20 March 2017 Draft (NOT FINAL)

White Paper Component

Transforming Special Collections Metadata to Linked Open Data: Working with Motley Collection of Theatre and Costume Design and Portraits of Actors

Transforming and migrating metadata for special collections to linked data requires metadata remediation, reconciliation, and mapping processes. This white paper summarizes the processes performed for the tow special digital collections, Motley Collection of Theatre and Costume Design[1] and Portraits of Actors[2], housed in the same digital assets management system CONTENTdm, including challenges encountered and solutions identified during the transforming process.

1. Metadata extraction

Since these two special digital collections (Motley Collection of Theatre and Costume Design and Portraits of Actors) are housed in the same digital asset management system, we exported collections metadata from CONTENTdm in a tab-delimited text file that has all local field names in a first row as shown in figure 1 below. In order to perform metadata remediation and reconciliation works, we then saved the exported metadata into an Excel spreadsheet.

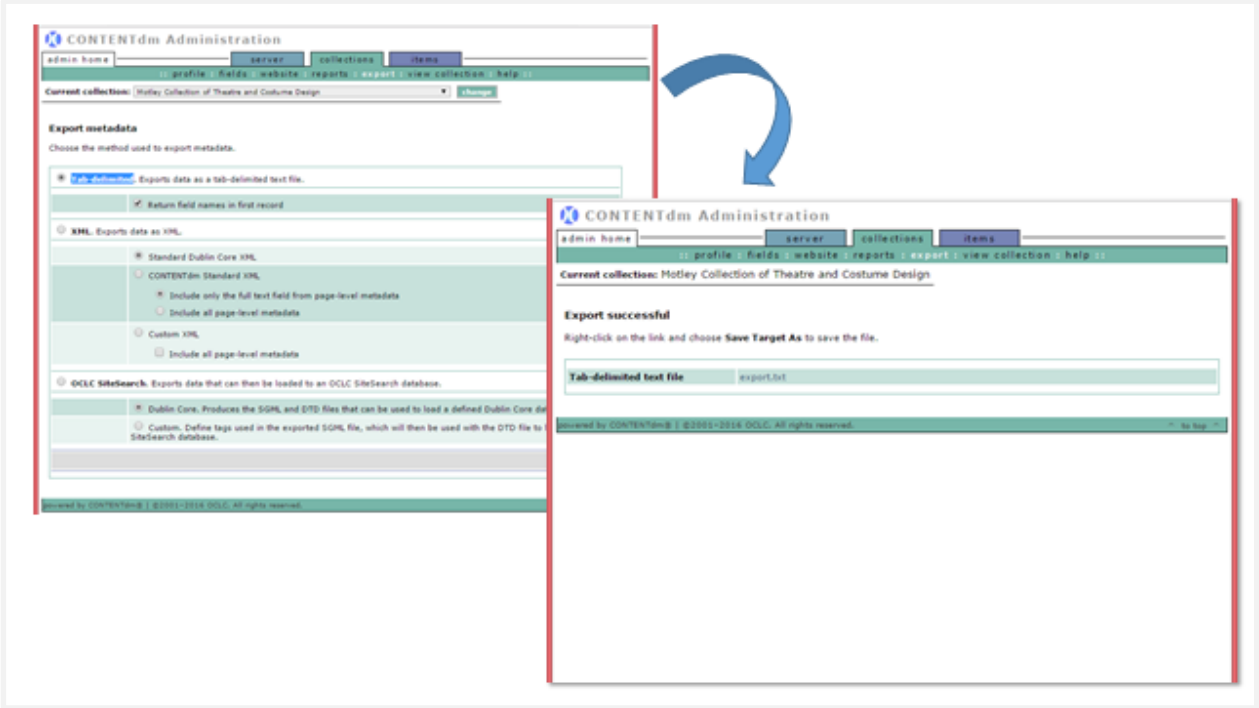

Figure 1: Metadata extraction from the CONTENTdm in a text file.

[1] http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/motley

[2] http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/actors

2. Metadata remediation

Although the metadata for these two collections has been cleaned up several times over the years, we identified that there were two areas where metadata remediation was needed before reconciliation and transformation into linked data could take place:

- The actual data cleanup/enhancement
- Changing of metadata fields in the Excel spreadsheet

Metadata cleanup and enhancement

The actual data cleanup is a very important task that is time-consuming as well as labor-intensive in the linked data preparation, since reconciliation only works when the values used for the metadata are in controlled terms. For this process, we focused on personal, corporate (theater), place names and subject terms since there are various established controlled vocabularies that support linked data, i.e., each term is represented with a unique URI that also includes information associated with the term. We examined terms used in the metadata with authority files and replaced them with controlled terms when it was applicable. For personal and theater names, we used following sources in addition to various print sources:

- Library of Congress (LC) Name Authority Files[3]
- Virtual International Authority File (VIAF)[4]
- Internet Movie Database (IMDb)[5]
- Internet Broadway Database (IBDB)[6]
- Wikipedia[7]
- Worldcat Identities[8]

There were a few names that required utilizing less common sources, such as:

- Canadian Theatre Encyclopedia[9]
- Encyclopedia Britannica[10]
- Turner Classic Movies[11]
- Goodreads[12]
- Obituaries in various digital newspapers
- Australian Dictionary of Biography[13]
- doollee.com[14]
- Opera Scotland[15]

---

[3] http://id.loc.gov/authorities/names.html
[4] http://viaf.org
[5] http://www.imdb.com/
[6] https://www.ibdb.com/
[7] https://en.Wikipedia.org/
[8] https://www.worldcat.org/identities/
[9] http://www.canadiantheatre.com/
[10] https://www.britannica.com/
[11] http://www.tcm.com/
[12] https://www.goodreads.com/
[13] http://adb.anu.edu.au/
[14] http://www.doollee.com/

- Copies of text on Amazon Books[16]
- Theatricalia[17]

Theatricalia was referenced for cross-checking individual's identities and their involvement with a particular performance. J.P. Wearing's book *The London Stage 1930-1939: A Calendar of Productions, Performers, and Personnel*, as well as the editions for 1940-1949 and 1950-1959, were also used to confirm cast and personnel lists of plays.

For locating sources for names, the LC Name Authority Files, Wikipedia, IMDb, and IBDB were the most useful in confirming the identity of an individual because generally the existence of one of these sources implied that the individual was established in the field of theater. The latter three sources were especially helpful for their extra contextual information. Additional sources were found as a result from simple Google searching, and from following links found on pages for particular performances. The surplus of individuals who never became established made confirming their roles difficult, but many were found in online encyclopedias and smaller databases. Finally, though Theatricalia is not one of our official sources and is an incomplete database, it was useful for confirming cast lists for specific productions, and gathering or confirming tentative birth and death dates.

Library of Congress Subject Headings,[18] Art & Architecture Thesaurus,[19] and Thesaurus for Graphic Materials[20] were used for matching subject terms. While doing this work, metadata enhancement was done as well, e.g., information that is included in the items but was previously missing from the metadata was added and capitalization, typos, and punctuation errors were also corrected.

Changing metadata fields

Metadata fields in the excel spreadsheet were also changed during this process. We identified that there were several fields that had more than one value separated with a semicolon, such as <Theater Name> and <Object>. This is a common practice in the CONTENTdm software, i.e., when there is more than one value for a field, each value is separated by a semicolon, instead of repeating the field. However, we realized that when more than one personal name or subject term is added into one field, it becomes harder to perform the automatic reconciliation work. So we divided these values into a separate field with the same field names. Some fields had been condensed to have two meanings, such as <Author/Composer>. These were divided into two fields, <Author> and <Composer> so the mapping to schema.org semantics would be clearer and more accurately represented. Also noticed was the information hidden within the text strings. For example, the field <Associated People> included a name with a role in a parenthesis such as Shaw, Glen Byam (director). In addition to <director>, there are other roles that appeared in the field, including <actor>, <producer>, <dancer>, and <translator>. All of

---

[15] http://www.operascotland.org
[16] https://www.amazon.com/books-used-books-textbooks/b?ie=UTF8&node=283155
[17] https://theatricalia.com/
[18] http://id.loc.gov
[19] http://www.getty.edu/research/tools/vocabularies/aat/
[20] http://www.loc.gov/pictures/collection/tgm/

these roles have matching semantics in schema.org, so we created new fields for them and moved each person with a unique role into the appropriate field. New information about Associated People and their roles is confirmed by cross-referencing existing metadata and consulting outside resources, and new roles for Associated People were added as needed. Names without any role information were kept in the <Associated People> field that later mapped to <schema:contributor>.

3. Metadata reconciliation

Metadata reconciliation work was performed against all personal names in <Author>, <Composer>, <Associated People>, as well as <Theater> names. Since a large portion of the names included in these two collections are people in performing arts, not authors of scholarly works, we extended our sources for the reconciliation work into not only the VIAF but also the IMDb, the IBDB, and Wikipedia as well. Reconciliation work with VIAF was performed both manually and in batch while the work with IMDb and IBDB were done manually. For the automatic reconciliation process with VIAF found 477 matches among 916 personal names (about 52%). Among the matches found in VIAF, 331 entries also have Wikipedia links (about 69.4%). The manual process complemented the automatic process. It found additional 25 Wikipedia entries and 533 entries from the Theatricalia that helped us to identify specific roles each person played for the certain performance mentioned in the metadata. In addition the Theatricalia, IBdd, and IMdb provide contextual information, including past productions and plays authored, of people those names do not available in LCNAF, VIAF, or Wikipedia. Because of the manual process, 195 of the names can have Theatricalia, IMDb or IBDB links. For theater names, while the automatic process found 6 matches, the manual process found 10 matches from the VIAF and 8 matches from the Wikipedia.

	Total No.	Sources	Manual	Automatic
Personal Names	916	VIAF	87	477
		LCNAF	197	n/a
		WorldCat	93	n/a
		Wikipedia	356	331
		IMDb	388	n/a
		Theatricalia	533	n/a
		IBDB	49	n/a
		Britannica	18	n/a
Theater Names	11	VIAF	10	6
		Wikipedia	8	2

Table 1: Name reconciliation results of both automatic and manual processes.

The process of cleaning the metadata and matching each individual name and performance with the collection of multiple sources took six months of graduate student work who worked 10 to 12 hours per week that totaled about 240 hours. As of this writing, we are in the middle of working on names appeared in the Portraits of Actors Collection.

4. Metadata mapping to schema.org

The Motley Collection of Theatre and Costume Design and Portraits of Actors collections consist of performing arts related items described in customized Dublin Core metadata with very specific display names which appear in the user interface. We tried to map the local field names that also include contextual information to schema.org semantics. (See Appendices 1 and 2 for Motley Collection of Theatre and Costume Design and Portraits of Actors collections field names.)

Metadata for Theater Collections: Creating Relationships between Item and Play

Metadata for the Motley Collection of Theatre and Costume Design uses 24 fields that describe the digitized item itself, the printed resource, the collection from which the item originated, and the original performance that the item was created for. As a first attempt, we created a mapping using <VisualArtWork> to describe the item itself and <TheaterEvent> to describe the original play and related information. As we moved forward, we realized that the base type <TheaterEvent> (a type of Event) was a poor match for the play since it was limited to actual performances, i.e., events, of it. This was a complication because the <VisualArtWork>, e.g., a costume drawing, was only indirectly linked to particular performances through an unrepresented <CreativeWork> entity. We consulted with the schema.org community and discovered that an Online Theater Ticket Sales Agency[21] had a similar use case to ours. The Online Theater Ticket Sales Agency and we presented our use cases and opened a discussion among the schema.org community which led to the proposal of a new <CreativeWork> type <StageWork>. This entity matched the actual representational situation: A <VisualArtWork> is part of some <StageWork> for which multiple <TheaterEvent>s are performed.

The current working mapping employs <CreativeWork> and two types of <CreativeWork>: <VisualArtWork> and <StageWork>. <VisualArtWork> has the property <isPartOf> to describe the collection that the item is a part of, and connects to <StageWork>. In order to make the fullest use of the enriched metadata a decision has been made to mint persistent identifiers (IRIs) for both the <VisualArtWork>s and the <StageWork>s. The current mappings use the URL of the CONTENTdm splashpage (i.e., its reference URL) for the <VisualArtWork>'s identity as a placeholder as we develop the infrastructure to mint persistent identifiers. The <StageWork> describes the author and composer of the play, and the original play from which the <StageWork> was based on, by using the <exampleOfWork> property. In some cases there are additional nesting recursive relationships where a <StageWork> is an <exampleOfWork> of another <StageWork> which itself is an <exampleOfWork> of a <Book>.

We also identify properties that should be included in schema.org that will further describe both <VisualArtWork> and <StageWork>. Under <VisualArtWork>, we will propose three properties; <artStyle> and <artPeriod>, which are included in Visual Resources Association (VRA) Core[22] and Categories for the Description of Works of Art (CDWA)[23] as one element, and <standardNumber> to describe a local item number. Under the <StageWork>, because the metadata includes several personal

[21] http://www.globetrottoirs.com
[22] https://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf
[23] https://getty.edu/research/publications/electronic_publications/cdwa/definitions.pdf

names, each having a specific role, we initially considered proposing each role as a separate property, e.g., director, choreographer, dancer, set designer etc., to clearly describe roles played in the <StageWork>. But we decided to use the recommendation of <contributor> with role information. Other than <productionVisual>, <StageWork> will have all properties allowed in <CreativeWork>, including <text>, <name>, <dateCreated>, <locationCreated>, and <exampleOfWork>. By this mapping, two different relationships between resources have been created, between an item and a collection the item belongs to, and between an item and a play for which the item was created. See the current mappings from table 1 below.

Field Name	**Mapping to schema.org – schema:VisualArtwork**[24]
Image Title	schema:name (Text)
Object	schema:genre (Text)
Type	schema:artform (Text or URL)
Material/Techniques	schema:artMedium (Text or URL)
Dimensions	schema:height & schema:width (schema:Distance or schema:QuantitativeValue)
Subject I (AAT)	schema:about (schema:Thing)
Subject II (TGMI)	schema:about (schema:Thing)
Subject III (LCSH)	schema:about (schema:Thing)
Rights	schema:copyrightHolder (schema:Organization)
Physical Location	schema:provider (schema:Organization)
Inventory Number	spc:standardNumber (Text or URL)
JPEG 2000 URL	schema:associatedMedia (schema:CreativeWork)
Collection Title	schema:isPartOf (schema:Collection)
[Design by]	schema:creator (schema:Organization) [always Motley (Organization) in this case]
[is part of Stage Production]	schema:isPartOf (schema:CreativeWork, spc:StageWork)

[24] Items for Character sketch, Costume design, Costume rendering, Costume sketch, Costume work drawing, Instrument rendering, Mask sketch, Prop design, Props, Sandals sketch, Set design, Set desing, Set detail, Set rendering, Sketch, Stage props, Working drawing are mapped to the <schema:VIsualArtWork>. Items for production notes and Cast notes, i.e., textual materials, are mapped to the <schema:CreativeWork>.

Field Name	Mapping to schema.org – schema:CreativeWork
Performance Title	schema:name
Theatre	schema:locationCreated (schema:Place)
Opening Performance Date	schema:dateCreated (Date)
Notes	schema:description or schema:mainEntityOfPage
[additional type]	schema:additionalType (URL) [spc:StageWork]
[production of]	schema:exampleOfWork (schema:Book, fabio:Play)
Field Name	**Mapping to schema.org – schema:Book**
Author/Composer	schema:author (schema:Person)
[additional type]	schema:additionalType (URL) [http://purl.org/spar/fabio/Play]
[Published Work]	schema:name
[publication date]	schema:datePublished (Date)
[part of]	schema:isPartOf (schema:CreativeWorkSeries) [when true]
[adaptation of]	schema:exampleOfWork (schema:Book or schema:CreativeWork) [when true]

Table 2: Mapping from Motley collection's local field names to schema.org.

Portraits of Actors: Creating Relationships between Item, Play, and Book

With lessons learned from the Motley collection mapping, we reviewed field names used for the Portraits of Actors collection (see appendix 2) and identified four different types of <CreativeWork>: <VisualArtWork>, <StageWork>, <Book>, and <Collection>. For the <VisualArtWork>, we mapped fields that described the visual image itself, such as ID Number, Title, Date, and Subject. Local fields that described the actual play performed were mapped to <StageWork>, and physical and digital collections information are mapped to <Collection> to show the relationships between item and collections (both physical and digital). We decided to use the <Book> for additional published work of a play with a <additionalType> that has 'Play' as the default value. Please see the current mapping for the Portraits of Actors collection from table 2 below.

Field name	schema.org mapping Thing > Creative work > VisualArtwork
ID Number	scp:standardNumber (Text or URL)
Title	schema:name (Text)
Date	schema:dateCreated (CreativeWork)

Role	schema:character (schema:Person)
Subject	schema:about (schema:Thing)
Type	schema:artform (Text or URL)
Dimensions	schema:height and schema:width
Technique	schema:artMedium (Text or URL)
Creator	schema:creator (schema:Person or schema:Organization)
Publisher	schema:publisher (schema:Organization)
Description	schema:description (Text)
Rights	schema:license and use URL. (The statement should be stored in somewhere, such as Project webpage.)
[copyright]	schema:copyrightHolder (schema:Organization and <rdf:about="http://viaf.org/viaf/123824539"> for UIUC Library as a default value.)
Collection	schema:isPartOf (schema:Collection)
Repository	schema:provider (schema:Organization)
[photo]	schema:isPartOf (scp:StageWork)
Field name	**schema.org mapping Thing > Creative work > StageWork**
Play	schema:name (Text)
[photo]	schema:image (URL) (Use the value of the reference URL from CONTENTdm)
Field name	**schema.org mapping Thing > Creative work > Book**
[published work]	schema:name (Text)
[additional type]	schema:additionalType (URL) [http://purl.org/spar/fabio/Play]
Field name	**schema.org mapping Thing > Creative work > Collection**
Collection	schema:name
Physical collection	schema:isPartOf (schema:Collection) [asserting that Portraits Collection is part of the physical collection]

Table 3: Mapping from Portraits of Actors collection's local field names to schema.org.

5. Preparing special digital collections metadata for linked data

Preparing special digital collections metadata for linked data conversion requires several processes including metadata cleanup, enhancement, and reconciliation. Although metadata for these two collections were created by subject specialists and had undergone several iterations of metadata cleanup, we found that linked data imposed a new set of challenges as below.

Working with unique local field names

While special collections could be described better with unique field names, when a local field name contains more than one meaning or multiple values with different roles in parentheses, those values AND the field name are better to be separated for the semantic mapping. As mentioned in the Motley collection mapping examples, the field name <Author/Composer> includes two distinct roles that schema.org can accommodate with two different properties. Another example is the local field <Associated People>, which includes values with name and role for which each role has its own property in schema.org, such as director, producer, and etc. Although these local field names work perfectly fine in the current digital collections user interface, collection owners and metadata professionals need to consider reviewing the way local field names are created and used in conjunction with possible linked data conversion work.

Working with metadata values

Metadata value cleanup and enhancement processes that ensure the values used in the metadata are controlled terms is the first step in moving toward linked data, because the reconciliation result depends on them. However, metadata cleanup is not as easy as it seems. Most of the names used in special collections are not well-known individuals whose names have established name authority files. Also some names have been changed over time, so tracing of those names and decisions on preferred names for display and use became a challenge. In addition, since many performers' names do not have authority files, there is no way we can use URIs for those names. We understand that there is discussion on establishing and developing a workflow for a local authority file that supports URIs. We hope that we will implement a system that hosts local name authority files for names without already established authority files.

Working with linked data sources that are not OPEN

We have learned that not all linked data sources are open and have the same services. As names included in our collections are associated with theater and performing arts, we found that the Internet Broadway Database and the International Standard Name Identifier (ISNI, www.isni.org/) include many names that are not found in VIAF. However, we quickly learned that the information available in the Internet Broadway Database is under copyright and cannot be used without prior permission. Also, although ISNI includes many performers' names in their database, in order to use their API service for batch searching, we have to be a member of the ISNI Community, which requires a fee. So when identifying linked data sources, it is also important to check whether the service is free to use and is closed or opened. Currently we are in the middle of communicating with the Internet Broadway Database to obtain permissions.

Metadata work requires manual process

Every step of metadata work - cleanup, enhancement, and reconciliation - requires manual process. We have learned that the batch process or automatic process can improve the metadata quality only so much that it may also cause unforeseen mistakes, i.e., wrong matches or adding wrong values into a wrong element. Although the metadata we have worked with for this project are in a fairly good quality and created by subject specialists, they required a new set of workflows for data cleanup and enhancement for the linked data transformation. Also for the reconciliation work, when there are multiple entries with the same name, we painfully learned that machine could not disambiguate and identify the exact match. Rather it usually picked up the first entry. For this reason, we have spent more than six months for data preparations for two theater collections, and the Kolb-Proust collection's metadata cleanup and enhancement work is still underway at this point.

Appendix 1: Local file names used for the Motley Collection of Theatre and Costume Design

Field name	DC map	Note	Controlled vocabulary
Image Title	title	Title of image	No
Dimensions	extent	Size of the physical item	No
Associated People	description	Add the qualifiers after the name, usually director, producer and actor's names are available	Yes (LC NAF)
Inventory Number	none	Inventory number used locally	No
Description	description	Additional information about the item or play	No
Inscriptions	description	Information inscribed on item	No
Repository	source	Holding library where the physical item is housed	Yes (LC NAF)
Collection	isPartOf	Collection title	No (one default value)
Author/Composer	contributor	Creator of the play or opera	Yes (LC NAF)
Production Notes	relation	Additional information of the performance (URL)	No
Performance Title	references	Title of the performance	Yes (LC NAF)
Theater	description	Name of the theater where the performance was held	Yes (LC NAF)
Opening Performance Date	date	Performance date	No (ISO 8601)
Materials/Techniques	medium	Materials/technique used for the item	Yes (TGM II)
Object	format	Describe the genre of the item	Yes (AAT)
Type	type	Type f the item	Yes (DCMI Type)
Subject I (AAT)	subject	Descriptions of costume and furnishings as well as concepts and style	Yes (AAT)
Subject II (TGM I)	subject	Descriptions of physical characteristics of the item	Yes (TGM I)

Appendix 2: Local file names used for the Portraits of Actors

Field name	DC map	Note	Controlled vocabulary
ID Number	identifier	Alpha-numeric code based on name of actor	not necessary- there is no duplication
Title	title	"Portrait of [name of actor]" or "Name of Actor as [role] in ["Play"]" or "Name of Actor 1 as [role] and Name of Actor 2 as [role] in a scene from ["Play"]	no
Date	created	4-digit year print was made, if known.	not necessary
Role	description	Controlled list of role names	local
Play	description	Controlled list of play titles (short titles, not the long titles many of these 18th century plays have)	local
Subject	subject	Name of actor from LC NAF- For those not in NAF, a name authority is created and recorded on the spreadsheet **Actors_portraits_data.xls**; some of these may not be detailed enough. Other LCSH subject headings, including headings that have general date ranges corresponding to when the actor was working [ie. eighteenth century] LCSH was chosen rather than Thesaurus for Graphic Materials because it had terms like "theatrical manager" and "breeches parts" that seemed necessary to describing this collection. The term "costume" was used whenever the actor is depicted in a role. It was difficult to decide whether to include a subject heading like "blackface entertainer" when a particular portrait was not in blackface, but in general we did so.	LCSH NAF
Type	type	Type of print, photograph, etc. *photomechanical prints* was used as a fairly catch-all category	AAT
Dimensions	extent	Indicate whether the measurement is: image sheet mounted sheet	

		plate marks …or whatever it says on card, ie.: image 3 x 2 ½ inches sheet 5 x 5 ¼ inches	
Technique	medium	Artistic/technical technique(s) used to create the print	AAT
Creator	creator	Name of artist whose painting/drawing the print was based on; name of printmaker; name of photographer	No – hard to control. Used the CV function of CONTENTdm just to catch typos
Publisher	publisher	Name of publisher; other corporate name responsible for making print available (i.e. name of lithography firm) – when difficult to distinguish between creator/publisher, an attempt was made to use publisher for corporate names, but there is some lack of consistency with this. Another issue: usually the publisher is an individual (i.e. John Bell) – and we used the name J. Bell rather than the title of the publication (*Bell's British Theatre*)	no
Description	description	Free-text description, including some details from costume, scenery. Including whether portrait is whole-length, half-length, bust, etc., for costume researchers who may only want certain portrait types. (I began to include the word "portrait" in the description because terms like "whole-length" seemed a little vague. This isn't consistent throughout the collection, ideally, it would be.)	no
Rights	rights	Not sure	blanket statement
Digital Collection	isPartOf	Portraits of Actors, 1720-1920 University of Illinois Digital Collections	blanket statement
Repository Collection	source	University of Illinois Theatrical Print Collection, #35. (It seems important to tie the actors portraits to the larger physical collection from which they were derived. Would be good to link to the rbml/archon finding aid when it is given a stable url.)	blanket statement
Digital File Creation	none	[administrative metadata about how scanning and conversion was done]	blanket statement
File Name	identifier	Name of the image file	