20 March 2017 Draft (NOT FINAL)

White Paper Component

Working with Kolb-Proust Collection's Metadata

About Kolb-Proust Collection and Its Metadata

Philip Kolb (1907 - 1992) was a pre-eminent Proust scholar and lead editor of the 21-volume edition of Proust's correspondence. The Kolb-Proust Archive's[1] collection consists of the transcribed research notes and documentation of Philip Kolb. The original textual data and metadata for the Kolb-Proust collection were encoded using TEI P5 v. 2.0.0. The collection maintains a local name database for all names that appear in the research notes, thus in the TEI file. The name database works as the local name authority file that includes authorized forms of all names and a wide range of various information about each name in the <note> field, such as dates of birth, wedding or divorce, family relationships (spouses, parents, children, and other relations) and information about professions. Metadata analysis and transformation for the Kolb-Proust Collection is being performed in two different parts: a name authority database and information contained in the original research notes.

1. Working with the Name Database

In order to better understand the social networks between people mentioned inside and outside of the collection, we reviewed the metadata about each name available in the <note> field and identified a set of relationships as well as specific information that might help to enhance building a linked name representation of the collection's name information that could be used for the visualization of the social network of people associated with Proust. To accomplish this we chose a linked open data-compliant vocabulary, Schema.org, to encode the information. One of Schema.org's entities, <Person>, has a list of properties that works well for representing relationships and other information available in the existing <note> field. We selected the following 10 properties to encode in our linked data representations:

- schema:familyName
- schema:givenName
- schema:birthDate
- schema:deathDate
- Schema:gender
- schema:nationality
- schema:spouse
- schema:children
- schema:parent
- schema:sibling
- schema:relatedTo
- schema:jobTitle.

[1] http://www.library.illinois.edu/kolbp/

Among these properties, following properties were added using automated scripts:

- schema:gender (by using specific terms included in the name (such as Mme=Mrs. and Mlle=Miss for female, and M=Mr. for male)
- schema:familyName
- schema:givenName (by using a 'comma' used in the name)
- schema:birthDate
- schema:deathDate (by using dates added after a personal name).

A graduate student was hired in late September 2016 and has been cleaning up the name database, sorting the information provided in the original <note> field into the 10 properties, as well as making explicit family relationship information contained in the form of some names that existed within the original <name> field. For example the entries ***"Daudet, Marthe Allard (daudet6) -- <note>1878-1960, cousine et 2ème femme de Léon"*** and ***"Daudet, Léon (daudet1) -- <note> 1868-1942, fils aîné d'Alphonse Daudet"*** include the husband/wife information within the name of the first entry. After encoded with schema.org properties, the first name entry will be represented as a graph below.

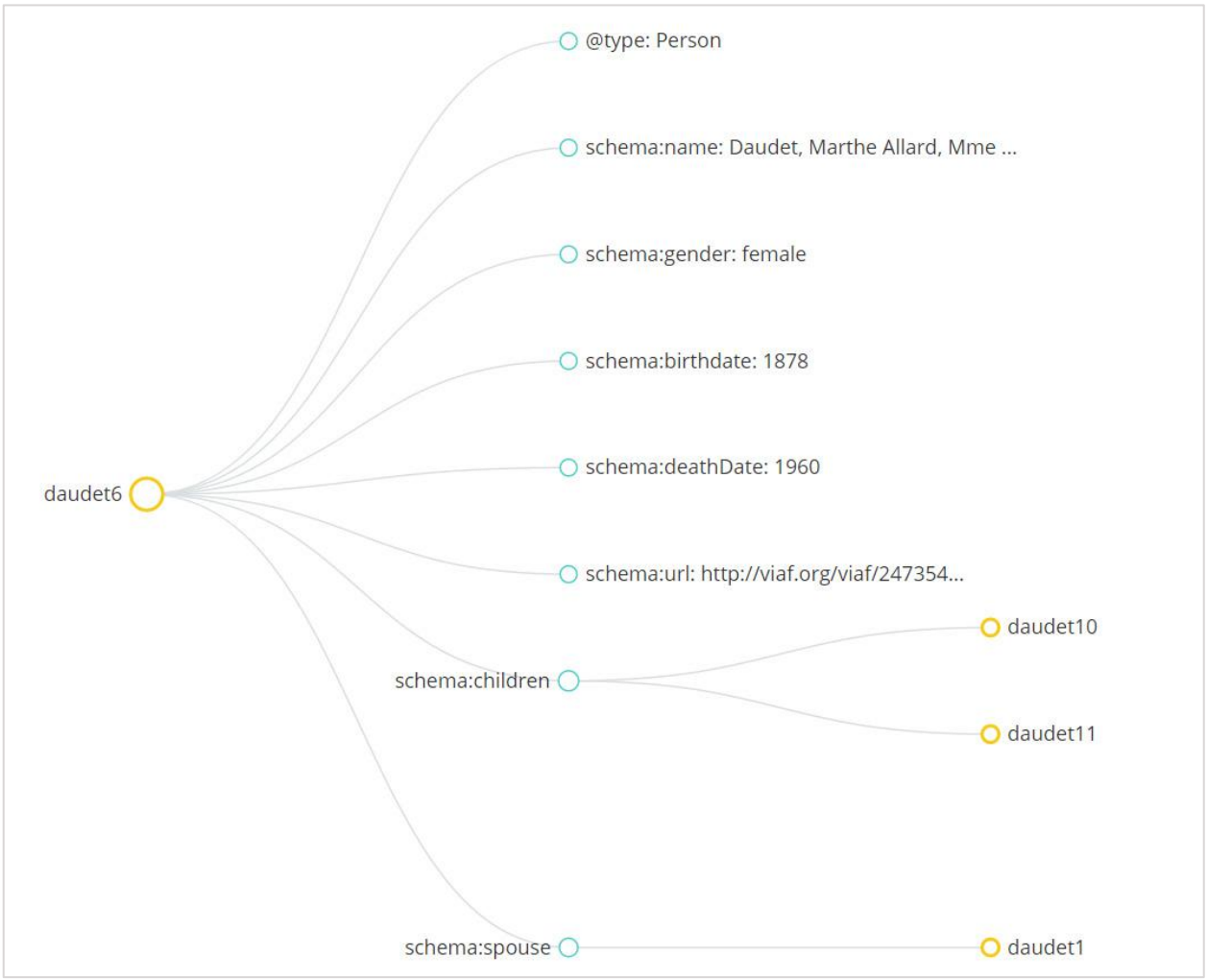

Figure 1: A name encoded with Schema.org properties

2. Entity Reconciliation

Names included in the collection's Name database were searched against VIAF, like the other two collections for the research. [Automated match counts to be added]. In addition to names, we have been looking at options for adding links for titles of journals (and exact volumes and page references) and books mentioned in the TEI documents. Many newspapers mentioned in the K-P Collection documents were digitized and made available by the National Library of France. However, we found that there is no easy way to find issue level links for these newspapers. As of now we are still looking at ways to generate issue- or page-level URLs for each citation.

3. Working with TEI Documents

The original research notes contain large amounts of information and citations collected in French historical newspapers and other publications. The original bibliographic citations are generally complete and follow a consistent style but their elements were not encoded, beyond the <title> and <name> element as below.

```
<div0 id="c20090" type="card">
   <head>
     <date value="18990000">1899</date>
   </head>
   <div1 type="subdiv">
     <bibl>Proust. <title type="es">La Peste de Vienne et le danger que peut faire
courir à l'Europe la peste du Turkestan.</title> <title>Comptes-rendus des séances
de l'Académie des sciences morales et politiques</title>, 59e année, p. 4
   </bibl>
   <bibl>Cf. 1897: Proust. <title type="es">La conférence sanitaire internationale de
Venise de 1897</title>, <title level="j">Revue d'Hygiène</title>, vol. XIX, p. 7
   </bibl>
   <bibl>Cf. 1897: Proust. <title type="es">La défense de l'Europe contre la
peste.</title> <title>Comptes-rendus des séances de l'Académie des sciences morales
et politiques</title>, 57e année, p. 4
   </bibl>
  </div1>
</div0>
```

Figure 2: A sample TEI data

As shown in the figure 2, each card is recorded separately in a <div0>, and each each <div0> has the <date> that records the date of the event and the <div1>. Each <div1> usually has one or more <bibl> and each <bibl> includes texts describing the event itself or mentioned in the publication. Detailed publication information is encoded with the <title> and sometimes with <name> with a proper role, such as <author>. We also found that some cards start with the <p> for texts and multiple <div1>s are added under the <listBibl> element. However, the context added into the <div1> and <bibl> are all the same.

Based on the TEI document structure and contents, we have come up with a mapping from TEI to the Schema.org semantics as below. By using this mapping the TEI XML document in the figure 2 can be encoded with Schema.org properties as below.

Field name	**Schema.org mapping**	**Remark**
div0	Thing > Event	A subcategory of Event can be used, if applicable.
div0@id	scp:standardNumber (text or URL)	
div0 > head > date	startDate (Event) endDate (Event)	
bibl	citation (CreativeWork) recordedAt (Event)	
bibl (datePublished)	datePublished (CreativeWork)	
bibl (pageStart)	pageStart (CreativeWork)	
bibl (pageEnd)	pageEnd (CreativeWork)	
title If title@level="a" If title@level="j" If title@level="s" If title@level="m"	name (CreativeWork) name (Article) name (Periodical) name (Series) name (Book)	
title@type= "es" (prose nonfiction) "ag" (graphic arts) "re" (prose fiction) "th" (theater) "sc" (sculpture) "sp" (variety show) "ds" (dance) "sp" (variety show) "ds" (dance) "po" (poetry) "mu" (music) title@type="op" (opera)	genre (CreativeWork) AND/OR typeOf="" typeOf="schema:VisualArtwork" typeOf="" typeOf="schema:StageWork" typeOf="schema:Sculpture" typeOf="schema:TheaterEvent" typeOf="schema:TheaterEvent" typeOf="schema:TheaterEvent" typeOf="schema:TheaterEvent" typeOf="" typeOf="" typeOf="schema:TheaterEvent"	
name	FamilyName; givenName (Thing > Person)	

Table 1: K-P TEI to schema.org mapping

```
{
    "@context": [
     "http://schema.org/",
    {
        "s": "http://schema.org/",
       "scp": "http://ns.library.illinois.edu/scp"
    }],
    "@id": "http://kolbproust.library.illinois.edu/proust/c20090",
    "@type": "Event",
    "startDate": "1899",
    "endDate": "1899",
    "name": "c20090",
    "url": "http://kolbproust.library.illinois.edu/proust/c20090",
    "location": "na",
    "recordedIn": [
       {
           "@type":"CreativeWork",
            "s:url": "http://kolbproust.library.illinois.edu/proust/c20090",
             "s:citation": "Proust. La Peste de Vienne et le danger que peut faire
             courir à l'Europe la peste du Turkestan. Comptes-rendus des séances de
             l'Académie des sciences morales et politiques, 59e année, p. 4",
           "s:genre": "prose nonfiction"
       },
       {
           "@type":"CreativeWork",
            "s:url": "http://kolbproust.library.illinois.edu/proust/c20090",
             "s:citation": "Cf. 1897: Proust. La conférence sanitaire internationale de
             Venise de 1897, Revue d'Hygiène, vol. XIX, p. 7",
             "s:genre": "prose nonfiction"
       },
       {
            "@type":"CreativeWork",
            "s:url": "http://kolbproust.library.illinois.edu/proust/c20090",
             "s:citation": "Cf. 1897: Proust. La défense de l'Europe contre la peste.
             Comptes-rendus des séances de l'Académie des sciences morales et
             politiques, 57e année, p. 4",
             "s:genre": "prose nonfiction"
       }
    ]
}
```


Figure 3: TEI document represented in JSON

4. Next Steps

As of this writing (March 10, 2017) we are in the middle of encoding information added in the <note> field of the name database and analyzing TEI documents to see how consistently TEI elements were used for citation information. Based on the analysis results, we will determine whether we will use a fully automatic transformation or relay on a hybrid automated-manual transformation from TEI to Schema.org semantics.