  - But they do contain metadata about named entities
- Identifying named entities
  - Need to differentiate when named entities appear in citations vs. only mentioned in passing
- Inconsistent structure of markup used to digitize cards insufficient to consistently differentiate the roles of various entities

Kolb-Proust Archive



1913 14 novembre

Swann paraît en librairie [illegible] le vendredi 14

Léon Daudet [illegible] "mon livre paraît le 14"

CG V à Flers — CG, IV, 49

[illegible] [9 décembre 13]

1750 [1er tirage] au lieu de 1250 [illegible]

cf. à Calmette du [12 novembre 913] [illegible]

p. 297-97 (296): "le volume paraissant après-demain"

[vendredi 14 nov.]

1 9 1 3 12 novembre Mercredi soir

Ecrit à Calmette pour faire annoncer Swann, se plaignant: "...mes derniers articles avaient paru vous avoir été à charge, si bien que j'ai renoncé à finir la série [en 1912]. D'autre part, M. Chevassu n'a pas pris ceux que j'avais envoyés cette année (ils étaient trop longs) pour le Supplément. C'est pourquoi je me suis contenté d'envoyer aux Annales et au Temps des extraits qui y paraîtront."

Souv 296 [12 novembre 1913]

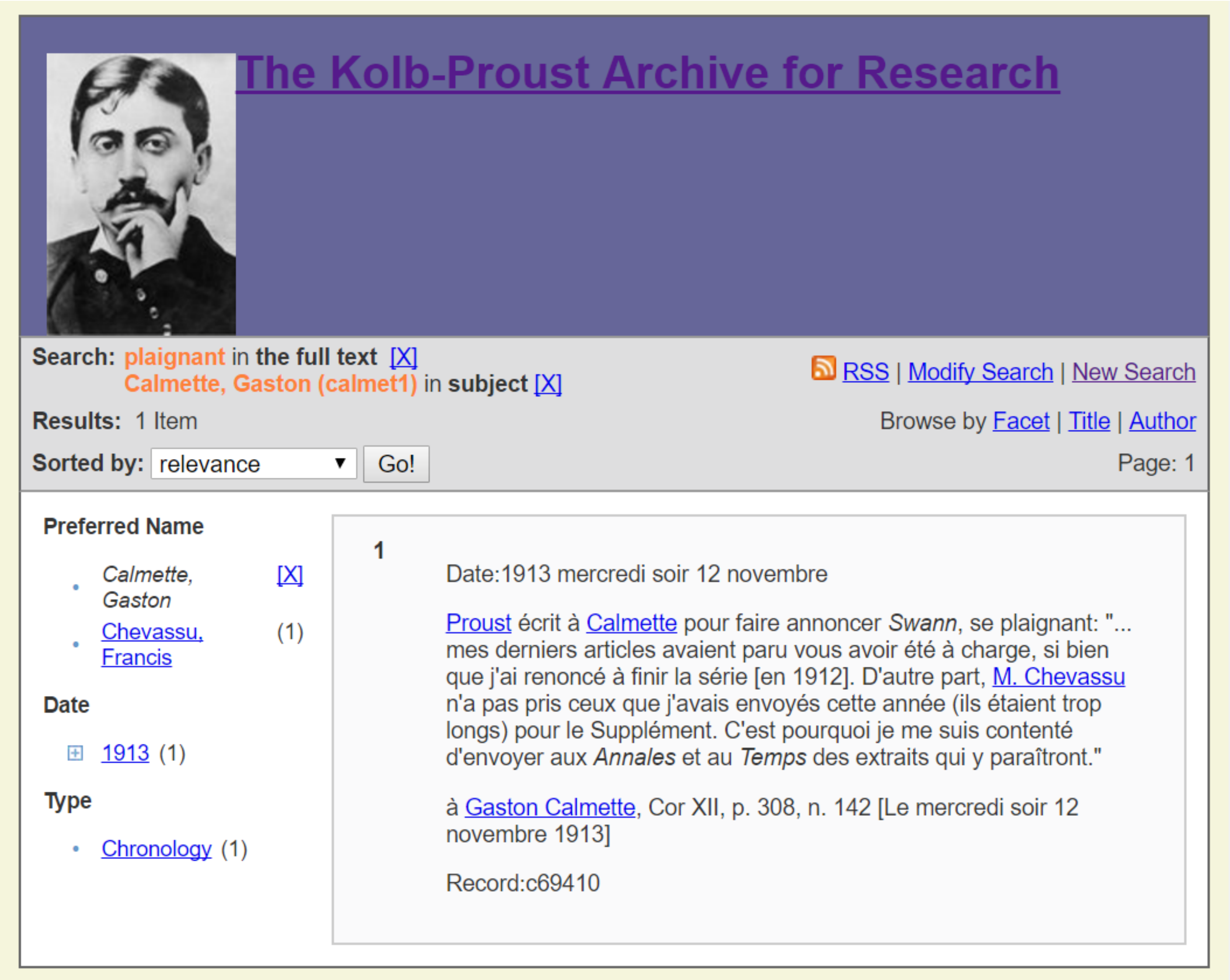
Bibesco (prince Antoine) (1878-1951)

[illegible]

Hahn (Reynaldo) (1874-1947)

né à Caracas, Venezuela

[illegible]



The Kolb-Proust Archive for Research

Search: plaignant in **the full text** [X]
Calmette, Gaston (calmet1) in **subject** [X]

RSS | Modify Search | New Search

Results: 1 Item

Browse by Facet | Title | Author

Sorted by: relevance ▼ Go!

Page: 1

Preferred Name

- *Calmette, Gaston* [X]
- Chevassu, Francis (1)

Date

⊞ 1913 (1)

Type

- Chronology (1)

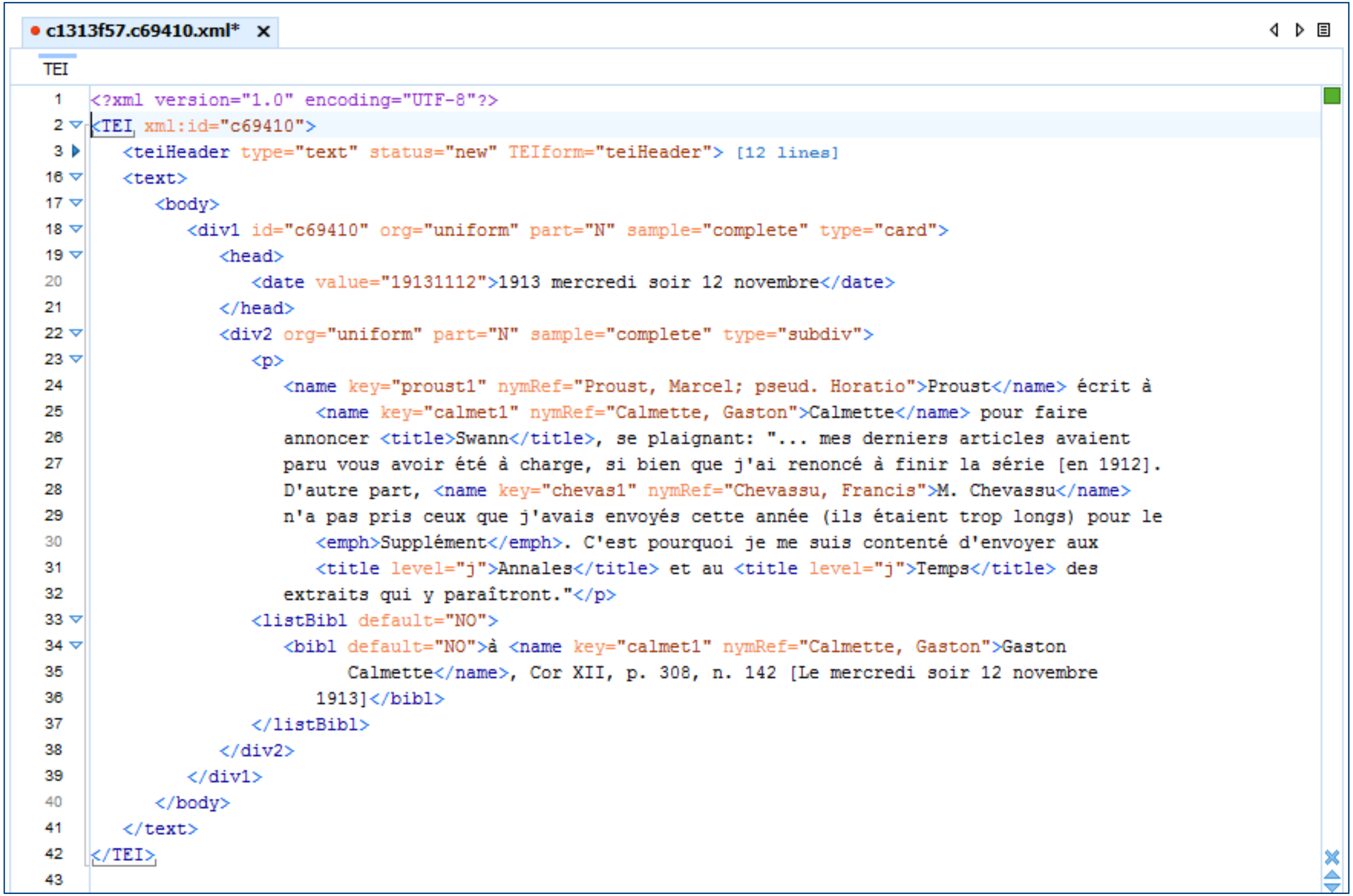
1

Date:1913 mercredi soir 12 novembre

Proust écrit à Calmette pour faire annoncer *Swann*, se plaignant: "... mes derniers articles avaient paru vous avoir été à charge, si bien que j'ai renoncé à finir la série [en 1912]. D'autre part, M. Chevassu n'a pas pris ceux que j'avais envoyés cette année (ils étaient trop longs) pour le Supplément. C'est pourquoi je me suis contenté d'envoyer aux *Annales* et au *Temps* des extraits qui y paraîtront."

à Gaston Calmette, Cor XII, p. 308, n. 142 [Le mercredi soir 12 novembre 1913]

Record:c69410



```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xml:id="c69410">
   <teiHeader type="text" status="new" TEIform="teiHeader"> [12 lines]
   <text>
      <body>
         <div1 id="c69410" org="uniform" part="N" sample="complete" type="card">
            <head>
               <date value="19131112">1913 mercredi soir 12 novembre</date>
            </head>
            <div2 org="uniform" part="N" sample="complete" type="subdiv">
               <p>
                  <name key="proust1" nymRef="Proust, Marcel; pseud. Horatio">Proust</name> écrit à
                     <name key="calmet1" nymRef="Calmette, Gaston">Calmette</name> pour faire
                  annoncer <title>Swann</title>, se plaignant: "... mes derniers articles avaient
                  paru vous avoir été à charge, si bien que j'ai renoncé à finir la série [en 1912].
                  D'autre part, <name key="chevas1" nymRef="Chevassu, Francis">M. Chevassu</name>
                  n'a pas pris ceux que j'avais envoyés cette année (ils étaient trop longs) pour le
                     <emph>Supplément</emph>. C'est pourquoi je me suis contenté d'envoyer aux
                     <title level="j">Annales</title> et au <title level="j">Temps</title> des
                  extraits qui y paraîtront."</p>
               <listBibl default="NO">
                  <bibl default="NO">à <name key="calmet1" nymRef="Calmette, Gaston">Gaston
                        Calmette</name>, Cor XII, p. 308, n. 142 [Le mercredi soir 12 novembre
                     1913]</bibl>
               </listBibl>
            </div2>
         </div1>
      </body>
   </text>
</TEI>
```

Partial Mapping of TEI document elements

TEI	Schema
div1 @id	schema:DataSet
	schema:author <http://viaf.org/44300868>
	schema:inLanguage “fr”
->head->date @value	schema:temporalCoverage [schema:DateTime]
->div2->p->name ->div2->note->name	schema:mentions [schema:Person]
->div2->p->title ->div2->note->title	schema:mentions [schema:CreativeWork]
->div2->(listBibl)->bibl	schema:citation [schema:CreativeWork]

Mapping Challenges Across all Collections

- Target vocabulary (Schema) still missing some key entities
  - Specifically no way to differentiate the production of a play from the individual performances
  - Solved by locally extending Schema
- Many entities are not currently listed in linked data sources
  - For Kolb-Proust we assigned URIs to every name and then linked the ones listed in authority control databases to those databases
  - Could do this for other collections

Enriching Metadata Links by Providing Linking or Canonical URIs for

- Persons
  - E.g., Peter Ustinov, Marcel Proust, etc.
- Venues
  - E.g., the Old Victoria Theatre, Alexandra Theatre, etc.
- Plays/Productions/Performances
  - E.g., *The Unknown Soldier & His Wife*, *Romeo & Juliet*, etc.
- Subject Headings/Terms
  - E.g., Theater—History, Costume Design, etc.
- Bibliographic References
  - E.g., *Figaro*, *Gaulois*, *Journal des Debats*, etc.

Entity Identification & Reconciliation

- Focused on VIAF and DBpedia, which also linked to additional resources like,
  - Library of Congress Name Authority File (LCNAF)
  - OCLC's WorldCat Identities
  - Wikipedia
- Also manually connected to non-linked data Web Resources
  - Theatricalia
  - International Broadway Database
  - International Movie Database
  - Gallica (BnF)

Count of Motley Person URI’s found through manual processes

Total persons identified in Motley metadata = 984 Links have been found for 624 names	Count of URIs Found
having Wikipedia / DBPedia links	311 (32%)
having VIAF links	218 (22%)
found by searching viaf.org directly	87**
found by searching LC Name Authority File	196**
found by searching WorldCat Identities	93**
*combined with automatic results	*582 (59%)
having Theatricalia links	475 (48%)
having IMDb links	353 (36%)
having IBDb links	42 (4%)
having more than 1 link	446 (45%)

*VIAF links for 476 persons (364 not found by manual search) were found using VIAF Auto Suggest

**Represents some overlapping results

Count of Motley Theater and Play/Performance URI’s found through manual searching

Total theaters identified in Motley metadata = 59 Links were found for 52 theaters	Count of URIs Found
having Wikipedia / DBPedia links	49 (83%)
having VIAF links	45 (76%)
having home page links	36 (61%)
having other links	16 (27%)
having more than 1 link	47 (80%)

Total plays / performances identified in Motley metadata = 127 Links were found for 105 plays / performances	Count of URIs Found
having Wikipedia / DBPedia links	95 (75%)
having Theatricalia links	45 (35%)
having other links	10 (8%)
having more than 1 link	44 (35%)

Late Breaking Figures for Kolb-Proust Archive Entities

Total number of names found in the Kolb-Proust dataset = 5,727 Links were found for 1,953 people	Count of URIs Found
having VIAF links	1,678 (29%)
having French Wikipedia links	1,236 (22%)
having English Wikipedia links	999 (17%)
having other links	264 (5%)

Total number of notecards in the Kolb-Proust dataset = 8,716	Count of URIs Found
Citations found on notecards	13,923 (~1.6 citations/card)
Links founds for citations	4,812 (35%)

Observations

- Additional name authority sources needed for special collections
  - Many current sources are focused on authors
- When searching manually for entity links we found:
  - Easiest to start in WorldCat Identities; Google Web Search next best
  - Googling with full names and birth dates usually insufficient, needed to include additional keyword instead for best results
- Manual clean up of metadata and manual search helps recall:
  - Different name spellings/maiden names/nicknames, slightly different birth/death dates, and looking for contextual clues

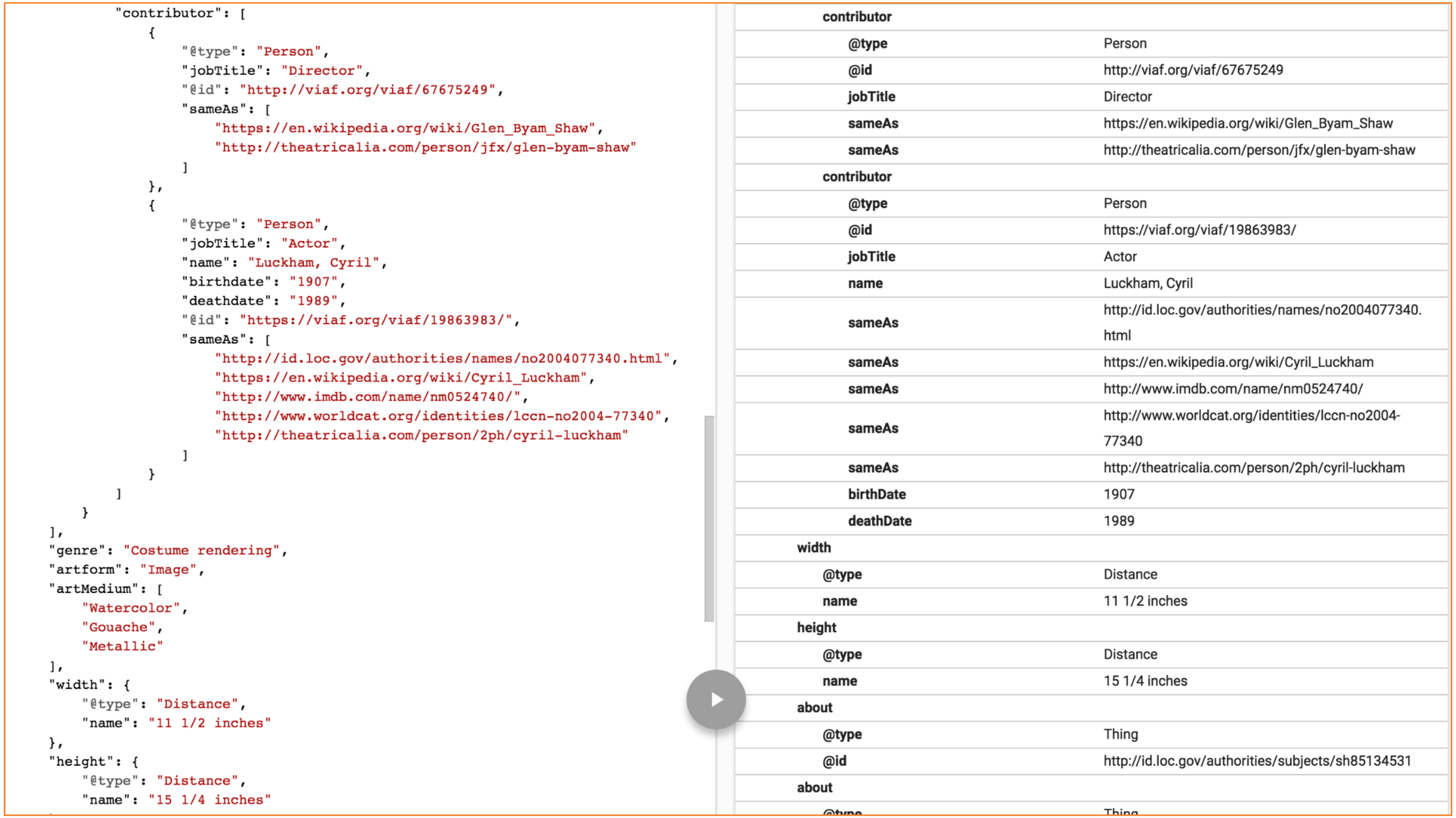
Tasks & Work Time Spent for Motley Linking

- Tasks
  - Cleanup
  - Enhancement
  - Reconciliation
  - Link collection
- Timeline
  - 6 months graduate hourly work at 10-12hrs/week
  - Total = ~240 hours

What's the benefit of all this?

How do we use it?

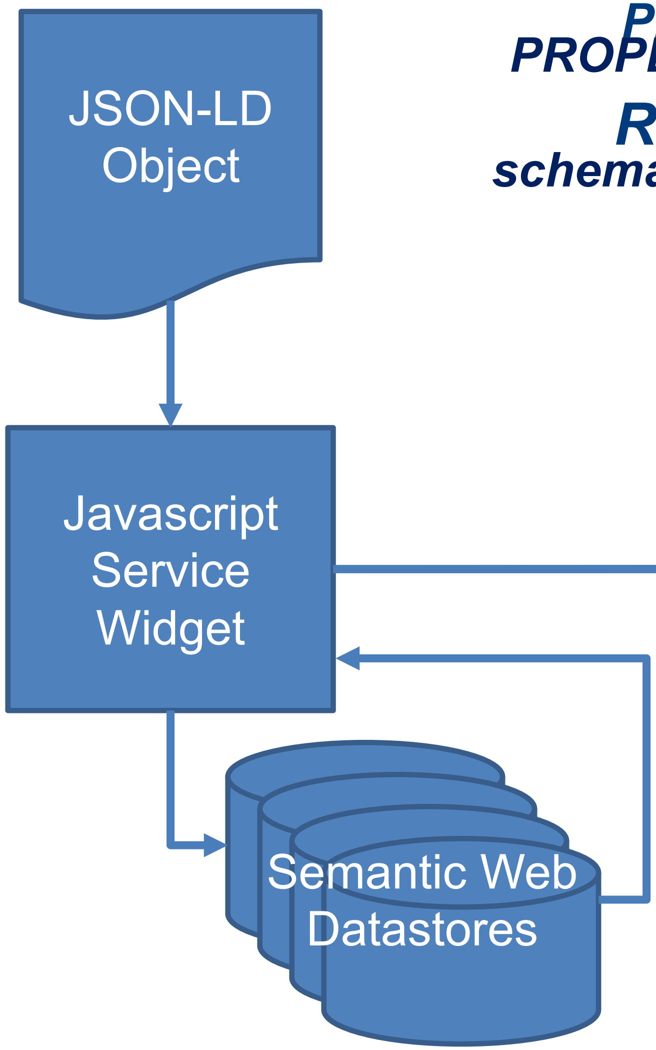
Google Structured Data Tool



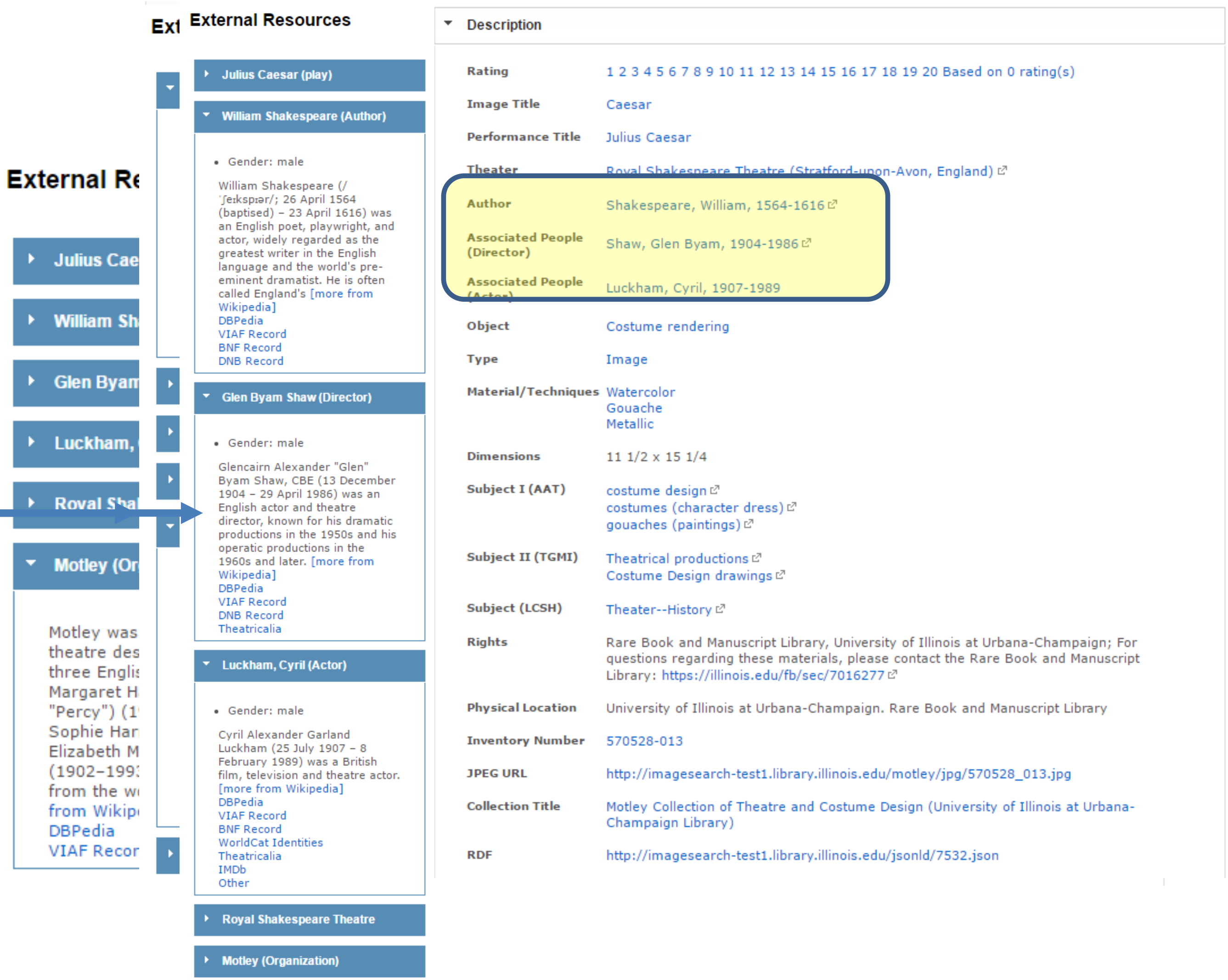
```
            "contributor": [
                {
                    "@type": "Person",
                    "jobTitle": "Director",
                    "@id": "http://viaf.org/viaf/67675249",
                    "sameAs": [
                        "https://en.wikipedia.org/wiki/Glen_Byam_Shaw",
                        "http://theatricalia.com/person/jfx/glen-byam-shaw"
                    ]
                },
                {
                    "@type": "Person",
                    "jobTitle": "Actor",
                    "name": "Luckham, Cyril",
                    "birthdate": "1907",
                    "deathdate": "1989",
                    "@id": "https://viaf.org/viaf/19863983/",
                    "sameAs": [
                        "http://id.loc.gov/authorities/names/no2004077340.html",
                        "https://en.wikipedia.org/wiki/Cyril_Luckham",
                        "http://www.imdb.com/name/nm0524740/",
                        "http://www.worldcat.org/identities/lccn-no2004-77340",
                        "http://theatricalia.com/person/2ph/cyril-luckham"
                    ]
                }
            ]
        }
    ],
    "genre": "Costume rendering",
    "artform": "Image",
    "artMedium": [
        "Watercolor",
        "Gouache",
        "Metallic"
    ],
    "width": {
        "@type": "Distance",
        "name": "11 1/2 inches"
    },
    "height": {
        "@type": "Distance",
        "name": "15 1/4 inches"
```

contributor	
@type	Person
@id	http://viaf.org/viaf/67675249
jobTitle	Director
sameAs	https://en.wikipedia.org/wiki/Glen_Byam_Shaw
sameAs	http://theatricalia.com/person/jfx/glen-byam-shaw
contributor	
@type	Person
@id	https://viaf.org/viaf/19863983/
jobTitle	Actor
name	Luckham, Cyril
sameAs	http://id.loc.gov/authorities/names/no2004077340.html
sameAs	https://en.wikipedia.org/wiki/Cyril_Luckham
sameAs	http://www.imdb.com/name/nm0524740/
sameAs	http://www.worldcat.org/identities/lccn-no2004-77340
sameAs	http://theatricalia.com/person/2ph/cyril-luckham
birthDate	1907
deathDate	1989
width	
@type	Distance
name	11 1/2 inches
height	
@type	Distance
name	15 1/4 inches
about	
@type	Thing
@id	http://id.loc.gov/authorities/subjects/sh85134531
about	
@type	Thing

Context Enhancement



PERFORMANCE
PROPERTIES
PERSONS
AND VENUES
REFERENCED
SKETCHES
ENTITIES
REFERENCED
schema:VisualArtwork



External Re

- Julius Cae
- William Sh
- Glen Byam
- Luckham, (
- Royal Sha
- Motley (Or

Motley was
theatre des
three Englis
Margaret H
"Percy") (1
Sophie Har
Elizabeth M
(1902–199
from the w
from Wikip
DBPedia
VIAF Recor

Ext

External Resources

- Julius Caesar (play)
- William Shakespeare (Author)
  - Gender: male

  William Shakespeare (/ˈʃeɪkspɪər/; 26 April 1564 (baptised) – 23 April 1616) was an English poet, playwright, and actor, widely regarded as the greatest writer in the English language and the world's pre-eminent dramatist. He is often called England's [more from Wikipedia]
  DBPedia
  VIAF Record
  BNF Record
  DNB Record
- Glen Byam Shaw (Director)
  - Gender: male

  Glencairn Alexander "Glen" Byam Shaw, CBE (13 December 1904 – 29 April 1986) was an English actor and theatre director, known for his dramatic productions in the 1950s and his operatic productions in the 1960s and later. [more from Wikipedia]
  DBPedia
  VIAF Record
  DNB Record
  Theatricalia
- Luckham, Cyril (Actor)
  - Gender: male

  Cyril Alexander Garland Luckham (25 July 1907 – 8 February 1989) was a British film, television and theatre actor. [more from Wikipedia]
  DBPedia
  VIAF Record
  BNF Record
  WorldCat Identities
  Theatricalia
  IMDb
  Other
- Royal Shakespeare Theatre
- Motley (Organization)

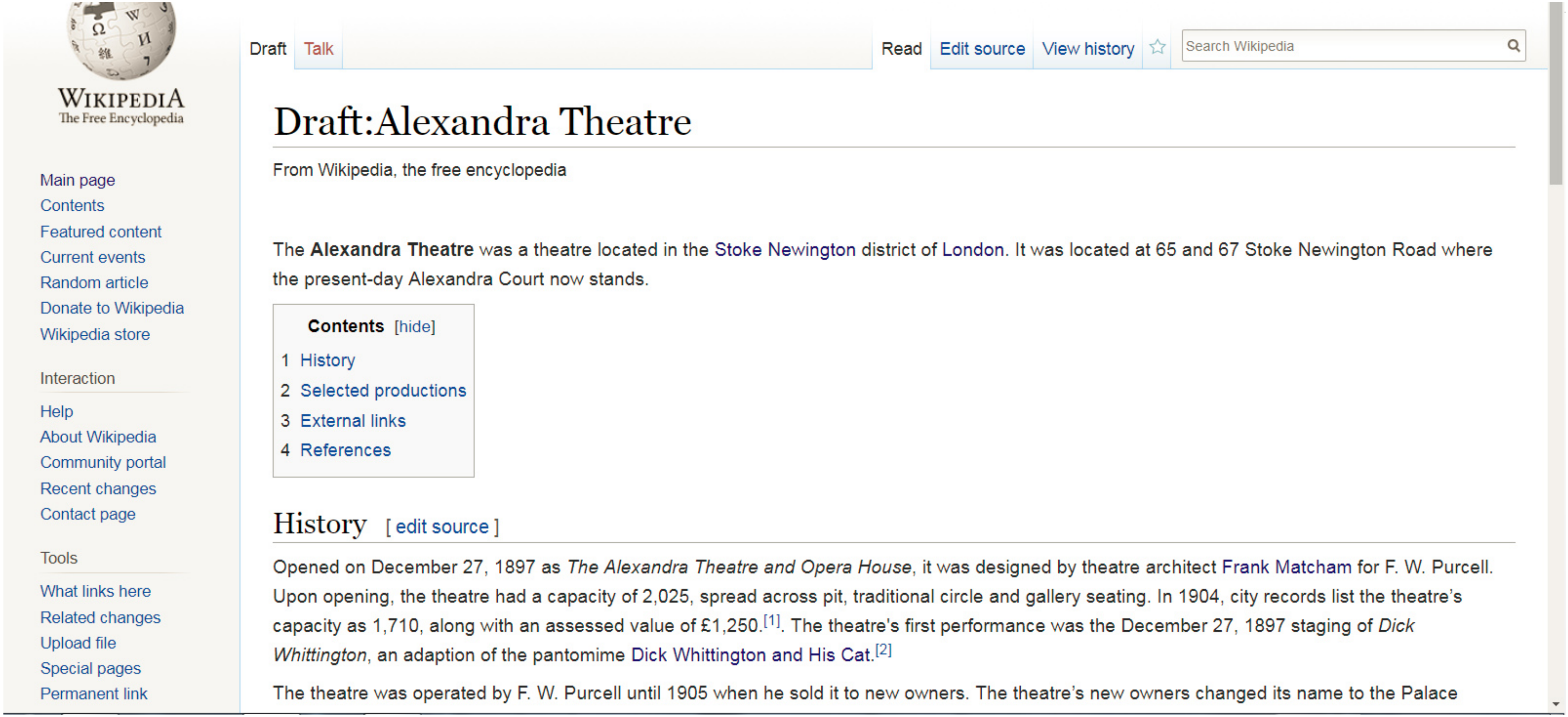
Description

Rating	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Based on 0 rating(s)
Image Title	Caesar
Performance Title	Julius Caesar
Theater	Royal Shakespeare Theatre (Stratford-upon-Avon, England)
Author	Shakespeare, William, 1564-1616
Associated People (Director)	Shaw, Glen Byam, 1904-1986
Associated People (Actor)	Luckham, Cyril, 1907-1989
Object	Costume rendering
Type	Image
Material/Techniques	Watercolor Gouache Metallic
Dimensions	11 1/2 x 15 1/4
Subject I (AAT)	costume design costumes (character dress) gouaches (paintings)
Subject II (TGMI)	Theatrical productions Costume Design drawings
Subject (LCSH)	Theater--History
Rights	Rare Book and Manuscript Library, University of Illinois at Urbana-Champaign; For questions regarding these materials, please contact the Rare Book and Manuscript Library: https://illinois.edu/fb/sec/7016277
Physical Location	University of Illinois at Urbana-Champaign. Rare Book and Manuscript Library
Inventory Number	570528-013
JPEG URL	http://imagesearch-test1.library.illinois.edu/motley/jpg/570528_013.jpg
Collection Title	Motley Collection of Theatre and Costume Design (University of Illinois at Urbana-Champaign Library)
RDF	http://imagesearch-test1.library.illinois.edu/jsonld/7532.json

Preliminary Findings & Conclusions

- Care needs to be taken when mapping legacy metadata to LOD-compliant vocabularies
  - May need to extend with additional entities and properties
  - However can sometimes be rewarded with additional linking properties (e.g., schema:mentions and schema:citation)
- User experiences enriched by adding contextual information
  - Through dynamically added sidebars and clickable links
    - Leverage existing Semantic Web sources
    - Provide opportunities for users to escape the siloed environments of traditional digital libraries
  - However, it is resource-intensive to manually add links, etc. to legacy metadata
- Additional Opportunities for leveraging the Semantic Web remain to be explored

Pushing Information Back Out to the Semantic Web



Draft:Alexandra Theatre

From Wikipedia, the free encyclopedia

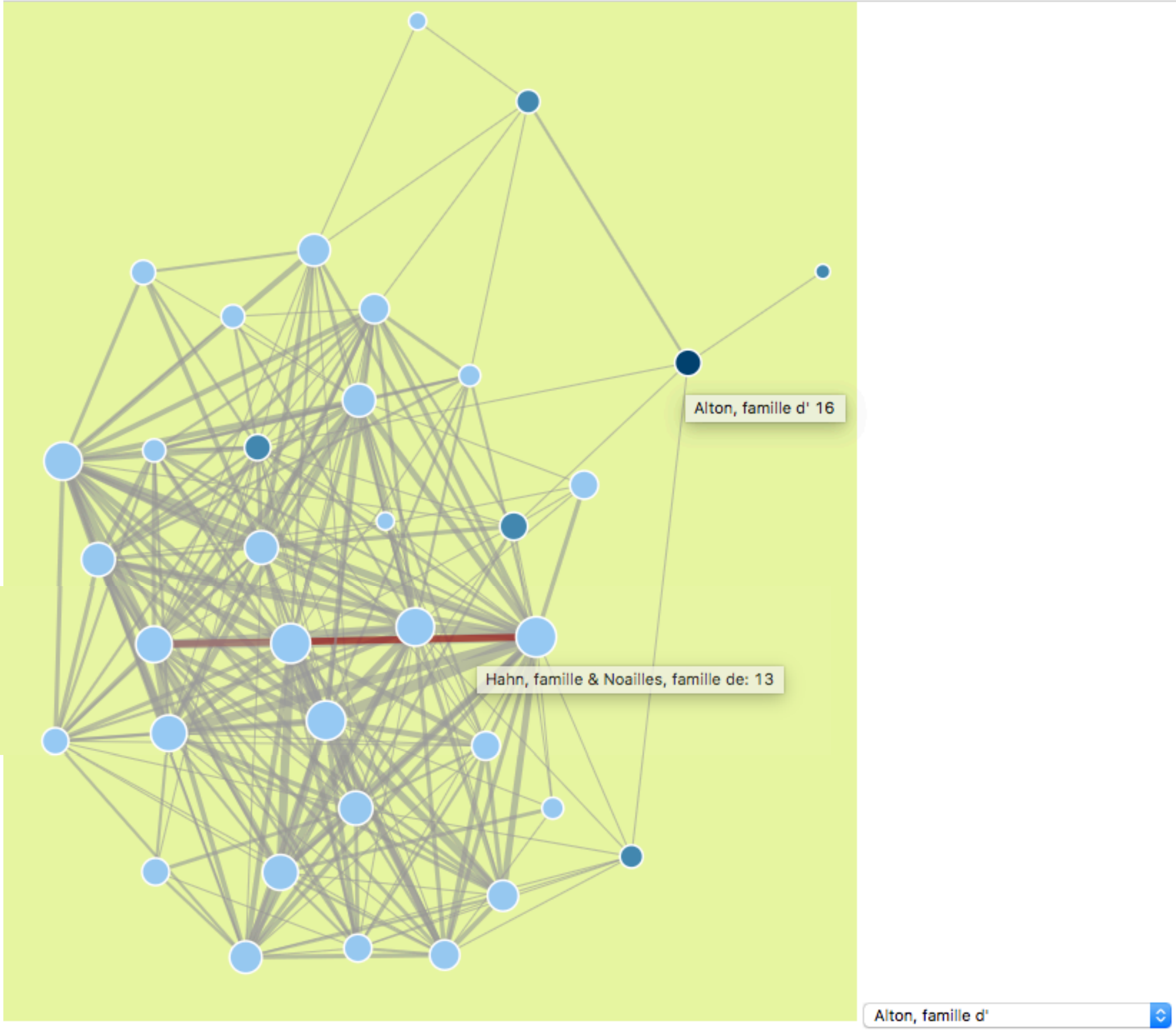
The **Alexandra Theatre** was a theatre located in the Stoke Newington district of London. It was located at 65 and 67 Stoke Newington Road where the present-day Alexandra Court now stands.

History [edit source]

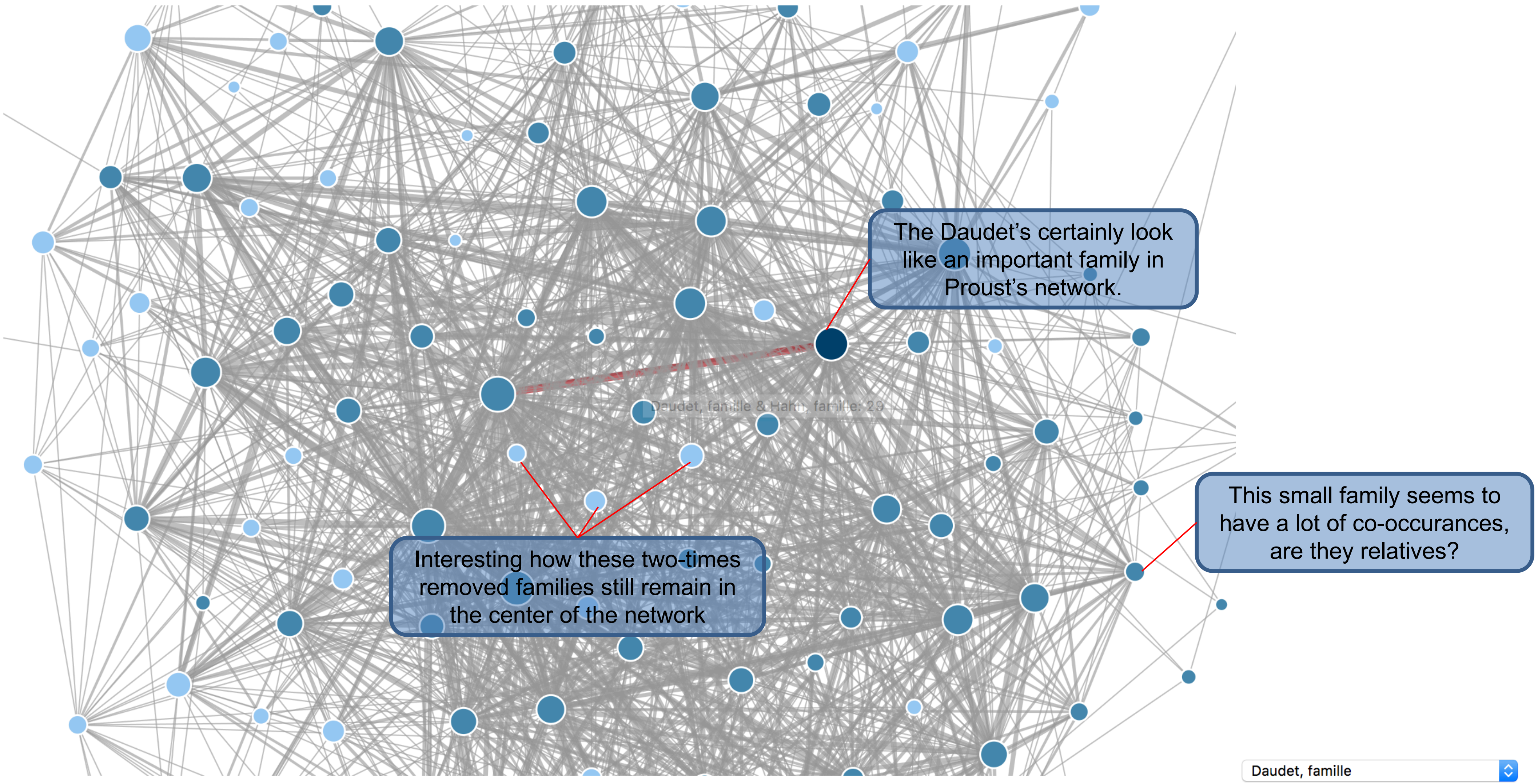
Opened on December 27, 1897 as *The Alexandra Theatre and Opera House*, it was designed by theatre architect Frank Matcham for F. W. Purcell. Upon opening, the theatre had a capacity of 2,025, spread across pit, traditional circle and gallery seating. In 1904, city records list the theatre's capacity as 1,710, along with an assessed value of £1,250.[1]. The theatre's first performance was the December 27, 1897 staging of *Dick Whittington*, an adaption of the pantomime Dick Whittington and His Cat.[2]

The theatre was operated by F. W. Purcell until 1905 when he sold it to new owners. The theatre's new owners changed its name to the Palace

Data Visualizations



Future Work – Allowing User Annotations

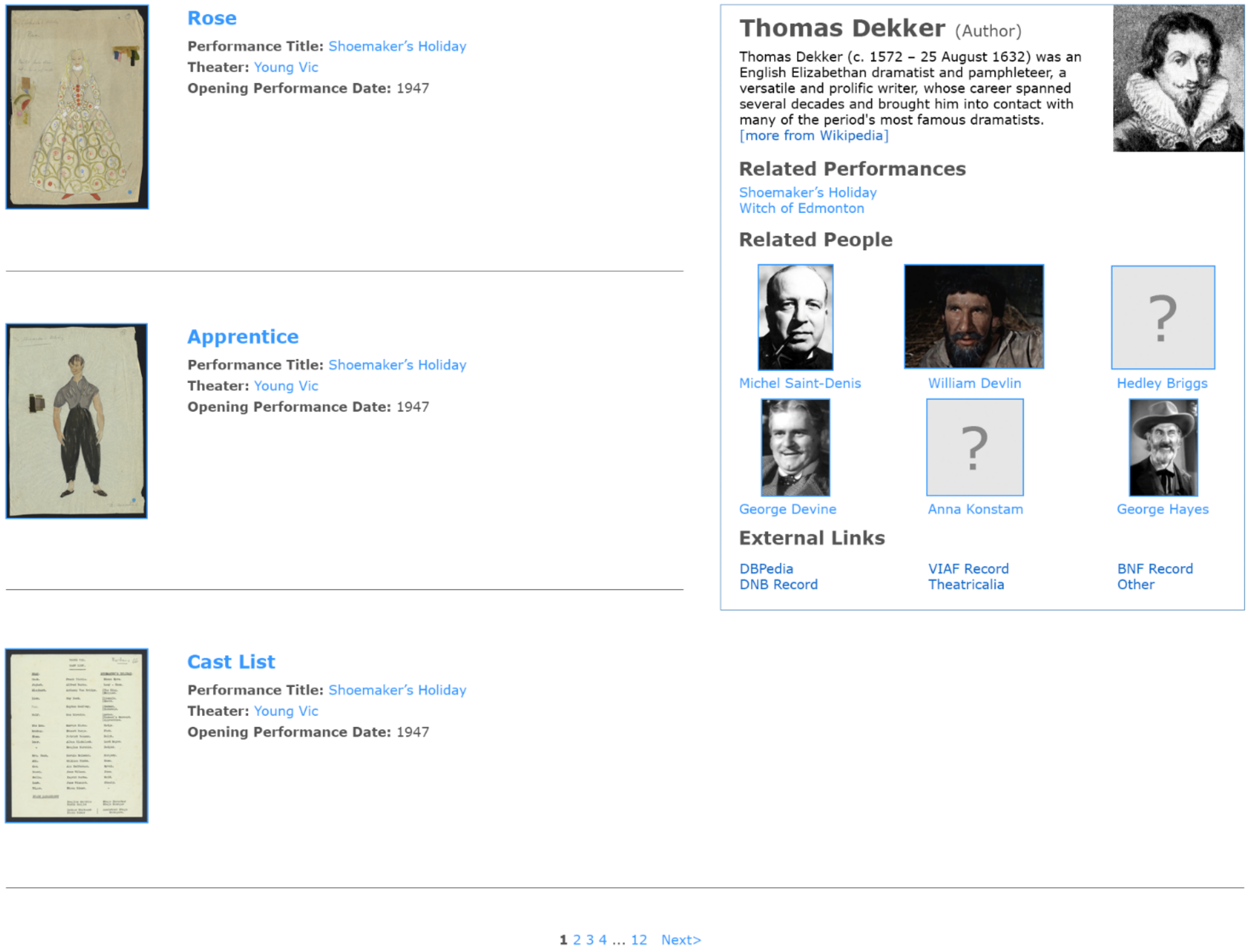


The Daudet’s certainly look like an important family in Proust’s network.

Interesting how these two-times removed families still remain in the center of the network

This small family seems to have a lot of co-occurances, are they relatives?

Future Work – Knowledge Cards on Search Results Pages



Rose

Performance Title: Shoemaker's Holiday
Theater: Young Vic
Opening Performance Date: 1947

Apprentice

Performance Title: Shoemaker's Holiday
Theater: Young Vic
Opening Performance Date: 1947

Cast List

Performance Title: Shoemaker's Holiday
Theater: Young Vic
Opening Performance Date: 1947

Thomas Dekker (Author)

Thomas Dekker (c. 1572 – 25 August 1632) was an English Elizabethan dramatist and pamphleteer, a versatile and prolific writer, whose career spanned several decades and brought him into contact with many of the period's most famous dramatists. [more from Wikipedia]

Related Performances

Shoemaker's Holiday
Witch of Edmonton

Related People

Michel Saint-Denis
William Devlin
Hedley Briggs
George Devine
Anna Konstam
George Hayes

External Links

DBPedia
DNB Record
VIAF Record
Theatricalia
BNF Record
Other

1 2 3 4 ... 12 Next>

Thank you for listening!

Questions?

Field Name	Mapping to schema.org – schema:isPartOf (schema:CreativeWork)
	schema:additionalType (URL) [spc:StageWork]
Performance Title	schema:name (Text)
Theater	schema:locationCreated (schema:Place)
Opening Performance Date	schema:dateCreated (Date)
Notes	schema:description (Text)

Field Name	Mapping to schema.org – schema:contributor (schema:Person, schema:Organization)
Associated People (...)	schema:name (Text)
(role)	schema:jobTitle (Text)

Field Name	Mapping to schema.org – schema:exampleOfWork (schema:Book)
Performance Title	schema:name (Text)
[publication date]	schema:datePublished (Date)
[part of]	schema:isPartOf (schema:CreativeWorkSeries) [when true]
Author/Composer	schema:author (schema:Person)

Add @id (URI), schema:sameAs (URI), schema:birthdate (Date), etc.
as found from LOD resources and services

Why Schema

- Industry-wide use by Web search engines
- Some promising schema's (e.g., Bibframe 2.0, etc.) where still under development at the time of the project's beginning
- Some existing schema's were considered to "heavy-weight" for the project's data needs and goals (e.g., FRBR_{OO}, CIDOC-CRM, etc.)
- Some existing schema's did not have wide-spread adoption (e.g., the SPAR family of ontologies)
- Were able to reuse previous library-oriented work (at UIUC and OCLC) with Schema.org