



**HAL**  
open science

# **Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes**

Simon Gabay, Lucie Rondeau Du Noyer, Mohamed Khemakhem

► **To cite this version:**

Simon Gabay, Lucie Rondeau Du Noyer, Mohamed Khemakhem. Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes. IX Convegno AIUCD, AIUCD, Jan 2020, Milan, Italy. <hal-02388407>

**HAL Id: hal-02388407**

**<https://hal.science/hal-02388407v1>**

Submitted on 1 Dec 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# Selling autograph manuscripts in 19th c. Paris: digitising the *Revue des Autographes*

**Simon Gabay**  
Université de Neuchâtel / ILF  
Neuchâtel, Suisse  
simon.gabay@unine.ch

**Lucie Rondeau du Noyer**

**Mohamed Khemakhem**  
INRIA / ALMAnaCH  
Université Paris Diderot, Paris 7  
mohamed.khemakhem@inria.fr

## Abstract

**English.** In Paris, the manuscript market appears in the early 20's of the 19th c. Fixed-price catalogues and auction catalogues are regularly published, describing each document in detail. Such descriptions being highly formalised, it is possible to extract and structure them (almost) automatically, and thus create a database of sold manuscripts in 19th c. Paris.

**Italiano.** Il mercato dei manoscritti appare a Parigi all'inizio degli anni '20 del XIX secolo. In questo contesto, mercanti specializzati vendono regolarmente cataloghi d'asta a prezzo fisso che descrivono minuziosamente ogni documento acquistabile. Tale testi hanno delle strutture ricorrenti che è possibile estrarre e strutturarle (quasi) automaticamente, creando così un database di tutti i manoscritti venduti nella Parigi del XIX secolo.

## 1 Introduction

The number of projects dealing with French related objects circulating on the private market is increasing: research is currently being carried out in art history (Saint-Raymond, 2018), book history (Montoya, 2018), medieval manuscripts (Wijsman, 2017). . . Following recent trends, the latter project, that is the most relevant to our own work, is now sharing its data with other teams (Burrows et al., 2019) in the USA<sup>1</sup> and in the UK<sup>2</sup> to trace the history of manuscripts over time and places, across national and linguistic borders.

Unfortunately, no similar survey has been conducted yet on modern French autographs. If Renaissance manuscripts and their history are better known thanks to the *Biblissima* project (Turcan-Verkerk and Bertrand, 2014), no systematic work has been carried out on 17th, 18th and 19th c. materials. However, sale catalogues are recognised as being useful, since they are, for instance, regularly used as sources for critical editions (Sévigé, 1978, p. 158) (Voltaire, 1960, p. 18) (Lamartine, 2001, p. 348).

Such a hole in our knowledge is due to the fact that it remains extremely tedious to extract information, because this task is either performed manually or with imperfect digital solutions (Cuadra and Michels, 2013; Barman, 2019). In the present paper, we therefore want to propose an (almost) automated workflow for the retroconversion of catalogues to transform images into structured information and create a database of sold items.

## 2 The corpus

### 2.1 The manuscript market

Since the beginning 19th c., rich collectors have been selling manuscripts on the private market (Bodin, 2000). Archives and libraries still keep sales catalogues that have been published on a regular basis (fixed price catalogues) or for special sales (auction catalogues) by dealers. Most of the manuscripts sold in these catalogues are modern and contemporary autograph manuscripts.

The information contained in these catalogues is crucial for at least four different reasons.

<sup>1</sup><https://sdbm.library.upenn.edu>

<sup>2</sup><http://mappingmanuscriptmigrations.org>

- It helps assessing the authenticity of autographs: it is unlikely that a document sold repeatedly on the private market, and therefore authenticated each time by an expert, is a forgery.
- It documents the reception of authors, via the history of collections (*i.e.* who collected what?) and prices (*i.e.* who costs how much?).
- It informs us on the distance between what has been sold and what is available in libraries (*i.e.* are there autographs we do not know about?).
- It provides us with images of documents which are still in private hands, because catalogues sometimes offer either facsimiles or pictures of autographs sold.

For a first test phase, we have concentrated our efforts on the *Revue des Autographes*, a journal published since 1860's in Paris by Gabriel Charavay.

## 2.2 The RDA collection

In the second half of the 19th c., the autograph market is mature and the first generation of dealers begins to retire. In 1865, Gabriel Charavay (1818-1879) ceases the opportunity of Auguste Laverdet's (1809-1867) retirement to take over his business, following the example of his elder brother Jacques (1809-1867), who opened his own shop in 1830. At that moment, Gabriel abandons his role of editor for *L'Amateur d'autographes*, a journal about the autograph market in Paris created in 1862, which keeps being published by his brother Jacques.

Realising the importance of a publication attached to his activities, Gabriel creates another journal one year after his installation, in 1866: the *Revue des Autographes, des curiosités de l'histoire et de la biographie (RDA)*. Two journals for such a small market is however too much: in December 1868, after eight months of interruption, the price of the publication is divided by two and part of the content consists now of a list based on the autographs for sale in Gabriel's stock. Over time, the proportion of articles keeps diminishing and the *RDA* becomes first a hybrid publication mixing news and items to be sold, and eventually a fixed-price catalogue with the name of a journal published (almost) monthly until 1936. In the meantime, Gabriel's shop is taken over by Gabriel's son Eugène (1879-1892), and then by Eugène's widow (1892-1918) and by Eugène's daughter (1918-1936).

The transformation of an hybrid journal into a disguised fixed-price catalogue under a journal's name is confirmed by a modification of the format: Eugène Charavay opts for a two-columns layout and a smaller font (cf. figure 1), harder to read but easier to browse for readers, who are now buyers, looking for the autograph of their dreams.

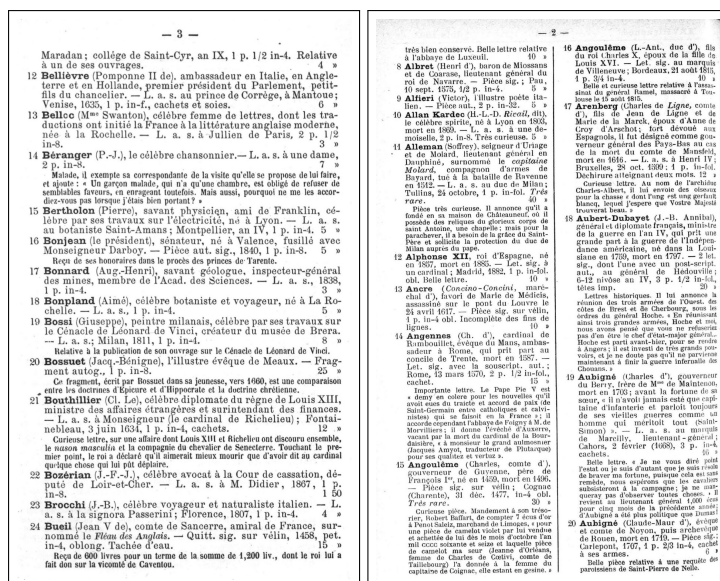


Figure 1: One-column layout (1873) vs two-columns layout (1893).

### 3 Encoding

#### 3.1 Entries

It is the images of these catalogues that we want to transform into minable data. Each of them generally contains a minimum of c.200 entries, all of them being extremely dense in information and always following the same structure:

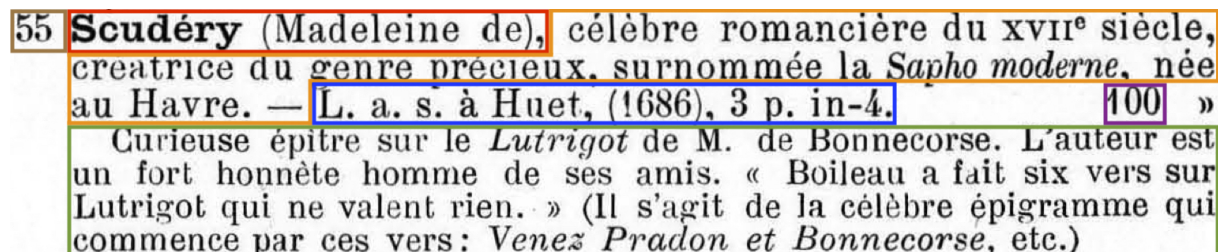


Figure 2: RDA, n°67 (March 1881), lot N°55.

We can clearly see the lot number (in brown), the name of the author (in red), a short biography (in orange), the material description of the autograph (in blue), the price (in purple) and an additional description (in green).

To render the structure of the document, we propose the following encoding in XML-TEI:

```
<item n="55" xml:id="CAT_000037_e55">
  <num>55</num>
  <name type="author">Scudéry (Madeleine de),</name>
  <trait>
    <p>célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la
      Sapho moderne, née au Havre.</p>
  </trait>
  <desc>L.a.s. à Huet, (1686), 3 p. in-4. 100 »</desc>
  <note>Curieuse épître sur le Lutrigot de M. de Bonnacorse. L'auteur est un fort
    honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent
    rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers: Venez
    Pradon et Bonnacorse, etc.)</note>
</item>
```

Such an encoding allows a simple disambiguation of the entry and increases the accuracy of the search. In our example, we have the names of three major 17th c. French writers: the novelist Madeleine de Scudéry (1607-1701), the bishop Daniel Huet (1630-1721), and the satirist and poet Nicolas Boileau-Despréaux (1636-1711). The three names are enclosed in three different tags (name, desc and note) reflecting their status in the document (author, addressee, mention): we can therefore easily narrow down our query to a name depending on its role.

#### 3.2 Workflow

The presented encoding can be compiled semi-automatically through a simple three steps workflow:

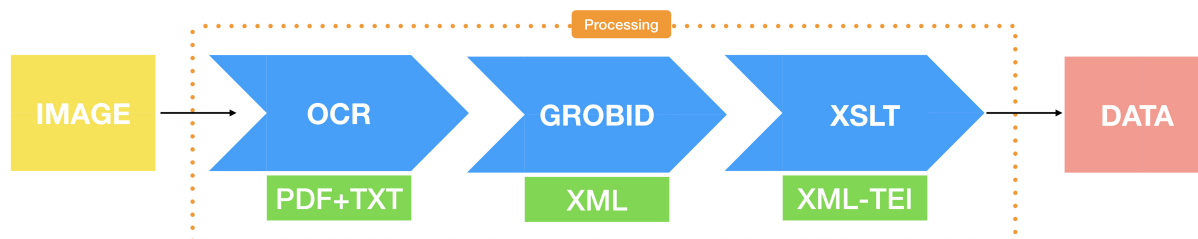


Figure 3: Workflow.

The scan is OCRised with Transkribus (Kahle et al., 2017), for which a substantive model of 125,000 words has been created (CER of 0.59%). The pdf with a text layer is then processed with GROBID-Dictionaries (Khemakhem et al., 2018b), a tool relying on text and layout features (cf. Figure 4) to perform a supervised classification of the parsed text and generate a TEI compliant encoding where the various segmentation levels are associated with an appropriate XML tessellation (Khemakhem et al., 2017). Preliminary designed for the retroconversion of dictionaries (Bohbot et al., 2018), it is also used to parse large bibliographical collections (Lindemann et al., 2018) or address directories (Khemakhem et al., 2018a).

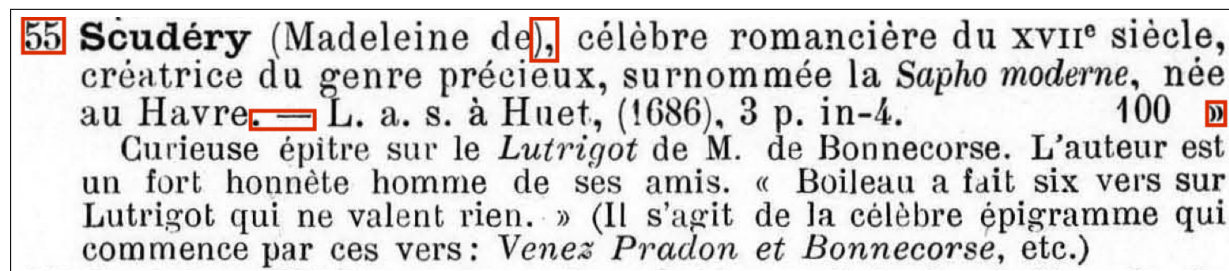


Figure 4: Features.

GROBID-dictionaries provides an answer to important issues left open by previous attempts, that do not produce standardised data (e.g. in XML-TEI), are not open source (Cuadra and Michels, 2013) or do not offer fine-grained encoding (Barman, 2019). On top of this, GROBID is a free, language agnostic, easily trainable solution compatible with other sub-projects of the GROBID galaxy, which leaves the door open to further analysis with complementary tools, such as GROBID-NERD (Named-Entity Recognition and Disambiguation).

After preliminary tests ensuring the compatibility of GROBID-Dictionaries with sale catalogues (Khemakhem et al., 2018c) models have been created for the RDA and many other catalogues (Rondeau du Noyer et al., 2019) with a new bigram template for the GROBID-Dictionaries models (Rondeau Du Noyer et al., 2019) to reinforce the parsing of the structure of each entry.

With GROBID-Dictionaries, the document is annotated using a cascading approach: several Conditional Random Fields (CRF) models are applied one after the other, each of them corresponding to a granularity level in the final XML hierarchy:

Levels	Tag(s)	Task
1	body	separates the content from running titles, page numbers, ...
2	entry	separates the entries in the <body>
3	num, form and sense	separates the lot n°, information on the author and the MS in <entry>
4	name and desc	separates the name of the author and its biography in <form>
5	subsense and note	separates the MS description and the additional note in <sense>

Table 1: GROBID-Dictionaries Segmentation levels

The GROBID-Dictionaries output for our example is therefore the following:

```
<entry>
  <num>55</num>
  <form type="lemma">
    <name type="author">Scudéry (Madeleine de</name>
    <pc>), </pc>
    <desc>célèbre romancière du XVIIe siècle, créatrice du genre précieux, ...</desc>
  </form>
  <pc>. -</pc>
```

```

<sense>
  <subSense>L.a.s. à Huet, (1686), 3 p. in-4. 100</subSense>
  <pc>></pc>
  <note>Curieuse épître sur le Lutrigot de M. de Bonnacorse. L'auteur est un ...</note>
</sense>
</entry>

```

GROBID-Dictionaries being developed for lexicographic purposes, its results are encoded in a TEI compliant output, but with tags reflecting the content of dictionaries rather than catalogues. Therefore, we automatically convert the output into a second TEI document whose tags are dedicated for the described catalogue elements. The consistency of the transformation output is controlled with a specific schema, prior to its final publication via an XML database.

#### 4 Future work

As for future work, four tasks will be undertaken. First, we will move towards a fully open source workflow and therefore abandon Transkribus for Kraken (Kiessling, 2019). Second, we will increase the size of our database by retroconverting the entire *RDA* collection (c. 500 catalogues), but also another important series of catalogues: the *Lettres autographes et documents* published by the other branch of the Charavay family up to the First World War (c. 500 catalogues). Third, we will improve our modeling and increase the granularity of our data collection to capture more informations regarding the document (size, format, length) and named entities (people, places). Fourth, we will use this additional information to reconcile entries and share our work with similar projects via an RDF export.

Ideally, we should eventually be able to go from the TEI digital edition of sales catalogues to a semantic dataset, described using controlled vocabularies where authors, places and manuscripts would be referred to using unique identifiers (ISNI, ISMI. . .). It would allow federated search with other databases of sold manuscripts, but also with catalogues of libraries in France and abroad.

#### Bibliografia

- Raphael Barman. 2019. Aucase - auction catalog segmentation. Type: dataset. <https://github.com/raphaelBarman/aucase-inha/>.
- Thierry Bodin. 2000. Les grandes collections de manuscrits littéraires. In *Les Ventes de livres et leurs catalogues: XVIIe-XXe siècle*, École des chartes, Paris, pages 169–190.
- Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem, and Laurent Romary. 2018. Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. In *GLOBALEX 2018 - Globalex workshop at LREC2018*. Miyazaki, Japan, pages 1–6. <https://hal.archives-ouvertes.fr/hal-01728328>.
- Toby Burrows, Eero Hyvönen, Lynn Ransom, and Hanno Wijsman. 2019. Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts. *Manuscript Studies* 3(1). [https://repository.upenn.edu/mss\\_sims/vol3/iss1/13](https://repository.upenn.edu/mss_sims/vol3/iss1/13).
- Ruth Cuadra and Suzanne Michels. 2013. Publishing German Sales, A Look under the Hood of the Getty Provenance Index. <https://blogs.getty.edu/iris/publishing-german-sales-a-look-under-the-hood-of-the-getty-provenance-index/>.
- Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Kyoto, Japan, volume 04, pages 19–24. <https://doi.org/10.1109/icdar.2017.307>.
- Mohamed Khemakhem, Carmen Brando, Laurent Romary, Frédérique Mélanie-Becquet, and Jean-Luc Pinol. 2018a. Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. In *JADH2018 "Leveraging Open Data"*. Tokyo, Japan. <https://hal.archives-ouvertes.fr/hal-01814189>.
- Mohamed Khemakhem, Luca Foppiano, and Laurent Romary. 2017. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. eLex, Leiden, the Netherlands. <https://hal.archives-ouvertes.fr/hal-01508868v2>.

- Mohamed Khemakhem, Axel Herold, and Laurent Romary. 2018b. *Enhancing Usability for Automatically Structuring Digitised Dictionaries*. In *GLOBALEX workshop at LREC 2018*. Miyazaki, Japan. <https://hal.archives-ouvertes.fr/hal-01708137>.
- Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, and Giancarlo Luxardo. 2018c. *Automatically Encoding Encyclopedic-like Resources in TEI*. Tokyo, Japan. <https://hal.inria.fr/hal-01819505>.
- Benjamin Kiessling. 2019. *Kraken - an universal text recognizer for the humanities*. Utrecht, The Netherlands. <https://dev.clariah.nl/files/dh2019/boa/0673.html>.
- Alphonse de Lamartine. 2001. *Correspondance*, volume 3. Honoré Champion, Paris.
- David Lindemann, Mohamed Khemakhem, and Laurent Romary. 2018. *Retro-digitizing and Automatically Structuring a Large Bibliography Collection*. In *European Association for Digital Humanities (EADH) Conference*. EADH, Galway, Ireland. <https://hal.archives-ouvertes.fr/hal-01941534>.
- Alicia C. Montoya. 2018. The MEDiate project. *Jaarboek voor Nederlandse Boekgeschiedenis / Yearbook for Dutch Book History* 25:229 – 232.
- Lucie Rondeau Du Noyer, Simon Gabay, Mohamed Khemakhem, and Laurent Romary. 2019. *Scaling up Automatic Structuring of Manuscript Sales Catalogues*. Graz, Austria. <https://hal.inria.fr/hal-02272962>.
- Lucie Rondeau du Noyer, Simon Gabay, Mohamed Khmakhem, and Laurent Romary. 2019. *Automatic TEI encoding of manuscripts catalogues with GROBID-Dictionaries*. Type: dataset. <https://doi.org/10.5281/zenodo.3383658>.
- Léa Saint-Raymond. 2018. *Le pari des enchères : le lancement de nouveaux marchés artistiques à Paris entre les années 1830 et 1939*. Corpus bibliographique des ventes aux enchères publiques considérées Type: dataset. <https://doi.org/10.7910/DVN/MZIBKB>.
- Marie de Rabutin-Chantal Sévigné. 1978. *Correspondance*, volume 3. Gallimard, Paris.
- Anne-Marie Turcan-Verkerk and Paul Bertrand. 2014. BIBLISSIMA : Bibliotheca bibliothecarum novissima, an observatory for written cultural heritage of the Middle Age and the Renaissance. In *Heritage and Digital Humanities. How should training practices evolve?*, Berlin, pages 129–139.
- Voltaire. 1960. *Correspondence*. 59. Institut et musée Voltaire, Genève.
- Hanno Wijsman. 2017. *The Bibale Database at the IRHT*. *Manuscript Studies* 1(2). [https://repository.upenn.edu/mss\\_sims/vol1/iss2/10](https://repository.upenn.edu/mss_sims/vol1/iss2/10).